
UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Ber dos Santos Neves

# Classificação de gênero social em padrões vocais utilizando redes neurais artificiais

Campinas

2024

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Ber dos Santos Neves

# Classificação de gênero social em padrões vocais utilizando redes neurais artificiais

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestra em Engenharia Elétrica, na Área de Engenharia da Computação.

Orientador: Prof. Dr. Romis Ribeiro de Faissol Attux

Co-orientador Profa. Dra. Ana Carolina Constantini

Este exemplar corresponde à versão final da dissertação defendida pela aluna Ber dos Santos Neves, e orientada pelo Prof. Dr. Romis Ribeiro de Faissol Attux

Campinas

2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca da Área de Engenharia e Arquitetura
Elizangela Aparecida dos Santos Souza - CRB 8/8098

Neves, Ber dos Santos, 1995-
N414c Classificação de gênero social em padrões vocais utilizando redes neurais artificiais / Ber dos Santos Neves. – Campinas, SP : [s.n.], 2024.

Orientador: Romis Ribeiro de Faissol Attux.
Coorientador: Ana Carolina Constantini.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Faculdade de Engenharia Elétrica e de Computação.

1. Aprendizado de máquina. 2. Aprendizado profundo. 3. Gênero. 4. Voz. I. Attux, Romis Ribeiro de Faissol, 1978-. II. Constantini, Ana Carolina, 1985-. III. Universidade Estadual de Campinas (UNICAMP). Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Informações Complementares

**Título em outro idioma:** Social gender classification in vocal patterns using artificial neural networks
**Palavras-chave em inglês:**
Deep learning
Machine learning
Gender
Voice
**Área de concentração:** Engenharia de Computação
**Titulação:** Mestra em Engenharia Elétrica
**Banca examinadora:**
Romis Ribeiro de Faissol Attux [Orientador]
Levy Boccato
Leonardo Wanderley Lopes
**Data de defesa:** 16-07-2024
**Programa de Pós-Graduação:** Engenharia Elétrica

**Identificação e informações acadêmicas do(a) aluno(a)**
- ORCID do autor: https://orcid.org/0009-0005-5889-0258
- Currículo Lattes do autor: https://lattes.cnpq.br/7613684099808678

COMISSÃO JULGADORA – DISSERTAÇÃO DE MESTRADO

**Candidato(a):** BER DOS SANTOS NEVES RA: 194768
**Data da defesa:** 16 de julho de 2024
**Título da Dissertação:** “CLASSIFICAÇÃO DE GÊNERO SOCIAL EM PADRÕES VOCAIS UTILIZANDO REDES NEURAIS ARTIFICIAIS.”

**Banca:**
Prof. Dr. Romis Ribeiro de Faissol Attux (Presidente)
Prof. Dr. Levy Boccato
Prof. Dr. Leonardo Wanderley Lopes

A Ata de Defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

*Dedico esta dissertação a todo mundo que não se sente parte da crueldade deste mundo.*

# Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Agradeço por todo suporte financeiro. Gostaria de agradecer a todes aqueles que em seus gestos vocais resistem e residem suas experiências mais profundas e subjetivas. Agradeço ao meu Pai, José, que deu seu sangue para que eu pudesse realizar meu sonho. Agradeço à minha mãe, Maria Terezinha, e às minhas irmãs Rafaela e Gabriela por estarem sempre presentes me dando apoio emocional. O amor da família foi essencial.

Ao prof. Dr. Romis Attux, agradeço pelo carinho e apoio, pelos ensinamentos e por ser um docente incrível e singular. Foi essencial na minha jornada profissional estar presente no DSPCom e conviver com o Romis. Ao Pedro, o Chefe, agradeço pelos ensinamentos em Machine Learning e por me ajudar a acreditar na minha capacidade.

Agradeço muito pela prof. Dra. Ana Carolina Constantini pela oportunidade de trabalhar lado a lado no CEPRE, o que me fez avançar muito em conteúdo teórico e humano. E neste mesmo ambiente, também convivi com o Ms. Diego Martinho, parceiro de pesquisa e de arte, que muito me apoiou e ensinou. Gratidão.

À prof. Dra. Regina machado, agradeço pelo apoio emocional e prático, pelos questionamentos riquíssimos e pela oportunidade de fazer parte do grupo de pesquisa Vox Mundi, o qual também devo muitas dessas reflexões. Agradeço muito à Dra. Maria Elisa Pompeu, que foi uma amiga de pesquisa essencial na minha jornada.

Agradeço ao professor Dr. Jônatas Manzolli, ao Dr. Micael Antunes e ao Lucas Bertoloto Pereira pelo auxílio em minhas pesquisas com voz e arduíno, que me possibilitaram pensar sobre as não-binariedades vocais associadas às ferramentas tecnológicas.

Agradeço também aos ex-alunes da Unicamp Gabriel Lopes de Cicco, Carolina Klingenberg, Victor Santos; aos alunes da UFScar Larissa Belafonte, Eric Escolástico e David Evaristo; e ao meus amigues Gabriel Zanetti e Rebeca de Nadai que estavam presentes na primeira vez que eu pensei na possibilidade de entender a voz LGBTQIA+, suas experiências pessoais e possibilidades psicoacústicas. Foi um momento muito incrível pra mim, que aconteceu em 2016 em Americana/SP e mudou minha vida para sempre.

# Resumo

Através da reflexão filosófica e linguística sobre os fundamentos da voz e do gênero, o presente trabalho propôs um problema de classificação inovador, o qual busca situar e comparar as vozes transgêneras e cisgêneras em relação aos seus atributos. Para tal, foi analisado um classificador de voz baseado no modelo Wav2Vec2, treinado para categorizar vozes em "male", "female"e "other". O modelo foi treinado por 7 épocas com CrossEntropyLoss e SGD. Avaliou-se o desempenho com matrizes de confusão, curvas ROC e AUC. A técnica de redução de dimensionalidade t-SNE foi usada para visualizar as representações internas do modelo. Os resultados mostram que o modelo classifica vozes de maneira satisfatória, com uma análise detalhada das suas capacidades discriminatórias, entretanto, não se obteve boa precisão para a classe "other".

**Palavras-chaves**: Voz e Gênero; Aprendizado Profundo; Redes Neurais Artificiais.

# Abstract

Through philosophical and linguistic reflection on the foundations of voice and gender, this work proposed an innovative classification problem that seeks to position and compare transgender and cisgender voices in relation to their attributes.

To achieve this, a voice classifier based on the Wav2Vec2 model was analyzed, trained to categorize voices as "male," "female," and "other." The model was trained for 7 epochs using CrossEntropyLoss and SGD. Its performance was evaluated using confusion matrices, ROC curves, and AUC. The t-SNE dimensionality reduction technique was used to visualize the model's internal representations. The results show that the model classifies voices robustly, with a detailed analysis of its discriminatory capabilities; however, it did not achieve good accuracy for the "other" class.

**Keywords**:Voice and Gender; Deep Learning; Artificial Neural Networks.

*"A voz é uma janela na qual você pode ver o que tem dentro"*
*(Marcela Sgavioli)*

# Lista de ilustrações

# Sumário

# 1 Introdução

Na atualidade, os estudos de gênero são essenciais para compreender e desafiar as normas e expectativas sociais que influenciam a vida das pessoas com base em sua expressão de gênero social. Essas normas podem perpetuar desigualdades, discriminação e violência, tornando crucial a análise crítica dessas estruturas para promover a igualdade de gênero e a inclusão social. Na era da tecnologia e da inteligência artificial (IA), esses estudos assumem uma nova relevância, especialmente quando consideramos o papel da escuta de máquina. A IA, como uma ferramenta poderosa, pode tanto perpetuar quanto desafiar desigualdades de gênero, dependendo de como é projetada, implementada e treinada para "ouvir"e interpretar os dados, o que ressalta a importância de uma abordagem sensível e inclusiva nesses processos.

É possível encontrar diversos modelos de *machine learning* e *deep learning* que desenvolvem classificação de áudio, os quais têm se mostrado promissores para tarefas de análise musical e vocal, por exemplo, dentro da área da engenharia do som (KAWAMURA *et al.*, 2019) (RAGAV *et al.*, 2023) (AGARWAL *et al.*, 2022) (AL-DHIEF *et al.*, 2021). Essas tarefas envolvem toda uma capacidade de escuta da máquina associada aos parâmetros acústicos que somos capazes de ouvir e descrever, ou seja, uma simbiose entre a desenvoltura da escuta humana e o desempenho da máquina em processar e inteligibilizar dados.

Do ponto de vista da máquina, assim como nossa percepção auditiva é subjetiva e culturalmente influenciada (CAO; GROSS, 2015) , a ferramenta computacional pode ser afetada por variações de sotaque, entonação, pelo contexto social, além do enviesamento do dataset utilizado (LYON, 2010). O encontro entre a escuta humana e a classificação de sons pode ofertar abordagens eficazes na resolução dos desafios enfrentados pelos sistemas de reconhecimento de voz e processamento de áudio na computação, além de que, no caso da presente pesquisa, pode se tornar uma ferramenta diagnóstica auxiliar na área da saúde.

## 1.1 Escuta humana versus. escuta de máquina

A habilidade de ouvir vai muito além da simples recepção de sons; ela envolve a interpretação e atribuição de significado a esses sinais auditivos, o que está diretamente relacionado tanto à psicoacústica quanto à linguística. A psicoacústica examina como os sons são percebidos pelo sistema auditivo humano, enquanto a linguística estuda como os

sons são organizados e compreendidos dentro do contexto da linguagem. Da mesma forma, o processo de classificação é uma atividade mental que organiza e categoriza informações para melhor compreensão. Quando aplicamos essa ideia à escuta, percebemos que nosso cérebro naturalmente classifica os sons, priorizando alguns e filtrando outros de acordo com a relevância do contexto. Portanto, a escuta e a identificação moldam nossa compreensão do mundo e influenciam como respondemos e interagimos com o som (PLACK, 2005).

Nossos ouvidos, ao serem afinados para perceber uma vasta gama de frequências e intensidades, permitem-nos experimentar o mundo sonoro de maneiras únicas. A capacidade inerente de identificar sons, que vai desde a distinção entre melodias até de alertas de perigo, é parte integrante de nossa natureza evolutiva. À medida que evoluímos, desenvolvemos uma sofisticada habilidade para atribuir significados emocionais e contextuais aos sons que ouvimos (PLACK, 2005). Essa semântica sonora, subjetiva e culturalmente influenciada, aclimata nossa percepção auditiva, além de criar conexões profundas com o ambiente ao nosso redor. Isso impacta diretamente na forma da produção vocal, que se integra diretamente à escuta em uma relação dialógica, ou seja, a voz conduz a escuta e vice versa (BARTHES, 1990).

A conexão entre a capacidade auditiva e a expressão vocal envolve um mecanismo contínuo (BARBOSA; MADUREIRA, 2015). Este mecanismo faz com que a estimulação do ouvido externo por um sinal acústico desencadeie imediatamente a resposta vocal do ouvinte. Esse processo leva à repetição vocal de certas linhas expressivas captadas durante a audição. Essa repetição pode se manifestar como uma vocalização convertida em energia acústica ou como a formação de uma imagem vocal mental, também conhecida como subvocalização (BIRAN, 2005).

A criação de uma representação vocal virtual, ou a ativação correspondente das "teclas do instrumento vocal", sem necessariamente gerar uma imagem acústica, torna-se uma etapa crucial no processo de escuta. Essa etapa funciona de maneira complementar ao sinal perceptivo que ocorre no ouvido quando este é estimulado por uma excitação sonora externa material.

Assim, a audição ocorre por meio da interação mútua de pelo menos dois meios distintos: o ouvido e a voz. Pode-se considerar essa interação como um curto-circuito sutil entre os actantes. Importante notar que isso não exclui a possibilidade da atuação simultânea de outros meios e sentidos corporais durante a escuta, dada a natureza polimodal ou multimodal da percepção. No entanto, é essencial destacar que, na audição, mesmo que vários meios perceptivos possam ser estimulados intensivamente em momentos específicos, esses meios sempre operam em conjunto com esse acoplamento dinâmico entre voz e ouvido (PENHA, 2021).

Ao considerar o acoplamento ouvido-voz como uma máquina, é importante entender o conceito de máquina de forma abrangente, não limitada ao sentido de uma máquina tecnológica. Na verdade, refere-se a uma forma mais ampla de conexão entre elementos e sistemas heterogêneos. Nessa concepção, a máquina é permeável ao exterior e estabelece diversos modos de relação com componentes orgânicos, sociais e subjetividades, os quais se manifestam através dos parâmetros acústicos que utilizamos para diferenciar os sons (PENHA, 2021).

Os parâmetros acústicos e os descritores de áudio são elementos da experiência auditiva do computador relevantes na transcrição da percepção humana para a construção da percepção da máquina. Enquanto os parâmetros acústicos se relacionam com o aspecto perceptivo do som, os descritores de áudio oferecem uma interpretação matemática, permitindo que o computador compreenda e reproduza sons de maneira significativa (CAETANO *et al.*, 2019). Resumidamente, temos que:

- Parâmetro Acústico: Características físicas diretas do som, como frequência, amplitude, e duração;
- Descritor de Áudio: Características extraídas e processadas, frequentemente usadas em tarefas de análise e reconhecimento de áudio, podendo incluir parâmetros acústicos e outras representações derivadas do sinal.

A habilidade do computador em discernir nuances tonais, distinguir entre diferentes intensidades e captar variações temporais é denominada escuta de máquina, neste caso sendo a máquina um produto tecnológico (LYON, 2010).

Essa ponte entre a percepção humana e a escuta de máquina é fundamental para a criação de interfaces de usuário mais intuitivas e sistemas de reconhecimento de voz avançados, melhorando significativamente a qualidade da percepção de escuta do computador. Ao integrar as análises precisas dos parâmetros acústicos com descritores de áudio robustos, o computador reproduz sons, interpreta e pode responder de maneira mais sofisticada, aproximando-se da complexidade da percepção auditiva humana (LYON, 2010).

Do ponto de vista da evolução da tecnologia de processamento de fala, os algoritmos de classificação baseados em aprendizado de máquina puderam extrair padrões complexos e relações sutis entre os parâmetros acústicos, o que permitiu uma identificação mais precisa e contextualizada das características vocais. Os descritores de áudio, nesse contexto, proporcionam ao sistema informações interpretativas cruciais para discernir entre diferentes falantes, emoções e contextos acústicos. Essa abordagem integrada não apenas aprimora a precisão da classificação de voz, mas também expande as capaci-

dades de interação homem-máquina, tornando os sistemas mais adaptáveis e sensíveis à diversidade da expressão vocal humana (LYON, 2010).

Tratando-se de tarefas classificação sonora, tais como classificação de gênero musical ou de emoções por exemplo, as redes neurais profundas oferecem uma via alternativa, especialmente quando se trata do processamento de dados brutos. Enquanto os descritores de áudio oferecem uma representação matemática abstrata e interpretativa das características sonoras, as redes neurais e o *deep learning* capacitam os sistemas a lidar diretamente com dados brutos, obtendo os atributos relevantes ao processar as amostras em densas camadas neuronais. Esse aspecto é particularmente relevante ao lidar com conjuntos de dados extensos e diversificados, onde as nuances intrínsecas às informações brutas podem ser cruciais para aprimorar a eficácia dos modelos. (AGGARWAL, 2021)

As redes neurais, por meio de suas camadas interconectadas de unidades de processamento, têm a capacidade de apreender padrões intricados e hierarquias complexas em dados de áudio, capturando nuances e atributos que são difíceis de expressar com descritores tradicionais. No entanto, alcançar uma alta precisão requer um número substancialmente maior de amostras em comparação com modelos classificadores que utilizam dados processados por descritores de áudio na entrada (AGARWAL *et al.*, 2022).

Quando dados filtrados por descritores de áudio são utilizados na entrada de uma rede neural, cria-se um ambiente sinérgico. Nesse contexto, a riqueza interpretativa dos descritores pode ser combinada com o poder de aprendizado profundo das redes neurais, o que, em algumas situações, leva à criação de sistemas de classificação de voz mais precisos, flexíveis e adaptáveis às complexidades do espectro auditivo. Essa abordagem é especialmente necessária quando se trabalha com um conjunto reduzido de amostras (GOODFELLOW *et al.*, 2016).

Essa capacidade de processamento direto de dados brutos via *deep learning* oferece uma vantagem significativa, mas é limitada sempre por fatores como a quantidade de amostras - que pode chegar à casa de milhares ou até milhões para que se obtenha um resultado relevante - e a quantidade de processamento demandada. Por conta disso, a utilização de sinais pré-processados por descritores de áudio na entrada do sistema pode aumentar a eficácia de uma rede neural que possui um campo amostral mais reduzido. Atualmente, a utilização de aprendizado de máquina como uma ferramenta de escuta passa por esses e outros fatores limitantes e, por conta disso, as transferências de aprendizado de um modelo para o outro têm sido muito usadas como forma de aumentar a precisão (GOODFELLOW *et al.*, 2016).

# 2 O que é voz?

## 2.1 Voz como fonte de comunicação

Atuando como uma ferramenta ativa, virtual e maleável da comunicação, a voz é um objeto de estudo que pode ser definido por diversos vieses: fisiológico, político-social, psicológico, artístico e o que mais convir. Isso acontece porque, para além da vibração das pregas vocais e da modulação do som no trato vocal, a voz se modifica através dos processos de interação entre sujeito, ambiente, sociedade e indivíduo, uma vez que ela mobiliza uma boa parte da comunicação social(MEY, 1998). Partindo desse pressuposto, a voz protagoniza o cruzo de uma encruzilhada, o encontro de vários caminhos percorridos por e com ela. Por definição, o cruzo é:

> *O devir, o movimento inacabado, saliente, não ordenado e inapreensível. O cruzo versa-se como atravessamento, rasura, cisura, contaminação, catalização, bricolagem [...]. O cruzo é a rigor uma perspectiva que mira e pratica a transgressão e não a subversão, ele opera sem a pretensão de exterminar o outro com que se joga, mas de engoli-lo, atravessá-lo, adicioná-lo como acúmulo de força vital* (RUFINO, 2019)

Segundo Roland Barthes, a voz materializa a transgressão da fronteira entre corpo e discurso (BARTHES, 1990): portanto, é um sinal acústico que representa essa simbiose. A natureza da produção vocal se dá através dessas duas dimensões: a natureza biológica constituída entre genótipo e fenótipo (corpo); e o que se pretende comunicar (discurso). Na alta variabilidade dos parâmetros que atravessam a produção vocal - tais como gênero, raça, classe social, território, língua - garante-se o comportamento não-linear da voz, com base nas duas dimensões que fundamentam o conteúdo vocal e localizam o sujeito/indivíduo para si e para os outros.

## 2.2 Teoria fonte-filtro da produção vocal: do acoplamento linear ao não-linear.

Tratando da parte relativa aos sinais, a teoria fonte-filtro foi adaptada para modelar a produção vocal. Os pioneiros nesse processo foram os japoneses Chiba e Kajiyama (CHIBA; KAJIYAMA, 1941); e Gunnar Fant, com a publicação da obra “Acoustic

theory of voice production" (FANT, 1970). A teoria fonte e filtro da produção da fala postula que a voz é produzida através da interação entre a fonte glótica e trato vocal, os quais funcionam, respectivamente, como oscilador (fonte do sinal) e filtro. Os pulmões têm a função de produzir a força motriz do sinal, comprimindo e gerando o fluxo de ar que passa pelas pregas vocais e, por meio da força de Bernoulli, gera o sinal que será modificado no trato vocal, o qual depende da resposta em frequência dada pelo ajuste dos articuladores para formar o som que ouvimos. (BARBOSA; MADUREIRA, 2015)

Matematicamente falando, essa relação pode ser convertida em uma convolução no tempo entre o sinal da fonte glótica e a resposta ao impulso do trato vocal:

$$s(t) = e(t) * h(t) \tag{2.1}$$

No domínio da frequência, através da transformada de Fourier de s(t), temos:

$$S(f) = E(f)H(f) \tag{2.2}$$

Um detalhe importante sobre a relação entre E(f) e H(f) é que, na teoria de Fant, fonte e filtro são actantes independentes um ao outro, ou seja, o ajuste de um não interfere no outro - assim, estão relacionados de forma linear.

A definição de cobertura no canto- técnica levantamento do palato para alargamento do trato vocal- já coloca em xeque essa teoria: o palato é um articulador que faz parte do filtro e interfere na fonte quando movimentado - quanto mais alto palato, mais baixa a laringe fica (SUNDBERG, 2015). Isso significa que fonte e filtro, nesse caso, são sistemas acoplados, o que complexifica as possibilidades e demanda um tratamento não linear para a teoria da produção da fala. Além disso, a teoria linear da fala não prevê bifurcações no filtro, então consoantes nasais e vogais nasais não podem ser representadas por esse modelo (TITZE, 2008).

Ingo Titze observou essas inconsistências e percebeu que a emissão mais favorável para a harmonicidade da voz têm relação direta com a inertividade da reatância supraglotal, que promove menor instabilidade no fluxo vocal por reforçar os harmônicos nessas regiões onde a reatância supraglótica tem valor positivo (TITZE, 2008). Essas regiões são chamadas de níveis de interação, e se dividem em dois níveis.

O nível 1 é referente à parte mais grave da extensão vocal (região de fala), na qual a vibração das pregas vocais não é influenciada pelas reatâncias supra e sub-glotais. Em vez disso, o que muda é o fluxo de pulsos glotais. O nível 2 é aquele representado pela dependência das reatâncias supra e sub-glotais em relação aos modos de fonação e aos padrões vibratórios das pregas vocais. No que tange o dinamismo da expressão vocal, o

modelo não linear da voz têm se mostrado como mais adequado para descrever a mecânica por trás da fluidez vocal. Esses detalhes são importantes para a presente pesquisa, porque é a partir da dinâmica e da plasticidade do sistema vocal humano que se obtém diversas possibilidades de gênero social na expressão vocal.

## 2.3 Relação entre gênero e voz

A expressão do gênero social na voz é um dos parâmetros que modificam os padrões entre as dimensões corpo e discurso. Tal como um viés do cruzo "voz", o gênero social se manifesta em diversos atributos vocais, como vistos em (MARTINHO; CONSTANTINI, 2024)(ZIMMAN, 2018), (CARTEI *et al.*, 2014), (BORSEL; MAESSCHALCK, 2008), (LATINUS; TAYLOR, 2012), (FOUQUET *et al.*, 2016), (HANCOCK *et al.*, 2014), (HANCOCK *et al.*, 2015), (PERNET; PASCAL, 2012), (CARTEI *et al.*, 2019), (AZUL *et al.*, 2020), (SKUK *et al.*, 2015), (DACAKIS *et al.*, 2012) e (BORSEL *et al.*, 2009).

Dentro da lógica do determinismo biológico, existe um binarismo homem x mulher e masculino x feminino nas vozes, proveniente da matriz da heterossexualidade compulsória. Segundo Butler (BUTLER, 2015), essa matriz social organiza as performances de gênero através da regulação política desses corpos. O gênero é um ato que se faz através da performance repetida ao longo do tempo e da vivência coletiva:

> *"O gênero não deve ser construído como uma identidade estável ou um locus de ação do qual decorrem vários atos; em vez disso, o gênero é uma identidade tenuemente constituída no tempo, instituído num espaço externo por meio de uma repetição de atos"* (BUTLER, 2015, p.242).

A virtualidade do tempo no gênero, sua característica atualizável, é compatível com a maleabilidade da voz no tempo, pois ambos são constituintes da vida e da comunicação humana: tudo depende do corpo que se tem e do que se deseja comunicar. Tal como um rizoma (DELEUZE; GUATTARI, 1996), a soma dessas perspectivas é uma tradução de linguagens originárias de diversas ciências, que possuem suas próprias formações semânticas e confluem para a produção do híbrido objeto migrado, fruto da relação entre corpo e discurso.

Para o entendimento dos parâmetros acústicos que fazem uma voz ser entendida como "de homem" ou "de mulher", é necessária a compreensão sobre como os seres humanos classificam os sons. Enquanto questões fundamentais das ciências cognitivas (MURPHY, 2004), a percepção e classificação de estímulos produzidos por vozes e rostos dizem muito a respeito de relações sociais e culturais. Então, a voz se torna um indicador

de como a relação entre sexo e gênero se manifesta, com parâmetros acústicos associados a essa percepção e à classificação de gênero social.

A identidade, em fragmentos, compacta na voz marcadores para além do binarismo de gênero, com a migração de características fixas dos papéis de gênero para uma posição mais híbrida. Segundo Herrero (HERRERO, 2009), a relação entre voz e identidade exprime marcadores que indicam aspectos físicos, sociais e psicológicos do indivíduo. Logo, a expressão de gênero binária oferece uma categorização desses marcadores entre masculinos e femininos, para homens cisgêneros e mulheres cisgêneras. Quando a combinação entre eles não oferece uma opção de timbre inteligível para essa categorização, esbarramos no que aqui se denomina como "vozes não-binárias".

Quando escutamos uma voz, a diferenciamos em relação a outras por suas particularidades acústicas: se é grave ou aguda, aveludada ou estridente, volumosa ou baixinha etc.. Assim, compõe-se o timbre, conceito associado à identificação de um som e suas particularidades perceptivas (PEETERS; AL., 2011). As descrições desses timbres se dão pela via da escuta — com os marcadores vocais que citamos, por exemplo, ou qualquer conjunto de parâmetros que se comprometa a descrever o som de forma sensível; o correlato físico-matemático desses parâmetros, estudados na presente pesquisa, são chamados de atributos ou features. Eles auxiliam na extração de dados acústicos precisos utilizando processamento digital de sinais, de forma a se constituírem possíveis valores para os aspectos perceptuais do timbre (PEREIRA, 2009).

Dessa forma, na presente pesquisa, também há de se relacionar os atributos acústicos com os marcadores sociais relevantes para a expressão do gênero social, de forma a transcrever a constituição dos signos estatisticamente, ou seja, seu significante acústico.

# 3 Aprendizado de Máquina e Aprendizado Profundo

Na presente pesquisa, os processos de extração de atributos e classificação serão realizadas por redes neurais artificiais (em inglês, *artificial neural networks*) sob o paradigma de aprendizado profundo (em inglês, *deep learning*). Por esse motivo, neste capítulo, partindo do conceito geral de aprendizado de máquina, será feita uma apresentação dos fundamentos técnicos subjacentes a essas redes, detalhando os elementos mais relevantes para a análise subsequente.

## 3.1 Breve Panorama Histórico

O conceito de aprendizado de máquina teve origem no século XX, havendo sido influenciado por estudos em teoria da computação, estatística e inteligência artificial. Durante as décadas de 1940 e 1950, figuras proeminentes como Alan Turing, Claude Shannon, Allen Newell, Herbert Simon e Warren McCulloch ajudaram a desenvolver teorias cruciais que moldaram o campo (OKHUNOV *et al.*, 2023).

Em 1950, Alan Turing propôs o célebre teste que hoje leva seu nome, um marco no campo da inteligência artificial. Ele postulou que uma medida pragmática da inteligência poderia ser a capacidade de uma máquina se comportar de forma indistinguível de um ser humano durante uma conversa(TURING, 1950). Embora o teste tenha suas limitações, o conceito subjacente destaca a natureza prática da inteligência artificial e do aprendizado de máquina, evidenciando a capacidade de realizar tarefas cognitivas de maneira autônoma. (TURING, 1937)

Essencialmente, o aprendizado de máquina refere-se à capacidade dos sistemas computacionais de aprenderem a partir de dados, em vez de serem explicitamente programados para realizar tarefas. Isso requer que o sistema se adapte, a partir da informação a que tem acesso, de modo a realizar proficientemente tarefas como reconhecimento de padrões, previsão e tomada de decisões. O aprendizado de máquina engloba uma variedade de técnicas e algoritmos, desde modelos simples de regressão linear até redes neurais com enorme quantidade de parâmetros (BISHOP, 2006).

Um dos pontos de inflexão no desenvolvimento do campo do aprendizado de máquina foi o advento do aprendizado profundo sobre redes neurais. O aprendizado profundo é uma subárea do aprendizado de máquina que se concentra no treinamento de

redes neurais profundas, ou seja, modelos compostos por número considerado elevado de camadas de unidades de processamento interconectadas, em analogia com aspectos da operação do sistema nervoso (GOODFELLOW *et al.*, 2016).

Pode-se considerar que a área de redes neurais surge com o trabalho pioneiro de McCulloch e Pitts no início da década de 1940 (MCCULLOCH; PITTS, 1943). O modelo proposto por eles foi marcante por ter expresso matematicamente o comportamento básico de um neurônio à luz da lógica booleana. Nesse contexto, o neurônio é modelado como um dispositivo que recebe múltiplas entradas, realiza uma agregação dessas entradas, e, em seguida, aplica uma função de ativação para determinar se o neurônio "dispara"ou não.

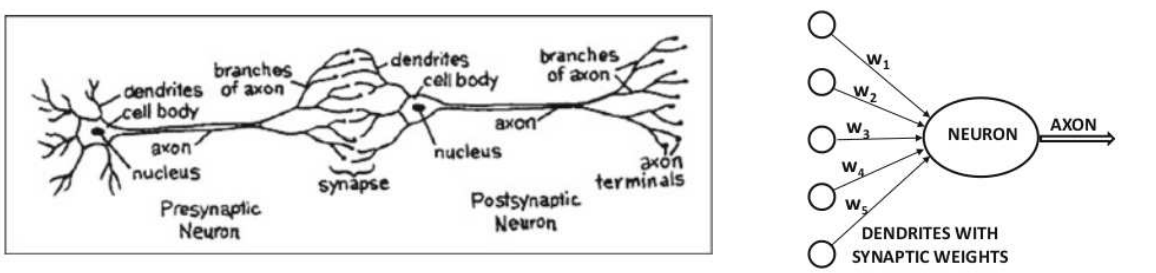
Porém, o primeiro trabalho que conseguiu trazer uma formulação madura para aprendizado dos parâmetros de uma rede neural no contexto de aprendizado supervisionado foi o de Frank Rosenblatt, em 1958 (ROSENBLATT, 1958). Rosenblatt introduziu o *perceptron*, um modelo que, guardando analogia com a formulação de McCulloch e Pitts, permitia a construção de redes neurais de camada única capazes de resolver problemas de classificação linear.

Um caminho para lidar com o caso não-linear geral foi através da concatenção de modelos tipo perceptron em múltiplas camadas, que deu origem à rede neural conhecida como MLP (do inglês *multilayer perceptron*) (GOODFELLOW *et al.*, 2016). A proposta de um processo de treinamento de redes MLP por meio do algoritmo de retropropagação de erro (em inglês, *error backpropagation*) (LECUN; FOGELMAN, 1987) (RUMELHART *et al.*, 1986), nos anos 1980, marcou uma primeira onda de popularização do estudo e da aplicação de redes neurais.

Após períodos de oscilação no interesse da comunidade por essas estruturas, no início da década de 2010, o paradigma neural ganhou muita força com o surgimento do arcabouço conceitual do que se convencionou denominar *aprendizado profundo* (em inglês, *deep learning*). Esse arcabouço envolve o emprego de redes neurais com número relativamente elevado de camadas, as quais, em tese, são capazes de realizar, a partir de dados brutos, a extração automática de características pertinentes a tarefas de classificação e regressão (GOODFELLOW *et al.*, 2016). O êxito de redes profundas, no entanto, depende da disponibilidade de vastas bases de dados e da possibilidade de empregar hardware paralelo de alto desempenho. De certa forma, ainda nos encontramos, atualmente, nessa nova onde de aplicação das redes neurais.

## 3.2 Redes Neurais

Conforme já mencionado, as redes neurais artificiais são técnicas de aprendizado de máquina que se inspiram nos mecanismos de aprendizagem em sistemas biológicos. Em certo sentido, elas têm por base modelos de neurônios como processadores de estímulos no sistema nervoso; esses neurônios, por sua vez, estão conectados uns aos outros através dos axônios e dendritos, formando a região denominada sinapse: é ali que os estímulos passam de um neurônio para outro com determinada eficiência, a qual é representada computacionamente como uma ponderação ou peso, como mostrado na Figura 3.1. Essa modelagem dos pesos das conexões sinápticas é crucial para o funcionamento das redes neurais, pois é responsável por determinar a importância relativa de cada entrada para a saída do neurônio (AGGARWAL, 2021).




(a) Biological neural network (b) Artificial neural network

Figura 3.1 – Conexões sinápticas entre neurônios. (AGGARWAL, 2021)

### 3.2.1 Perceptron

O perceptron, proposto por Frank Rosenblatt em 1958 (ROSENBLATT, 1958), é um modelo de neurônio artificial originalmente concebido para lidar com problemas binários de classificação linear. Ele recebe entradas $x_1, x_2, ..., x_n$, aplica pesos $w_1, w_2, ..., w_n$ a essas entradas e produz uma saída $y$ de acordo com a equação 1:

$$y = f\left(\sum_{i=1}^{n} w_i \cdot x_i + b\right)$$

[1]

Onde:

- $x_i$ são as entradas.
- $w_i$ são os pesos associados às entradas.
- $b$ é o viés (bias).

- $f(.)$ é a função de ativação, que determina se o neurônio deve ser ativado ou não com base na soma ponderada das entradas.

A função de ativação clássica do perceptron é a função degrau (*step function*), que retorna 1 se o resultado da soma ponderada mais o viés for maior ou igual a zero, e 0 caso contrário. Matematicamente, pode-se escrever:

$$f(z) = \begin{cases} 1 & \text{se } z \geq 0 \\ 0 & \text{se } z < 0 \end{cases}$$

onde $z$ recebe o nome de *ativação interna*:

$$z = \sum_{i=1}^{n} w_i \cdot x_i + b$$

Entretanto, outras funções de ativação são mais comumente utilizadas na atualidade, algumas das quais serão expostas na seção 3.2.3.5.

Durante o treinamento do perceptron, os pesos $w_i$ e o viés $b$ são ajustados de acordo com o erro cometido pela rede. A presença de uma referência caracteriza um processo de aprendizado supervisionado, e a abordagem mais usual é lançar mão de uma lei adaptação baseada nas derivadas da função custo, como a do gradiente descendente (AGGARWAL, 2021).

### 3.2.2 Redes Neurais MLP (Multilayer Perceptron)

Uma rede neural MLP (Multi-Layer Perceptron) é uma rede composta por múltiplas camadas de neurônios tipo perceptron, incluindo uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada neurônio em uma camada está conectado a todos os neurônios na camada seguinte, formando uma estrutura de alimentação adiante (em inglês, feedforward) (GOODFELLOW *et al.*, 2016).

Matematicamente, a saída de um neurônio em uma camada oculta pode ser representada como na equação 1 do perceptron. Essas equações são aplicadas a todos os neurônios em uma camada, e as saídas são então propagadas para a próxima camada oculta ou até a camada de saída, onde são produzidas as previsões ou classificações finais.

A função de ativação é um componente crucial no MLP, introduzindo não linearidades que permitem ao modelo capturar relações complexas nos dados. Funções de ativação comuns incluem a função sigmoide, a tangente hiperbólica (tanh) e a ReLU (Rectified Linear Unit), que serão detalhadas quais serão na seção 3.2.3.5. . A escolha da função de ativação pode influenciar significativamente o desempenho da rede neural,

uma vez que determina como os sinais são transformados e propagados através da rede, afetando a capacidade do modelo de capturar e representar padrões complexos nos dados.

A principal aplicação do MLP está na resolução de problemas de classificação e regressão. Em tarefas de classificação, a rede neural pode distinguir entre diferentes categorias, como reconhecimento de dígitos manuscritos ou classificação de imagens. Em regressão, o MLP é usado para prever valores contínuos, como preços de casas ou temperaturas.

O treinamento da rede neural MLP é realizado através do algoritmo de retropropagação (backpropagation), que ajusta os pesos das conexões entre os neurônios para minimizar o erro da previsão. Esse processo envolve a computação do gradiente do erro em relação a cada peso usando o método de gradiente descendente. Técnicas como a regularização e o dropout são frequentemente empregadas para prevenir o sobreajuste e melhorar a generalização do modelo.

### 3.2.3 Redes Neurais Convolucionais

Uma Rede Neural Convolucional (CNN) é um tipo de rede neural especialmente projetada para processar dados que possuem uma estrutura de grade, como imagens e séries. É composta por pelo menos uma camada convolucional, ou seja, uma camada em que se executem convoluções para extrair / reconhecer características importantes dos dados de entrada (GOODFELLOW *et al.*, 2016). Discutiremos brevemente, a seguir, os principais componentes de uma CNN e suas operações matemáticas.

#### 3.2.3.1 Camada de Convolução

Uma camada de convolução é um componente essencial em uma Rede Neural Convolucional (CNN). Ela opera aplicando filtros ou *kernels* aos dados de entrada segundo operações de convolução. Matematicamente, isso envolve o cálculo do produto escalar entre o kernel e diferentes regiões locais da imagem (AGGARWAL, 2021).

Suponha que tenhamos uma imagem de entrada $I$ e um kernel $K$ de dimensões $m \times n$. Uma versão computacionamente simples da operação de convolução pode ser realizada deslizando o kernel sobre a imagem e calculando o produto escalar entre os elementos do kernel e a região correspondente da imagem em cada posição. Seja $I_{ij}$ a região da imagem correspondente à posição $(i, j)$, então o cálculo da convolução $C$ na posição $(i, j)$ é dado por:

$$C_{ij} = \sum_{p=1}^{m} \sum_{q=1}^{n} I_{ij}(p, q) \times K(p, q)$$

O resultado $C_{ij}$ é conhecido como elemento do mapa de características ou *feature map*, que destaca áreas de importância na imagem. A aplicação de múltiplos filtros resulta em múltiplos mapas de características, capturando diferentes aspectos da imagem.

Esses mapas de características são, então, sujeitos a uma função de ativação e processados por camadas subsequentes da CNN para realizar tarefas como classificação ou detecção de objetos. A capacidade das camadas de convolução de extrair e aprender características importantes diretamente dos dados de entrada é uma das razões para o sucesso das CNNs em várias tarefas de visão computacional.

#### 3.2.3.2 Kernel

O kernel é uma matriz de pesos que é convoluída com a imagem de entrada durante a operação de convolução. Cada neurônio na camada de convolução possui seu próprio conjunto de pesos, representado pelo kernel, que é compartilhado em toda a imagem. Isso permite que a CNN aprenda a detectar padrões relevantes em diferentes partes da imagem (GOODFELLOW *et al.*, 2016).

#### 3.2.3.3 Camada de Pooling

Após a aplicação das camadas de convolução, é comum adicionar camadas de *pooling*. Estas camadas têm o objetivo de reduzir a dimensionalidade dos mapas de características gerados pelas camadas convolucionais, além de poderem introduzir certas invariâncias que podem robustecer o desempenho da rede (AGGARWAL, 2021).

Uma técnica comum de pooling é o Max Pooling, que mantém, para cada região, o valor máximo dentre os valores considerados. Em termos mais formais, o Max Pooling é realizado dividindo a imagem em regiões não sobrepostas e mantendo apenas o valor máximo de cada região. Outra técnica popular é o Average Pooling, que calcula a média dos valores dentro de uma região e atribui esse valor à saída correspondente. O Average Pooling é útil quando desejamos reduzir a dimensionalidade da imagem enquanto mantemos uma representação mais suave das características. Além disso, existe o Pooling Global, no qual a operação de pooling é aplicada à imagem inteira, resultando em um único valor para cada mapa de características. Esse valor representa uma síntese global das características presentes no mapa de características.

Ao adicionar camadas de pooling após as camadas de convolução, a CNN se torna capaz de aprender características mais abstratas e robustas, o que é crucial para o seu desempenho em uma variedade de tarefas de visão computacional, como reconhecimento de objetos, segmentação de imagens e detecção de características (GOODFELLOW

*et al.*, 2016).

#### 3.2.3.4 Camada Totalmente Conectada

Após a extração de características pelas camadas convolucionais e de pooling, os mapas de características resultantes são achatados em um vetor unidimensional. Este vetor é, então, passado a uma ou mais camadas totalmente conectadas (em inglês, *fully-connected*), também conhecidas como camadas densas (GOODFELLOW *et al.*, 2016).

Nas camadas totalmente conectadas, cada neurônio está conectado a todos os neurônios da camada anterior, formando uma rede densa de conexões. Essas camadas realizam operações de multiplicação de matriz entre os valores de entrada e os pesos da camada, seguidas de uma operação de ativação não linear. O resultado é a saída final da rede (AGGARWAL, 2021).

As camadas também são responsáveis por combinar as características aprendidas nas camadas convolucionais e de pooling, permitindo que a rede aprenda representações mais abstratas e complexas dos dados de entrada. Através do ajuste dos pesos durante o treinamento, as camadas totalmente conectadas são capazes de aprender a mapear os padrões de entrada para as saídas desejadas.

As camadas totalmente conectadas são frequentemente utilizadas nas partes finais das redes neurais convolucionais, antes da camada de saída, para realizar tarefas como classificação, regressão ou segmentação (AGGARWAL, 2021).

#### 3.2.3.5 Funções de Ativação

Como já exposto, após as operações de combinação linear ou convolução em cada camada, aplica-se uma função de ativação não-linear. Esse passo é vital para que a rede tenha um caráter não-linear, permitindo que ela aprenda relações mais complexas subjacentes aos dados. Uma função de ativação comumente utilizada no âmbito de redes profundas é a ReLU (Rectified Linear Unit), que retorna zero para valores negativos e o próprio valor para valores positivos (GOODFELLOW *et al.*, 2016).

Matematicamente, a função ReLU é definida como:

$$f(x) = \max(0, x)$$

,

onde $x$ é a entrada para a função. Portanto, se $x$ for positivo, a função retorna o próprio valor de $x$; caso contrário, retorna zero.

Outra função de ativação popular e historicamente importante é a função sigmoide, mapeia o argumento de entrada para o intervalo entre 0 e 1 da seguinte maneira:

$$f(x) = \frac{1}{1 + e^{-x}}$$

,

onde $e$ é a base do logaritmo natural e $x$ é a entrada para a função. A função tangente hiperbólica (tanh) é semelhante à sigmoide, possuindo a forma de um degrau suavizado, mas sua imagem é o intervalo entre -1 e 1:

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

Esses são os principais componentes de uma CNN. Através da repetição dessas camadas em profundidade, uma CNN é capaz de aprender representações hierárquicas de características nos dados de entrada, tornando-se especialmente eficaz em tarefas de visão computacional, como classificação de imagens, detecção de objetos e segmentação semântica. Passaremos, agora, à discussão de uma formulação mais recente, a qual vem recebendo enorme atenção por parte da comunidade: os *transformers*.

### 3.2.4 Transformers

Os *transformers* compõem uma classe de arquiteturas de rede neural que tiveram grande impacto em diversos domínios práticos desde sua introdução em 2017(VASWANI *et al.*, 2023), e causaram uma verdadeira revolução no processamento de linguagem natural (PLN).

A essência dos transformers reside na sua capacidade de processar sequências de entrada e saída de forma paralela e eficiente, sem depender de estruturas recorrentes como LSTMs (Long Short-Term Memory) ou GRUs (Gated Recurrent Units), que podem ser computacionalmente intensivas e têm dificuldade em capturar relações de longo alcance em sequências. Em vez disso, os Transformers utilizam mecanismos de atenção para entender a relação entre todas as partes de uma sequência, permitindo uma modelagem mais robusta e eficaz das interações entre os elementos da mesma, como é possível observar na figura 3.2 (VASWANI *et al.*, 2023).

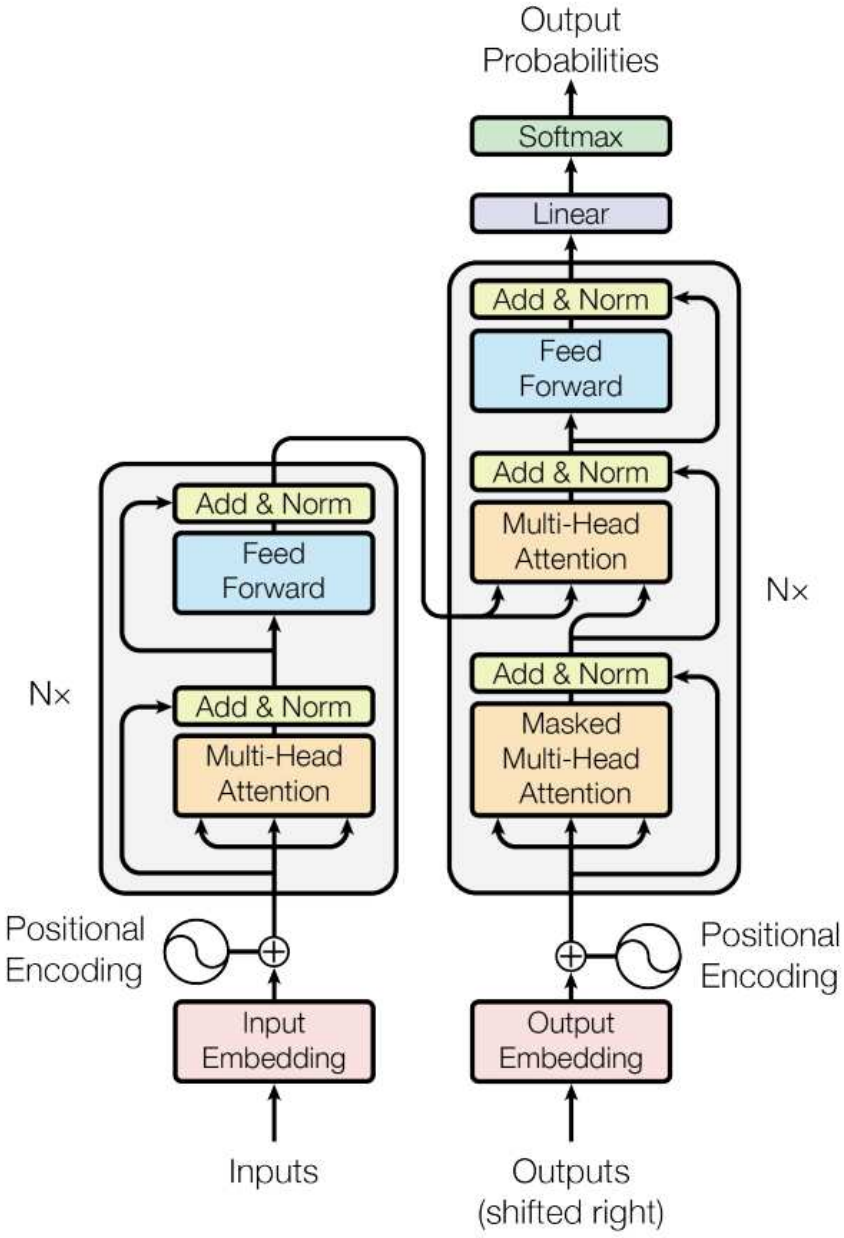



Figura 3.2 – Estrutura completa de um transformer (VASWANI *et al.*, 2023)

A estrutura essencial de um Transformer pode ser utilizada de três maneiras diferentes (LIN *et al.*, 2022):

Codificador-Decodificador: A arquitetura completa do Transformer (figura 3.2), é empregada. Isso é tipicamente usado em modelagem de sequência para sequência (por exemplo, tradução neural de máquina);

Apenas Codificador: Apenas o codificador é utilizado e as saídas do codificador são empregadas como representação para a sequência de entrada. Isso é frequentemente utilizado para tarefas de Compreensão de Linguagem Natural (NLU) (por exemplo, classificação de texto e marcação de sequência);

Apenas Decodificador: Apenas o decodificador é utilizado, onde o módulo de atenção cruzada codificador-decodificador também é removido. Isso é tipicamente usado para geração de sequência (por exemplo, modelagem de linguagem).

A parte fundamental de um bloco de Transformer é a camada de autoatenção, que permite que o modelo atribua diferentes pesos a diferentes partes da entrada, focando nas informações mais relevantes para a tarefa em questão. A autoatenção funciona calculando um conjunto de pesos para cada par de palavras na sequência de entrada, indicando a importância relativa de cada palavra em relação às outras palavras na sequência. Isso permite que o modelo capture as relações de dependência e contextualização entre as palavras em uma sequência, o que é fundamental para o entendimento semântico(LIN *et al.*, 2022).

Além da camada de autoatenção, um bloco de Transformer também pode conter camadas totalmente conectadas e camadas de normalização, como é possível observar na figura 3.2. As camadas totalmente conectadas são responsáveis por combinar as informações capturadas pela autoatenção em representações mais abstratas e compactas, enquanto as camadas de normalização ajudam a estabilizar o treinamento do modelo, garantindo que as ativações da rede permaneçam dentro de uma faixa razoável durante o treinamento (VASWANI *et al.*, 2023).

A arquitetura em cascata dos blocos de codificação e decodificação permite que os Transformers capturem informações em diferentes níveis de abstração e contextos temporais, fornecendo uma representação detalhada e informada das sequências de entrada. Além disso, a capacidade dos Transformers de processar sequências de entrada e saída em paralelo os torna altamente eficientes e escaláveis, permitindo que sejam aplicados a conjuntos de dados de grande escala com facilidade. (VASWANI *et al.*, 2023)(LIN *et al.*, 2022)

Um dos exemplos mais conhecidos de uso prático de Transformers é o modelo GPT (do inglês *Generative Pre-Trained Transformer*). Os modelos GPT são treinados em grandes quantidades de texto de forma não supervisionada, aprendendo a prever a próxima palavra em uma sequência dada uma série de tokens anteriores. Depois de treinados, os modelos GPT podem ser ajustados para tarefas específicas de PLN, como tradução automática, sumarização de texto e geração de texto, por exemplo (RAY, 2023).

Outro exemplo é o BERT (Bidirectional Encoder Representations from Transformers). Este modelo é treinado em grandes *corpora* de texto de forma supervisionada, aprendendo a prever palavras ocultas em uma sequência dadas as palavras circundantes. O BERT alcançou resultados de estado da arte em uma ampla gama de tarefas de PLN, incluindo classificação de texto, análise de sentimentos, perguntas e respostas, entre outras

(Devlin *et al.*, 2018).

Além dos exemplos mencionados, os Transformers também são amplamente utilizados em aplicações de reconhecimento de fala e processamento de áudio. Modelos como o Wav2Vec2 usam arquiteturas baseadas em Transformers para extrair recursos de áudio de forma eficiente, permitindo a transcrição precisa de fala em texto e uma série de outras aplicações de processamento de áudio (BAEVSKI *et al.*, 2020).

Em resumo, os Transformers representam uma evolução significativa no campo da inteligência artificial e do processamento de linguagem natural. Sua capacidade de capturar relações complexas em dados sequenciais de forma eficiente e escalável os torna uma ferramenta indispensável na atualidade para uma variedade de aplicações práticas, desde a tradução automática até a geração de música. A estrutura modular e em cascata dos blocos de codificação e decodificação permite que os Transformers capturem informações em diferentes níveis de abstração e contextos temporais, fornecendo uma representação detalhada e informada das sequências de entrada (BILAN *et al.*, 2020).

Essa habilidade de transgredir por entre os níveis de abstração está diretamente interligada ao aprendizado profundo, o que faz dos transformers ferramentas importantes para a área. Na próxima seção, será discutido o sentido filosófico e ontológico deste conceito.

## 3.3 Aprendizado Profundo

A constituição semântica dos dados, fornecida pelo logos, é a base pela qual a inteligência se desenvolve e, segundo Aristóteles, é o que diferencia os humanos dos animais(ARISTóTELES, 1998). Essa afirmação nos leva a uma jornada pelo mundo da semiótica e da metafísica do pensamento, explorando não apenas a mecânica das redes neurais, mas também suas implicações mais profundas sobre a natureza do conhecimento e da realidade. A capacidade única dos seres humanos de raciocinar e usar a linguagem para expressar pensamentos complexos, o que Aristóteles chamou de 'zoon logon echon', é crucial nesse contexto, pois é ela que nos permite transcender os limites da experiência sensorial e adentrar o reino do abstrato e do simbólico.

Na semiótica, o estudo dos signos e símbolos nos ajuda a entender como os dados são interpretados e processados pela mente humana na medida em que este se comunica. Por extensão, essa lógica é aplicada aqui para questões de interpretabilidade das redes neurais. Os dados não são apenas pontos de dados isolados, mas sim elementos que carregam significado e contexto(SANTAELLA, 2003). Analogamente, a morfologia dos dados representa a forma física dos símbolos, enquanto a sintaxe reflete as regras que

governam sua organização e estrutura. Finalmente, a semântica mergulha na essência do significado subjacente aos dados, relacionando-os com conceitos e ideias mais abstratas.

Na metafísica do pensamento, aprofundamos ainda mais nossa investigação sobre a natureza dos dados e da inteligência. O logos, nesse contexto, representa não apenas a lógica ou razão, mas também a ordem subjacente ao universo. É essa ordem que permeia os dados e os torna inteligíveis(ARISTóTELES, 1998). O aprendizado profundo, então, é uma jornada em direção à compreensão dessa ordem mais profunda, uma busca pela verdade subjacente aos dados e à realidade que eles representam.

Ao longo dessa jornada, encontramos a ideia de que a profundidade da compreensão dos dados está diretamente ligada à capacidade de processamento computacional e ao número de amostras disponíveis (GOODFELLOW *et al.*, 2016). Isso nos leva a refletir sobre as implicações filosóficas dessas limitações. Ao focar no caso dos dados, o aprendizado profundo depende diretamente da forma de captação e transmissão dos mesmos para produzir qualquer tipo de análise.

A proliferação exponencial de dados acessíveis só se tornou possível quando começamos a reproduzir a realidade. O surgimento de dispositivos como o gramofone, a câmera fotográfica, a filmadora, a gravação elétrica e uma variedade de sensores capacitados para capturar, armazenar e reproduzir sinais, foi fundamental para a emergência da reprodutibilidade técnica da arte (BENJAMIN, 2000), e consequentemente, da realidade. A revolução promovida pela tecnologia digital desempenhou um papel crucial no avanço dessa reprodutibilidade técnica, especialmente no contexto do hardware e software, possibilitando o processamento de alto desempenho, o qual é essencial para técnicas de gravação de áudio, fotografia, filmagem, edição de vídeo, entre outros.

Contudo, além da mera reprodução técnica, a transposição da aura da experiência humana, mediada pela tecnologia digital e pela sensibilidade do indivíduo que a manipula, deve ser contemplada dentro do universo dos dados. O ato de ouvir uma gravação de voz difere substancialmente da experiência de ouvir a voz sem a intermediação de um microfone; da mesma forma, visualizar uma fotografia de uma paisagem é uma experiência distinta de estar presente no exato momento em que a imagem foi capturada. Os dados, em sua maioria, são mediados pela percepção e perspectiva tanto do indivíduo quanto das máquinas envolvidas no processo de captura e reprodução.

Essa mediação da realidade através da tecnologia introduz nuances significativas na interpretação e compreensão dos dados. As interpretações e significados atribuídos aos dados são inevitavelmente influenciados pela subjetividade humana e pelas limitações técnicas dos dispositivos de captura e reprodução. Portanto, ao lidar com dados, é crucial ter em mente essa mediação, reconhecendo que os dados não são uma representação

puramente objetiva da realidade, mas sim uma construção complexa e multifacetada que reflete tanto a realidade quanto as percepções e interpretações humanas.

Portanto, ao considerar as redes neurais e o aprendizado profundo, não devemos nos limitar apenas às questões técnicas e práticas. Devemos também explorar as implicações mais amplas dessas tecnologias para nossa compreensão do conhecimento, da realidade e de nós mesmos. É somente ao fazer isso que podemos verdadeiramente apreciar o poder e o potencial das redes neurais como ferramentas não apenas para processar dados, mas também para expandir nossas mentes e nossa compreensão do universo.

# 4 O problema e a Metodologia

## 4.1 Vozes Subversivas

A partir das reflexões sobre a binariedade de gênero obtidas no estudo "Vozes Subversivas"(NEVES; MACHADO, 2023), no qual se propôs a compreender as vozes das intérpretes Pabllo Vittar e Cássia Eller, objetivou-se classificar vozes faladas em relação ao gênero social, buscando entender as fronteiras entre as vozes "masculinas", "femininas"e, principalmente, as que estão no limiar entre uma classe e outra. Essas vozes abjetas, que não se encaixam no padrão normativizado pela matriz da heterossexualidade compulsória (BUTLER, 2015), são intituladas aqui como subversivas por conta de suas próprias condições: resistem à um sistema binarista de gênero que sistematicamente silencia vivências que não se encaixam no padrão homem/mulher.

Ao considerarmos a Ontologia Vocálica da Unicidade, elaborada por Adriana Cavarero (CAVARERO, 2011), somos convidados a valorizar e celebrar a diversidade de vozes que compõem o espectro de gênero humano, reconhecendo a sua complexidade e singularidade. Segundo essa perspectiva, cada voz é única e singular, carregando consigo uma existência e uma ímpar identidade encarnada. Ao aplicar essa abordagem ao estudo das vozes subversivas, somos levados a reconhecer não apenas a binariedade de gênero, mas também a multiplicidade de expressões vocais que desafiam essa dicotomia. Essas vozes, ao resistirem à categorização estrita como "masculinas"ou "femininas", revelam a riqueza e a fluidez das identidades de gênero, desafiando a imposição normativa da heterossexualidade compulsória.

A partir dessas reflexões, buscou-se aqui criar um classificador de voz que fosse capaz de delimitar três grupos diferentres: ***homem***, ***mulher*** e ***outro***. Nesse caso, ***homem*** e ***mulher*** são cisgênero; então a categoria ***outro*** se refere ao grupo das pessoas transgêneras. É importante salientar que não foi encontrado na literatura científica um problema de classificação tal como esse, sendo, então, esta uma contribuição inovadora associada ao presente trabalho.

## 4.2 Metodologia

A presente pesquisa possui uma amálgama qualitativa e quantitativa em sua abordagem. A primeira parte refere-se aos estudos de gênero e as possíveis motivações sociais para as vozes produzirem os ajustes fisiológicos que são manifestados; a segunda

parte se refere aos correlatos acústico-vocais, fisiológicos e às metodologias baseadas em aprendizado profundo e processamento digital de sinais.

Fazendo uso parcial do modelo Wav2Vec2 (BAEVSKI *et al.*, 2020), objetivou-se a criação um classificador que fosse capaz de separar o banco de dados em três grupos: ***homem***, ***mulher*** e ***outro***. Na figura 4.1, ilustra-se como o Wav2Vec2 funciona na transformação de fala para texto, função para a qual o modelo foi desenvolvido.

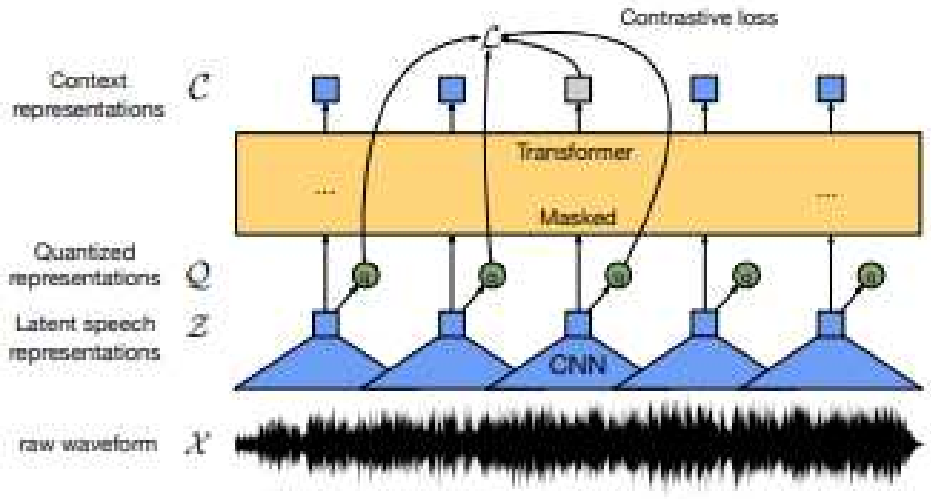



Figura 4.1 – Estrutura do modelo Wav2Vec2 (BAEVSKI *et al.*, 2020)

Na figura 4.1 é possível notar que o modelo recebe o sinal de áudio bruto (sem extração de atributos), processa-o usando uma rede neural convolucional densa, que cria um campo de representações latentes da fala, e o envia para o transformer, que dá como saída as representações contextuais e as compara com as representações quantizadas, calculando a perda contrastiva. Essa escolha foi feita para realizar a transferência de aprendizado do Wav2Vec2, que foi treinado para fala, então é esperado que ele tenha absorvido padrões de gênero social, onde teve que reconhecer e diferenciar vozes graves e agudas para entender o significante (palavras) das sentenças.

Neste trabalho, foi retirada a parte da máscara (transformer) da estrutura, e, em seu lugar, foi utilizada uma rede MLP feed-foward para classificar os dados, utilizando retropropagação como base de minimização dos erros, baseada no gradiente estocástico descendente (SGD). Nos pontos de gradação entre uma classe e outra pretendeu-se compreender o que foi definido nesse trabalho como "vozes subversivas", ou seja, vozes que não são inteligíveis para a matriz da heterossexualidade compulsória (BUTLER, 2015). Após o treinamento, validação e teste, foram extraídos os atributos relevantes para a rede aproximar ou distanciar os dados entre si. Todo o processo analítico *a posteriori* foi rea-

lizado de forma supervisionada. Os bancos de dados que serão utilizados nesse trabalho estão em inglês e são:

- LibreSpeech, o qual foi utilizado para treinar o modelo Wav2Vec2 a priori; (PANAYOTOV *et al.*, 2015);
- Mozilla Common Voice Dataset, com gravações de voz de frases prontas. (ARDILA *et al.*, 2019);

### 4.2.1 Dados Utilizados

Os dados foram processados para uniformizar a frequência de amostragem em 16KHz para todas as amostras. Abaixo, segue a quantidade de amostras em cada conjunto:

- **Quantidade de dados no conjunto de treinamento:** 367507
- **Quantidade de dados no conjunto de validação:** 108836
- **Quantidade de dados no conjunto de teste:** 4628

### 4.2.2 Pseudocódigo

Abaixo, segue o pseudocódigo com cada passo dado pelo algoritmo.

1. **Importe todas as bibliotecas necessárias**: Isso inclui bibliotecas como PyTorch, scikit-learn, etc., que oferecem suporte a operações matemáticas, carregamento de dados, otimização e avaliação de modelos.
2. **Defina a classe `Wav2Vec2Config`**: Esta classe é responsável por definir os hiperparâmetros e configurações do modelo Wav2Vec2, como o tamanho do vocabulário, o número de camadas ocultas, o tamanho da camada oculta, o número de cabeças de atenção, etc. Esses hiperparâmetros afetam a arquitetura e o comportamento do modelo.
3. **Defina a classe `COMMONVOICE`**: Esta classe é responsável por carregar os dados do conjunto de dados CommonVoice. Ele lê os metadados do arquivo TSV (tab-separated values), carrega os arquivos de áudio correspondentes e retorna os pares de amostras de áudio e seus rótulos.
4. **Crie instâncias dos conjuntos de dados**: Aqui, instanciamos objetos dos conjuntos de dados de treinamento, validação e teste usando a classe `COMMONVOICE`, fornecendo o caminho para os arquivos de áudio e os arquivos TSV correspondentes.

5. **Defina a classe do modelo principal `FrankModel`**: Esta classe define a arquitetura do modelo. Ele inclui a extração de características do Wav2Vec2, seguida por camadas lineares adicionais para processar as características extraídas e fazer previsões. A arquitetura do modelo é definida em termos de operações matemáticas, como convoluções, linearizações e ativações.

6. **Inicialize o otimizador e a função de perda**: O otimizador (como SGD ou Adam) é usado para ajustar os pesos do modelo durante o treinamento, enquanto a função de perda (como entropia cruzada) é usada para calcular a discrepância entre as previsões do modelo e os rótulos verdadeiros.

7. **Defina o número de épocas e o parâmetro de acumulação de lote**: As épocas indicam quantas vezes o modelo passará por todo o conjunto de treinamento durante o treinamento. O parâmetro de acumulação de lote controla quantos gradientes são acumulados antes de atualizar os pesos do modelo, o que pode ajudar a estabilizar o treinamento em GPUs com memória limitada.

8. **Inicialize o melhor valor de perda de validação**: Este valor é usado para comparar a perda de validação atual com a melhor perda de validação vista até agora. Se a perda atual for menor, o modelo atual é salvo como o melhor modelo até agora.

9. **Inicie o loop de treinamento e o loop de validação para cada época**: Os loops de treinamento e validação iteram sobre os dados de treinamento e validação, respectivamente, durante cada época.

10. **Dentro do loop de treinamento**: Para cada lote de dados, o modelo faz uma previsão, calcula a perda, calcula os gradientes usando a retropropagação e atualiza os pesos do modelo usando o otimizador.

11. **No loop de validação**: Para cada lote de dados de validação, o modelo faz uma previsão sem calcular gradientes, apenas para avaliar o desempenho do modelo. A perda e a precisão são calculadas para monitorar o desempenho do modelo.

12. **Se a perda de validação atual for menor que a melhor perda de validação**: O modelo atual é salvo como o melhor modelo até agora, para ser usado posteriormente para inferência.

13. **Exiba a perda de validação e precisão após cada época**: Isso fornece feedback sobre o desempenho do modelo durante o treinamento.

14. **Compute a matriz de confusão para os dados de validação**: A matriz de confusão mostra o número de previsões corretas e incorretas para cada classe, permitindo uma análise mais detalhada do desempenho do modelo.

15. **No loop de teste**: O modelo é avaliado nos dados de teste da mesma maneira que nos dados de validação, calculando a perda e a precisão.

16. **Compute a matriz de confusão para os dados de teste**: Isso fornece uma avaliação final do desempenho do modelo em dados não vistos.

17. **Exiba a perda e a precisão do teste**: Isso fornece uma medida final do desempenho do modelo em um conjunto de dados independente.

# 5 Aplicação do Modelo e Análise dos Resultados

Neste capítulo, serão apresentados e discutidos os testes experimentais e os resultados obtidos com a aplicação da metodologia discutida no capítulo anterior.

## 5.1 Hiperparâmetros Utilizados

- **Função de perda:** Entropia Cruzada (*Cross Entropy Loss*)

  A função de perda baseada na entropia cruzada é amplamente utilizada em problemas de classificação, especialmente quando se lida com múltiplas classes. Esta função mede a divergência entre as probabilidades previstas pelo modelo e as classes reais, penalizando previsões incorretas de maneira significativa. Ao minimizar a entropia cruzada, garantimos que o modelo aprenda a prever as classes corretas com maior precisão.

- **Otimizador:** Stochastic Gradient Descent (SGD)

  O Stochastic Gradient Descent (SGD) é um dos algoritmos de otimização mais utilizados para treinamento de modelos de aprendizado de máquina. Ele atualiza os pesos do modelo iterativamente usando pequenas amostras do conjunto de dados (mini-batches) a cada iteração, o que estabelece um compromisso entre complexidade e desempenho de busca.

  - **Taxa de aprendizado (lr):** 0.001

    A taxa de aprendizado deve ser escolhida de modo a permitir que se obtenha uma adequada velocidade de convergência sem que se incorra em instabilidade ou em excessivas flutuações estocásticas. Após testes preliminares, chegou-se a uma taxa de 0.001.

- **Número de épocas:** 7

  O número de épocas indica quantas vezes o algoritmo percorre todo o conjunto de dados de treinamento. Este número foi escolhido para equilibrar o tempo computacional de treinamento e o nível de desempenho atingível. Chegou-se a um total de sete épocas, um valor relativamente pequeno, mas que permitiu a execução de testes de duração factível e a obtenção de resultados iniciais adequados.

- **Batch accumulation parameter:** 3

  O parâmetro de acumulação de batches (*batch accumulation parameter*) de 3 é usado para simular um tamanho de batch maior, acumulando gradientes ao longo de várias iterações antes de atualizar os pesos do modelo. Isso é útil em cenários onde a capacidade de memória é limitada, permitindo que o modelo aproveite os benefícios de um tamanho de batch maior, como a suavização dos gradientes, sem exigir uma grande quantidade de memória por iteração.

Esses hiperparâmetros foram selecionados para otimizar o processo de treinamento, balanceando a precisão e a eficiência do modelo, e garantindo que ele aprenda os padrões dos dados de maneira eficaz sem overfitting.

## 5.2 Processo de Treinamento

O processo de treinamento foi realizado em ciclos (épocas), em que cada iteração consistiu em:

1. **Forward pass:** As entradas são passadas pelo modelo para obter as predições.
2. **Cálculo da perda:** A perda entre as predições e os rótulos reais é calculada utilizando a função de perda CrossEntropyLoss.
3. **Backward pass:** A perda é retropropagada para atualizar os pesos do modelo.
4. **Atualização de pesos:** A cada acumulação de 3 batches, os pesos do modelo são atualizados pelo otimizador.

## 5.3 Avaliação e Resultados

Durante a fase de avaliação, o modelo foi testado no conjunto de validação, calculando a perda de validação e a acurácia:

- **Perda de Validação:** A média da perda calculada no conjunto de validação.
- **Acurácia de Validação:** A proporção de predições corretas no conjunto de validação.

O modelo com a menor perda de validação foi salvo como o melhor modelo.

## 5.4 Matriz de Confusão

Uma matriz de confusão é uma tabela que permite a visualização do desempenho de um algoritmo de classificação. Para classificação binária, cada coluna da matriz representa as instâncias de uma classe prevista, enquanto cada linha representa as instâncias de uma classe real (ou verdadeira).

No caso do presente trabalho, a matriz de confusão é multiclasse, com três rótulos distintos. A tabela construída permite a visualização do desempenho de um algoritmo de classificação ao disponibilizar uma visualização com múltiplas classes. Ao invés de ter apenas duas dimensões (positiva e negativa), a matriz terá uma dimensão para cada classe, tanto nas linhas quanto nas colunas. Levando em consideração os rótulos utilizados-***homem***, ***mulher*** e ***outro*** - a matriz de confusão possui a seguinte estrutura:

- **Acertos:** A diagonal principal da matriz (Homem-Homem, Mulher-Mulher, Outro-Outro) indica o número de instâncias corretamente classificadas para cada classe.
- **Erros:** As células fora da diagonal principal indicam erros de classificação. Por exemplo, a célula Homem-Mulher mostra quantas instâncias da classe homem foram incorretamente classificadas como mulher, e a célula Mulher-Outro mostra quantas instâncias da classe mulher foram incorretamente classificadas como outro.

A matriz de confusão foi utilizada para avaliar a performance do modelo nos conjuntos de treinamento, teste e validação do presente trabalho. À direita, nas figuras 5.1, 5.2 e 5.3, está disponível uma escala de cores: quanto mais próximo se estiver do azul escuro, maior será a densidade de dados.

É possível observar, na figura 5.1 [1], que o modelo conseguiu separar bem as classes entre ***homem*** e ***mulher*** ; entretanto, como existe a classe ***outro***, ocorreram muitos erros entre todas as classes, principalmente nas células Homem-Mulher e Mulher-Homem.

[1] A matriz de confusão contou todos os dados das 7 épocas de treinamento, o que não é muito usual, mas mostrou quantas vezes obteve-se o acerto na classe outro.

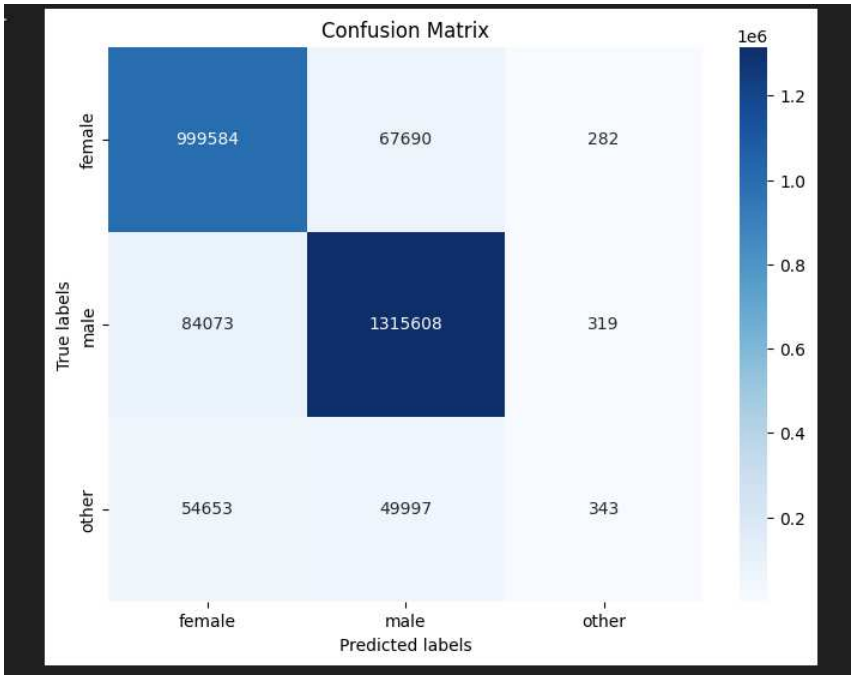



Figura 5.1 – Matriz de Confusão das 7 épocas de treinamento

Já na figura 5.2, que representa a etapa de validação, o caso mais expressivo foi confundir a classe ***outro*** com a ***mulheres*** (célula Outro-Mulher), além de nenhum acerto na classe ***outro***. Apesar disso, manteve-se o cenário de clasificação mais precisa para as classes ***homem*** e ***mulher***.

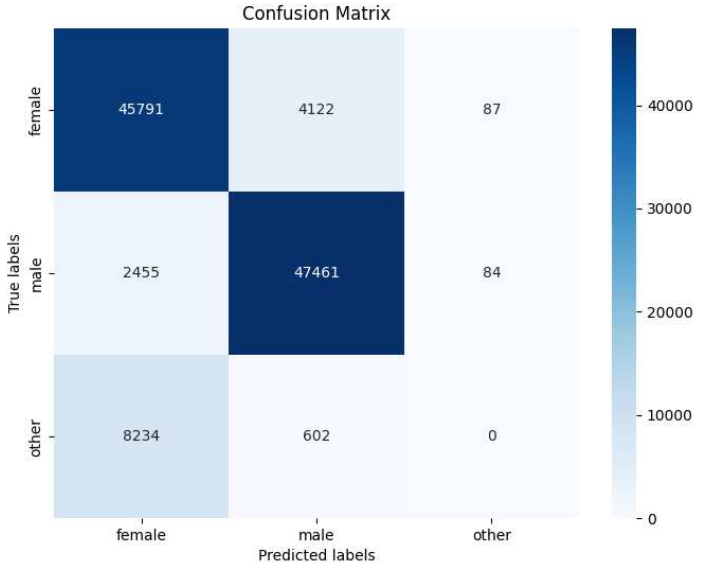



Figura 5.2 – Matriz de Confusão do conjunto de validação

Na etapa de teste do modelo, expresso na figura 5.3, houve maior confusão dos rótulos ***homem*** com ***mulher*** e vice versa (células Homem-Mulher e Mulher-Homem), como durante o treinamento; além disso, não houve acerto na classe ***outro*** (célula Outro-

Outro).

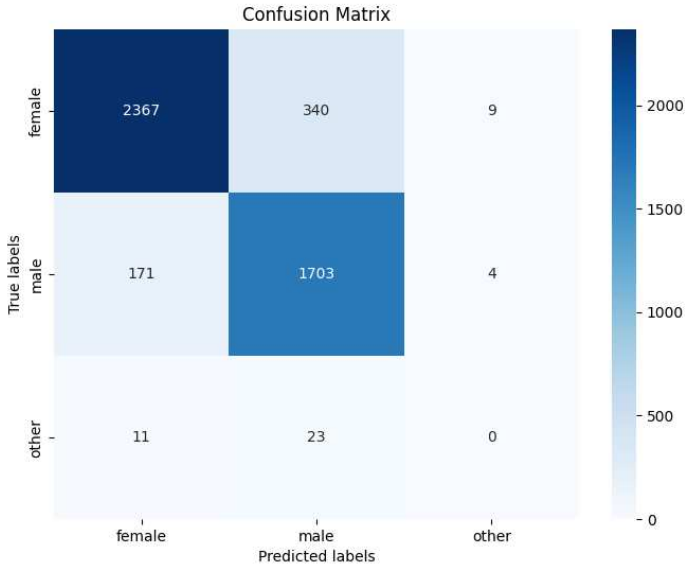



Figura 5.3 – Matriz de Confusão do conjunto de teste

## 5.5 Curva ROC

A curva ROC (do inglês *Receiver Operating Characteristic*) foi utilizada para avaliar a capacidade do modelo de distinguir entre as classes. A curva ROC é um gráfico que ilustra o desempenho de um classificador binário conforme seu limiar de decisão é alterado. Cada ponto na curva representa um par de valores (taxa de verdadeiros positivos, taxa de falsos positivos) para um determinado limiar.

Para problemas multiclasse, como neste caso, foram geradas curvas ROC individuais para cada classe contra todas as outras (*one-vs-rest*). A área sob a curva ROC (AUC-ROC) foi calculada para quantificar a capacidade discriminatória do modelo. Um AUC de 1 indica um classificador perfeito, enquanto um AUC de 0.5 indica um classificador sem poder discriminatório.

As curvas ROC das etapas de treinamento (figura 5.4) e validação (figura 5.5) foram de extrema importância para o desenvolvimento do treinamento do modelo. Na figura 5.4 obteve-se o melhor resultado da classe ***outro*** (AUC= 0.71), além de uma boa resposta para as classes ***homem*** e ***mulher***. Na figura 5.5, por outro lado, verifica-se que o classificador não conseguiu realizar a discriminação da classe ***outro*** de maneira adequada.

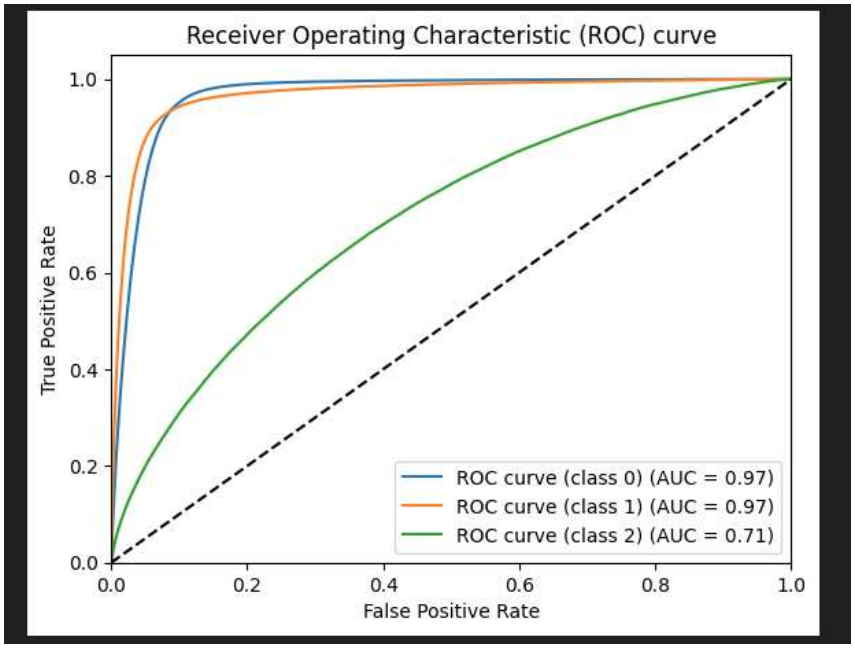



Figura 5.4 – Curva ROC do treinamento

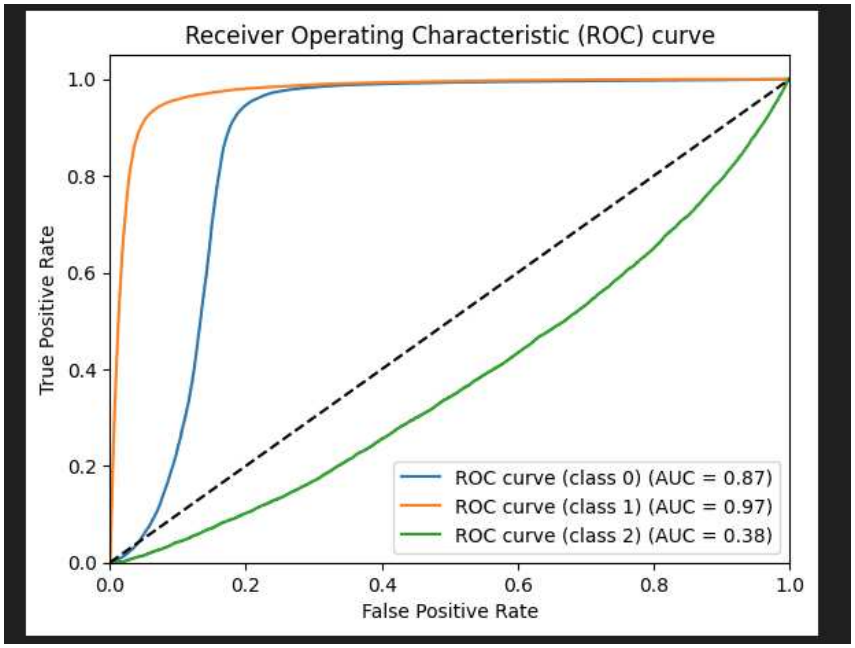



Figura 5.5 – Curva ROC da validação

Já na figura 5.6 (teste), com o reduzido campo amostral (4628 amostras), não foram obtidos bons resultados para a classe ***outro*** , mas, para as classes ***homem*** e ***mulher***, obteve-se boa precisão. Estes dados podem indicar que mesmo com pouco dados no conjunto de teste, é possível obter boa precisão para os casos binários de gênero. Por outro lado, o modelo não conseguiu resultados satisfatórios com a classe ***outro***, composta por vozes não binárias.

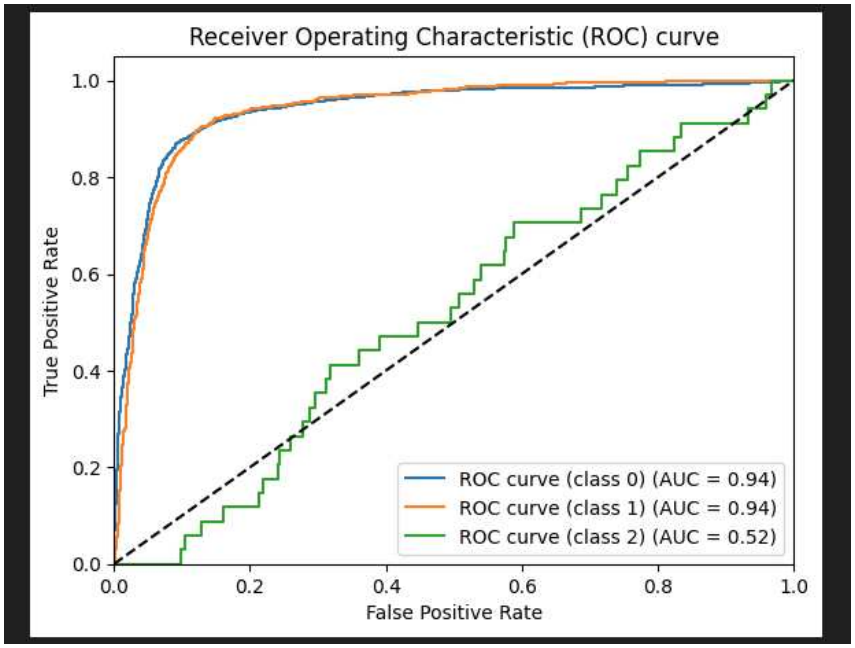



Figura 5.6 – Curva ROC do teste

## 5.6 Análise de Redução de Dimensionalidade com t-SNE

Para visualizar a distribuição das representações internas do modelo, utilizamos o t-SNE (*t-Distributed Stochastic Neighbor Embedding*), uma técnica de redução de dimensionalidade que é particularmente útil para visualização de dados de alta dimensão. O t-SNE projeta os dados em um espaço de duas ou três dimensões, mantendo a estrutura das vizinhanças locais.

A seguir, apresentamos a visualização dos dados no espaço bidimensional, onde cada ponto representa uma amostra e a cor indica sua classe verdadeira.

É possível observar, na figura 5.7 (etapa de treinamento), que as classes ***homem*** e ***mulher*** ficaram visivelmente separadas, que são representadas pelas cores roxo e verde, respectivamente; e a classe ***outro*** (representada pelos pontos amarelos) está centralizada e sobreposta entre as duas classes, o que corrobora com a ideia de neutralidade de gênero nesta classe e ajuda a compreender a dificuldade de classificação verificada. Observa-se, também, que há um acumulo maior de dados da classe ***outro*** na borda superior da imagem 5.7, o que sugere que a componente 2 influencia a percepção de gênero neutro. Essa etapa foi essencial para confirmar que o problema de classificação era factível.

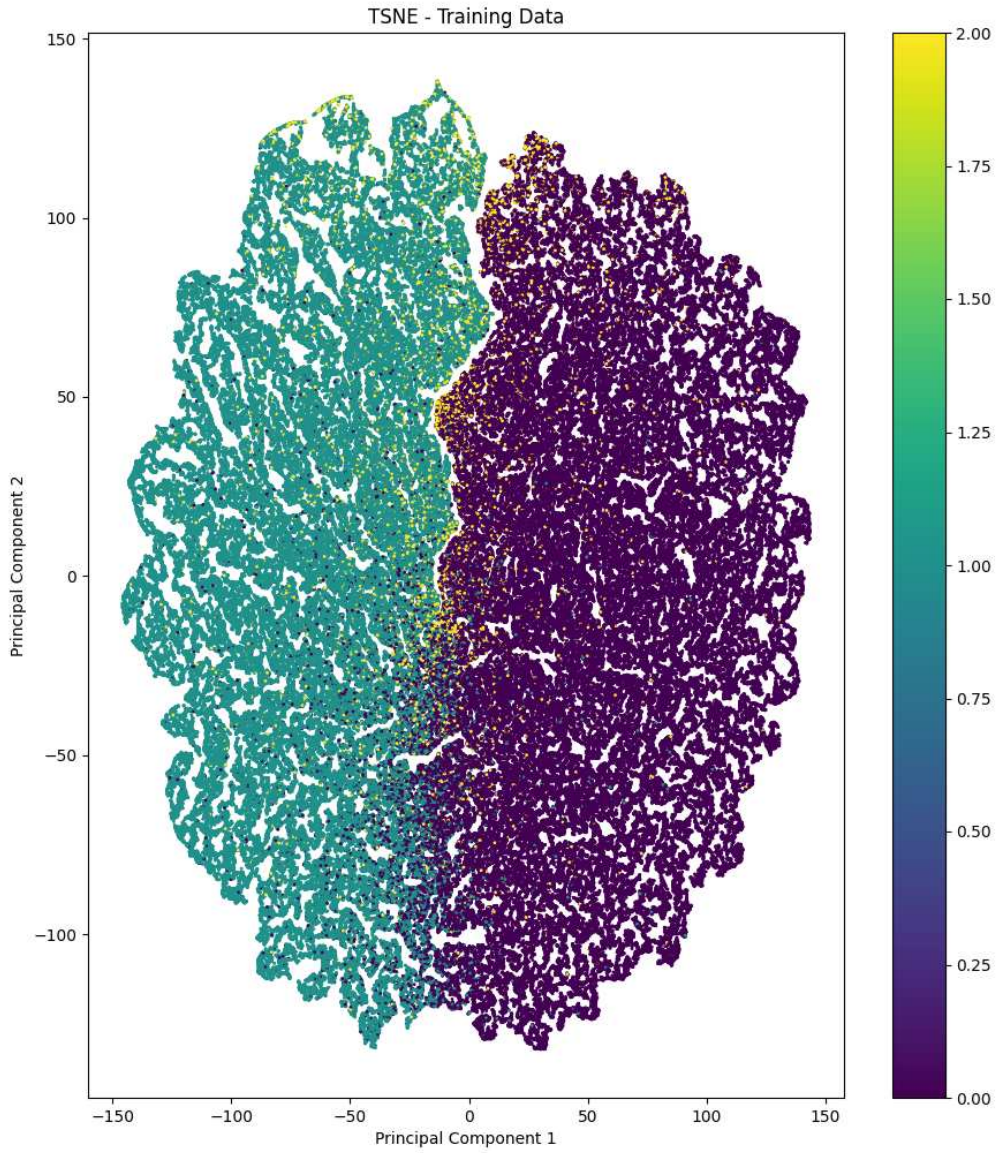



Figura 5.7 – Visualização t-SNE dos dados de treinamento

Na figura 5.8, que representa os dados da etapa de validação, as amostras da classe ***outro*** ficaram no extremo da classe das mulheres. Para tal comportamento, mantém-se a hipótese de que as amostras da classe ***outro*** estão em um espectro mais "feminino"da voz. As classes ***homem*** e ***mulher*** ficaram visivelmente separadas, assim como nas figuras 5.7 (etapa de treinamento) e 5.9 (etapa de teste), que será explicada a seguir.

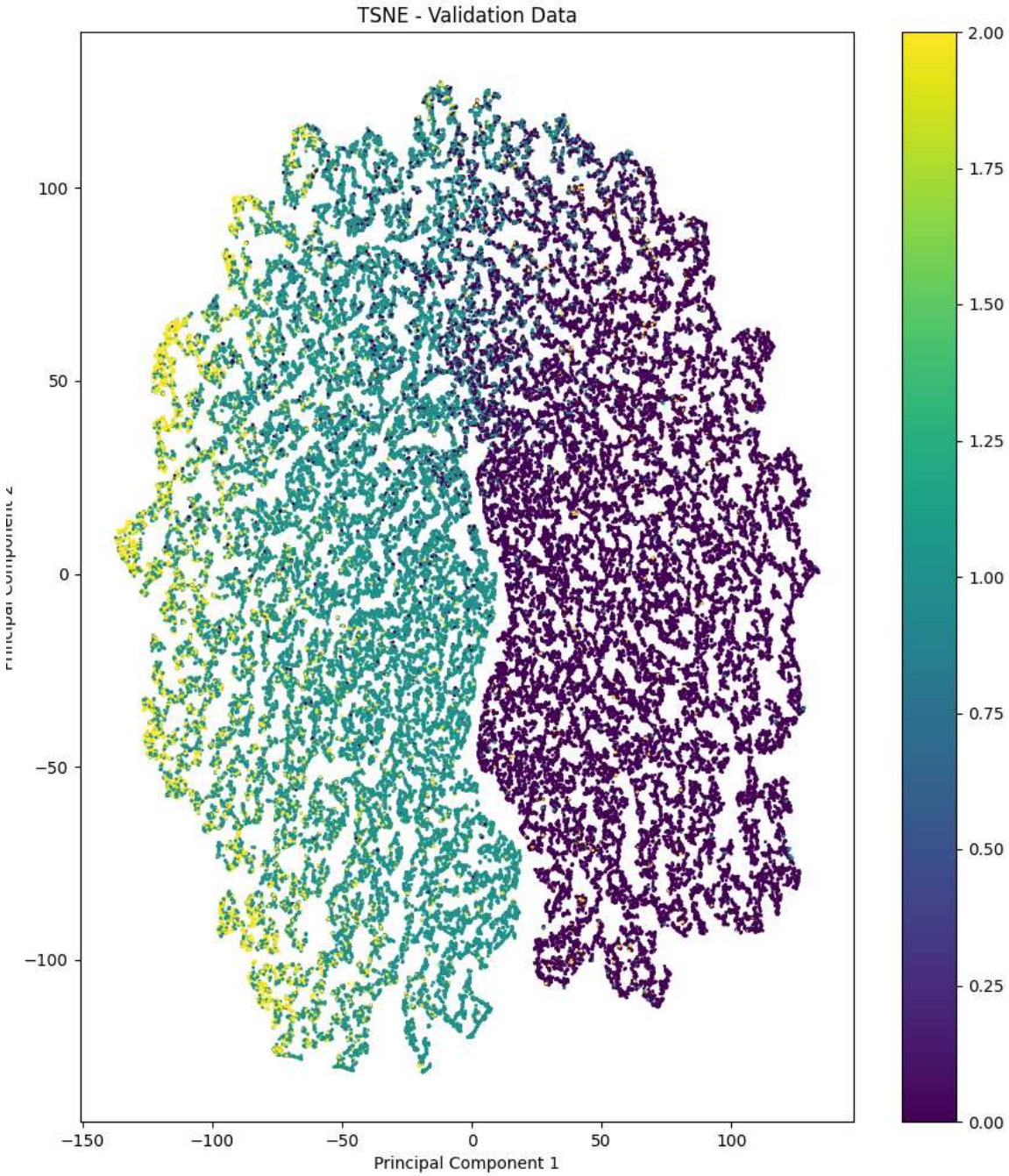



Figura 5.8 – Visualização t-SNE dos dados de validação

Na figura 5.9 há uma quantidade menor de dados, o que é possível de visualizar nos espaços em branco entre os dados. Mesmo assim, é identificável um padrão de separação entre as classes ***homem*** e ***mulher***; além de uma dificuldade em incorporar a classe ***outro*** por conta da baixa densidade de amostras.

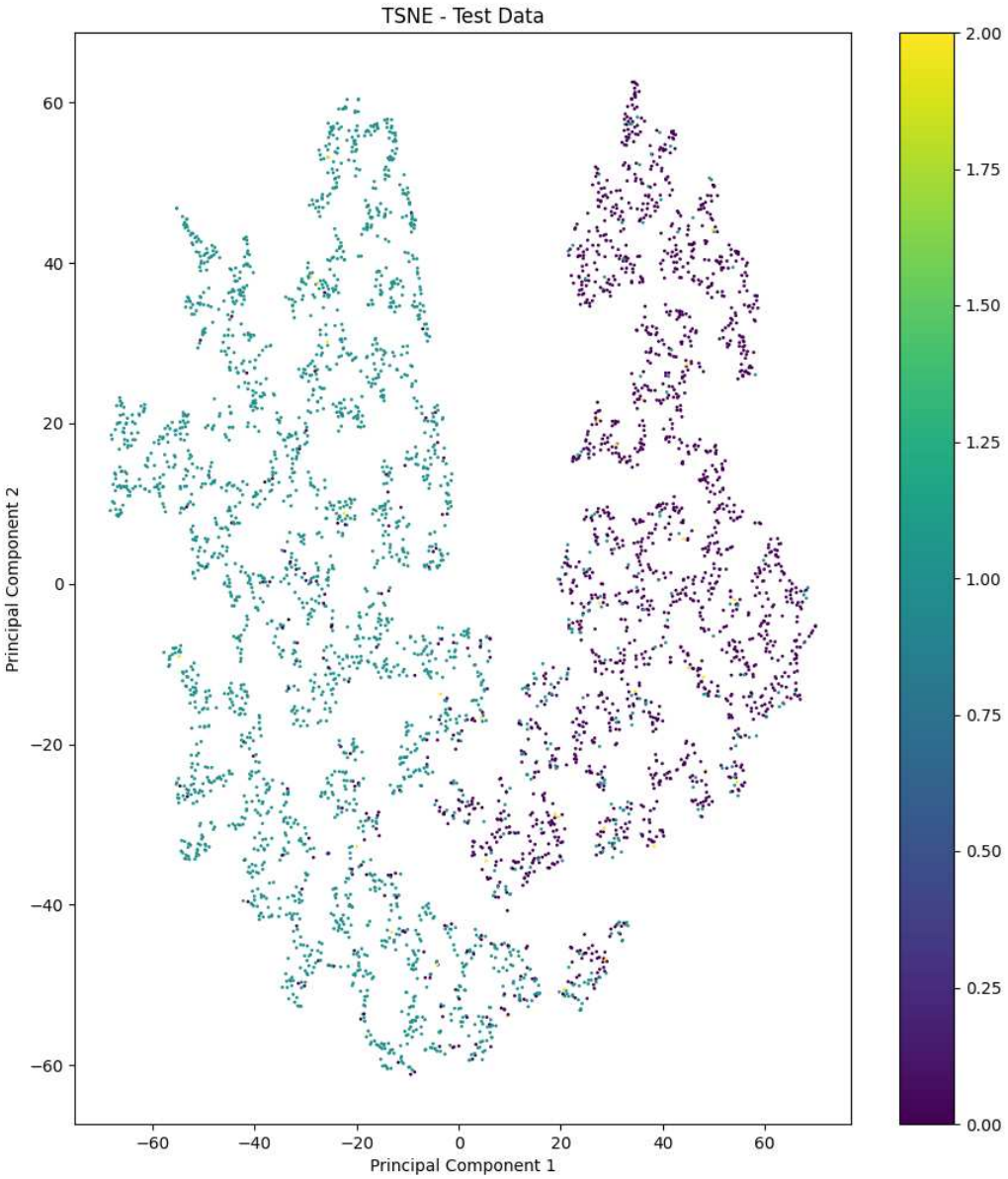



Figura 5.9 – Visualização t-SNE dos dados de teste

# 6 Conclusão

Neste trabalho, realizou-se uma investigação do problema de classificação de gênero social a partir de padrões de voz, abrangendo tanto referenciais teóricos quanto um estudo experimental preliminar baseado em aprendizado de máquina. Não obstante o caráter inicial do estudo, foi possível chegar a conclusões que nos parecem relevantes para a compreensão do problema abordado em diferentes dimensões.

Verificou-se que o modelo baseado no Wav2Vec2 se mostrou eficaz para a tarefa de classificação de vozes, principalmente no que diz respeito a vozes binárias (homem e mulher). No entanto, para o rótulo 'outro', parece-nos necessário realizar uma análise mais aprofundada do desempenho de classificação atingível e buscar a análise de um conjunto mais amplo de arquiteturas e dados.

Por outro lado, a análise das curvas ROC forneceu insights sobre a capacidade do modelo em distinguir entre as diferentes classes, e sobre o caráter peculiar da classificação da classe 'outro' em comparação com as outras duas.

A visualização com t-SNE permitiu uma compreensão mais intuitiva de como o modelo separa as diferentes classes em um espaço de características visualizável, o que, principalmente no gráfico do treinamento, mostrou a gradação entre as classes e a aproximação entre os dados da classe "outro", o que é coerente com a questão central do trabalho, as existência das "vozes subversivas"citadas no começo do trabalho, as quais representam toda uma ontologia vocal não binária. Cada voz presente neste trabalho é um som vivo e encarnado, a compreensão de suas realidades é um passo importante para o mapeamento da existência de pessoas que não estão coerentes com a binariedade de gênero, tema que é extremamente relevante para a fonoaudiologia e ciências humanas.

## 6.1 Perspectivas

O presente trabalho traz várias perspectivas de trabalhos futuros:

- **Análise do Potencial de Classificação:** Realizar uma investigação estatística aprofundada sobre o potencial de discriminação das classes de voz, com especial atenção à classe 'outro'.
- **Aumento do Conjunto de Dados:** Buscar a inclusão de mais dados para treinamento, validação e teste, de modo a permitir a construção de modelos mais robustos.

- **Experimentação com Arquiteturas Diferentes:** Testar um maior leque de arquiteturas, abrangendo modelos de diversas inspirações e graus de flexibilidade. Também parece interessante a ideia de testar modelos nebulosos (*fuzzy*) de modo a incorporar pertinências não-binárias e um maior escopo semântico.

- **Filosofia do uso e construção de inteligência artificial:** Discutir e aprofundar o conceito da "reprodutibilidade técnica da realidade"e das capacidades de processamento dos modelos em um artigo científico.

- **Criação de banco de dados:** Gravação de uma base de vozes em português para teste deste modelo e mesmo para abertura à comunidade acadêmica para desenvolvimento da área de pesquisa.

# Referências

AGARWAL, A. R.; TIWARI, S.; PATAGE, V. V.; S, S. G.; S, S. M. A method for voice activity detection using k-means clustering. In: *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. [S.l.: s.n.], 2022. p. 1–5. Citado 2 vezes nas páginas 13 e 16.

AGGARWAL, C. C. *Neural Networks and Deep Learning: A Textbook*. New York, NY: Springer, 2021. Citado 7 vezes nas páginas 9, 16, 23, 24, 25, 26 e 27.

AL-DHIEF, F. T.; BAKI, M. M.; LATIFF, N. M. A.; MALIK, N. N. N. A.; SALIM, N. S.; ALBADER, M. A. A.; MAHYUDDIN, N. M.; MOHAMMED, M. A. Voice pathology detection and classification by adopting online sequential extreme learning machine. *IEEE Access*, v. 9, p. 77293–77306, 2021. Citado na página 13.

ARDILA, R. A.; BRANSON, M.; DAVIS, K.; HENRETTY, S.; KOHLER, J.; MOELLER, S.; MULLIGAN, A.; STADERMANN, J.; TOWNSEND, C.; WEINSTEIN, Y. Common voice: A massively-multilingual speech corpus. In: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. [S.l.: s.n.], 2019. p. 3513–3517. Citado na página 36.

ARISTóTELES. *A Política*. Belo Horizonte: Vega, 1998. Tradução disponível em Português. ISBN 9726995612. Citado 2 vezes nas páginas 31 e 32.

AZUL, D.; HANCOCK, A. B.; NYGREN, U. Forces affecting voice function in gender diverse people assigned female at birth. *Journal of Voice*, 2020. Citado na página 19.

BAEVSKI, A.; ZHOU, H.; MOHAMED, A.; AULI, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Not specified*, 2020. Preprint. Disponível em: <https://arxiv.org/pdf/2006.11477.pdf>. Citado 3 vezes nas páginas 9, 31 e 35.

BARBOSA, P. A.; MADUREIRA, S. *Manual de fonética acústica experimental: aplicações a dados do português*. São Paulo: Editora Cortez, 2015. Citado 2 vezes nas páginas 14 e 18.

BARTHES, R. A escuta. In: *O óbvio e o obtuso: ensaios críticos III*. Rio de Janeiro: Nova Fronteira, 1990. Citado 2 vezes nas páginas 14 e 17.

BENJAMIN, W. A obra de arte na época de sua reprodutibilidade técnica. In: ADORNO; AL. et (Ed.). *Teoria de cultura de massa*. São Paulo: Paz e Terra, 2000. p. 221–254. Citado na página 32.

BILAN, I.; BORCHMANN, ; NAMYSłO, M.; TRZCIńSKI, T. *Attention in Natural Language Processing*. [S.l.]: Springer, 2020. Citado na página 31.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. Citado na página 21.

BORSEL, J. V.; JANSSENS, J.; BODT, M. D. Breathiness as a feminine voice characteristic: A perceptual approach. *Journal of Voice*, v. 23, n. 3, p. 291–294, 2009. Citado na página 19.

BORSEL, J. V.; MAESSCHALCK, D. D. Speech rate in males, females, and male-to-female transsexuals. *Clinical Linguistics & Phonetics*, v. 22, n. 9, p. 679–685, 2008. Citado na página 19.

BUTLER, J. *Problemas de gênero: feminismo e subversão da identidade*. 8th. ed. Rio de Janeiro: Civilização Brasileira, 2015. Citado 3 vezes nas páginas 19, 34 e 35.

CAETANO, M.; SAITIS, C.; SIEDENBURG, K. Audio content descriptor of timbre. In: _____. *Timbre: Acoustics, Perception, and Cognition*. Cham: Asa Press & Springer, 2019. p. 389. Citado na página 15.

CAO, L.; GROSS, J. Cultural differences in perceiving sounds generated by others: Self matters. *Frontiers in Psychology*, v. 6, p. 1865, 12 2015. Citado na página 13.

CARTEI, V.; BANERJEE, R.; HARDOUIN, L.; REBY, D. The role of sex-related voice variation in children's gender-role stereotype attributions. *British Journal of Developmental Psychology*, 2019. Citado na página 19.

CARTEI, V.; BOND, R.; REBY, D. What makes a voice masculine: Physiological and acoustical correlates of women's ratings of men's vocal masculinity. *Hormones and Behavior*, 2014. Citado na página 19.

CAVARERO, A. *Vozes plurais: filosofia da expressão vocal*. Belo Horizonte: UFMG, 2011. 312 p. Citado na página 34.

CHIBA, T.; KAJIYAMA, M. The vowel, its nature and structure. *The Journal of the Acoustical Society of America*, v. 13, n. 1, p. 75–86, 1941. Citado na página 17.

DACAKIS, G.; OATES, J.; DOUGLAS, J. Beyond voice. *Current Opinion in Otolaryngology & Head and Neck Surgery*, v. 20, n. 3, p. 165–170, 2012. Citado na página 19.

DELEUZE, G.; GUATTARI, F. *Mil Platôs*. 2nd. ed. São Paulo: Editora 34, 1996. v. 1. Citado na página 19.

Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, p. arXiv:1810.04805, out. 2018. Citado na página 31.

FANT, G. *Acoustic Theory of Voice Production*. The Hague: Mouton, 1970. Citado na página 18.

FOUQUET, M.; PISANSKI, K.; MATHEVON, N.; REBY, D. Seven and up: individual differences in male voice fundamental frequency emerge before puberty and remain stable throughout adulthood. *Royal Society Open Science*, v. 3, n. 160395, 2016. Citado na página 19.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, MA: MIT Press, 2016. Citado 7 vezes nas páginas 16, 22, 24, 25, 26, 27 e 32.

HANCOCK, A. B.; COLTON, L.; DOUGLAS, F. Intonation and gender perception: Applications for transgender speakers. *Journal of Voice*, v. 28, n. 2, p. 203–209, 2014. Citado na página 19.

HANCOCK, A. B.; STUTTS, H. W.; BASS, A. Perceptions of gender and femininity based on language: Implications for transgender communication therapy. *Language and Speech*, p. 1–19, 2015. Disponível em: <https://las.sagepub.com>. Citado na página 19.

HERRERO, B. P. *Voice and Identity: a contrastive study of identity perception in voice*. 224 p. Dissertação (Master's thesis) — Ludwig-Maximilians-Universität, München, 2009. Citado na página 20.

KAWAMURA, S.; LIU, Z.; YOSHIDA, H. Estimation of the kansei information obtained from musical scores via machine learning algorithms : - classification of tempo into two classes using only information available in musical scores -. In: *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*. [S.l.: s.n.], 2019. p. 1–5. Citado na página 13.

LATINUS, M.; TAYLOR, M. J. Discriminating male and female voices: Differentiating pitch and gender. *Brain Topography*, v. 25, n. 2, p. 194–204, 2012. Citado na página 19.

LECUN, Y.; FOGELMAN, F. S. Modeles connexionnistes de l'apprentissage. *Intellectica, special issue apprentissage et machine*, v. 2, 01 1987. Citado na página 22.

LIN, T.; WANG, Y.; LIU, X.; QIU, X. A survey of transformers. *AI Open*, v. 3, p. 111–132, 2022. ISSN 2666-6510. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666651022000146>. Citado 2 vezes nas páginas 29 e 30.

LYON, R. F. Machine hearing: An emerging field. *IEEE Signal Processing Magazine*, v. 27, p. 131–139, 2010. Citado 3 vezes nas páginas 13, 15 e 16.

MARTINHO, D. H. d. C.; CONSTANTINI, A. C. Auditory-perceptual assessment and acoustic analysis of gender expression in the voice. *Journal of Voice*, 2024. Published online February 08, 2024. Disponível em: <https://doi.org/10.1016/j.jvoice.2023.12.024>. Citado na página 19.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, v. 5, p. 115–133, 1943. Citado na página 22.

MEY, J. L. As vozes da sociedade: letramento, consciência e poder. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, Pontifícia Universidade Católica de São Paulo - PUC-SP, v. 14, n. 2, p. 331–348, 1998. ISSN 0102-4450. Disponível em: <https://doi.org/10.1590/S0102-44501998000200003>. Citado na página 17.

MURPHY, G. L. *The Big Book of Concepts*. Cambridge: MIT Press, 2004. Citado na página 19.

NEVES, B. d. S.; MACHADO, R. Vozes subversivas: análise perceptivo-auditiva dos padrões vocais de pabllo vittar e cássia eller e suas relações de descontinuidade com a binaridade de gênero. In: *XXXIII Congresso da ANPPOM*. São João del-Rei: [s.n.], 2023. v. 33. Disponível em: <https://anppom.org.br/congressos/anais/v33/>. Citado na página 34.

OKHUNOV, D. M.; OKHUNOV, M. H.; MINAMATOV, Y. E. The use of machine learning and neural networks in the digital economy and international digital integration.

*Journal of Ethics and Diversity in International Communication*, v. 3, n. 2, p. 79–84, 2023. Disponível em: <https://openaccessjournals.eu/index.php/jedic/article/view/1859>. Citado na página 21.

PANAYOTOV, V.; CHEN, G.; POVEY, D.; KHUDANPUR, S. Librispeech: An asr corpus based on public domain audio books. In: IEEE. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2015. p. 5206–5210. Citado na página 36.

PEETERS, G.; AL. et. The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, v. 130, n. 4, p. 2902, 2011. Citado na página 20.

PEREIRA, E. M. *Estudos Sobre uma Ferramenta de Classificação Musical*. Not specified p. Dissertação (Master's thesis) — Faculdade de Engenharia Elétrica e Computação, Unicamp, Campinas, 2009. Citado na página 20.

PERNET, C. R.; PASCAL, B. The role of pitch and timbre in voice gender categorization. *Frontiers in Psychology*, v. 3, 2012. Citado na página 19.

PLACK, C. *The Sense of Hearing*. Lawrence Erlbaum Associates, 2005. ISBN 9780805848847. Disponível em: <https://books.google.com.br/books?id=DoGzm3soUoMC>. Citado na página 14.

RAGAV, T. R.; B, D.; A, K. Musical genre classification using support vector machines and convolutional neural networks. In: *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*. [S.l.: s.n.], 2023. p. 389–393. Citado na página 13.

RAY, P. P. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, v. 3, p. 121–154, 2023. ISSN 2667-3452. Disponível em: <https://www.sciencedirect.com/science/article/pii/S266734522300024X>. Citado na página 30.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, v. 65, n. 6, p. 386–408, 1958. ISSN 0033-295X. Disponível em: <http://dx.doi.org/10.1037/h0042519>. Citado 2 vezes nas páginas 22 e 23.

RUFINO, L. *Pedagogia das encruzilhadas*. 1 ed. ed. [S.l.]: Mórula Editorial, 2019. Citado na página 17.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature*, v. 323, p. 533–536, 1986. Citado na página 22.

SANTAELLA, L. *Semiótica e Tecnologia: Interfaces entre a Filosofia, a Linguagem e a Comunicação*. [S.l.]: Cengage Learning, 2003. Citado na página 31.

SKUK, V. G.; DAMMANN, L. M.; SCHWEINBERGER, S. R. Role of timbre and fundamental frequency in voice gender adaptation. *The Journal of the Acoustical Society of America*, v. 138, n. 2, p. 1180–1193, 2015. Citado na página 19.

SUNDBERG, J. *A ciência da Voz: Fatos sobre a voz na fala e no Canto*. 1st. ed. São Paulo: Editora da Universidade de São Paulo, 2015. 323 p. Citado na página 18.

TITZE, I. R. Nonlinear source-filter coupling in phonation: theory. *The Journal of the Acoustical Society of America*, v. 123, n. 5, p. 2733–2749, May 2008. Citado na página 18.

TURING, A. Computing machinery and intelligence. *Mind*, LIX, n. 236, p. 433–460, 1950. Citado na página 21.

TURING, A. M. On computable numbers, with an application to the entscheidungs-problem. *Proceedings of the London Mathematical Society*, v. 42, n. 2, p. 230–265, 1937. Citado na página 21.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. *Attention Is All You Need*. 2023. Citado 4 vezes nas páginas 9, 28, 29 e 30.

ZIMMAN, L. Transgender voices: Insights on identity, embodiment, and the gender of the voice. *Language and Linguistics Compass*, 2018. Citado na página 19.