Rede Transformers Aplicada na Detecção de Voz Cantada

Arthur Ruan Bizerra Florentino[1], **Yuri de Almeida Malheiros Barbosa**[1]

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brasil

arthurflorentino@eng.ci.ufpb.br, yuri@ci.ufpb.br

Abstract. *Singing voice detection (SVD) aims to identify vocal excerpts in songs and is essential in areas of musical information retrieval (MIR) related to vocals, such as singer identification and in pre-processing processes, being used in the separation of singing voices, automatic transcription of song lyrics, among other applications. Although singing detection seems like a simple task for humans, it proves to be extremely demanding for machines, with its main challenges being the complexity of sound patterns, frequency overlap and variations in singing style. However, with the advancement of deep learning models and audio manipulation techniques, the results have improved significantly over the years. This work aims to demonstrate the impact on SVD of using Transformers in two datasets: the Jamendo Corpus and DALI. The results showed that the model was more difficult to train with the second, presenting a difference of about 10% in the F1-score compared to the first, which was expected due to the lower accuracy and balance of this set. Furthermore, using Demucs, a music source separation, to separate voices from instrumental accompaniment improved vocal detection accuracy. The comparison highlights how different data volumes and qualities influence the model's performance in the SVD task.*

Resumo. *A detecção de voz cantada (SVD) visa identificar trechos vocais em músicas e é fundamental em áreas de recuperação de informações musicais (MIR) relacionadas à vocais, como na identificação de intérpretes e em processos de pré-processamento, sendo utilizada na separação de vozes cantadas, transcrição automática de letras de músicas, entre outras aplicações. Embora a detecção de canto pareça uma tarefa simples para humanos, ela se revela extremamente desafiadora para máquinas, tendo como principais desafios a complexidade dos padrões sonoros, a sobreposição de frequências e as variações no estilo de canto. Não obstante, com o avanço dos modelos de aprendizagem profunda e técnicas de manipulação de áudio, os resultados têm melhorado significativamente ao longo dos anos. Este trabalho tem como objetivo demonstrar o impacto em SVD da utilização de Transformers em dois conjuntos de dados: o Jamendo Corpus e o DALI. Os resultados mostraram que o modelo enfrentou mais dificuldade ao treinar com o segundo, apresentando uma diferença de cerca de 10% no F1-score em relação ao primeiro, o que era esperado devido à menor acurácia e balanceamento desse conjunto. Além disso, o uso do Demucs, um separador de fonte musical, para separar as vozes do acompanhamento instrumental melhorou a precisão na detecção de vocais. A comparação destaca como diferentes volumes e qualidades de dados influenciam o desempenho do modelo na tarefa de SVD.*

Catalogação na publicação
Seção de Catalogação e Classificação

F633r Florentino, Arthur Ruan Bizerra.
Rede transformers aplicada na detecção de voz cantada / Arthur Ruan Bizerra Florentino. - João Pessoa, 2024.
15 f. : il.

Orientação: Yuri de Almeida Malheiros Barbosa.
TCC (Graduação) - UFPB/CI.

1. Detecção de voz cantada. 2. Rede transformers. 3. Banco de dados DALI. 4. Pré-processamento. I. Barbosa, Yuri de Almeida Malheiros. II. Título.

UFPB/CI CDU 004.6

Elaborado por Michelle de Kassia Fonseca Barbosa - CRB-738

1. Introdução

A detecção de voz cantada (*Singing Voice Detection* - SVD, do inglês) desempenha um papel central em diversas aplicações dentro da recuperação de informações musicais (*Music Information Retrieval* - MIR, do inglês). O componente vocal é um dos aspectos mais importantes em músicas populares, sendo crucial em tarefas como a identificação de intérpretes, transcrição de músicas [McVicar et al. 2014], alinhamento automático de letras e separação de vozes. Com isso, SVD tornou-se um tópico de pesquisa bastante ativo na área de MIR, recebendo crescente atenção nos últimos anos [Zhang et al. 2020, Bonzi et al. 2023, Lehner et al. 2015, Leglaive et al. 2015, Choi et al. 2017], devido à sua relevância prática e ao impacto direto que pode ter em várias áreas de análise musical [Lee et al. 2018].

Sabe-se que, a voz cantada difere da fala em vários aspectos fundamentais, fazendo o uso da respiração de maneira controlada para ajustar o tom e a duração, o que resulta em uma intensidade média maior que a da fala. Além disso, sua faixa dinâmica é mais ampla e seu tom frequentemente varia em relação ao da fala [Vijayan et al. 2019]. Essas diferenças tornam a tarefa de detecção de vocais particularmente desafiadora, exigindo modelos que consigam capturar as nuances vocais, em meio a uma complexa mistura de áudio.

Tradicionalmente, três abordagens foram largamente exploradas na detecção de voz cantada. A primeira explora a similaridade entre a voz cantada e a fala, aproveitando as semelhanças acústicas entre esses dois tipos de emissão vocal [Berenzweig e Ellis 2001, Kim e Whitman 2002]. Já a segunda abordagem utiliza classificadores de aprendizado de máquina, que suportam modelos ocultos de Markov (*Hidden Markov Models*, do inglês), empregando grandes conjuntos de descritores de áudio, para modelar as características vocais [Regnier e Peeters 2009]. No entanto, mais recentemente, houve um avanço significativo com o uso de redes neurais profundas, incluindo redes convolucionais (*Convolutional Neural Networks* - CNNs, do inglês) e redes neurais recorrentes (*Recurrent Neural Networks* - RNNs, do inglês) [Zhang et al. 2020, Bonzi et al. 2023].

Essas técnicas de aprendizado profundo, especialmente a combinação de CNNs e redes recorrentes de longa memória (*Long Short-Term Memory* - LSTM, do inglês), juntamente com aplicação de técnicas de separação de fonte musical (*Music Source Separation* - MSS, do inglês), obtiveram resultados que atigiram o estado da arte na detecção de voz cantada, superando as abordagens tradicionais [Zhang et al. 2020].

Assim como em [Bonzi et al. 2023], este trabalho propõe o uso de uma das técnicas mais atuais de aprendizado profundo, a rede Transformer, aplicada a dois conjuntos de dados: Jamendo Corpus [Ramona et al. 2008] e DALI [Meseguer-Brocal et al. 2018], sendo este último, conforme nossas pesquisas, inédito para a tarefa de detecção de voz cantada. Para demonstrar os benefícios do pré-processamento, realizamos dois tipos de treinamentos: o primeiro usando as bases de dados sem qualquer processamento, ou seja, com as músicas cruas contendo a mistura entre vocais e instrumentais; e o segundo, utilizando os dados pré-processados pelo Demucs [Rouard et al. 2022], atualmente o estado da arte em MSS, isolando os vocais. Dessa forma, avaliamos como a separação de fontes pode melhorar o desempenho do modelo ao eliminar interferências instrumentais. A comparação entre diferentes conjuntos de dados e abordagens busca explorar como a diversidade e a qualidade dos dados influenciam os resultados obtidos.

2. Trabalhos Relacionados

Neste trabalho, foi ultizada como base referencial diversos artigos que focam na aplicação de aprendizado profundo, especificamente com o uso de CNNs e ou redes recorrentes, para solucionar o problema da detecção de voz cantada. O objetivo foi analisar os dados empregados, os tipos de características acústicas, ou atributos (*features*, do inglês) extraídas e os resultados alcançados pelos autores, a fim de identificar abordagens eficazes e os desafios enfrentados nessa tarefa.

Uma das primeiras contribuições significativas para a redução do tempo de latência e aumento da resolução temporal foi o método introduzido por [Lehner et al. 2015], que utilizou redes neurais recorrentes de longa memória (LSTM-RNN). Este método foi projetado para oferecer um desempenho em tempo real, tornando-o especialmente adequado para aplicações de baixa latência.

Outro trabalho relevante é o de [Leglaive et al. 2015], que propõe o uso de redes recorrentes bidirecionais (*Bidirectional Long Short-Term Memory* - BLSTM, do inglês). Este modelo consegue considerar tanto o contexto temporal passado, quanto o futuro, ao classificar a presença de canto, melhorando significativamente a precisão do modelo. Dessa forma, [Lee et al. 2018, Lehner et al. 2015] evidenciam a importância dos modelos recorrentes, que conseguem capturar as dependências temporais de forma mais completa, reforçando o papel das RNNs, em particular das variantes LSTM e BLSTM, no avanço das técnicas de detecção de voz cantada.

Além dos avanços em redes recorrentes, a extração de características de áudio, juntamente com as CNNs, têm sido essencial para o sucesso de muitos modelos. Trabalhos como [Choi et al. 2016, Choi et al. 2017], demonstram a efetividade de representações como espectrogramas, mel-espectrogramas e coeficientes cepstrais de frequência mel (MFCC), na captação dos principais aspectos sonoros, essenciais para a detecção de voz cantada. Atualmente, essas representações têm sido amplamente utilizadas em modelos de aprendizagem profuda para manipulação de músicas e áudios.

Na fronteira da pesquisa em SVD, [Zhang et al. 2020] atingiu o estado da arte, ao combinar CNN com redes LSTM em um modelo denominado *Long-Term Recurrent Convolutional Network* (LRCN). O diferencial desse estudo foi a aplicação de pré-processamento para filtrar a voz cantada, a partir da mistura de áudio e a validação de diferentes conjuntos de características, como espectrogramas e MFCC, em dados do conjunto Jamendo. Esse trabalho evidenciou a eficácia da fusão entre CNNs, para a extração de características, e LSTMs, para capturar a relação temporal, estabelecendo um novo padrão de desempenho na área.

Ampliando o trabalho de [Zhang et al. 2020], o estudo de [Bonzi et al. 2023] demonstrou a eficácia do uso do Demucs, uma técnica de separação de fontes musicais (MSS), para isolar a voz cantada da música de fundo. Utilizando o Demucs (segunda versão), o estudo aplicou essa técnica no modelo LRCN, além de testar um novo modelo que combinava CNNs com Transformers. Este foi o primeiro trabalho a introduzir Transformers na tarefa de SVD, segundo nossa pesquisa na literatura até então. Alguns resultados superaram os obtidos por [Zhang et al. 2020], tanto em termos de precisão, quanto de F1-score, mostrando os benefícios do uso de Transformers para esse problema. O estudo também sugere que pesquisas futuras devem explorar formas de refinar as técnicas de pré-

processamento de dados, e abordar desafios relacionados à diversidade e qualidade das anotações nos conjuntos de dados utilizados, para potencialmente melhorar ainda mais o desempenho dos modelos.

Esses estudos, detalhados no Quadro 1, refletem o progresso contínuo na área de SVD, destacando tanto o papel das RNNs e CNNs, como o uso de técnicas de extração de atributos. Embora os avanços tenham sido consideráveis, [Lee et al. 2018] sugere que ainda há desafios a serem resolvidos.

Deste modo, um dos principais diferenciais do nosso trabalho é o uso do banco de dados DALI, um conjunto de dados consideravelmente maior do que o Jamendo, que apesar de sua alta qualidade e balanceamento, não representa a diversidade de um cenário real. O DALI, por outro lado, por ter sido gerado automaticamente por inteligência artificial, possui anotações de rótulos de canto menos precisos e uma qualidade geral inferior, o que torna a tarefa mais desafiadora. Além disso, utilizamos a versão mais recente do Demucs para pré-processamento aplicada no modelo CNN-Transformer, proposto por [Bonzi et al. 2023]. Permitiu-se, com isso, avaliar como o modelo se comporta em um ambiente mais desafiador, aproximando-se de situações do mundo real.

Quadro 1. Resumo sobre os trabalhos relacionados analisados.

Autor	Objetivo	Algoritmo	Dataset	Métrica
[Lehner et al. 2015]	Validar o modelo proposto de Redes Neurais Recorrentes do tipo Long Short-Term Memory (LSTM-RNN).	LSTM-RNN	Conjunto de dados interno, Jamendo e RWC	Acurácia, Precisão, Sensibilidade e F1-score
[Leglaive et al. 2015]	Propor um novo método para detecção de voz cantada utilizando uma RNN do tipo BLSTM.	BLSTM-RNN	Jamendo	Acurácia, Precisão, Sensibilidade e F1-score
[Choi et al. 2016]	Desenvolver um algoritmo de rotulamento automático de músicas com base em conteúdo, utilizando redes neurais convolucionais totalmente conectadas (FCNs).	FCN	MagnaTagATune e *Million Song Dataset* (MSD)	Área abaixo da curva (*Area under the curve* - AUC, do inglês)
[Choi et al. 2017]	Demonstrar que esse recurso extraído pela convnet pode servir como uma representação musical de propósito geral.	Convnet	Jamendo	Acurácia
[Zhang et al. 2020]	Desenvolver um modelo utilizando uma rede *Long-term Recurrent Convolutional Network* (LRCN), que combina a extração de características por meio de uma CNN e o aprendizado de relações temporais com camadas LSTM.	LRCN	RWC, Jamendo, MIR1k, iKala e MedleyDB	Acurácia, Precisão, Sensibilidade e F1-score
[Bonzi et al. 2023]	O trabalho propõe um novo sistema de SVD combinando o Demucs para pré-processamento, com dois modelos LRCN e Transformer, afim de destacar o impact do Demucs e mostrar o potencial de aprendizagem profunda.	LRCN e CNN-Transformer	Jamendo, MedlyDB e MIR-1K	Acurácia, Precisão, Sensibilidade e F1-score
Este estudo	Validar o desempenho de um modelo de aprendizagem profunda baseado em Transformer, aplicando uma nova versão do Demucs para separar fontes musicais, utilizando o conjunto de dados DALI.	CNN-Transformer	Jamendo e DALI	Acurácia, Precisão, Sensibilidade e F1-score

3. Metodologia

Nesta seção, será apresentada a metodologia aplicada para alcançar os resultados deste estudo. Inicialmente, abordamos os conjuntos de dados utilizados e as etapas de pré-processamento essenciais para preparar os dados para o modelo. Em seguida, descrevemos o ambiente de desenvolvimento configurado para a implementação do projeto, as métricas escolhidas para avaliação de desempenho e detalhes dos modelos empregados. Por fim, explicamos o processo de treinamento e a etapa de validação.

3.1. Base de dados

O Jamendo Corpus [Ramona et al. 2008], utilizado neste estudo, é um conjunto de dados aberto que oferece 93 músicas, totalizando aproximadamente 7 horas de áudio em estéreo, com uma taxa de amostragem de 44,1kHz. As faixas foram rotuladas manualmente, garantindo a alta qualidade dos dados. Os arquivos são de diferentes artistas e representam uma variedade de gêneros da música comercial populares. A separação do conjunto já foi realizada pelos autores, com 61 músicas reservadas para treinamento, 16 para validação e 16 para teste. Além disso, o conjunto de dados é balanceado, contendo um número quase igual de quadros com canto e sem canto.

O segundo conjunto de dados utilizado foi o DALI [Meseguer-Brocal et al. 2018], uma base de dados multimodal, que inclui informações adicionais como gênero, idioma, músico, capas de álbuns e links para vídeoclipes. Composto por 5.358 faixas de áudio, o DALI oferece notas de melodia vocal e letras alinhadas temporalmente, como mostrado na Figura 1, organizadas em quatro níveis de granularidade: notas, paralavras, linhas e parágrafos. O DALI foi criado automaticamente, por meio do paradigma de aprendizagem de máquina professor-aluno (*teacher-student*, do inglês). Devido ao seu tamanho, foram selecionadas 1.278 músicas para este experimento, com uma separação aleatória de 80% para treinamento, 10% para validação e 10% para testes. Todas as músicas foram baixadas com uma taxa de amostragem de 16kHz (taxa de amostragem padrão do modelo). Por ser um conjunto de dados mais extenso que o Jamendo, tiraremos proveito de sua maior quantidade para avaliar o desempenho do nosso modelo em bases de dados maiores. No entanto, em comparação ao Jamendo, que foi rotulado manualmente, a qualidade do DALI pode ser inferior, o que nos permitirá também testar a robustez do modelo em um cenário com esse tipo de dado.

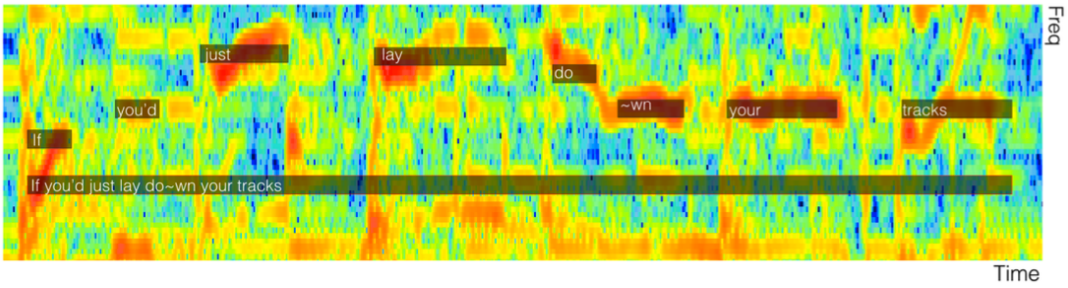

Figura 1. Exemplo de espectrograma com alinhamento temporal das melodias vocais e letras de um pequene pedaço de uma música do dataset DALI. Adaptado de [Meseguer-Brocal et al. 2018]

3.2. Pré-processamento

Na etapa de pré-processamento, aplicamos uma redução de amostragem nos dois conjuntos de dados, Jamendo e DALI para uma taxa de 16kHz. Além disso, todas as faixas foram convertidas para formato monofônico. Utilizamos a quarta versão do Demucs [Rouard et al. 2022], atualmente considerado o estado da arte na tarefa de separação de fontes musicais, para isolar exclusivamente os vocais. Dessa forma, estudos anteriores, como [Zhang et al. 2020, Bonzi et al. 2023] mostraram que essa abordagem é essencial para melhorar o desempenho em tarefas de detecção de canto, pois a separação de vozes pode eliminar ou atenuar significativamente o acompanhamento instrumental. Isso resulta em uma extração de voz mais pura, permitindo que o modelo se concentre de maneira mais eficaz nas seções vocais, aumentando a precisão da detecção.

Para alimentar o modelo de detecção de voz cantada, utilizamos uma estratégia consolidada de extração de atributos, que é amplamente aplicada em tarefas envolvendo áudio ou voz. A abordagem envolve o concatenamento de diferentes tipos de características acústicas, permitindo ao modelo capturar aspectos diversos do sinal de áudio.

Neste estudo, foram extraídos os seguintes características: Coeficientes Cepstrais de Frequência Mel (MFCC), estatísticas espectrais (centroide, *roll-0ff*, *band-width*, contraste e *flatness*) e predição linear perceptual (*Perceptual Linear Prediction* - PLP, do inglês). Essa estratégia foi validada em trabalhos anteriores, como o de [Zhang et al. 2020], e demonstrou ser eficiente em tarefas relacionadas à análise de voz cantada.

O processo de extração começa com a divisão do sinal de áudio em pequenos quadros sobrepostos, facilitando a análise de suas características temporais e espectrais. O tamanho do quadro foi definido como 2.048 amostras (1,28 segundos) e a sobreposição entre quadros foi de 1.536 amostras (0,96 segundos). Essa segmentação permite capturar detalhadamente as variações de frequência e amplitude do sinal, além de garantir que as informações nas bordas dos segmentos não sejam perdidas.

3.3. Ambiente de desenvolvimento

Na implementação deste trabalho, utilizamos a linguagem de programação Python 3.9 e fazendo o uso de diversas bibliotecas amplamente conhecidas na área de aprendizado de máquina e processamento de áudio. As bibliotecas utilizadas foram:

- `PyTorch` - versão 2.4.0: Biblioteca *open-source* para o desenvolvimento de modelos e processamento via GPU;
- `Numpy` - versão 2.1.2: Biblioteca *open-source* para processamento de dados multidimensionais e matrizes;
- `Scipy` - versão 1.14.1: Biblioteca *open-source* para computação científica e análise de dados;
- `Lightning` - versão 2.4.0: Framework *open-source* para simplificar o treinamento de modelos PyTorch e gerenciar rotinas de treinamento complexas;
- `Librosa` - versão 0.10.2.post1: Biblioteca *open-source* especializada em processamento de áudio, especialmente para a extração de características sonoras;
- `Scikit-learn` - versão 1.5.2: Biblioteca *open-source* para o desenvolvimento de modelos de aprendizagem de máquina.

Para o desenvolvimento e execução do código, utilizamos o ambiente na nuvem fornecido pelo Google Colab. A máquina disponibilizada contou com as seguintes especificações de hardware:

- Processador: Intel® Xeon® CPU @ 2.30GHz;
- Memória RAM: 51 GB;
- Placa de vídeo: NVIDIA Tesla T4, 15 GB de VRAM.

3.4. Métricas

As métricas desempenham um papel fundamental neste estudo, pois fornecem uma maneira objetiva de avaliar o desempenho do modelo. Para a avaliação, utilizamos as seguintes métricas: sensibilidade, acurácia, precisão e F1-score [Mesaros et al. 2016]. Cada uma delas é descrita a seguir, juntamente com suas fórmulas matemáticas.

3.4.1. Acurácia

A acurácia, descrita pela Equação (1), é a proporção de previsões corretas, tanto para as classes positivas, quanto para as negativas, em relação ao total de exemplos avaliados, sendo uma métrica geral de desempenho.

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Onde:

- TP = verdadeiros positivos (*True Positives*, do inglês)
- TN = verdadeiros negativos (*True Negatives*, do inglês)
- FP = falsos positivos (*False Positives*, do inglês)
- FN = falsos negativos (*False Negatives*, do inglês)

3.4.2. Precisão

A precisão, ilustrada na Equação (2), indica quantas das previsões positivas feitas pelo modelo realmente pertencem à classe positiva. Essa métrica é especialmente importante quando os falsos positivos têm um custo elevado.

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{2}$$

3.4.3. Sensibilidade

Também conhecido como *recall*, do inglês, sensibilidade, ou taxa de verdadeiros positivos, essa métrica avalia a capacidade do modelo de identificar corretamente as amostras positivas, ou seja, a proporção de casos verdadeiramente positivos que foram corretamente classificados. Sua definição é apresentada na Equação (3).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

3.4.4. F1-Score

O F1-score, conforme definido na Equação(4), é a média harmônica entre precisão e *recall*, equilibrando as duas métricas. Ela é útil quando existe um desequilíbrio entre as classes, pois reflete tanto a precisão, quanto o *recall*, em uma única métrica.

$$\text{F1 Score} = 2 \times \frac{Precisão \times Recall}{Precisão + Recall} \tag{4}$$

Cada uma dessas métricas oferece uma visão complementar sobre o desempenho do modelo, sendo o F1-score particularmente relevante em cenários com classes desbalanceadas, enquanto a acurácia pode ser útil em conjuntos de dados mais balanceados.

3.5. Modelo

Nesta seção, abordamos o modelo desenvolvido para o estudo, inspirado nas abordagens de [Zhang et al. 2020] e [Bonzi et al. 2023], que combinam CNN e Transformers, para realizar a tarefa de detecção de voz cantada.

As redes CNNs são amplamente reconhecidas por sua capacidade de extrair características espaciais com eficácia. Para esta tarefa, elas desempenham um papel crucial, ao identificar padrões e atributos acústicos no espectrograma de áudio, que são fundamentais para melhorar o desempenho do modelo.

Por outro lado, os Transformers têm se mostrado altamente eficazes no processamento de sequências de dados [Lin et al. 2021, Vaswani et al. 2023], o que os torna ideais para capturar as relações temporais entre diferentes trechos do áudio. Sua arquitetura, baseada em mecanismos de *self-attention*, do inglês, permite que o modelo foque nas partes mais relevantes da sequência de dados, melhorando a precisão da classificação.

Sendo assim, o modelo desse estudo combina essas duas abordagens. Ele recebe como entrada as características acústicas extraídas dos dados de áudio, que são processados por quatro camadas de CNNs, responsáveis pela extração espacial das características. Em seguida, essas características são passadas por duas camadas do Encoder do Transformer, que explora as dependências temporais do áudio, por meio de mecanismos de *self-attention*. Finalmente, o modelo conta com três camadas lineares completamente conectadas, que fazem a classificação final de cada trecho de áudio em uma classe binária (1, para voz cantada e 0, para ausência de voz cantada). A topologia do modelo pode ser vista na Figura 2.

3.6. Etapa de Treinamento

Na etapa de treinamento, utilizamos dois métodos distintos para avaliar o impacto do pré-processamento na detecção de voz cantada. O primeiro treinamento foi realizado com os conjuntos de dados em seu estado bruto, com as músicas contendo a mistura de vocais e instrumentais. Já no segundo treinamento, utilizamos os dados pré-processados, apenas com os vocais. Ambas as abordagens foram aplicadas aos dois conjuntos de dados escolhidos, Jamendo Corpus e DALI, permitindo uma análise comparativa dos resultados obtidos em cenários com e sem pré-processamento.

O modelo foi treinado utilizando um *batch size*, do inglês, de 32, o que proporcionou um equilíbrio adequado entre eficiência computacional e qualidade de aprendizado.

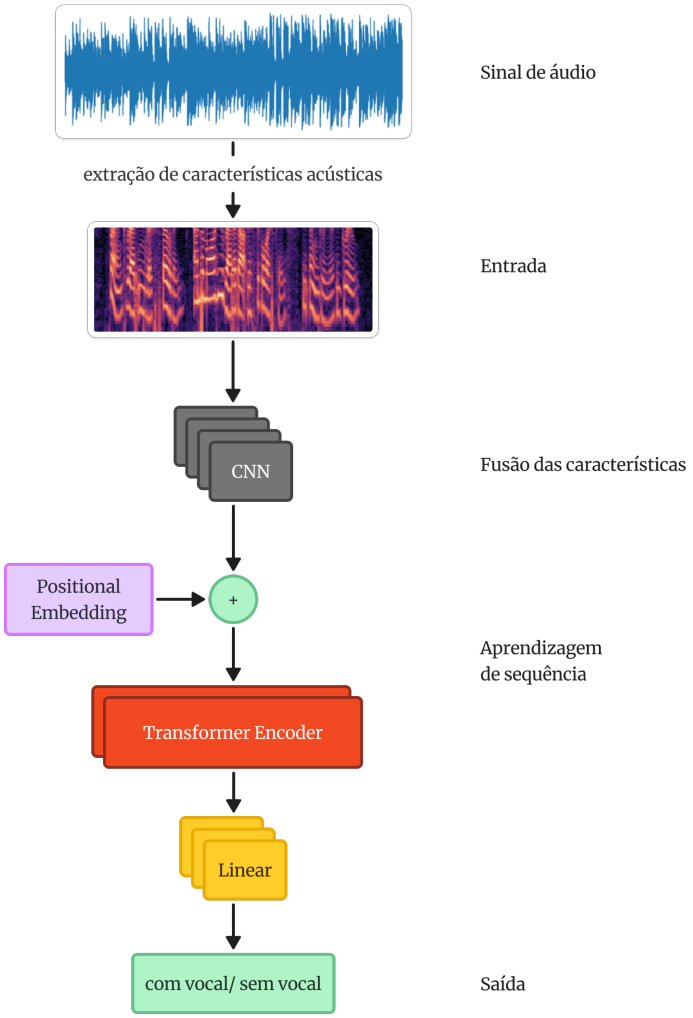

Figura 2. Topologia do modelo CNN-Transformer. Adaptado de [Zhang et al. 2020, Bonzi et al. 2023]

A taxa de aprendizagem inicial foi definida em 0,0001, uma taxa baixa que garante a estabilidade do processo de treinamento, prevenindo saltos bruscos nas atualizações dos pesos e ajudando o modelo a convergir de forma mais controlada. Porém, uma estratégia de ajuste dinâmica dessa taxa foi incorporada, monitorando a curva de perda ao longo do tempo. Quando o modelo atingia um platô na redução da perda, a taxa de aprendizado era reduzida em 50%, permitindo que o modelo encontrasse um equilíbrio ideal entre estabilidade e velocidade de convergência.

Além disso, foi aplicada uma taxa de *drop-out*, do inglês, de 0,5 nas camadas, ajudando a regularizar o treinamento e prevenir o sobreajuste (*overfitting*, do inglês). Essa técnica permite que o modelo generalize melhor, desativando temporariamente algumas conexões durante o treinamento.

Por fim, a quantidade de épocas para o treinamento foi ajustada com base na observação das curvas de perda após cada experimento, visando encontrar o ponto em que o modelo estabilizava seu aprendizado sem *overfitting*. Para o conjunto de dados Jamendo, o treinamento com a versão misturada foi realizado por 35 épocas, enquanto

a versão contendo apenas vocais foi executada por 60 épocas. Já para o DALI, o treinamento com os dados misturados foi conduzido por 40 épocas, enquanto a versão contendo apenas vocais foi submetida a 50 épocas.

3.7. Etapa de validação

A avaliação do modelo foi realizada através dos métodos do `PyTorch Lightning`, que facilitaram o processo de validação. Para isso, o *Data Loader* de validação foi utilizado, garantindo que o modelo fosse testado periodicamente durante o treinamento.

Durante a validação, o modelo fez previsões com base nas amostras de entrada e comparou essas previsões com os valores separados de validação, utilizando a função de perda binária cruzada (*binary cross-entropy*, do inglês). Essa função foi responsável por calcular a perda de validação em cada etapa. Além disso, foram calculadas as métricas de acurácia, F1-score, precisão e sensibilidade.

4. Resultados e Discussões

A seção de Resultados e Discussões apresenta uma análise comparativa do desempenho do modelo utilizando duas bases de dados distintas, Jamendo e DALI, e considerando dois tipos de entradas de áudio: a música a "crua"(misturada com instrumentais) e a versão pré-processada, contendo apenas os vocais. Os resultados são detalhados na Tabela 2, e a seguir discutimos as diferenças e implicações dos achados.

Tabela 2. Resultados obtidos pelo modelo em dois tipos de entrada de dados

Entrada de Audio	Base de dados	Acurácia	Precisão	Sensibilidade	F1-score
Mistura	Jamendo	,677	,663	,930	,774
Vocais	Jamendo	,872	,910	,872	,900
Mistura	DALI	,685	,609	,838	,705
Vocais	DALI	,757	,703	,912	,794

Nos experimentos com o conjunto de dados Jamendo, observamos uma diferença significativa no desempenho do modelo, ao comparar a entrada de áudio misturada com a versão contendo apenas vocais. Ao usar apenas vocais, o modelo alcançou uma acurácia de 0,872, uma precisão de 0,910 e um F1-score de 0,900, superando os resultados da versão misturada, que teve uma acurácia de 0,677, cerca de 20% de melhoria, e um F1-score de 0,774. Isso sugere que a separação de fontes, ao remover o acompanhamento instrumental, permitiu ao modelo focar de maneira mais eficaz nos trechos de voz cantada, melhorando sua capacidade de identificar corretamente os vocais.

As curvas de perda para o Jamendo, mostradas na Figura 3, refletem esse comportamento. Para o conjunto de dados pré-processado, contendo apenas vocais, a curva de validação mostrou uma queda consistente ao longo de várias épocas, com um aumento apenas ao redor da época 60, indicando que o modelo continuava a melhorar, até as últimas etapas do treinamento. Em contraste, no conjunto de dados com as músicas cruas, a curva de validação começou a subir mais rapidamente, por volta da época 30, sinalizando um menor ajuste do modelo e uma maior dificuldade em alcançar uma estabilidade semelhante.

No conjunto de dados DALI, a diferença entre as versões misturadas (vocais e instrumentos) e com apenas vocais também foi evidente, mas menos pronunciada. Para a versão contendo apenas vocais, o modelo obteve uma acurácia de 0,757 e um F1-score de 0,794, enquanto a versão misturada apresentou uma acurácia de 0,685 e um F1-score de 0,705. Apesar do desempenho ser inferior ao do Jamendo, o DALI ainda demonstra que a separação de fontes beneficia a detecção de vocais, especialmente em termos de sensibilidade, que alcançou 0,912 na versão de vocais.

Ao analisar as curvas de perda para o conjunto DALI, Figura 4, observou-se uma diferença na estabilidade do aprendizado. A curva de validação para o áudio misturado foi consideravelmente mais instável, sugerindo dificuldades maiores para o modelo em distinguir corretamente entre vocais e instrumentais. Além disso, ambas as curvas atingiram um platô por volta da época 35, indicando que o modelo havia alcançado sua capacidade de aprendizado máxima para esses dados.

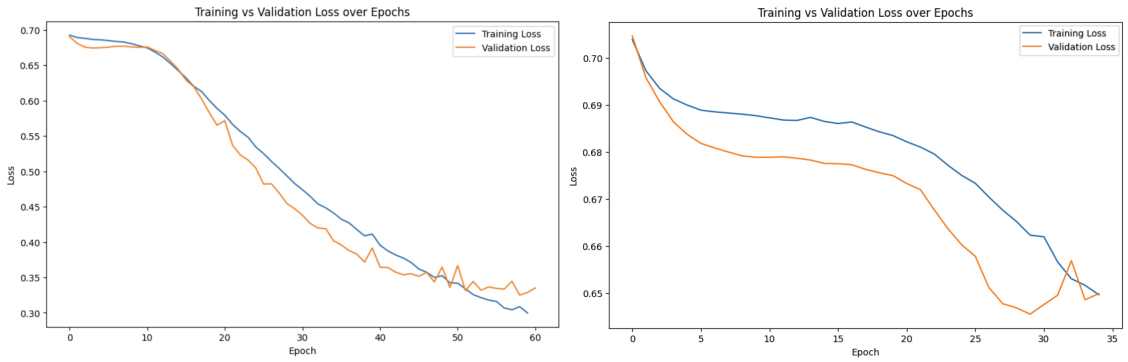

Figura 3. Curvas de perda para o dataset Jamendo: à esquerda, o gráfico representa o conjunto de dados pré-processado contendo apenas os vocais isolados; à direita, as curvas correspondem às músicas com vocais e instrumentais misturados.

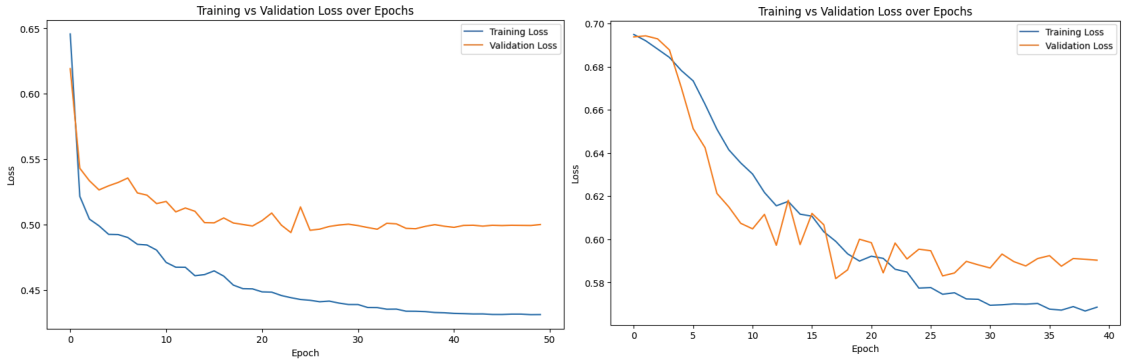

Figura 4. Curvas de perda para o dataset DALI: à esquerda, o gráfico representa o conjunto de dados pré-processado contendo apenas os vocais isolados; à direita, as curvas correspondem às músicas com vocais e instrumentais misturados.

O Jamendo apresentou consistentemente melhores resultados, devido, provavelmente, à sua maior qualidade de anotação. Esse conjunto de dados foi rotulado manualmente e balanceado, o que significa que as características dos cantores e gêneros musicais

foram equilibradas, oferecendo uma base sólida para treinamento e validação. Esse processo cuidadoso de curadoria pode ter contribuído para o desempenho superior do modelo, quando comparado ao DALI, particularmente em termos de precisão e F1-score.

Por outro lado, o DALI, apesar de ter sido rotulado automaticamente por inteligência artificial e possuir uma qualidade inferior, oferece vantagens significativas para a avaliação do modelo. Ele contém uma grande quantidade de músicas em termos de idioma, gênero e características vocais, além de ser um conjunto de dados muito mais extenso. Avaliar o modelo com o DALI permitiu testar sua robustez em cenários mais desafiadores, onde a qualidade das anotações e o equilíbrio das classes não são garantidos. Embora o desempenho do modelo tenha sido inferior no DALI, o uso dessa base de dados é benéfico para verificar como o modelo generaliza em um ambiente mais variado e realista, sugerindo que sua robustez pode ser ampliada com um volume maior de dados, mesmo que não sejam tão refinados quanto os do Jamendo.

5. Conclusão

Os resultados deste estudo indicam que a combinação de redes CNN e Transformers, aliada a técnicas avançadas de separação de fontes como o Demucs, é uma abordagem eficaz para a detecção de voz cantada. O modelo apresentou um desempenho satisfatório, especialmente em conjuntos de dados de maior qualidade, com destaque para os treinamentos onde os vocais foram isolados. A qualidade superior do Jamendo, com anotações manuais e balanceamento de características dos vocais, contribuiu para os melhores resultados gerais, enquanto o DALI também se mostrou valioso para testar a robustez do modelo, oferecendo um panorama mais realista das condições de implementação prática.

5.1. Trabalhos futuros

Há várias direções promissoras para pesquisas futuras. Uma possibilidade é explorar técnicas de aumento de dados (*data augmentation*, do inglês) para aumentar a diversidade de amostras de treino, o que pode melhorar ainda mais a robustez do modelo, especialmente em cenários com dados limitados ou desequilibrados. Além disso, o uso de outros conjuntos de dados, como o banco de dados DAACI-VoDAn [Cuesta et al. 2023], que foi desenvolvido especificamente para a melhoria na detecção de voz cantada e contém 706 músicas, pode trazer um impacto significativo. Esse conjunto de dados especializado poderia elevar ainda mais os resultados, fornecendo ao modelo um treinamento com dados de alta qualidade.

Referências

[Berenzweig e Ellis 2001] Berenzweig, A. e Ellis, D. P. W. (2001). Locating singing voice segments within music signals. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, pages 119–122.

[Bonzi et al. 2023] Bonzi, F., Mancusi, M., Deo, S. D., Melucci, P., Tavella, M. S., Parisi, L., e Rodolá, E. (2023). Exploiting music source separation for singing voice detection. In *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

[Choi et al. 2016] Choi, K., Fazekas, G., e Sandler, M. (2016). Automatic tagging using deep convolutional neural networks.

[Choi et al. 2017] Choi, K., Fazekas, G., Sandler, M., e Cho, K. (2017). Transfer learning for music classification and regression tasks.

[Cuesta et al. 2023] Cuesta, H., Kroher, N., Pikrakis, A., e Djordjevic, S. (2023). Daacivodan: Improving vocal detection with new data and methods. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 136–140.

[Kim e Whitman 2002] Kim, Y. E. e Whitman, B. (2002). Singer identification in popular music using warped linear prediction. In *International Society for Music Information Retrieval Conference*.

[Lee et al. 2018] Lee, K., Choi, K., e Nam, J. (2018). Revisiting singing voice detection: a quantitative review and the future outlook.

[Leglaive et al. 2015] Leglaive, S., Hennequin, R., e Badeau, R. (2015). Singing voice detection with deep recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 121–125.

[Lehner et al. 2015] Lehner, B., Widmer, G., e Bock, S. (2015). A low-latency, real-time-capable singing voice detection method with lstm recurrent neural networks. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 21–25.

[Lin et al. 2021] Lin, T., Wang, Y., Liu, X., e Qiu, X. (2021). A survey of transformers.

[McVicar et al. 2014] McVicar, M., Ellis, D. P. W., e Goto, M. (2014). Leveraging repetition for improved automatic lyric transcription in popular music. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3117–3121.

[Mesaros et al. 2016] Mesaros, A., Heittola, T., e Virtanen, T. (2016). Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6).

[Meseguer-Brocal et al. 2018] Meseguer-Brocal, G., Cohen-Hadria, A., e Peeters, G. (2018). Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm.

[Ramona et al. 2008] Ramona, M., Richard, G., e David, B. (2008). Vocal detection in music with support vector machines. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1885–1888.

[Regnier e Peeters 2009] Regnier, L. e Peeters, G. (2009). Singing voice detection in music tracks using direct voice vibrato detection. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1685–1688.

[Rouard et al. 2022] Rouard, S., Massa, F., e Défossez, A. (2022). Hybrid transformers for music source separation.

[Vaswani et al. 2023] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., e Polosukhin, I. (2023). Attention is all you need.

[Vijayan et al. 2019] Vijayan, K., Li, H., e Toda, T. (2019). Speech-to-singing voice conversion: The challenges and strategies for improving vocal conversion processes. *IEEE Signal Processing Magazine*, 36(1):95–102.

[Zhang et al. 2020] Zhang, X., Yu, Y., Gao, Y., Chen, X., e Li, W. (2020). Research on singing voice detection based on a long-term recurrent convolutional network with vocal separation and temporal smoothing. *Electronics*, 9(9).