UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelo de conversão de áudio de vídeo em texto para geração de legenda e dublagem

Simone Rodrigues da Silva

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Simone Rodrigues da Silva

Modelo de conversão de áudio de vídeo em texto para geração de legenda e dublagem

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Profa. Dra. Solange Oliveira Rezende
Co-orientador: Brucce Neves dos Santos

Versão original

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S586m

Silva, Simone Rodrigues da
Modelo de conversão de áudio de vídeo em texto para geração de legenda e dublagem / Simone Rodrigues da Silva ; orientadora Profa. Dra. Solange Oliveira Rezende; co-orientador Brucce Neves dos Santos. – São Carlos, 2023.
44 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Rezende, Solange Oliveira, orient. II. dos Santos, Brucce Neves, co-orient. II. Título.

Simone Rodrigues da Silva

Audio to Video Text Conversion Model for Captioning and Dubbing

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Profa. Dra. Solange Oliveira Rezende
Co-Advisor: Brucce Neves dos Santos

Original version

São Carlos

2023

AGRADECIMENTOS

A professora Dra. Solange Oliveira Rezende e Brucce Neves dos Santos, por terem sido meus orientadores e terem desempenhado tal função com dedicação e amizade. Aos professores, pelas correções e ensinamentos que me permitiram apresentar um melhor desempenho no meu processo de formação profissional ao longo do curso. Em especial aos alunos Arosti Iskandar Nahas, Ézio José de Oliveira Rego e Letícia Cruz da Silva (Gang da 2° turma do MBA) pela parceria para superar os obstáculos. Agradeço também a minha família que foram compreensivos, entendendo a necessidade de dedicar uma parcela do tempo disponível nesta empreitada.

"Infeliz quem amou apenas corpos, formas e aparências. A morte vai tirar tudo dele. Tente amar as almas e um dia você as encontrará novamente."

Wei Wuxian, Mo Dao Zu Shi

RESUMO

Silva, S. R **Modelo de conversão de áudio de vídeo em texto para geração de legenda e dublagem**. 2023. 44p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

O Brasil está vivenciando uma nova onda de crescimento no consumo de produtos e serviços oriundos de países asiáticos, como o Japão, Coreia do Sul e China, que incluem música, filmes, séries e jogos, popularizando assim a cultura pop asiática. O número de consumidores de serviços de *streaming* especializados em conteúdos asiáticos tem aumentado consideravelmente. Por exemplo, a plataforma Crunchyroll, especializada nesse segmento, possui um amplo acervo com cerca de 800 séries, mais de 200 doramas e mais de 50 títulos de mangás. Até 2021, a plataforma já tinha ultrapassado 5 milhões de assinantes em todo o mundo. Esse aumento no consumo é evidenciado pela disponibilidade de mais títulos em plataformas não especializadas, como Netflix e Amazon, seja por meio de produções próprias ou licenciamentos. No entanto, apenas uma parte desses produtos possui legendas ou opções de dublagem em português, normalmente reservadas para os títulos de maior audiência. Isso significa que a maioria dos assinantes precisa assistir a esses conteúdos com legendas ou dublagens em outros idiomas, sendo o inglês a opção mais comum. Vale ressaltar que, no Brasil, o nível de proficiência em inglês, de acordo com o relatório English Proficiency Index - EPI de 2022 da Education First - EF, é classificado como moderado, ocupando a 58ª posição entre 111 países e territórios e a 12ª posição entre os países latinos. No entanto, houve um declínio em relação ao relatório anterior de 2021, onde o Brasil estava na 60ª posição entre 112 países e territórios pesquisados. Com base nesse cenário, surge a necessidade de acelerar a disponibilização de conteúdos legendados e dublados para o público brasileiro. Nesse contexto, propõe-se o uso de algoritmos de Inteligência Artificial (IA) para a geração automática de legendas e dublagem para os áudios desses títulos. O foco deste trabalho está na transcrição e tradução do áudio do idioma original para o português, bem como na geração do áudio da tradução em português.

Palavras-chave: *Streaming*, Reconhecimento Áudio, Conversão de áudio para texto, Conversão de texto para áudio.

ABSTRACT

Silva, S. R **Audio to Video Text Conversion Model for Captioning and Dubbing**. 2023. 44p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Brazil is experiencing a new wave of growth in the consumption of products and services from Asian countries such as Japan, South Korea, and China. This wave includes popular culture products and services, such as music, movies, series, and games. The number of consumers of streaming services specializing in Asian content has grown significantly. For instance, Crunchyroll, a platform specialized in this segment, offers an extensive catalog of approximately 800 series, over 200 doramas, and the option to read more than 50 manga titles. By 2021, it had already surpassed 5 million subscribers worldwide. This increase in consumption is evidenced by the availability of more titles on non-specialized platforms, like Netflix and Amazon, whether through their own productions or licensing agreements. However, some of these products have subtitles or dubbing options in Portuguese, typically reserved for the most popular titles. This means most subscribers must watch this content with subtitles or dubbing in other languages, with English being the most common option. It is worth noting that, in Brazil, English proficiency is classified as moderate, according to the 2022 English Proficiency Index (EPI) report by Education First (EF). Brazil ranks 58th out of 111 countries and territories and 12th among Latin American countries. However, there has been a decline compared to the previous year's report in 2021, when Brazil was ranked 60th out of 112 countries and territories surveyed. Given this scenario, there is a need to expedite the availability of subtitled and dubbed content for the Brazilian audience. In this context, the proposal is to use Artificial Intelligence (AI) algorithms to generate subtitles and dubbing automatically for the audio of these titles. This work focuses on the transcription and translation of the original language audio into Portuguese and the generation of audio for the Portuguese translation.

Keywords: Streaming, Audio Recognition, Audio to text conversion, Text to audio conversion.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

Abral	Associação Brasileira de Licenciamento de Marcas e Personagens
CCXP	Comic Con Experience
EF	Education First
EPI	English Proficiency Index
IA	Inteligência Artificial
BSD	Berkeley Software Distribution
ASR	Automatic Speech Recognition
MFCCs	Mel-Frequency cepstral coeficients
TTS	Text to Speech
STT	Speech to Text
RNC	Rede Neural Convolucional
RNR	Rede Neural Recorrente
PLN	Processamento de linguagem natural
API	Application Programming Interface
MVP	Mínimo Produto Viável
AVI	Audio Video Interleave
MP4/MPEG-4	Formato padrão para container de áudio e vídeo
MP3/MPEG-1/2	Formato para áudio
WAV	Windows Wave formato para áudio
UTF-8	Lista de código universal para codificar todo o conjunto de caracteres Unicode
ISO	International Organization for Standardization
IEC	International Electrotechnical Commission
CD	Compact disk

PC	Personal computer
MIT	Massachusetts Institute of Technology
KHz	Unidade de frequência, equivalente a um milhão de Hertz ou mil kilohertz
MB	Unidade de medida de informação, equivalente a 1024 quilobytes
GIF	Graphics Interchange Format
MPEG	Motion Picture Experts Group
USB	Universal Serial Bus
NLTK	Natural Language Toolkit

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

No Brasil, durante a década de 80, a TV aberta passou a transmitir animes (que são desenhos animados japoneses), o que foi muito bem aceito pelo público e durante a década de 90 teve a sua consolidação, ao qual podemos chamar neste contexto, de primeira onda de popularização de consumo deste tipo de produto, porém durante o ano de 2000 a oferta em massa deste tipo de produção teve um declínio significativo.

Passados alguns anos, a população brasileira passou a ter acesso à internet em massa com velocidade constante e estável o que propiciou o consumo de serviços de *streaming*, o que por sua vez, viabilizou o acesso a estes produtos do universo Pop Asiático, principalmente os animes no Japão, *manhwas* na Coreia do Sul e *manhuas* na China e dramas conhecidos como J-drama (Japão), K-dramas (Coreia do Sul) e os C-dramas (China).

Em eventos dedicado ao tema, o público aproximadamente chegou a 120 mil pessoas no *Anime Friends* (edição de 2022) e a 280 mil no *Comic Con Experience* - CCXP (edição de 2022) por dia conforme informações divulgadas pelos organizadores e estes eventos contam com concursos de *cosplay*, convidados nacionais e internacionais (atores, diretores, dubladores, escritores, mangakás, editoras e etc.), apresentação de grupos musicais e fornecedores de diversos produtos relacionados ao tema. (CODOGNO, 2022) e (RIBNEIRO, 2022).

Este comércio girou em torno de 21,5 bilhões em 2021, segundo dados da Associação Brasileira de Licenciamento de Marcas e Personagens - Abral, em 2020 foi de 21 bilhões e em 2019 de 19 bilhões. O *ticket* médio deste segmento, segundo levantamento realizado pela Rakuten Digital Commerce (multinacional japonesa que é uma das cinco maiores empresas do ramo *e-commerce* no mundo) entre maio de 2018 e abril de 2019, mostrou que este consumidor em seus *tickets* médios gastam R$ 548,00 enquanto a média nacional é de R$ 329,00. A maior representatividade de consumo deste setor se concentra em São Paulo, Minas Gerais, Rio de Janeiro e Distrito Federal mas regiões como Amapá, Maranhão e Pará registraram os maiores *tíckets* médio. (TOLEDO, 2022).

Empresas como o Mercado Livre, identificaram que a tendência é que este mercado siga aquecido, mesmo tendo como ponto negativo para alguns produtos, as altas taxas de impostos de importação e licenciamentos, e uma de suas apostas foi investir em eventos do segmento, como o CCXP 2022 dedicando uma página para os produtos desta categoria contando com 26 mil itens disponíveis. Segundo informações de Fernanda Schmid - diretora de Field Marketing Mercado Livre e Mercado Pago, a página da CCXP teve crescimento

de 188% de maio a julho de 2022 em comparação com o período de outubro a dezembro de 2021. (SCHNAIDER, 2022).

Sobre o aumento no consumo dos animes, uma das justificativas é que estes títulos são fiéis a obra que os originaram (mangás, *manhwas* e *manhuas*) e considerando que a chegada de um título em formato mangá no Brasil é burocrática, existe uma demora muito maior na disponibilidade das obras, ficando este mercado fora dos principais lançamentos e o preço final de cada edição é considerada elevada (média de R$ 35,00 por volume e ser for importado a média é de R$ 150,00 mais impostos), o que torna difícil a aquisição de produtos pelo grande público. Este mesmo cenário não ocorre nas plataformas de *streaming*, ou seja, o mercado recebe primeiro o anime para depois ter disponível (se houver condições favoráveis para a comercialização do título) em formato de mangá, além de que o custo custo médio da mensalidade gira em torno de R$ 35,00 ao mês.

1.2 Justificativa e Objetivos

Neste contexto, o uso deste serviços especializado é a forma mais abrangente de atender ao público, porém algumas plataformas demoram para disponibilizar as legendas ou dublagens no idioma português do Brasil. Em alguns serviços de *streaming*, o tempo de espera em ter estas opções não é informado, seja puramente por falta de informações em seus sites ou ainda existem casos que a legenda é criada de forma colaborativa e não possui um planejamento divulgado. Com a falta desta informação, o público acaba assistindo o título legendado ou dublagem em outro idioma que normalmente em língua inglesa que é a primeira disponibilizada.

Sabendo desta dificuldade do público brasileiro em ter proficiência no idioma do título e também na língua inglesa e pensando no crescimento do consumo destes serviços, o objetivo do trabalho é criar um processo automatizado para legendar e dublar os títulos utilizando a técnica de Inteligência Artificial.

Este mesmo modelo poderá ser reavaliado quanto a sua aplicação em serviços de transcrição de áudios - tais como: *chatbot*, transcrição de denúncias, depoimentos, *call centers* e podem ser expandido incluindo a análise de sentimentos.

1.3 Organização do trabalho

O referencial teórico é apresentado no capítulo 2 com a introdução do modelo de trabalho a ser utilizado para atender aos objetos propostos neste trabalho, incluindo conceitos gerais ou diretrizes que viabilizam a conclusão desta entrega.

2 FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

Neste capítulo, serão apresentados os fundamentos teóricos necessários para a construção desta proposta, assim como os dados de pesquisa, que envolvem a análise de publicações para viabilizar a execução do trabalho proposto, a extração de texto a partir do áudio de um vídeo e sua subsequente tradução para o idioma português do Brasil.

2.1 Transformando áudio em texto

A conversão de áudio em texto, também conhecida como ASR (*Automatic Speech Recognition*), é um processo que envolve a utilização de algoritmos de aprendizado de máquina para transcrever o conteúdo falado em um arquivo de áudio para texto (YU; DENG, 2016). Essa técnica tem sido amplamente aplicada em uma variedade de campos, como assistentes virtuais, legendagem automático de vídeos, transcrição de reuniões e muito mais (MALIK *et al.*, 2021; LI *et al.*, 2022).

Existem várias abordagem para realizar a conversão de áudio em texto usando algoritmo de aprendizado de máquina, sendo uma das mais populares a utilização de modelos de aprendizado profundo, como Redes Neurais Convolucionais (RNC) (ABDEL-HAMID *et al.*, 2014; NASSIF *et al.*, 2019) e Redes Neurais Recorrentes (RNR) (NASSIF *et al.*, 2019; ORUH; VIRIRI; ADEGUN, 2022).

Redes Neurais Artificiais que está ilustrada na Figura 1, são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência. Uma grande Rede Neural Artificial pode ter centenas ou milhares de unidades de processamento; já o cérebro de um mamífero pode ter muitos bilhões de neurônios (SILVA, 2010).

Com base na pesquisa conduzida durante este MBA na área de TI de bancos e varejo, os modelos de Redes Neurais Artificiais utilizados para a conversão de áudio em texto são, em grande parte, fundamentados em RNC ou RNR. Esses modelos servem como a base tanto das aplicações comerciais quanto dos projetos internos. No caso dos projetos internos eles fazem uso da biblioteca ***SpeechRecognition*** na linguagem Python. Enquanto que as aplicações comerciais usadas incluem:

- O ***Filmora*** da Wondershare é um software para edição de vídeos que oferece a funcionalidade de conversão de áudio para texto. Essa opção está disponível nas plataformas *Windows* e *MAC*, permitindo a geração automática de legendas em 16 idiomas.

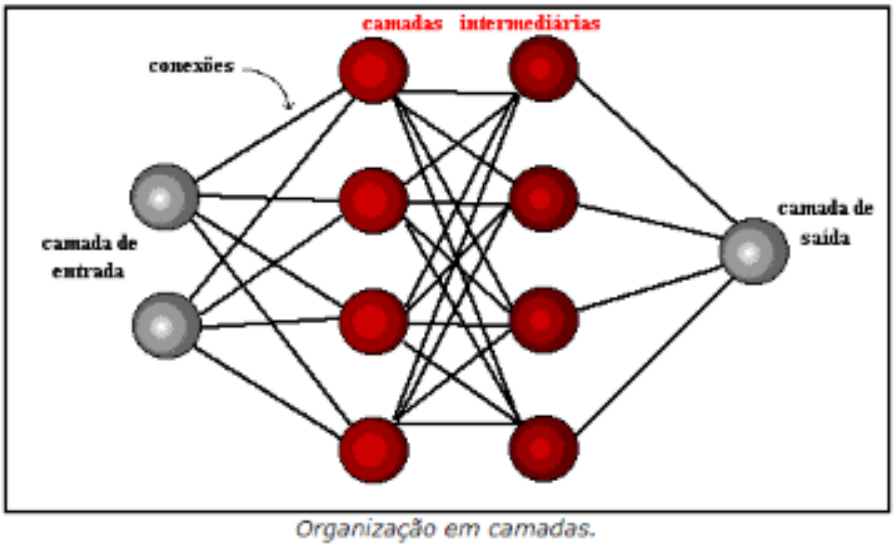

Figura 1 – Representação do esquema de uma Rede Neural Artificial.
Fonte: <https://sites.icmc.usp.br/andre/research/neural/>.

- O ***SubtitleBee*** da SubtitleBee é um *software* voltado para edição de vídeos que oferece a conveniência de gerar legendas automaticamente em até 120 idiomas.
- O ***Veed.IO*** da Veed é um *software* que faz uso da inteligência artificial para criar legendas com uma precisão quase perfeita. Ele funciona como uma ferramenta de geração automática de legendas.

As **RNC** possuem arquiteturas de rede similares às Redes Neurais Artificiais, mas diferem significativamente por serem projetadas especificamente para processar dados em grade, como imagens, usando camadas convolucionais para extrair características. Além disso, elas podem ter uma maior profundidade, e a informação flui através de cada camada, com a saída de uma alimentando a entrada da próxima camada, como é comum em redes neurais (FELTRIN, 2021; LI *et al.*, 2021). Dentro desse tipo de rede, algumas camadas importantes incluem:

Convolução: é responsável por extrair, bem como mapear o conteúdo da informação transformando-o em dados. Esse processo ocorre por meio da aplicação de pequenos blocos, conhecidos como filtros, que permitem a obtenção das informações de sub-blocos do dado;

Pooling: esta camada recebe os blocos que contém as informações extraídas na etapa de convolução. Ela simplifica a informação, resumindo os dados do sub-bloco do dado em um único valor e os repassa para uma camada totalmente conectada;

Camada totalmente conectada: onde é iniciado o processo para classificar as informações extraídas pelas camadas anteriores. A camada totalmente conectada achata o sub-bloco contendo os dados extraídos, ou seja, o bloco é transformado em uma única linha que contém todas as informações extraídas;

Na composição de uma Rede Neural Convolucional, além das camadas de convolução, *pooling* e da camada totalmente conectada, existem outros dois elementos fundamentais: a camada de *dropout* e a função de ativação.

A camada de ***dropout*** tem a função de combater o *overfitting*, que ocorre quando a rede memoriza os dados de treinamento, sendo incapaz de aplicar o conhecimento adquirido a novos dados. Nesse cenário, o modelo resultante da rede treinada não consegue generalizar bem e, portanto, não apresenta um desempenho satisfatório quando usado em produção.

Por outro lado, a função de **ativação** é responsável pelo processo de aprendizado da rede, determinando as relações entre as variáveis e ativando os neurônios relevantes.

Na Figura 2 está ilustrado um exemplo das camadas presentes em uma Rede Neural Convolucional aplicada a um algoritmo de identificação de imagens.

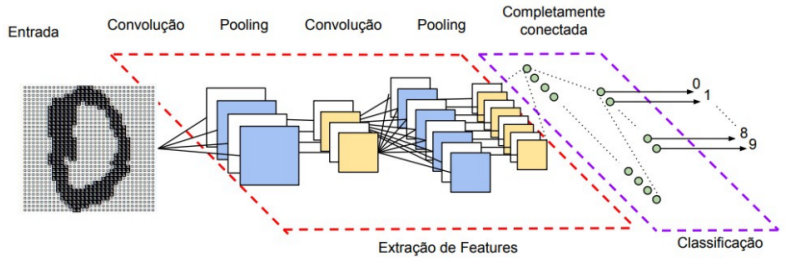

Figura 2 – Exemplo de uma Rede Neural Convolucional.
Fonte: extraída de (VARGAS; PAES; VASCONCELOS, 2016).

As **RNR** diferem-se pela presença de ao menos uma conexão de *feedback* entre os neurônios. Este *loop* permite que o resultado processado no passo anterior componha e contribua com o resultado seguinte, por isso são consideradas redes com memória. Elas também podem ser parcial ou totalmente recorrentes, onde todos os neurônios são conectados entre si. Isso permite que as redes façam processamento temporal, aprendam sequências, realizem reconhecimento de padrões, associação ou predição temporal. Na Figura 3, está ilustrado uma representação gráfica de uma RNR (FELTRIN, 2021).

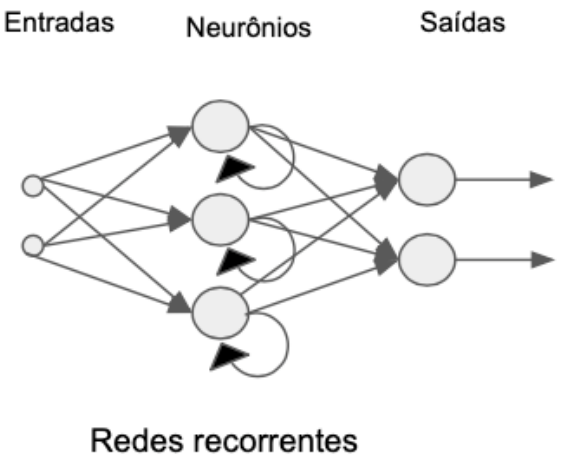

Figura 3 – Exemplo de uma Rede Neural Recorrente.
Fonte: <https://ateliware.com/blog/redes-neurais-artificiais>.

O processo de conversão de áudio em texto geralmente envolve as seguintes etapas:

1. **Pré-processamento de áudio**: o áudio de entrada é normalizado e transformado em uma representação adequada para o modelo de aprendizado de máquina. Isso pode incluir a extração de recursos, como espectrogramas, MFCC's (*Mel-Frequency Cepstral Coefficients*) ou outros descritores relevantes.

2. **Treinamento do modelo de ASR**: nesta etapa, o modelo de aprendizado de máquina é treinado com dados rotulados, que consistem em pares de áudio e texto correspondente. O objetivo é otimizar os parâmetros do modelo para que ele possa mapear o áudio para o texto de forma precisa.

3. **Decodificação após o treinamento**: o modelo é capaz de fazer previsões sobre o texto correspondente a um determinado áudio. A decodificação envolve o uso de algoritmos, como o ***beam search*** - que é um algoritmo de busca heurística que explora um grafo expandindo o nó mais promissor em um conjunto limitado, ou seja, constrói sua árvore de pesquisa e em cada nível da árvore , são gerados todos os sucessores para encontrar a sequência mais provável de palavras dado o áudio de entrada.

4. **Pós-processamento**: a saída da decodificação pode ser refinada por meio de técnicas de pós-processamento, ao qual aplicam modelos de linguagem para melhorar a fluência e coerência do texto gerado, além de correções ortográficas e gramaticais.

É importante destacar que a conversão de áudio em texto usando algoritmos de aprendizado de máquina é um campo ativo de pesquisa, e há constantes avanços visando

melhorar a precisão e a robustez desses sistemas (NASSIF *et al.*, 2019; VAJJALA *et al.*, 2020; LI *et al.*, 2022). Além disso, a disponibilidade de grandes conjuntos de dados de treinamento e o poder computacional cada vez maior têm impulsionado o desenvolvimento de modelos mais sofisticados e eficientes.

Em resumo, a conversão de áudio em texto utilizando algoritmos de aprendizado de máquina é um processo complexo que envolve o treinamento de modelos com grandes quantidades de dados de áudio e texto correspondente. Essa abordagem tem mostrado resultados promissores e é amplamente utilizada em diversas aplicações em que a transcrição automática de áudio se faz necessária.

2.2 Transformando texto para áudio

A conversão de texto para áudio, também conhecida como síntese de fala, é um processo no qual um algoritmo de aprendizado de máquina é utilizado para transformar texto em voz humana artificial. Esse processo envolve a aplicação de técnicas de processamento de linguagem natural (PLN) e síntese de fala e isto envolve a criação de um banco de dados de unidades sonoras individuais (como fonemas, sílabas ou palavras) gravadas por falantes humanos. O algoritmo então seleciona e concatena essas unidades sonoras de acordo com o texto fornecido para gerar a fala sintetizada (NING *et al.*, 2019; VAJJALA *et al.*, 2020).

Uma outra maneira de fazer é usando a síntese de fala baseada em modelos generativos, como Redes Neurais Recorrentes (RNR) ou Redes Neurais Convolucionais (RNC) (CHO *et al.*, 2021; KUMAR; KOUL; SINGH, 2023). Esses modelos são treinados em grandes conjuntos de dados de áudio e texto correspondente, aprendendo a mapear sequências de texto para sequências de áudio. Com base nesse aprendizado, o algoritmo de aprendizado de máquina é capaz de gerar áudio sintetizado com base no texto de entrada.

Além disso, avanços recentes em algoritmos de aprendizado profundo, como redes neurais de *Transformers*, têm sido aplicados com sucesso na síntese de fala, tais como o *Whisper* da OpenAi e IA da *Text-to-Speech* da Google. Esses modelos são capazes de capturar relações complexas entre palavras e gerar uma fala mais natural e fluente.

É importante ressaltar que a qualidade da conversão de texto para áudio depende da qualidade dos dados de treinamento utilizados, bem como do algoritmo de aprendizado de máquina selecionado para o trabalho. Quanto mais dados de fala disponíveis para treinamento e quanto mais avançado for o algoritmo, melhor será a qualidade do áudio sintetizado.

Resumidamente, a conversão de texto para áudio usando modelos de aprendizado profundo (RNC e RNR) que tem possibilitado avanços significativos na geração de fala sintetizada de alta qualidade, com aplicações em assistentes virtuais, sistemas de navegação

por voz, livros falados entre outras aplicações.

2.3 Considerações Finais

Mediante a disponibilidade de ferramentas que possibilitam a conversão de áudio para texto e de texto para áudio, a proposta deste trabalho é realizar a conversão ***Text to Speech* - TTS** e ***Speech to Text* - STT** com o objetivo de acelerar o processo de criar legendas e áudio o idioma português do Brasil para os episódios de animes.

Para embasar este trabalho, foi utilizado diversas fontes de pesquisa, incluindo livros e artigos relacionados ao tema. A seguir, é destacado algumas das referências que serviram como base para esta fundamentação:

- HUANG *et al.* (1993) abordou a complexidade da tarefa de reconhecimento de fala.
- YNOGUTI; VIOLARO (2000) discutiu a complexidade fonética no reconhecimento de fala.
- CAMASTRA; VINCIARELLI (2015) apresentou técnicas de aprendizado de máquina aplicadas à análise de áudio, vídeo e imagem.
- MENEZES NEY COUTINHO (2019) ofereceu ensinamentos relacionados à linguagem Python.
- REIMERS; GUREVYCH (2019) e THAKUR *et al.* (2020) exploraram modelos SBERT, focando na semelhança textual semântica e no desempenho.
- YIN; HENTER (2020) disponibilizou um tutorial sobre a biblioteca Translate.
- ROSA; SILVA (2021) abordou o tema das redes neurais.
- GUNDAVARAPU *et al.* (2022) tratou do reconhecimento de escrita.

3 TRABALHO PROPOSTO

Neste capítulo, são descritos as etapas necessárias para atingir o objetivo deste trabalho, incluindo os conceitos associados e as ferramentas utilizadas. Nosso objetivo principal é selecionar um vídeo, transcrever o áudio em seu idioma original, traduzi-lo e, finalmente, gerar o áudio na língua alvo. Para alcançar isso, propomos um processo de quatro etapas que cada áudio passará antes de ser traduzido.

A primeira etapa consiste na **conversão de tipo de mídia** (Seção 3.1), esta etapa deve ser realizada caso seja necessário adequar o tipo de mídia e garantir a compatibilidade com as bibliotecas utilizadas neste estudo. A segunda etapa é a **transcrição do áudio para texto** (Seção 3.2) aqui é aplicado o modelo de transcrição no idioma original. A terceira etapa é a **tradução do texto** (Seção 3.3) em que realizado a tradução do texto extraído para o idioma selecionado. Por fim temos a etapa de **transformar o texto traduzido em áudio** (Seção 3.4) com a tradução concluída, o modelo irá gerar o áudio no idioma traduzido.

As etapas foram definidas para criar um processo sequencial e facilitar o monitoramento e avaliação dos resultados. As bibliotecas selecionadas foram escolhidas com base na facilidade de implementação e na disponibilidade de versões gratuitas.

Todas as etapas foram desenvolvidas em *Python* que é uma linguagem criada em 1989 por Guido Van Rossum, inicialmente chamada de Modula-3 e posteriormente renomeada. Hoje, *Python* é amplamente adotado por sua legibilidade, versatilidade e aplicabilidade em diversos cenários, incluindo *Data Science*, desenvolvimento *web*, *Back-End*, criação de jogos e *scripts*.

3.1 Conversão de tipo de mídia

Esta etapa é opcional e tem como objetivo adaptar a mídia para a biblioteca a ser utilizada. Após a seleção do arquivo de vídeo, é essencial verificar seu formato (por exemplo: MP4, MP3 ou WAV) e determinar se é necessário realizar uma transcodificação para alcançar o formato esperado por cada ferramenta de conversão.

A operação de **transcodificação** envolve a conversão direta de uma codificação para outra, abrangendo uma variedade de tipos de dados, como arquivos de vídeo (por exemplo, AVI, MP4), arquivos de áudio (por exemplo, MP3, WAV) e codificação de caracteres (por exemplo, UTF-8, ISO/IEC 8859). Geralmente, isso é realizado quando o dispositivo de destino ou o fluxo de trabalho não suporta o formato original, tem restrições de capacidade de armazenamento que exigem um tamanho de arquivo menor ou precisa-se converter dados incompatíveis ou obsoletos em um formato mais moderno e amplamente

suportado. Isso permite a disponibilização do conteúdo em várias mídias que o formato original não permitiria.

A transcodificação é frequentemente um processo que pode resultar em perda de qualidade, mas existem técnicas disponíveis para minimizar essas perdas, seja na compressão ou descompressão do arquivo. É importante notar que a transcodificação é um processo irreversível.

O procedimento típico para realizar a transcodificação envolve a conversão de um arquivo com formato MPEG-4, frequentemente identificado pela extensão oficial .mp4, para o formato WAV (*Waveform Audio File Format*). O formato WAV é um padrão de áudio sem compressão e perdas.

O formato **MPEG-4** foi desenvolvido pela Organização Internacional para Padronização (ISO) e pelo Grupo de Especialistas em Imagens com Movimento. Ele é uma parte do padrão MPEG-4, especificamente a Parte 14, que é uma implementação da especificação ISO/IEC 14496-14:2004. O MPEG-4 é projetado para servir como um contêiner que pode armazenar não apenas vídeos, mas também imagens estáticas e legendas. É amplamente utilizado para armazenar e exibir vídeos com legendas, com aplicações significativas em reprodutores de mídia portáteis e serviços de *streaming*.

O formato **MP3** representa uma eficiente forma de compressão de arquivos de áudio, sem uma perda substancial de qualidade. A sigla MP3 é a abreviação de MPEG-1/2 *Audio Layer* 3, ou simplesmente *Layer*-3 MPEG. A compressão de um arquivo MP3 pode reduzir arquivos de áudio digital sem degradar significativamente a qualidade, uma vez que o sistema auditivo humano tende a não perceber a faixa comprimida ou removida do áudio original.

Quando foi criado na Alemanha em 1987, o MP3 surgiu com o objetivo de reproduzir o som com a qualidade dos CDs, mas com uma taxa de compressão razoável. Para gravar um CD, a taxa de gravação é de aproximadamente 1,4 Megabit por segundo. O MP3 conseguiu reduzir essa taxa para 128 KB/s. Mesmo com uma redução de tamanho de 10 vezes, o MP3 manteve praticamente inalterada a qualidade sonora do arquivo de áudio.

Essa preservação da qualidade sonora dos arquivos MP3 foi possível graças às técnicas de codificação perceptual, um método que comprime apenas as frequências sonoras imperceptíveis ao ouvido humano. Os arquivos de CDs e áudio *Wave* contêm mais dados do que o MP3, incluindo informações que não são captadas pelo sistema auditivo humano. O MP3 descarta essas frequências que não são percebidas pelo ouvido humano, mantendo a qualidade do som, já que elimina apenas o que não seria percebido de qualquer maneira.

O formato WAV, que significa *Windows Wave*, é um formato de áudio e extensão de arquivo criado em colaboração entre a IBM e a Microsoft. Tornou-se o padrão de arquivo

de áudio para PCs e continua sendo o melhor padrão para CDs de áudio. Os arquivos WAV não sofrem perda de qualidade e não são compactados, o que os torna maiores em tamanho em comparação com outros formatos de arquivo de áudio, pois armazenam o áudio bruto. A indústria fonográfica profissional frequentemente utiliza esses arquivos para garantir a mais alta qualidade de áudio possível.

Em relação às diferenças entre os formatos MP4, MP3 e WAV:

- MP4 e MP3 são tecnologias diferentes, apesar de nomes semelhantes. O MP3 (MPEG-2 *Audio Layer* 3) é um *codec* de áudio, enquanto o MP4 é um contêiner que pode conter áudio e vídeo. Em um arquivo MP4, você pode encontrar, por exemplo, o *codec* de áudio MP3 e o *codec* de vídeo MPEG-4 combinados em um único arquivo.

- A principal diferença entre MP3 e WAV é a compressão. Em média, um minuto de música em formato WAV pode ocupar cerca de 10 MB para uma gravação de som em 16 bits estéreo com uma taxa de amostragem de 44.1 kHz. Em contraste, um minuto de música em formato MP3 ocupa cerca de 1 MB. A proporção entre WAV e MP3 é de aproximadamente 10 para 1, o que significa que os arquivos MP3 são significativamente menores em tamanho. Isso ocorre porque o MP3 utiliza algoritmos de compressão que removem informações de áudio imperceptíveis ao ouvido humano, enquanto o WAV mantém o áudio não comprimido e, portanto, maior em tamanho.

Na etapa de conversão de mídia, você pode utilizar as seguintes bibliotecas:

- **MoviePy** é uma biblioteca para edição de vídeo: cortes, concatenações, inserções de títulos, composição de vídeo (também conhecida como edição não linear), processamento de vídeo e criação de efeitos personalizados e que pode ler e gravar todos os formatos mais comuns de áudio e vídeo, inclusive GIF. [1]

- **Pydub** manipula o áudio com uma interface de alto nível. [2]

- **TQDM** que faz tratamento de *loops* e mostra seu progresso. [3]

- **Imageio-ffmpeg** fornece leitura e gravação para uma ampla variedade de formatos de filme, como AVI, MPEG, MP4, etc., bem como a capacidade de ler fluxos de *webcams* e câmeras USB. Ele se baseia na biblioteca ffmpeg e é inspirado e baseado na biblioteca MoviePy. [4]

[1] Licença: MIT License (MIT). <https://zulko.github.io/moviepy/>
[2] Licença: MIT License (MIT). <http://pydub.com/>
[3] Licença: MIT License (MIT). <https://github.com/tqdm/tqdm>
[4] Licença: BSD License (BSD-2-Clause). <https://github.com/imageio/imageio-ffmpeg>

Com o arquivo no formato correto, podemos iniciar o processo de transcrição. Neste TCC, utilizaremos duas bibliotecas para transcrição: a *SpeechRecognition* e a *Whisper*. No caso da *SpeechRecognition*, é necessário converter o vídeo para o formato WAV, como mencionado anteriormente. Já a *Whisper* realiza essa conversão automaticamente, dispensando intervenção adicional.

Na Figura 4 está ilustrado os processos necessários para a conversão de tipo de mídia, realizados para atender aos requisitos da biblioteca *SpeechRecognition*, que trabalha com arquivos no formato WAV. Inicialmente, os arquivos selecionados estavam nos formatos de mídia MP4 e MP3. Para torná-los compatíveis com a biblioteca *SpeechRecognition*, foi necessária uma operação de transcodificação. Para os arquivos em formato MP4, a transcodificação foi realizada em duas etapas, de MP4 para MP3 e de MP3 para WAV. Para os arquivos do tipo MP3 a transcodificação foi realizada em uma única etapa de MP3 para WAV.

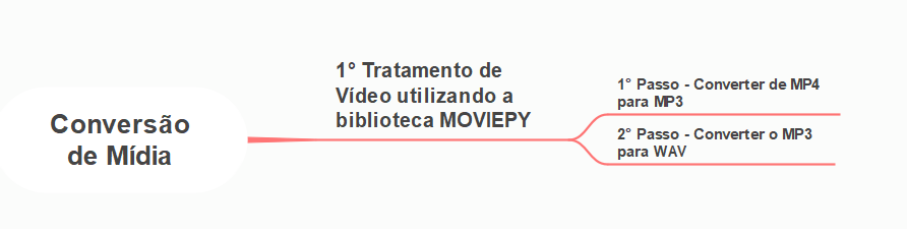

Figura 4 – Conversão de tipo de mídia.
Fonte: a autora.

3.2 Transcrição do áudio para texto

Nesta etapa, realizaremos a operação de transcrição do áudio para texto no idioma original, utilizando algoritmos baseados em redes neurais artificiais. Propomos duas abordagens para essa tarefa:

- A primeira abordagem utiliza a biblioteca *SpeechRecognition*[5], sendo de código aberto. Essa biblioteca oferece acesso por meio de APIs a diversos sistemas de síntese de fala (TTS), como Google e IBM.
- A segunda abordagem utiliza a biblioteca *Whisper*[6] da empresa OpenAI. O *Whisper* utiliza um modelo "*encoder-decoder Transformer*", que é uma rede neural projetada para aprender contextos e significados por meio da análise das relações entre os dados sequenciais, como palavras em uma frase. Esse modelo emprega um conjunto de técnicas matemáticas para detectar as complexas interações entre elementos de dados distantes em uma sequência.

[5] Licença BSD. <https://github.com/Uberi/speech_recognition>.
[6] Licença: MIT License. <https://openai.com/research/whisper>.

Dependendo da abordagem escolhida, é necessário segmentar o arquivo de áudio em partes menores, geralmente com duração de 30 a 60 segundos. Neste estudo, optamos por segmentar a cada 30 segundos usando a biblioteca Pydub[7]. Esses segmentos serão submetidos ao processo de transcrição, seja por meio do Google *Speech Recognition* com a biblioteca *SpeechRecognition*, ou pela biblioteca *Whisper*. A diferença fundamental é que, ao utilizar o *Whisper*, não é necessário realizar a segmentação prévia do arquivo, pois ele efetua essa etapa durante o processo de transcrição. O fluxo de transcrição está ilustrado na Figura 5.

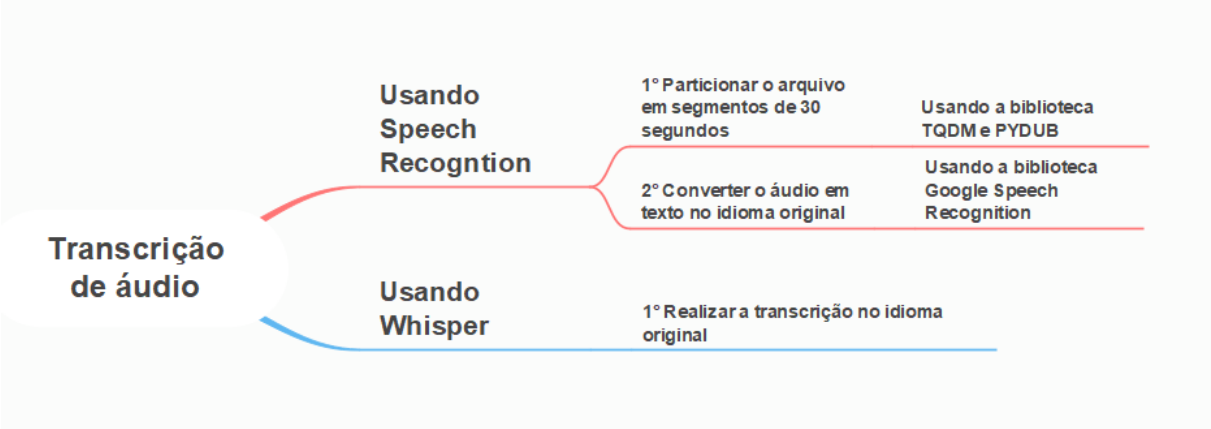

Figura 5 – Etapa de transcrição de áudio para texto.
Fonte: a autora.

3.3 Tradução do texto

Após a conclusão das etapas de transcrição do áudio original, prossegue-se com a tradução desse texto para o idioma português do Brasil. No caso de produtos audiovisuais japoneses, é comum encontrar palavras no idioma inglês nos textos, áudios ou músicas. Para aprimorar o desempenho dessa tarefa, foi adotado a estratégia de traduzir do idioma original para o idioma inglês, criando uma primeira tradução intermediária, e depois traduzir essa versão para o idioma português do Brasil. Essa abordagem visa garantir uma tradução mais precisa e contextualizada.

Para a tradução, será utilizada a biblioteca ***Translate***[8] que tem suporte para vários provedores de traduções tais como: de Microsoft, Google, a *API Translated MyMemory*, o *LibreTranslate* e DeepL utilizando API´s. Na Figura 6 está ilustrado as tarefas para a etapa de tradução. Vale destacar que, ao adotar a abordagem do *Whisper*, não é necessário realizar a tradução para o idioma inglês por meio da biblioteca *Translate*, pois o *Whisper* já incorpora essa funcionalidade internamente.

[7] Licença: MIT License (MIT). <http://pydub.com/>
[8] Licença: MIT License (MIT). <https://github.com/terryyin/translate-python>

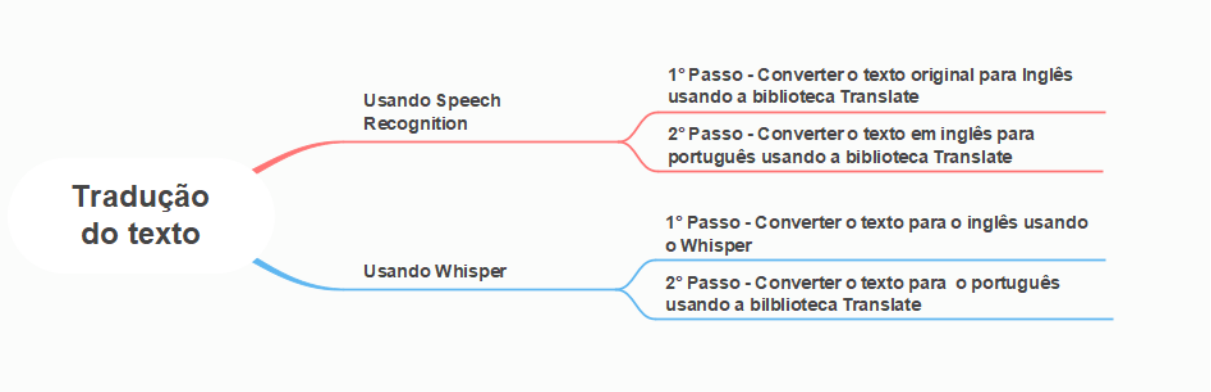

Figura 6 – Etapa de tradução.
Fonte: a autora.

3.4 Transformar texto traduzido em áudio

Nesta etapa as transcrições e traduções são encerradas e a partir deste subproduto gerado é realizada a conversão dos textos em áudio. Uma vez que o áudio estiver gerado, ele será gravado em formato MP3 usando a biblioteca gTTS.

A biblioteca **gTTS**[9] (Google *Text-to-Speech*) é uma biblioteca *Python* que permite interagir com a API de conversão de texto em fala do Google. Ela é capaz de gravar dados de fala em formato MP3 e armazená-los em um arquivo. As tarefas desta etapa estão descritas na Figura 7.

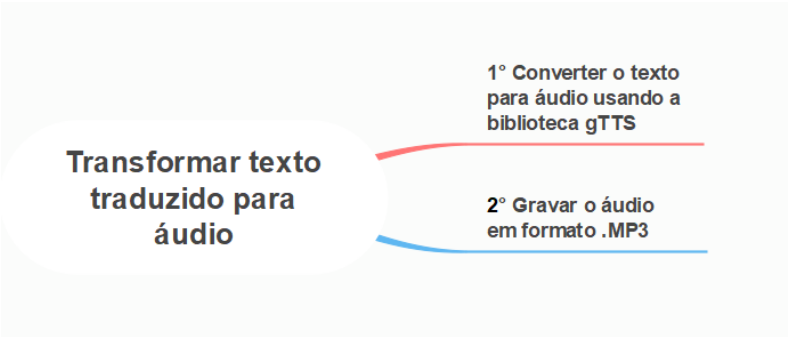

Figura 7 – Etapa de transformar o texto em áudio.
Fonte: a autora.

[9] Licença: MIT License (MIT). <https://github.com/pndurette/gTTS>.

4 EXPERIMENTOS

Para validar a proposta de geração de legendas e áudio, foram conduzidos experimentos utilizando os conjuntos de dados detalhados na Seção 4.1. Os resultados obtidos com esses conjuntos de dados foram avaliados de duas maneiras, conforme detalhado na Seção 4.2. Os resultados da primeira avaliação foram discutidos na Seção 4.3, enquanto a segunda avaliação será abordada na Seção 4.4. Por fim, foi discutido as limitações da abordagem proposta na Seção 4.5.

4.1 Conjuntos de Dados

Foram selecionados seis arquivos de vídeos e áudio no idioma japonês para a realização deste experimento, sendo divididos em:

- Dois arquivos de vídeo, um contendo mais diálogos e outro contendo mais músicas (arquivos 1 e 2). Devido ao tamanho desses vídeos, foram usados os primeiros 6 minutos.
- Dois arquivos de vídeo musicais (arquivos 3 e 4).
- Dois arquivos de vídeo musicais convertidos em áudio somente com a parte vocal, foram utilizados os mesmos arquivos do item anterior (arquivos 5 e 6).

Para futuras referências, utilizaremos a nomenclatura “Arquivo” seguida pelo número correspondente, conforme ilustrado na Figura 8.

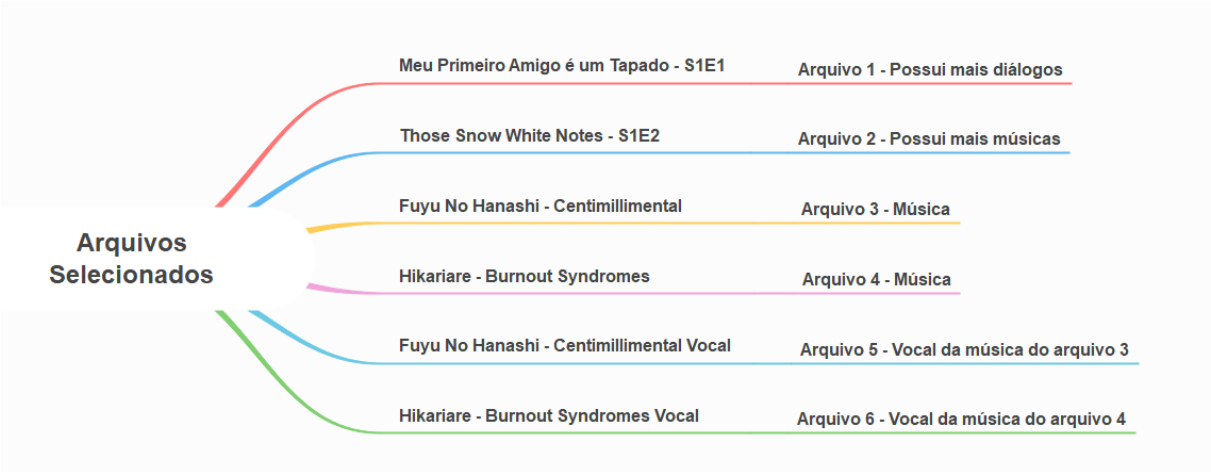

Figura 8 – Lista de arquivos selecionados para o experimento.
Fonte: a autora.

4.2 Configuração Experimental

Conforme descrito na Seção 3.2, foi avaliado duas abordagens de transcrição: o *SpeechRecognition* e o *Whisper*. O objetivo principal da avaliação é verificar a qualidade da tradução para o português. Para isso, foi realizado a transcrição dos áudios do conjunto de dados selecionado, conforme ilustrado na Figura 5, e em seguida, foi executado a tradução, como ilustrado na Figura 6. Após a geração da tradução para o português do Brasil, foi feita uma comparação com o texto original para avaliação.

Dado que o processo de tradução envolve a interpretação do tradutor e que podem ocorrer ruídos durante o processo de transcrição, torna-se inviável realizar uma comparação direta entre o texto original e os resultados obtidos pela abordagem proposta. Para lidar com essa complexidade, adotou-se uma avaliação em duas abordagens distintas:

1. Na primeira avaliação (Seção 4.3), foi feita uma comparação na quantidade de *tokens* identificados pela biblioteca *spaCy* nos textos originais e nas transcrições geradas por ambas as abordagens. Isso possibilita realização de avaliações mais objetivas e robustas dos resultados obtidos.

   O spaCy[1], conhecido como "*Industrial-Strength Natural Language Processing*", é uma biblioteca para processamento avançado de linguagem natural. Ele oferece suporte a várias tarefas, incluindo tokenização, análise sintática, reconhecimento de entidades nomeadas, classificação de texto e muito mais. O *spaCy* utiliza modelos de rede neural para realizar essas tarefas e é capaz de aproveitar modelos pré-treinados, como o BERT, para aprendizado multitarefa.

2. Na segunda avaliação (Seção 4.4), foi conduzida uma análise de similaridade de cosseno entre dois vetores de informações. Essa medida calcula o cosseno do ângulo entre esses vetores para determinar o quão semelhantes são os textos comparados. Essa abordagem foi adotada para lidar com as diferenças resultantes da transcrição e tradução em comparação com os textos originais. Os vetores foram gerados usando as bibliotecas SBert e NLTK.

   O SBert[2], ou *SentenceTransformers*, é uma biblioteca que se destina a criar representações semânticas de texto usando modelos de linguagem pré-treinados, como o BERT e o RoBERTa. Essas representações são vetores numéricos que capturam o significado e o contexto das palavras e frases em um texto. O SBert é amplamente utilizado em tarefas de Processamento de Linguagem Natural (PLN) para melhorar a eficácia de algoritmos de busca de informações, agrupamento de texto, recuperação de informações e muito mais. Ele é particularmente útil quando se deseja medir a

[1] Licença: MIT License (MIT). <https://pypi.org/project/spaCy/>
[2] Licença: Cornell University. <https://www.sbert.net/index.html>

similaridade semântica entre frases ou documentos, permitindo que as máquinas compreendam o significado subjacente do texto.

O NLTK[3] (*Natural Language Toolkit*) é uma poderosa biblioteca *Python* amplamente utilizada no processamento de linguagem natural (PLN). Ela oferece uma ampla gama de ferramentas para a análise de linguagem humana, incluindo tokenização, *stemming*, identificação de categorias gramaticais das palavras, análise sintática e muito mais. Uma de suas características distintivas é o acesso ao WordNet, um banco de dados lexical que facilita a exploração de relações semânticas entre palavras, tornando mais simples a busca por sinônimos, antônimos e aprofundando a análise de texto.

4.3 Avaliação da Quantidade de *Tokens* nos textos

Para a análise comparativa entre os resultados do *SpeechRecognition* e do *Whisper*, foi utilizado a contagem de *tokens* em relação à legenda original. Nesse contexto, a biblioteca spaCy foi empregada para tokenizar tanto o texto original quanto os textos resultantes do processo de tradução. Além disso, foi realizado uma tokenização manual do texto original, permitindo assim fazer uma comparação com a abordagem humana para esse processo. A Figura 9 ilustra os resultados obtidos por meio desse processo.

Arquivo	Quantidade de tokens no texto original		Quantidade de tokens no texto após as etapas de transcrição e tradução	
	MANUAL	SPACY	SR	Whisper
Arquivo 1 - Possui mais diálogos	550	547	443	487
Arquivo 2 - Possui mais músicas	511	473	459	677
Arquivo 3 - Música	161	157	114	154
Arquivo 4 - Música	342	277	206	358
Arquivo 5 - Vocal da música do arquivo 3	161	157	88	150
Arquivo 6 - Vocal da música do arquivo 4	342	277	196	625

Figura 9 – Avaliação comparativa entre os arquivos transcritos pelas bibliotecas.
Fonte: a autora.

Ao comparar a quantidade de *tokens* do texto original extraídos pela biblioteca spaCy com a extração manual, é possível notar que o spaCy identificou, em média, **9%** a menos de *tokens* do que um humano. No entanto, em alguns casos, essa diferença foi bem pequena, variando entre 3 a 4 *tokens*. A principal razão para essa diferença reside na maneira como as regras de tokenização são definidas. O spaCy utiliza um conjunto

[3] Licença: Apache Software License (Apache License, Version 2.0). <https:https://www.nltk.org/book_1ed>

predefinido de regras e modelos de linguagem para segmentar o texto em *tokens*, e essas regras podem não ser tão flexíveis quanto o julgamento humano em determinados contextos.

Ao analisar a diferença na quantidade de *tokens* entre o texto original, conforme identificada pelo spaCy, e a quantidade de *tokens* encontrados nos textos resultantes da etapa de tradução, foi observado que os resultados do *SpeechRecognition* são, em média, **20%** menores do que os originais, enquanto os resultados do *Whisper* são **30%** maior.

Por outro lado, a quantidade de *tokens* obtidos dos textos traduzidas com base nas abordagens *SpeechRecognition* e *Whisper*, foi observado que o segundo apresentou até **39%** mais *tokens* do que a primeira abordagem. Essa diferença foi avaliada, e a causa encontrada deve-se ao fato de que o *Whisper*, quando não consegue identificar o áudio ou não encontra a informação, retorna uma mensagem informando a ocorrência, e essas mensagens são incorporadas ao texto gerado. Neste estudo, optou-se por não remover essas informações do texto. Levando esses fatores em consideração, o modelo que obteve maior sucesso foi o texto gerado pela biblioteca *SpeechRecognition*.

Em relação aos arquivos de 2 a 6, que contêm mais músicas ou palavras cantadas, observou-se que, de forma geral, o desempenho na transcrição das palavras cantadas foi inferior em comparação aos vídeos que não possuíam música ou palavras cantadas. A diferença média foi de **21%** a menos ao usar a biblioteca *SpeechRecognition*. Por outro lado, com a *Whisper*, essa diferença chegou a ser até **46%** maior do que o texto original, devido à inclusão das mensagens que indicam falha na compreensão daquele trecho.

4.4 Avaliação de Similaridade entre Textos

Nesta seção, foi realizada uma análise de similaridade entre os textos gerados após as etapas de transcrição e tradução com os textos originais. Para isso, foi utilizado uma métrica de similaridade de cosseno, que calcula o ângulo entre os vetores de informações desses textos. Essa abordagem é importante para avaliar quão semelhantes são os textos resultantes das etapas de processamento em relação aos textos originais. Conforme mencionado na Seção 4.2, os vetores de informações foram gerados utilizando as bibliotecas SBert e NLTK. Os resultados dessa análise estão ilustrados na Figura 10.

Similaridade de Cosseno	SpeechRecognotion		Whisper	
	SBERT	NLTK	SBERT	NLTK
Arquivo 1 - Possui mais diálogos (original versus Transcrito)	0,9364	0,7064	0,9503	0,6418
Arquivo 2 - Possui mais músicas (Original versus Transcrito)	0,9580	0,7458	0,8658	0,7307
Arquivo 3 - Original versus Transcrição Full	0,9441	0,7327	0,9068	0,6934
Arquivo 5 - Original versus Transcrição Vocal	0,9316	0,5555	0,8763	0,4146
Arquivo 3 e 5 - Transcrição Full versus Transcrição Vocal	0,9755	0,7120	0,9820	0,6727
Arquivo 4 - Original versus Transcrição Full	0,9291	0,7608	0,8203	0,9702
Arquivo 6 - Original versus Transcrição Vocal	0,9302	0,7256	0,6381	0,9618
Arquuvo 4 e 6 - Transcrição Full versus Transcrição Vocal	0,9723	0,7650	0,8094	0,7678
	0,9472	**0,7130**	**0,8561**	**0,7316**

Figura 10 – Similaridade de cosseno obtida através do *SpeechRecognotion* e *Whisper*. Fonte: a autora.

Além de calcular a similaridade entre o texto original e o texto transcrito e traduzido, uma análise adicional foi realizada nos casos dos arquivos que contêm músicas (Arquivos 3-6). Nesses casos, também foi avaliada a similaridade entre o texto da música completa e o texto da mesma música que contém apenas a parte vocal, ambos gerados pela mesma abordagem. Esses arquivos foram identificados como "Arquivo 3 e 5 - Comparação entre Transcrição Completa e Transcrição Vocal" e "Arquivo 4 e 6 - Comparação entre Transcrição Completa e Transcrição Vocal". Essa análise visa avaliar especificamente a qualidade da transcrição da parte vocal da música, separada do restante da composição musical. Isso permite compreender como as abordagens se comportaram na transcrição de elementos específicos, como a letra da música, em contraste com outros sons e instrumentos presentes na gravação, fornecendo assim *insights* adicionais sobre o desempenho das abordagens em contextos específicos de transcrição.

É evidente que, em geral, o *SpeechRecognition*, quando combinado com o SBert, alcançou uma similaridade média superior em relação às outras abordagens. No entanto, o *Whisper* superou ele quatro vezes por uma pequena margem, sendo duas vezes em comparação com o SBert e outras duas vezes em comparação com o NLTK. No entanto, vale destacar que o *SpeechRecognition* pareceu manter uma estabilidade ao longo de todas as análises, o que pode ser um fator relevante a considerar.

4.5 Evoluções e Limitações

Existem várias questões que ainda precisam ser aprimoradas nesta proposta, consideradas como limitações neste MVP (Produto Mínimo Viável). A seguir, é apresentado uma breve descrição das principais limitações identificadas em várias etapas do processo proposto:

- Como identificar e marcar o tempo de fala e silêncio no áudio.
- Como tratar os demais sons (por exemplo - música de fundo, aparelhos elétricos e etc).
- Distinção entre diálogos e outros sons ambiente.
- Tratamento de sobreposição de falas (conversas ao fundo).
- Como identificar e separar a fala de cada personagem.
- Como tratar palavras diferente do idioma original.
- Aprimoramento no sistema de tradução.
- Geração de áudio para dublagem com tratamento e interpretação de personagens.

Essas limitações representam oportunidades de desenvolvimento futuro e aprimoramento da abordagem proposta.

5 CONCLUSÕES

Neste TCC, o objetivo principal foi realizar a conversão de áudios de vídeos de animes para o idioma português do Brasil, visando agilizar o processo de criação de legendas e dublagem.

Ao longo da pesquisa e desenvolvimento, foi observado que a tecnologia de modelos de linguagem natural está evoluindo rapidamente, oferecendo avanços significativos em diversas áreas, incluindo processamento de linguagem natural, tradução automática e sumarização de texto. Cada modelo possui suas vantagens e desafios específicos, o que levou a considerar cuidadosamente qual abordagem seria a mais adequada para o contexto de aplicação. No entanto, ficou claro que esses modelos ainda não estão prontos para criar de forma independente e isolada legendas e dublagens de áudio, como demonstrado nesse trabalho.

Com este trabalho, foi obtido uma compreensão detalhada do processo de transcrição de áudio para texto, sua subsequente tradução e a conversão do texto traduzido para áudio. Além disso, foi avaliado a aderência do texto gerado e traduzido para o idioma português do Brasil, comparando-o com as legendas originais. É importante mencionar que este estudo não realizou a avaliação da qualidade do áudio gerado a partir do texto traduzido, mas sim gerou o áudio da tradução.

Para realizar todas essas atividades, foi utilizado bibliotecas gratuitas, de fácil implementação e amplamente reconhecidas em suas respectivas áreas de atuação. No geral, este estudo proporcionou *insights* sobre as possibilidades e limitações das atuais tecnologias de processamento de linguagem natural aplicadas à tradução e criação de legendas para vídeos de animes em português do Brasil.

Durante o estudo, optou-se por conduzir a análise das informações e resultados em duas etapas distintas. A primeira etapa consistiu na verificação da quantidade de *tokens*, enquanto a segunda focou na avaliação da similaridade de cosseno.

1. **Avaliação dos *tokens***

   Nessa avaliação, foi considerado tanto as legendas originais extraídas manualmente dos arquivos quanto as legendas resultantes do processo de transcrição e tradução. O foco foi analisar a quantidade de *token*s nesses arquivos, utilizando a biblioteca Spacy.

   Após realizar a contagem de *tokens*, foi comparado a quantidade de *tokens* extraídos por cada uma das abordagens propostas de transcrição. Ao examinar a tradução gerada a partir da transcrição realizada pelo *SpeechRecognition*, foi identificado que

o texto resultante era, em média, **20%** menor em comparação com a contagem do arquivo original. Já em relação ao *Whisper*, o resultado médio foi aproximadamente **30%** maior. Essa diferença na quantidade de *tokens* ocorreu porque o *Whisper*, quando não consegue identificar o áudio, inclui um texto padrão informando que a transcrição não foi possível. Neste estudo, optamos por não remover essas frases padrão.

Observamos que nos arquivos de vídeo que continham músicas ou palavras cantadas, o desempenho na identificação dessas palavras foi inferior. A média de *tokens* identificados foi cerca de **21%** menor no *SpeechRecognition* e aproximadamente **46%** maior no *Whisper* em comparação com o texto original. A diferença maior no caso do *Whisper* ocorreu devido à inclusão das frases padrão.

2. **Avaliação da similaridade por cosseno**

   Para avaliação de similaridade, foi utilizada a métrica de similaridade por cosseno com base nas bibliotecas SBERT e a NLTK. O modelo usando a biblioteca **Speech-Recognation** foi o que obteve a maior pontuação nos dois modelos de avaliação, obtendo a média de **0,95% no SBERT** e **0,71% no NLTK** contra **0,88% no SBert** e **0,73% no NLTK** usando a biblioteca **Whisper**.

Com base na análise comparativa, o uso do *SpeechRecognition* para realizar a transcrição e tradução do áudio de vídeos se mostrou mais eficiente e alinhado com os objetivos deste estudo, sendo a melhor escolha para implementação. Além disso, possui a vantagem de facilitar a avaliação e verificação mais precisa do resultado de cada etapa do processo, uma vez que as etapas são executadas sequencialmente.

A implementação bem-sucedida desse modelo tem o potencial de acelerar o trabalho de transcrição de texto a partir de vídeos, o que pode ser aplicado em uma variedade de cenários além deste estudo, como transcrição de atendimento em *call centers*, *chatbots*, acessibilidade, entre outros. Isso abre caminho para futuros avanços nesse campo.

No entanto, é importante destacar que este trabalho não esgota todas as possibilidades e variações de aplicação de modelos de linguagem natural disponíveis. À medida que novas técnicas e avanços surgem, é fundamental explorar abordagens inovadoras para atender melhor às demandas em constante evolução.

REFERÊNCIAS

ABDEL-HAMID, O. *et al.* Convolutional neural networks for speech recognition. **IEEE/ACM Transactions on audio, speech, and language processing**, IEEE, v. 22, n. 10, p. 1533–1545, 2014.

CAMASTRA, F.; VINCIARELLI, A. **Machine learning for audio, image and video analysis: theory and applications**. [*S.l.: s.n.*]: Springer, 2015.

CHO, Y.-P. *et al.* A survey on recent deep learning-driven singing voice synthesis systems. *In*: IEEE. **2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)**. [*S.l.: s.n.*], 2021. p. 319–323.

CODOGNO, Y. **CCXP divulga balanço de 2022 de dezembo de 2022**. 2022. <https://www.exibidor.com.br/noticias/mercado/13087-ccxp-apresenta-balanco-de-2022-bate-proprio-recorde-e-confirma-edicao-de-2023>. Acesso em: 27 set 2023.

FELTRIN, F. **Redes Neurais Artificiais**. [*S.l.: s.n.*]: ebook Amazon, 2021.

GUNDAVARAPU, M. R. *et al.* Smart bot for handwritten digit string recognition. *In*: IEEE. **2022 International Conference for Advancement in Technology (ICONAT)**. [*S.l.: s.n.*], 2022. p. 1–5.

HUANG, X. *et al.* The sphinx-ii speech recognition system: an overview. **Computer Speech & Language**, Elsevier, v. 7, n. 2, p. 137–148, 1993.

KUMAR, Y.; KOUL, A.; SINGH, C. A deep learning approaches in text-to-speech system: A systematic review and recent research perspective. **Multimedia Tools and Applications**, Springer, v. 82, n. 10, p. 15171–15197, 2023.

LI, J. *et al.* Recent advances in end-to-end automatic speech recognition. **APSIPA Transactions on Signal and Information Processing**, Now Publishers, Inc., v. 11, n. 1, 2022.

LI, Z. *et al.* A survey of convolutional neural networks: analysis, applications, and prospects. **IEEE transactions on neural networks and learning systems**, IEEE, 2021.

MALIK, M. *et al.* Automatic speech recognition: a survey. **Multimedia Tools and Applications**, Springer, v. 80, p. 9411–9457, 2021.

MENEZES NEY COUTINHO, N. **Introdução à Programação com Python: Algoritmos e Lógica de Programação**. [*S.l.: s.n.*]: Novatec Editora, 2019.

NASSIF, A. B. *et al.* Speech recognition using deep neural networks: A systematic review. **IEEE access**, IEEE, v. 7, p. 19143–19165, 2019.

NING, Y. *et al.* A review of deep learning based speech synthesis. **Applied Sciences**, MDPI, v. 9, n. 19, p. 4050, 2019.

ORUH, J.; VIRIRI, S.; ADEGUN, A. Long short-term memory recurrent neural network for automatic speech recognition. **IEEE Access**, IEEE, v. 10, p. 30069–30079, 2022.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *In*: **Conference on Empirical Methods in Natural Language Processing**. [*S.l.: s.n.*], 2019. Available at: <https://api.semanticscholar.org/CorpusID:201646309>.

RIBNEIRO, P. H. **Anime Friends 2022 é marcado por reencontros e nova geração de otakus de julho de 2022**. 2022. <hhttps://www.omelete.com.br/mangas-animes/anime-friends-sp-2022>. Acesso em: 27 set 2023.

ROSA, R. K.; SILVA, D. Conversão texto-fala para o português brasileiro utilizando tacotron 2 com vocoder griffin-lim. *In*: **Anais do XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais**. Sociedade Brasileira de Telecomunicações, 2021. Available at: <https://doi.org/10.14209/sbrt.2021.1570727280>.

SCHNAIDER, A. **Mercado Livre e Omelete levarão pequenos empreendedores à CCXP de agosto de 2022**. 2022. <https://www.meioemensagem.com.br/marketing/mercado-livre-e-omelete-levarao-pequenos-empreendedores-a-ccxp>. Acesso em: 27 set 2023.

SILVA, I. N. S. D. H. F. R. A. da. **Redes Neurais Artificiais para engenharia e ciências aplicadas**. [*S.l.: s.n.*]: Artliber Editora, 2010.

THAKUR, N. *et al.* Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. **arXiv preprint arXiv:2010.08240**, 10 2020. Available at: <https://arxiv.org/abs/2010.08240>.

TOLEDO, M. **De olho no mercado geek: faturamento com produtos licenciados cresce e chega a R$ 21,5 bi no Brasil de agosto de 2022**. 2022. <https://istoedinheiro.com.br/mercado-geek-faturamento-com-produtos-licenciados-cresce-e-chega-a-r-215-bi-no-brasil/>. Acesso em: 27 set 2023.

VAJJALA, S. *et al.* **Practical Natural Language Processing: A Comprehensive Guide to Building Real-world NLP Systems**. [*S.l.: s.n.*]: O'Reilly Media, 2020.

VARGAS, A. C. G.; PAES, A.; VASCONCELOS, C. N. Um estudo sobre redes neurais convolucionais e sua aplicação em detecção de pedestres. O'Reilly Media, 2016.

YIN, T.; HENTER, R. **Translate python documentation**. 2020.

YNOGUTI, C. A.; VIOLARO, F. Sobre a importância da transcrição fonética em sistemas de reconhecimento de fala. **Revista da Sociedade Brasileira de Telecomunicações**, v. 15, n. 1, 2000.

YU, D.; DENG, L. **Automatic speech recognition**. [*S.l.: s.n.*]: Springer, 2016. v. 1.