MODELO DE SÍNTESE DE VOZ CANTADA COM BASE EM *MACHINE LEARNING*

Fernando Parra Cano[1]; Luis Antonio Galhego Fernandes[2]; Otávio dos Santos Gaijutis[3]; Daniel Soares e Marques[4]

Resumo

Esta pesquisa apresenta um modelo de síntese de voz cantada baseado em *machine learning*. Aborda a interseção entre tecnologia e música, destacando a importância da síntese de voz cantada para a inovação na indústria musical e sua relevância. O objetivo geral do estudo é desenvolver um modelo que utilize aprendizado de máquina para gerar vozes cantadas. Os objetivos específicos incluem a pesquisa sobre técnicas de síntese de voz e a definição das etapas necessárias para a criação do modelo. A metodologia abrange a elaboração de um *dataset*, o treinamento do modelo e a conversão da voz, utilizando redes neurais profundas (DNN) para aprimorar a naturalidade e expressividade da síntese. A pesquisa aborda os avanços na síntese de voz, desde modelos tradicionais como os Modelos Ocultos de Markov (HMM) até abordagens modernas que incorporam inteligência artificial (IA). Os resultados demonstram que o uso de DNNs permite uma representação mais natural das nuances vocais, superando limitações das técnicas anteriores. O trabalho conclui que o desenvolvimento deste modelo abre novas possibilidades criativas para artistas, contribuindo assim para o avanço da pesquisa em inteligência artificial aplicada à música.

Palavras-chave: síntese de voz cantada; aprendizado de máquina; áudio digital.

Abstract

This research presents a model for singing voice synthesis based on machine learning. It addresses the intersection of technology and music, highlighting the importance of singing voice synthesis for innovation in the music industry and its relevance. The overall goal of the study is to develop a model that utilizes machine learning to generate sung voices. Specific objectives include researching voice synthesis techniques and defining the necessary steps for model creation. The methodology encompasses the development of a dataset, model training, and voice conversion, using deep neural networks (DNNs) to enhance the naturalness and expressiveness of the synthesis. The research discusses advances in voice synthesis, from traditional models like Hidden Markov Models (HMMs) to modern approaches that incorporate artificial intelligence. The results demonstrate that the use of DNNs allows for a more natural representation of vocal nuances, overcoming limitations of previous techniques. The study concludes that the development of this model opens new creative possibilities for artists, thereby contributing to the advancement of research in artificial intelligence applied to music.

Keywords: singing voice synthesis; machine learning; digital audio.

[1] Graduado em Produção Fonográfica pela Faculdade de Tecnologia de Tatuí-FATEC; E-mail: ferpc.work@outlook.com.

[2] Mestre em Engenharia Química pela Universidade Estadual de Campinas-UNICAMP; professor da Faculdade de Tecnologia de Tatuí - FATEC; E-mail:galhegofernandes@hotmail.com.

[3] Mestre em Engenharia Química pela Universidade Estadual de Campinas-UNICAMP; professor da Faculdade de Tecnologia de Tatuí - FATEC; E-mail:osgaijutis@engineer.com.

[4] Mestre pelo Programa de Pós-Graduação em Informática (PPGI) da Universidade Federal da Paraíba-UFPB; professor da Faculdade de Tecnologia de Tatuí - FATEC; E-mail:daniel.marques12@fatec.sp.gov.br.

1 INTRODUÇÃO

A síntese de voz cantada, uma interseção entre música e tecnologia, tem ganhado destaque nas últimas décadas, especialmente com o avanço das técnicas de inteligência artificial e *machine learning*. Este campo não apenas transforma a maneira como a música pode ser criada e produzida, mas também abre novas possibilidades para compositores e artistas. O desenvolvimento de modelos de síntese de voz baseados em redes neurais profundas, DNNs, tem possibilitado a geração de vozes cantadas que imitam a naturalidade e expressividade da vocal humana.

Segundo Brum (2023), os sintetizadores de voz cantada permitem a geração artificial de vocais a partir de letras e notas musicais. Isso é feito através de métodos computacionais que podem replicar a performance vocal humana, oferecendo uma nova ferramenta para compositores e produtores musicais.

Alguns sintetizadores também servem como ferramentas educacionais, permitindo que cantores e estudantes analisem suas performances vocais, auxiliando os usuários a visualizar aspectos da sua voz, como a afinação, facilitando o aprendizado e aprimoramento das habilidades vocais (Fiuza, 2017).

Conforme Tomkinson (2024), os sintetizadores do trato vocal também estão relacionados ao fenômeno dos vocalistas virtuais, como os *Vocaloids*. Sua importância cultural reside na tradição japonesa de personificar entidades não-humanas, já que os *Vocaloids* frequentemente recebem personagens fictícios que promovem conexões emocionais com os ouvintes.

O uso de tecnologia avançada de projeção em performances ao vivo de vocalistas virtuais cria uma experiência que cativa os fãs, ao mesmo tempo em que levanta questões sobre autenticidade e a natureza das celebridades na era digital (Lam, 2016).

Conforme Walden (2023), o *software* Synthesizer V é capaz de gerar vozes realistas e expressivas, tendo em vista ser projetado para replicar o canto humano com realismo notável.

Conversão de Voz baseada em Recuperação (RVC) é uma tecnologia de clonagem de voz que permite a transformação da voz de um falante na voz de outro, utilizando um mecanismo de recuperação em vez de modelagem estatística tradicional. Este método tem ganhado destaque devido à sua eficiência, versatilidade e capacidade de produzir saídas de voz de alta qualidade com dados de treinamento relativamente mínimos (Cochard, 2024).

Neste contexto, o presente artigo visa explorar e desenvolver um modelo de síntese de voz cantada utilizando técnicas avançadas de *machine learning*. A proposta se fundamenta na

utilização de sintetizadores modernos, como o *Dreamtonics Synthesizer* V e o *Retrieval Voice Conversion* (RVC).

2 REVISÃO BIBLIOGRÁFICA

2.1 Síntese de voz cantada

De acordo com Choi (2022), a síntese de voz cantada é uma tarefa computacional focada na geração de sons vocais a partir de partituras musicais e letras usando computadores. Essa tecnologia tem uma história que remonta ao início da década de 1960, quando o primeiro modelo analógico digitalizado do trato vocal foi usado para sintetizar a música 'Daisy Bell Bicycle Built for Two', disponível para escuta em: https://www.youtube.com/watch?v=41U78QP8nBk. Desde então, avanços significativos foram feitos na imitação de vozes cantadas naturalmente.

Singing voice synthesis (SVS) envolve a criação de áudio de alta qualidade, expressões vocais controláveis e até mesmo síntese de voz cantada multilíngue. Os pesquisadores exploram vários aspectos do SVS, com o objetivo de desenvolver sistemas que possam gerar automaticamente vozes cantadas naturais e expressivas a partir de determinadas melodias e letras (Ardaillon, 2017).

Segundo Bock (2024), dentre os métodos de síntese de voz cantada, destacam-se: síntese por formantes, *vocoder*, resíntese vocal, geradores de voz com base em inteligência artificial e sistemas híbridos. A síntese por formantes simula as frequências ressonantes do trato vocal, em sintetizadores como o Yamaha FS1R, enquanto os *vocoders*, originários das telecomunicações, modulam sinais portadores com características vocais, criando sons distintos na produção musical. A ressíntese vocal analisa gravações para remapear informações de frequência, permitindo simulações vocais realistas. Geradores de voz baseados em inteligência artificial utilizam aprendizado de máquina para replicar qualidades únicas de vozes, gerando saídas realistas. Sistemas híbridos modernos combinam essas abordagens, integrando IA e métodos tradicionais para aprimorar a qualidade e expressividade vocal, manipulando elementos naturais e sintéticos.

Conforme Kim *et al*. (2018), os sistemas de SVS baseados em IA são compostos de duas partes: compor os recursos linguísticos e musicais e sintetizar a voz cantada a partir do modelo aprendido usando os recursos de entrada.

2.2 Cadeias de Markov

Os modelos ocultos de Markov, *hidden markov model* (HMMs), são uma técnica geral de modelagem estatística para problemas lineares, como sequências ou séries temporais. A ideia principal é que um HMM é um modelo finito que descreve uma distribuição de probabilidade sobre um número infinito de sequências possíveis (Eddy, 1996).

De acordo com Ghahramani (2001), HMMs são ferramentas estatísticas usadas para modelar sequências de observações em processos de Markov com estados ocultos. Com aplicações em áreas como reconhecimento de fala, biologia molecular, compressão de dados e visão computacional, os HMMs representam distribuições de probabilidade sobre observações, definindo estados ocultos e suas transições. O aprendizado envolve a estimação de parâmetros a partir de dados e a inferência sobre estados ocultos, utilizando técnicas como Expectativa-Maximização. Apesar de suas limitações em modelar relações temporais complexas, HMMs e suas generalizações oferecem uma estrutura robusta para entender e prever sequências de dados.

Embora ofereçam uma abordagem poderosa para prever sequências baseadas em dados observados, os HMMs enfrentam desafios de complexidade computacional que exigem algoritmos eficientes. O treinamento envolve resolver problemas como identificar sequências de estados mais prováveis, e técnicas de normalização são empregadas para garantir a estabilidade numérica durante os cálculos (Kohlschein, 2006).

Conforme Skansi (2018), os Modelos Ocultos de Markov podem ser integrados com outras técnicas, como redes neurais, para aprimorar o desempenho em tarefas complexas. Sua capacidade de simular processos cognitivos humanos também contribui para o desenvolvimento de sistemas de IA que simulam o raciocínio e a tomada de decisão.

2.3 Inteligência artificial

De acordo com Skansi (2018), a IA é um campo amplo que busca criar sistemas capazes de realizar tarefas que normalmente exigem inteligência humana, incluindo aprendizado, raciocínio e autocorreção. Divide-se em subcampos como processamento de linguagem natural, representação do conhecimento, planejamento, agendamento e visão computacional. O aprendizado de máquina, uma parte da inteligência artificial, permite que algoritmos melhorem com a exposição a mais dados. Influenciada por filosofia e ciência cognitiva, a IA evoluiu de sistemas baseados em regras para abordagens sofisticadas como aprendizado profundo *(deep learning)*, abrangendo uma variedade de aplicações e continuando a se desenvolver com o avanço da tecnologia.

2.3.1 *Machine learning*

Segundo Skansi (2018), o aprendizado de máquina *(machine learning)* é um subcampo da inteligência artificial que se dedica ao desenvolvimento de algoritmos que permitem aos computadores aprender com dados e fazer previsões. Ele se divide em três ramos principais: aprendizado supervisionado, que utiliza dados rotulados para treinar modelos; aprendizado não supervisionado, que identifica padrões em dados sem rótulos; e aprendizado por reforço.

Um modelo de *machine learning* é um programa computacional que aprende a reconhecer padrões e tomar decisões com base em dados. Ele é criado a partir da aplicação de algoritmos a um conjunto de dados, permitindo que o modelo identifique relações e faça previsões sobre novos dados que não foram vistos durante o treinamento (Radich; Jenks; Cowley, 2024).

Segundo Pereira (2023), o treinamento de um modelo de *machine learning* consiste em fornecer a ele um conjunto de dados rotulados, que contém exemplos de entradas e suas respectivas saídas esperadas. O objetivo é que o modelo aprenda a mapear as entradas para as saídas corretas. Este processo é realizado através de um algoritmo que ajusta os parâmetros do modelo com base nos dados apresentados.

De acordo com Aggarwal *et al*. (2022), sistemas de tecnologia da informação baseados em *machine learning* aprendem automaticamente padrões a partir dos dados sem serem explicitamente programados; esse aprendizado pode produzir conhecimento automaticamente, treinar algoritmos e reconhecer padrões desconhecidos. Os padrões identificados podem ser utilizados para um conjunto de dados novo e desconhecido, a fim de fazer previsões e otimizar processos.

O aprendizado de máquina é amplamente utilizado em áreas como reconhecimento de imagens e processamento de linguagem natural, e combina uma base matemática sólida com técnicas avançadas, como o *deep learning*, que usa redes neurais complexas. Essa combinação torna o aprendizado de máquina uma ferramenta essencial para a tomada de decisões baseada em dados na tecnologia moderna (Skansi, 2018).

2.3.2 *Deep learning*

De acordo com Skansi (2018), *deep learning* é um subsetor do aprendizado de máquina e inteligência artificial que utiliza redes neurais profundas para modelar padrões complexos em dados. Ele se diferencia das técnicas tradicionais por automatizar a extração de características, permitindo que o modelo aprenda diretamente de dados brutos. Suas aplicações abrangem diversas áreas, como processamento de linguagem natural e visão computacional, e tem se

tornado cada vez mais importante na IA. O processo de aprendizado pode ser comparado ao domínio de habilidades, envolvendo progresso contínuo e refinamento. Assim, o aprendizado profundo representa um avanço significativo nas capacidades da IA.

Machine learning refere-se ao uso de algoritmos que permitem que os computadores analisem dados, aprendam e façam previsões ou decisões com base no que aprenderam. Esse processo geralmente requer intervenção humana para ajustar os algoritmos e melhorar a precisão das previsões. Por outro lado, *deep learning* é um subconjunto de *machine learning* que utiliza redes neurais artificiais com múltiplas camadas. Essas redes são projetadas para imitar o funcionamento do cérebro humano, permitindo que as máquinas aprendam de maneira mais autônoma e complexa, sem necessidade de intervenção constante (Kleina, 2023).

Segundo Evangelista (2022), os modelos de *deep learning* permitem a personalização de voz sintetizada para diferentes estilos musicais e gêneros, sendo particularmente útil em aplicações onde a identidade vocal é crucial, como em jogos, filmes e assistentes virtuais.

2.3.3 Redes neurais

De acordo com Aggarwal (2018), as redes neurais são modelos computacionais inspirados no funcionamento das redes neurais biológicas do cérebro humano, compostas por neurônios interconectados, organizados em camadas. Sua principal função é aprender a partir de dados, ajustando os pesos das conexões para minimizar a diferença entre saídas previstas e reais, por meio de um processo chamado retropropagação. Com a capacidade de modelar relações complexas e aprender representações hierárquicas, as redes neurais são eficazes em tarefas como reconhecimento de imagem e processamento de linguagem natural. Apesar dos desafios, como *overfitting*, quando o modelo fornece previsões precisas para dados de treinamento, mas não para novos dados, avanços em algoritmos têm contribuído para amenizar possíveis falhas, proporcionando maior sucesso em diversas aplicações.

2.3.4 Redes neurais profundas

Nos últimos anos, as redes neurais profundas artificiais, denominadas *deep neural networks* (DNN), inspiradas na estrutura e funcionamento do cérebro humano, têm inovado a área de visão computacional, além de outros campos, como o processamento de linguagem natural. Essas redes são compostas por unidades de processamento simples, análogas a neurônios, que operam de forma paralela e estão arranjadas em camadas interconectadas. As redes neurais simples constituem-se de uma camada de entrada e uma de saída. À medida que

mais camadas são empilhadas, essas redes são designadas como profundas (Cichy e Kaiser, 2019).

Tradicionalmente, os sistemas de síntese de voz cantada são baseados em HMMs, que podem modelar diversas características acústicas de uma voz cantada simultaneamente. Embora os modelos ocultos de Markov precisem de recursos de composição fluente, a qualidade sonora é conhecida por não ser natural porque os HMMs têm um problema crônico de suavização excessiva tanto na frequência quanto nos domínios de tempo (Kim *et al*., 2018).

De acordo com Nishimura *et al*. (2016), DNNs melhoraram amplamente as abordagens convencionais, como HMM, em várias áreas de pesquisa, incluindo reconhecimento de fala, reconhecimento de imagem, síntese de fala, etc.

As redes neurais profundas são capazes de modelar complexas relações entre os dados de entrada (como texto e informações melódicas) e a saída desejada (a voz cantada). Isso resulta em vozes sintéticas que não apenas soam mais naturais, mas também capturam nuances emocionais e estilísticas que eram difíceis de reproduzir com técnicas anteriores (Evangelista, 2022).

Segundo Montavon, Samek e Müller (2018), as técnicas de aprendizado de máquina, em especial as DNNs, emergiram como uma ferramenta essencial em diversas áreas de aplicação, tais como classificação de imagens, reconhecimento de fala e processamento de linguagem natural. A precisão preditiva alcançada por essas técnicas é notavelmente elevada, frequentemente equiparável ao desempenho humano.

De acordo com Nishimura *et al*. (2016), uma base de dados de áudio de voz cantada e notação musical devem ser utilizados para a realização do treinamento de um modelo de voz por meio de DNNs. Na fase preliminar do treinamento, uma partitura musical específica é inicialmente transformada em uma sequência de recursos de entrada antes de ser utilizada no treinamento do modelo de voz baseado em DNN. Esses recursos de entrada incluem valores binários que descrevem contextos linguísticos como a identidade do fonema atual, a quantidade de fonemas na sílaba atual e as durações do fonema atual, e, contextos musicais como o tom do compasso musical atual e o tom da nota atual.

3 MATERIAL E MÉTODOS

A pesquisa abordada neste projeto, caracterizada como desenvolvimento experimental, objetiva a criação de um modelo de voz de síntese de voz cantada com base em *machine learning* e DNN por meio do *software* RVC.

Os materiais utilizados para o desenvolvimento deste projeto estão listados no quadro 1.

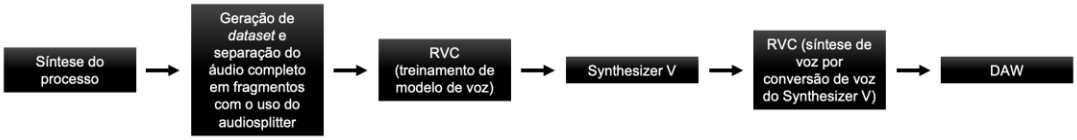
Quadro 1 - **Recursos utilizados no projeto**

ID	Descrição	Marca
01	MacBook Pro com chip M1 e sistema operacional macOS Sonoma 14.5	Apple
02	Computador com processador Intel i5-13600KF, placa de vídeo NVIDIA Geforce RTX 4060, memória RAM de 64GB DDR4, com sistema operacional Microsoft Windows 11	Intel, NVIDIA, XPG, WD
03	Word 2016	Microsoft
04	DAW Logic Pro X 10.7.9	Apple
05	Interface de áudio U-PHORIA UMC204HD	Behringer
06	Microfone SAS57	Santo Angelo
07	Synthesizer V Studio Basic 1.10.0	Dreamtonics
08	Retrieval-based-Voice-Conversion-WebUI 1006	RVC-Project

Fonte: Elaborada pelos autores.

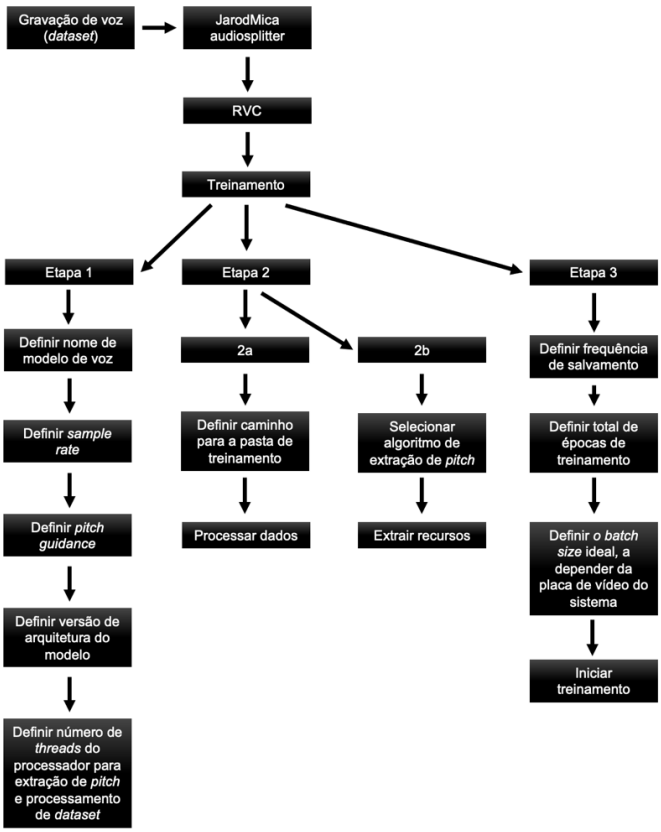
Para a realização deste trabalho, os passos utilizados podem ser vistos nas figuras 1, 2 e 3.

Figura 1 - Síntese do processo



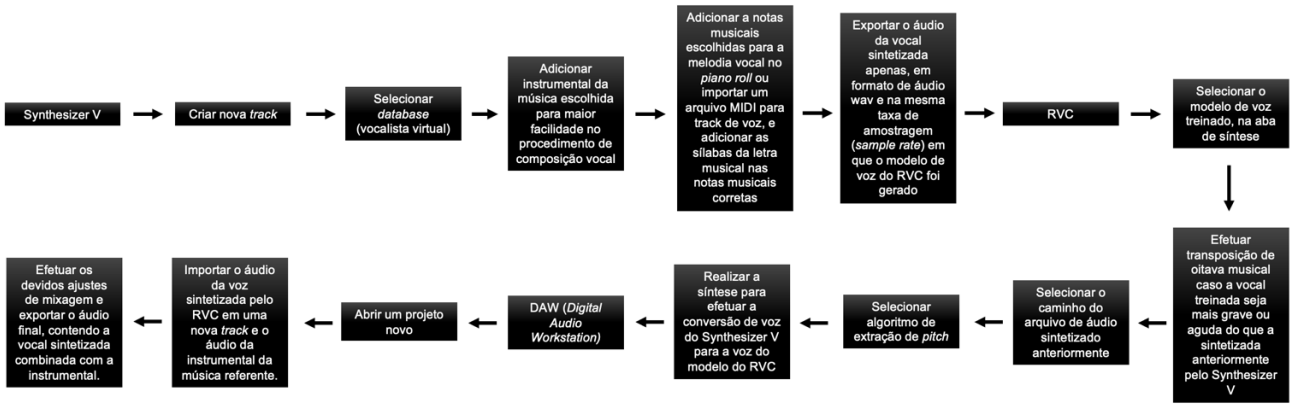
Fonte: Elaborada pelos autores.

Figura 2 - Diagrama com etapas detalhadas referentes ao RVC



Fonte: Elaborada pelos autores.

Figura 3 - Diagrama com etapas detalhadas referentes ao Synthesizer V e DAW



Fonte: Elaborada pelos autores.

Os métodos adotados neste trabalho estão detalhados em etapas, a seguir:

3.1 Primeira etapa: *dataset* e separação de áudio em fragmentos

A primeira etapa refere-se à elaboração de um *dataset*, gravação de voz falada e/ou cantada para aplicação na etapa de treinamento, e, separação do áudio completo em fragmentos de dez segundos por meio do uso do *software* Audiosplitter, para evitar problemas de falta de memória VRAM.

3.1.1 Elaboração de *dataset*

Para a elaboração do *dataset* foi efetuada gravação de voz falada e cantada em inglês, para isso utilizou-se como *hardware* para gravação um microfone dinâmico Santo Angelo SAS57 combinado com uma interface de áudio Behringer U-PHORIA UMC204HD e um notebook Apple MacBook Pro M1; a *Digital Audio Workstation* (DAW) escolhida para a gravação foi o Logic Pro X 10.7.9.

A gravação foi efetuada em uma sala com pouca reverberação, contou com uma taxa de amostragem de 48KHz e 24 *bits* de profundidade, totalizando em 14 minutos e 23 segundos de áudio gravado.

Durante a gravação foram efetuados ajustes de edição de áudio como mesclagem de *takes* e, após finalizado esse procedimento, foram realizados ajustes de mixagem como compressão de áudio, para equilibrar o nível de volume sonoro da gravação, cujos parâmetros ideais foram -22.5 dB de *threshold*, 3.7:1 de *ratio*, 11.5dB de *make up*, 18.5 ms de *attack* e 190 ms de *release*, como pode ser observado na figura 4.

Figura 4 - Compressor do Logic Pro X

Fonte: Elaborada pelos autores.

3.1.2 Separação do áudio em fragmentos

Após a renderização do arquivo de áudio em formato WAV, o áudio da voz gravada foi separado em fragmentos de 10 segundos por meio do uso do *software* Audiosplitter (disponível para *download* em: https://github.com/JarodMica/audiosplitter), do desenvolvedor Jarod Mica.

Logo após aberto o *software* Audiosplitter em um computador com o sistema operacional Microsoft Windows 11 foi apresentada uma janela do *prompt* de comando questionando a funcionalidade desejada a ser utilizada, foi digitado o número 2 e confirmado com a tecla *enter*

para selecionar a função de separação de áudio em fragmentos de 10 segundos, em seguida foi apresentado um navegador de arquivo e foi escolhida a pasta do arquivo de áudio renderizado do *dataset*.

Após isso, o Audiosplitter separou o áudio total em 87 seguimentos, sendo 86 destes de 10 segundos de duração e o último de 3 segundos.

3.2 Segunda etapa: treinamento de modelo de voz

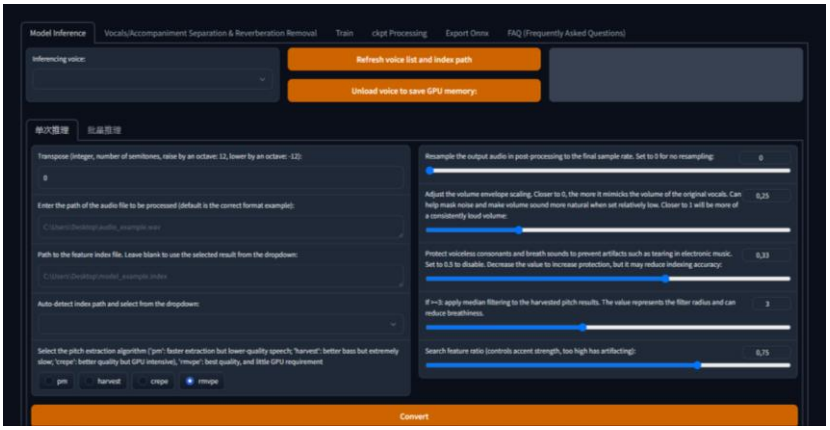
A segunda etapa consiste no treinamento do modelo de voz com o uso do RVC.

Para a execução dessa etapa foi utilizado um computador com o sistema operacional Microsoft Windows 11, onde anteriormente foi instalado o *software* da linguagem de programação Python (disponível em: https://www.python.org/downloads/), tendo em vista ser pré-requisito para o funcionamento do RVC.

Em seguida, foi efetuado o *download* da versão 1006 para NVIDIA do *software* RVC (disponível em: https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI) e descompactado o arquivo 'RVC1006Nvidia.7z' por meio do uso do *software* WinRAR (disponível em: https://www.rarlab.com/).

Após isso, dentro da pasta descompactada do RVC, foi executado o arquivo 'go-web.bat', e, uma caixa de diálogo do *firewall* foi apresentada questionando a permissão de redes públicas e privadas para acesso ao Python, foi selecionada a opção permitir, na sequência uma janela do *prompt* de comando foi aberta e a página do RVC foi apresentada no navegador da *web*, neste caso o Google Chrome, como pode ser observado na figura 5.

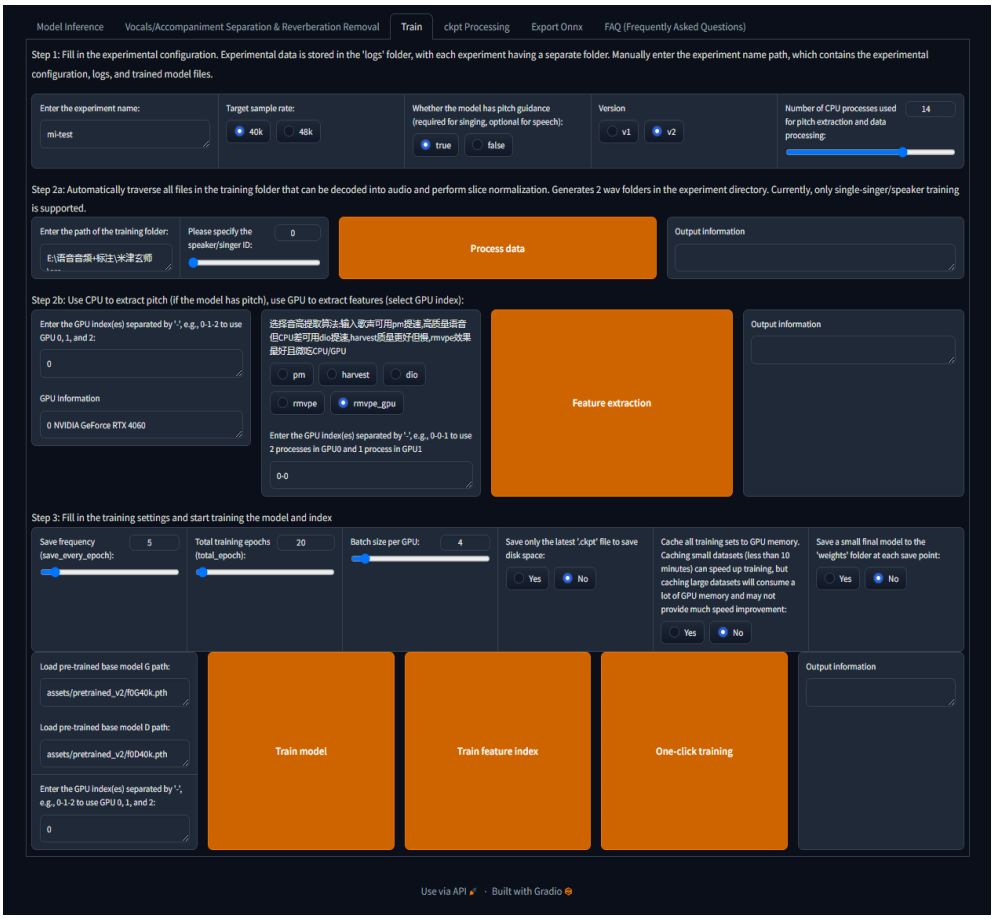
Figura 5 - Página web do RVC



Fonte: Elaborada pelos autores.

Em seguida, foi selecionada na página do RVC a seção *'train'*. Selecionada esta opção, foi apresentada uma nova seção separada por etapas de treinamento a serem efetuadas e configuradas, como pode ser visto na figura 6.

Figura 6 - Seção *train* do RVC



Fonte: Elaborada pelos autores.

Na sequência, na etapa 1 do treinamento, no campo *'enter the experimente name'* foi definido um nome para o modelo de voz que foi gerado, neste caso 'FPC'. Na opção *'target sample rate'* foi definida a taxa de amostragem de 48KHz, tendo em vista que foi a escolhida previamente durante a gravação da voz. Na configuração *'whether the model has pitch guidance'*, foi mantida a opção *'true'*, considerando que é necessária para síntese de voz cantada. Na opção *'version'* foi selecionada 'v2' e na configuração *'number of CPU processes used for pitch extraction and data processing'* foi mantida a configuração padrão.

Em seguida, na etapa 2a, no campo *'enter the path of the training folder'*, foi inserido o caminho da pasta dos seguimentos de áudio. Na configuração *'please specify the speaker/singer ID'* foi mantida a definição padrão. Após isso, foi selecionado o botão *'process data'* e efetuado o processamento de dados.

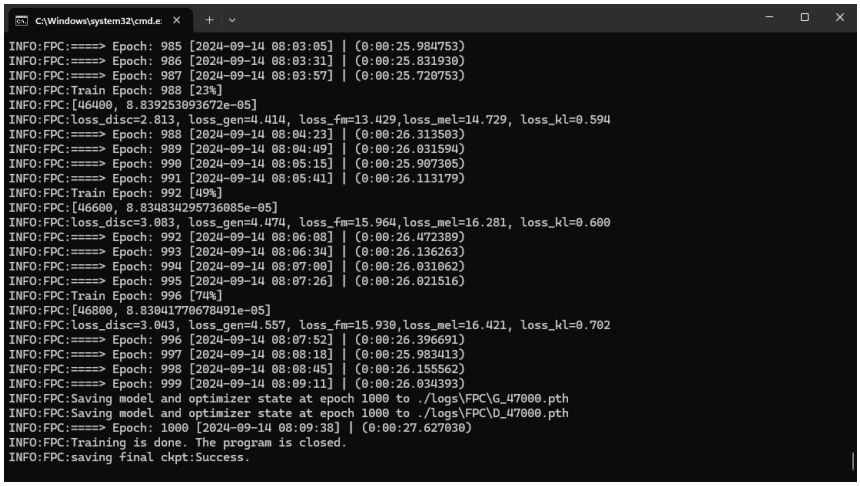
Logo após, na etapa 2b, no campo *'enter the gpu index(es) separated by '-', e.g., 0-1-2 to use gpu 0, 1, and 2'*, foi mantida a configuração padrão, na configuração *'gpu information'* foi mantida a opção padrão e em '*enter the gpu index(es) separated by '-', e.g., 0-0-1 to use 2 processes in gpu0 and 1 process in gpu1'*, também foi mantida a configuração padrão. Na opção

de seleção do algoritmo de extração de *pitch*, foi selecionado ‘rmvpe_gpu’ e em seguida foi selecionado o botão *‘feature extraction’* para extrair o contorno melódico *(pitch)* por meio do processamento do processador e extração de recursos, como os coeficientes de frequência mel-cepstrais (MFCC), por meio do processamento da placa de vídeo (GPU).

Na etapa 3 do treinamento, as opções *‘save only the latest '.ckpt' file to save disk space’*, *‘cache all training sets to gpu memory. caching small datasets (less than 10 minutes) can speed up training, but caching large datasets will consume a lot of gpu memory and may not provide much speed improvement’* e *‘save a small final model to the 'weights' folder at each save point’* foram mantidas com as configurações padrões. O parâmetro *‘save frequency (save_every_epoch)’* foi definido para 50. A configuração *‘total training epochs (total_epoch)’* foi definida para 1000. O parâmetro *‘batch size per GPU’* foi definido para 7, de acordo com a GPU do sistema. Após isso, foi selecionado o botão *‘one-click training’* e iniciou-se o processamento da última etapa. Durante o processo foi possível acompanhar o andamento do treinamento por meio da janela do *prompt* de comando do RVC.

Ao final do treinamento foi apresentada uma mensagem de conclusão e o arquivo ckpt final foi gerado com sucesso, como pode ser visto na figura 7. A duração do procedimento de treinamento foi de 7 horas, 13 minutos e 41 segundos, e, contou com 1000 épocas.

Figura 7 - Janela do *prompt* de comando do RVC ao final do treinamento



Fonte: Elaborada pelos autores.

3.3 Terceira etapa: Synthesizer V

Essa etapa refere-se à síntese da voz fonte, tendo como informação de entrada as notas musicais juntamente à letra musical em um *piano roll*, com a utilização do *software* Synthesizer V. Durante essa etapa o computador utilizado foi um notebook MacBook Pro com chip M1 e sistema operacional macOS Sonoma 14.5.

Inicialmente foi efetuado o *download* do *software* Synthesizer V (disponível em: https://dreamtonics.com/synthesizerv/), e, foi escolhido e feito o *download* de um banco de voz (disponível em: https://resource.dreamtonics.com/download/English/Voice%20Databases/Lite%20Voice%20Databases/), neste caso o banco de voz escolhido foi 'ANRI (Lite)'.

Após isso, o *software* Synthesizer V foi instalado e aberto, e, um novo projeto em branco foi apresentado na interface do *software*.

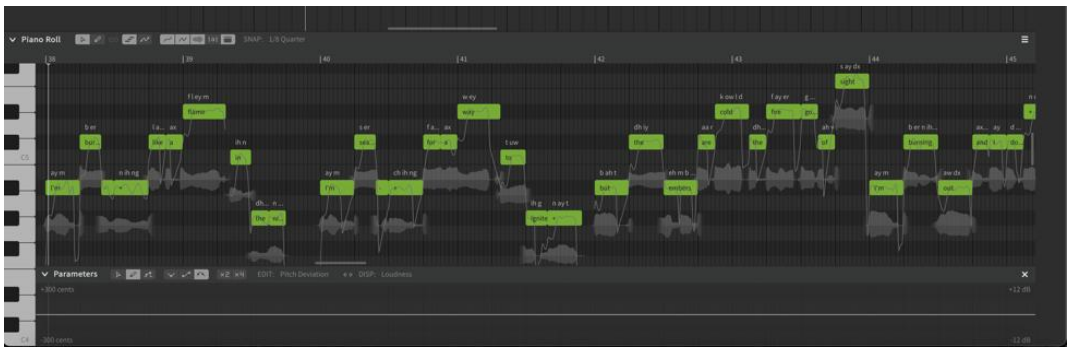
Em seguida, foi aberto o arquivo SVPK de instalação do banco de voz do Synthesizer V e foi apresentada uma janela de instalação no Synthesizer V. Na sequência, foi efetuada a instalação do banco de voz.

Logo após, foi definido um nome para a *track* de voz, tendo sido substituído o nome padrão 'Unnamed Track' para 'VOZ SYNTH V'. Em seguida foi selecionado o modelo de voz da *track* na opção *'current database (track)'*.

Na sequência, foi arrastado o arquivo da instrumental da música, nomeado 'INSTRUMENTAL', para dentro da janela do Synthesizer V, e, foi gerada uma *track* de áudio contendo a instrumental da música.

Depois foram adicionadas as notas musicais escolhidas para a melodia vocal no *piano roll*, combinadas com as sílabas das palavras da letra musical nas notas musicais correspondentes, como pode ser visto na figura 8.

Figura 8 - *Piano roll* do *software* Synthesizer V



Fonte: Elaborada pelos autores.

Após isso, foi efetuada a renderização da vocal sintetizada do Synthesizer em uma taxa de amostragem de 48KHz, 24 *bits* de profundidade e em mono. Para isso, foi selecionado o quinto menu da barra lateral direita da interface do *software*, em seguida, foi definida uma pasta de destino do arquivo renderizado na opção *'destinination folder'*, foi definido o nome 'VOZ SYNTH V' no campo *'file name'*, selecionada somente a *track* 'VOZ SYNTH V' para ser renderizada na seção *'tracks'*, na seção *'format'* foram definidas para a renderização as

configuração mono, 24-bit de *bit depth* e 48000Hz (48KHz) de *sample rate*, por último foi selecionado o botão *'bounce to files'* para renderizar o áudio.

3.4 Quarta etapa: conversão de voz

A quarta etapa refere-se à conversão da voz fonte sintetizada pelo Synthesizer V para a voz alvo do modelo treinado pelo RVC. Para a execução dessa etapa foi utilizado um computador com o sistema operacional Microsoft Windows 11.

Inicialmente foi criada uma pasta nomeada *'songs'* dentro da pasta do RVC, em seguida, foi copiado o arquivo da voz sintetizada pelo Synthesizer V para esta pasta.

Após isso, na seção *'model inference'* da página *web* do RVC, selecionou-se a opção *'refresh voice list and index path'*.

Na sequência, na opção *'inferencing voice'*, foi selecionado o modelo de voz treinado anteriormente.

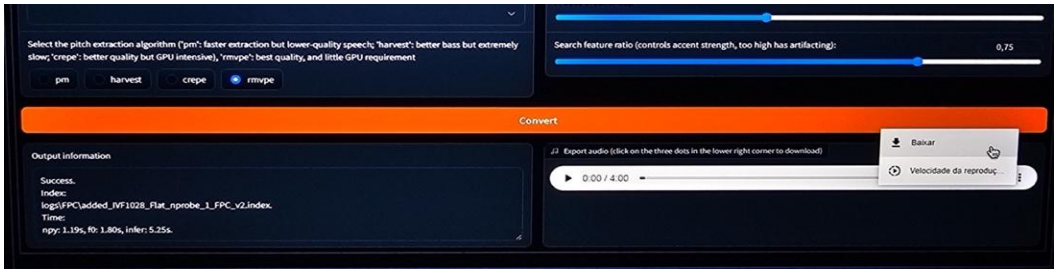
Em seguida, na configuração *'transpose (integer, number of semitones, raise by an octave: 12, lower by an octave: -12)'*, tendo em vista que a vocal do banco de voz escolhido anteriormente no Synthesizer V é uma vocal feminina e a vocal do modelo de voz treinado no RVC é uma vocal masculina, e, pelo fato de vocais masculinas geralmente apresentarem um timbre mais grave, foram reduzidos 12 semitons musicais, tendo sido digitado '-12', resultando na redução de uma oitava musical.

Logo após, foi inserido o caminho do arquivo na opção *'enter the path of the audio file to be processed (default is the correct format example)'*, que neste caso foi E:\Bibliotecas\Documentos\projeto-modelo-voz\RVC1006Nvidia\songs\VOZ SYNTH V_VOZ_SYNTH_V.wav.

Em seguida, definiu-se como algoritmo de extração de *pitch*, o algoritmo 'rmvpe' e após isso foi selecionado o botão *'convert'*.

Após finalizada a conversão de voz, foi efetuado o *download* do arquivo da voz convertida pelo RVC, conforme a figura 9.

Figura 9 - Procedimento de *download* do arquivo da voz convertida pelo RVC



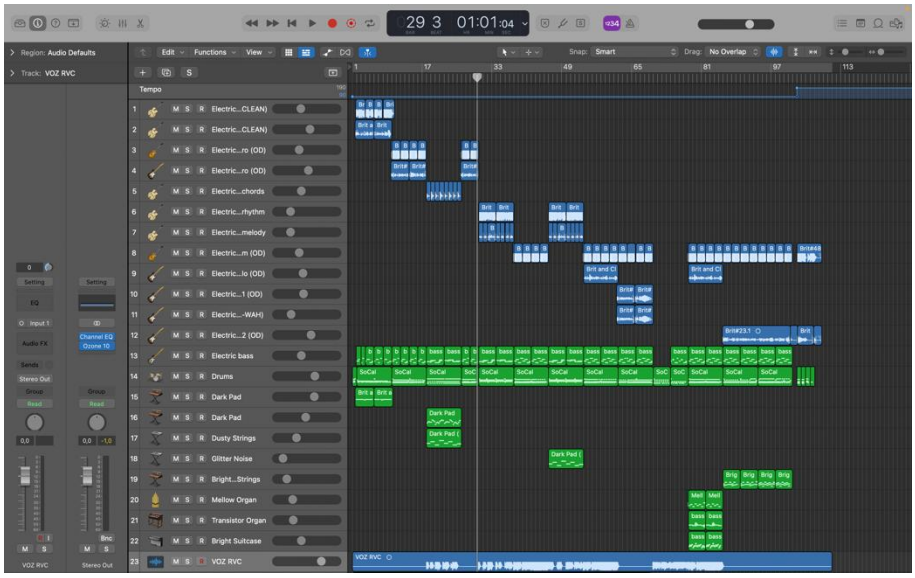
Fonte: Elaborada pelos autores.

3.5 Quinta etapa: alinhamento da vocal e instrumental

Nessa etapa, em uma DAW, alinhou-se a voz sintetizada pelo RVC com a instrumental da música referente e efetuaram-se os devidos ajustes de mixagem de áudio. Durante essa etapa o computador utilizado foi um notebook MacBook Pro com chip M1 e sistema operacional macOS Sonoma 14.5, e, a DAW escolhida foi o Logic Pro X 10.7.9.

Inicialmente foi aberto no Logic Pro X o projeto instrumental da música e importado o áudio da vocal convertida pelo RVC em uma nova *track*, conforme apresentado na figura 10.

Figura 10 - Projeto do Logic Pro X



Fonte: Elaborada pelos autores.

Na sequência, os trechos vocais do interlúdio e do refrão final da música foram separados em 2 novas *tracks* de áudio.

Na primeira *track* de voz foram efetuados ajustes de mixagem como os de volume, compressão e foi adicionada reverberação.

Na segunda *track* vocal, referente ao trecho do interlúdio da música, foram efetuados ajustes de mixagem como os de volume, compressão, e, foi adicionada reverberação e *delay* estéreo, como apresentado na figura 11.

Figura 11 - Janelas dos *plugins* utilizados na segunda *track* de voz

Fonte: Elaborada pelos autores.

Na terceira *track* de voz, referente ao trecho do refrão final da música, foram efetuados ajustes de mixagem como os de volume, compressão, e, foi adicionada reverberação e distorção.

Após isso, foi aberta a janela de renderização *(bounce)*, mantendo pressionada a tecla *'command'* e clicando na tecla 'b', na sequência foi efetuada a renderização do projeto do Logic Pro X em formato WAV à uma taxa de amostragem de 48KHz e 24 *bits* de profundidade.

4 RESULTADOS E DISCUSSÃO

As pesquisas e experimentos realizados durante a execução deste projeto resultaram na síntese de voz cantada para uma música. Os resultados foram subdivididos em quatro arquivos (disponíveis em: https://is.gd/fpcsamplemusic1), sendo três arquivos de áudio em formato WAV e um arquivo SVP do Synthesizer V.

O primeiro arquivo gerado durante os experimentos foi o projeto SVP do Synthesizer V, que contou com a letra musical dividida por sílabas juntamente à melodia vocal, tendo sido uma ferramenta essencial para efetuar o procedimento decorrido ao longo dos experimentos.

O segundo arquivo gerado foi o arquivo de áudio WAV da voz sintetizada pelo RVC, tendo sido crucial para efetuar o processo de conversão de voz para o modelo treinado no RVC.

O terceiro arquivo gerado foi o arquivo WAV da voz convertida e sintetizada pelo RVC com base na voz sintetizada pelo Synthesizer V, tendo sido efetuada a conversão e síntese de voz com êxito, assim como a geração do arquivo de áudio.

O quarto também foi gerado com êxito, tendo sido um arquivo de áudio de formato WAV renderizado pelo Logic Pro X, onde foi alinhada a voz convertida e sintetizada pelo RVC com a instrumental, e, foram efetuados os devidos ajustes de mixagem de áudio.

Após uma análise minuciosa dos arquivos gerados e resultados obtidos, ficou constatado que, tanto a geração do arquivo SVP e síntese de voz do Synthesizer V, assim como a conversão e síntese do RVC, e, alinhamento da vocal à instrumental, obtiveram bons resultados, apresentando alta precisão e qualidade na síntese de vozes cantadas naturais e expressivas.

5 CONSIDERAÇÕES FINAIS

Infere-se que as pesquisas relacionadas à síntese de voz cantada, HMMs e às áreas da IA, e, o desenvolvimento de procedimentos a serem adotados para alcançar resultados no desenvolvimento e aplicação prática de um modelo de síntese de voz cantada é fundamental para averiguar e aperfeiçoar a qualidade da criação de um modelo de voz, utilizando treinamento de modelo e conversão de voz fonte para voz alvo por meio do *software* RVC, combinado ao uso do *software* Synthesizer V para síntese de voz fonte, e, uso da DAW Logic Pro X para os devidos ajustes de edição e mixagem de áudio.

Após compreendido o assunto, traçaram-se as etapas necessárias para a geração e aplicação prática do modelo de voz, elaborando um *dataset* e separando em fragmentos de áudio, efetuando treinamento do modelo de voz por meio do RVC, gerando e sintetizando um projeto do Synthesizer V referente a música escolhida, convertendo a voz fonte do Synthesizer V para a voz fonte do modelo por meio do RVC, e, alinhando a voz convertida e efetuando devidos ajustes de mixagem por meio da DAW Logic Pro X. A voz sintetizada obteve bons resultados, apresentando alta precisão e qualidade na síntese de vozes cantadas naturais e expressivas.

REFERÊNCIAS

AGGARWAL, C. C. **Neural networks and deep learning: a textbook**. Springer, 2018. Disponível em: https://dlib.hust.edu.vn/bitstream/HUST/24439/1/OER000003177.pdf. Acesso em: 26 set. 2024.

AGGARWAL, K. *et al*. Has the future started? The current growth of artificial intelligence, machine learning, and deep learning. **Iraqi Journal for Computer Science and Mathematics**, v. 3, n. 1, p. 115-123, 2022. Disponível em: https://www.iasj.net/iasj/download/cefbfd60eb11898a. Acesso em: 25 maio 2024.

ARDAILLON, L. **Synthesis and expressive transformation of singing voice**. 2017. 249 f. Tese (Doutorado) – Université Pierre et Marie Curie – Paris VI, Paris, 2017. Disponível em: https://hal.science/tel-01710926/file/2017PA066511.pdf. Acesso em: 19 maio 2024.

BOCK, C. **Vocal Synthesizers: 4 Ways to Create Synthetic Voices and How to Use Them**. 2024. Disponível em: https://babyaud.io/blog/vocal-synthesizers. Acesso em: 26 set. 2024.

BRUM, L. A. Z. **Patricia: um sintetizador de canto em tempo real para o português brasileiro**. 2023. 113 f. Trabalho de Conclusão de Curso (Graduação em Música) – Universidade Federal de Sergipe, São Cristóvão, 2023. Disponível em: https://ri.ufs.br/bitstream/riufs/19472/2/LEONARDO_ARAUJO_ZOEHLER_BRUM.pdf. Acesso em: 17 jul. 2026.

CHOI, S. **Singing voice Synthesis**. Disponível em: https://mac.kaist.ac.kr/singing_voice_synthesis.html. Acesso em: 19 maio 2024.

CICHY, R. M.; KAISER, D. Deep neural networks as scientific models. **Trends in cognitive sciences**, v. 23, n. 4, p. 305-317, 2019. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1364661319300348. Acesso em: 25 set. 2024.

COCHARD, D. **RVC: An AI-Powered Voice Changer**. 2024. Disponível em: https://medium.com/axinc-ai/rvc-an-ai-powered-voice-changer-39927cc83bee. Acesso em: 13 out. 2024.

EVANGELISTA, L. A. de O. **Uso de redes neurais generativas para síntese de voz: estudo e revisão de literatura**. 2022. 34 f. Monografia (Bacharelado em Ciência da Computação) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2022. Disponível em: https://www.linux.ime.usp.br/~luneomena/mac0499/TCC_monografia-10.pdf. Acesso em: 17 jul. 2026.

EDDY, S. R. Hidden markov models. **Current opinion in structural biology**, v. 6, n. 3, p. 361-365, 1996. Disponível em: https://courses.cs.duke.edu/fall14/compsci260/resources/HMM/eddy96.pdf. Acesso em 22 set. 2024.

FIUZA, M. **A tecnologia no auxílio do canto**. 2017. Disponível em: https://www.estudiodevoz.com.br/2017/09/a-tecnologia-no-auxilio-do-canto.html. Acesso em: 13 out. 2024.

GHAHRAMANI, Z. An introduction to hidden Markov models and Bayesian networks. **International journal of pattern recognition and artificial intelligence**, v. 15, n. 01, p. 9-42, 2001. Disponível em: https://i2pc.es/coss/Docencia/SignalProcessingReviews/Ghahramani2001.pdf. Acesso em 24 set. 2024.

KOHLSCHEIN, C. An introduction to hidden Markov models. *In*: **Probability and Randomization in Computer Science. Seminar in winter semester**. 2006. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c04443dfd58a816e821fc5278d3e0d5857f7d800. Acesso em: 22 set. 2024.

KIM, J. *et al*. Korean singing voice synthesis system based on an LSTM recurrent neural network. *In*: **Proc. Interspeech**. 2018. p. 1551-1555. Disponível em: https://www.isca-archive.org/interspeech_2018/kim18b_interspeech.pdf. Acesso em: 22 maio 2024.

KLEINA, O. **A diferença entre machine learning e deep learning**. 2023. Disponível em:https://posdigital.pucpr.br/blog/machine-learning-deep-learning. Acesso em: 26 set. 2024.

LAM, K. Y. The Hatsune Miku Phenomenon: More Than a Virtual J-Pop Diva. **Journal of Popular Culture**, v. 49, n. 5, 2016. Disponível em: https://scholars.cityu.edu.hk/en/publications/the-hatsune-miku-phenomenon-more-than-a-virtual-j-pop-diva/. Acesso em: 17 jul. 2026.

MONTAVON, G.; SAMEK, W.; MÜLLER, K-R. Methods for interpreting and understanding deep neural networks. **Digital signal processing**, v. 73, p. 1-15, 2018. Disponível em: https://www.sciencedirect.com/science/article/pii/S1051200417302385. Acesso em: 22 maio 2024.

NISHIMURA, Masanari *et al*. Singing Voice Synthesis Based on Deep Neural Networks. *In*: **Interspeech**. 2016. p. 2478-2482. Disponível em: https://www.isca-archive.org/interspeech_2016/nishimura16_interspeech.pdf. Acesso em: 22 maio 2024.

PEREIRA, F. **Machine Learning:** Um guia atualizado! 2023. Disponível em: https://www.dataex.com.br/machine-learning-um-guia-atualizado/. Acesso em: 26 set. 2024.

RADICH, Q.; JENKS, A.; COWLEY, E. **O que é um modelo de machine learning?** 2024. Disponível em: https://learn.microsoft.com/pt-br/windows/ai/windows-ml/what-is-a-machine-learning-model. Acesso em: 26 set. 2024.

SKANSI, S. **Introduction to Deep Learning:** from logical calculus to artificial intelligence. Springer, 2018.

TOMKINSON, S. This kind of life has no meaning: Neru's Vocaloid Album CYNICISM. **M/C Journal**, v. 27, n. 2, 2024. Disponível em: https://journal.media-culture.org.au/index.php/mcjournal/article/view/3037. Acesso em 29 set. 2024.

WALDEN, J. **Dreamtonics Synthesizer V**. 2023. Disponível em: https://www.soundonsound.com/reviews/dreamtonics-synthesizer-v. Acesso em: 29 set. 2024.