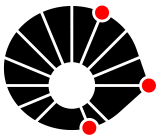

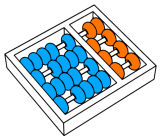

Transcrição Automática de Vídeos para Textos em Linguagem Natural

Julio Cesar dos Reis *Enrico Delbuono*

Relatório Técnico - IC-PFG-23-54
Projeto Final de Graduação
2023 - Dezembro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Transcrição Automática de Vídeos para Textos em Linguagem Natural

Julio Cesar dos Reis Enrico Delbuono*

Resumo

A utilização de vídeos como ferramenta educacional tem se consolidado como uma prática cada vez mais comum, podendo potencializar a assimilação de informações e proporcionando uma experiência de aprendizado mais envolvente e eficaz ao estudante. Soluções na criação de quizzes para apoiar o treinamento e avaliação do conhecimento de estudantes, portanto, limitam o seu potencial ao se utilizarem apenas arquivos textuais para a construção de bases de dados, dado que o conhecimento na atualidade é amplamente disseminado por meio de diversos meios de comunicação distintos. Este trabalho investiga técnicas e ferramentas que auxiliem na transcrição automática de arquivos de vídeo para o formato de texto. Para isso, estudamos a *OpenAI Whisper*, uma ferramenta open-source responsável pelo reconhecimento de voz e pela conversão do áudio obtido para o formato textual. O modelo permite uma transcrição de voz multilíngue. Conduzimos avaliações experimentais nas línguas inglêsa, Português e espanhol. Adicionalmente, estruturada uma plataforma Web desenhada como um espaço para a comunidade criar, compartilhar e realizar quizzes colaborativamente.

1 Introdução

A evolução das tecnologias educacionais tem desempenhado um papel fundamental na forma como os indivíduos buscam e desenvolvem conhecimento. Quizzes são um conjunto de perguntas e respostas geradas para treinamento e avaliação de estudantes[1]. A integração de quizzes como ferramenta complementar ao ensino tradicional e o crescente papel dos vídeos como fonte de estudo adicional aos livros convencionais são aspectos cruciais desse panorama educacional contemporâneo. Quizzes oferecem uma abordagem interativa, proporcionando aos alunos uma oportunidade não apenas de testar, mas também de consolidar seus conhecimentos de maneira dinâmica. Paralelamente, vídeos tornaram-se recursos valiosos, explorando a sinergia entre elementos

*Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.

visuais e auditivos para proporcionar uma experiência de aprendizado mais rica e envolvente.

No contexto deste trabalho, exploramos uma dimensão específica da integração entre vídeos e texto: a transcrição de vídeos para facilitar o acesso e a revisão do conteúdo. Essa transcrição envolve duas etapas essenciais - a conversão do conteúdo visual para áudio e, subsequentemente, a transcrição precisa do áudio para o formato de texto.

Este trabalho visa definir uma estrutura capaz de transcrever vídeos educativos para o formato de textos em diversas línguas, que posteriormente serão utilizados para a construção de quizzes. Contribuímos igualmente da estruturação preliminar de uma plataforma Web, desenvolvendo um ambiente colaborativo de criação, execução e compartilhamento de quizzes. Nosso objetivo é explorar conteúdo textual gerado pela transcrição dos vídeos em ferramentas automatizadas de geração de quizzes.

Em nossa proposta, estudamos e exploramos como método de solução o uso de um modelo especializado em transcrição. Em particular, experimentamos o modelo *Whisper*[2] em nossas avaliações experimentais. O objetivo destes experimentos com o uso da ferramenta é entender a efetividade deste modelo para a geração de textos que possam ser usados como base na criação de quizzes. Para testar esta efetividade, realizamos uma série de avaliações de modo a identificar o impacto de algumas condições dos vídeos sobre o resultado final da transcrição. Os experimentos foram realizados em três linguagens distintas e organizados de forma a englobar tanto situações favoráveis, como uma fala mais lenta, ausência de ruídos e sons de fundo e sotaques com maior base de dados para compreensão, quanto situações adversas, em que uma ou mais destas características se distancia do ideal.

Neste trabalho, desenhamos e desenvolvemos uma versão de uma plataforma para instrumentalizar a jornada de usuários na criação de quizzes a partir de vídeos online. Visamos oferecer uma iniciativa abrangente que visa proporcionar uma experiência de aprendizado completa. Esta plataforma incorpora um ambiente interativo pela qual a comunidade pode interagir, compartilhar conhecimento e participar quizzes. O desenvolvimento dessa plataforma envolveu a implementação tanto do front-end quanto do back-end, proporcionando um espaço virtual coeso para a comunidade colaborar e se aprimorar coletivamente. Este trabalho representa, assim, uma convergência inovadora entre tecnologias educacionais, oferecendo uma abordagem holística para aprimorar a experiência de aprendizado.

Este trabalho está organizado da seguinte forma: A Seção 2 aborda os fundamentos incluindo conceitos chaves no trabalho. A Seção 3 descreve a proposta de solução organizada em duas partes: a primeira descreve o processo de concepção da solução de modo geral; e a segunda detalha a implementação da solução concebida com o uso das técnicas e tecnologias exploradas. A Seção 4 aborda as avaliações experimentais realizadas incluindo os resultados obtidos. A Seção 5 detalha o funcionamento das transcrições obtidas na plataforma Web. A Seção 6 desenvolve uma discussão sobre

os resultados. Por fim, apresentamos uma conclusão na Seção 7.

2 Fundamentos

De modo a auxiliar a compreensão geral deste trabalho e de cada etapa, abordaremos os principais conceitos e técnicas utilizados para resolver o problema em estudo. Esta etapa de fundamentos está organizada em duas partes principais: o detalhamento da biblioteca *Python Pytube*[5], sua importância e usabilidade em (Subseção 2.1); e o modelo de reconhecimento automático de voz *OpenAI Whisper*[2], juntamente com suas principais funcionalidades e entendimento de sua arquitetura em (Subseção 2.2).

2.1 Biblioteca Pytube

Uma das ferramentas que serão utilizadas ao longo deste estudo é a biblioteca *Pytube*. Esta biblioteca do Python é leve, livre de dependências e tem como propósito o download de vídeos do *YouTube*®. A utilização da biblioteca *Pytube* emerge como uma ferramenta essencial no contexto de extração de dados audiovisuais, oferecendo funcionalidades simplificadas e intuitivas para realizar o download de vídeos do *YouTube*®. A ferramenta permite que sejam inseridos uma série de filtros que customizem o seu funcionamento de acordo com a necessidade do usuário, tal como o download somente do áudio presente no vídeo em questão.

A biblioteca demonstra flexibilidade ao lidar com diversos formatos de vídeo e qualidades de resolução, garantindo uma ampla gama de opções para os usuários. Sua capacidade de contornar as limitações impostas por políticas de restrição de acesso a vídeos, aliada à facilidade de integração em projetos mais amplos, faz com que a biblioteca seja a melhor opção para a obtenção de arquivos de áudio por meio de vídeos de forma rápida e fácil.

A partir dessa biblioteca, é possível de forma automática inserir um link para um vídeo do *YouTube*® e receber como saída um arquivo contendo o áudio completo do vídeo em questão, já devidamente baixado na máquina.

2.2 Modelo de Reconhecimento Automático de Voz

Para a realização da transcrição de um áudio para o formato textual, é necessário o uso de um modelo de *Automatic Speech Recognition* (ASR)[4], ou reconhecimento automático de fala. O modelo que será estudado e utilizado ao longo deste trabalho é o *OpenAI Whisper*[2].

Whisper é um modelo de reconhecimento de fala que pode ser utilizado para os mais variados propósitos. Ele foi treinada a partir de um grande conjunto de dados de áudio - 680 mil horas de dados coletados da Web em diversas línguas. Por ser um

modelo robusto, ele permite realizar transcrições e entendimento de falas sob as mais variadas condições, como diferentes sotaques, presença de ruídos no áudio, termos mais técnicos, diferentes velocidades de interlocução, *etc.*

Este modelo de sequência é treinado em várias tarefas de processamento de fala, incluindo reconhecimento de fala multilíngue, tradução de fala, identificação de idioma falado e detecção de atividade vocal. Essas tarefas são representadas conjuntamente como uma sequência de *tokens* a serem previstos pelo decodificador, permitindo que um único modelo substitua muitas etapas de uma pipeline de processamento de fala tradicional. Este conjunto de *tokens* especiais funciona como especificadores de tarefa ou alvos de classificação. A Figura 4 apresenta arquitetura do modelo *Whisper* de forma detalhada.

A arquitetura do *Whisper* possui um conjunto de 5 modelos responsáveis por realizar todas as suas atividades: Tiny, Base, Small, Medium e Large. Cada um dos modelos possui tamanhos e taxas de aprendizado diferentes, o que impacta diretamente no desempenho das transcrições geradas. O modelo Tiny, o menor e mais simples, possui uma menor quantidade de camadas, profundidade e parâmetros que todos os outros, mas é mais leve e simples de ser executado em qualquer máquina. De forma oposta, o modelo Large é o mais complexo e que possui mais parâmetros e camadas, mas requer uma estrutura muito mais resistente. Para efeito de experimentação ao longo deste projeto, todas as etapas e testes serão realizadas utilizando o modelo Base.

3 Transcrição Automatizada de Vídeos

3.1 Concepção

Para que o processo de transcrição possa ser realizado, é necessário que inicialmente o usuário insira um *link* de um vídeo do *YouTube*®. A partir disso, o vídeo é convertido para o formato de áudio e salvo. Com o arquivo de áudio, usamos um modelo de reconhecimento automático de voz (ASR model), que analisa cada *token* do áudio e o converte para o formato de texto. Ao fim do processo para cada token, é gerado um texto completo e correspondente ao áudio ou vídeo inserido pelo usuário. Caso seja de interesse do usuário, é possível que seja feita de forma direta a submissão de um arquivo de áudio para que seja feita a transcrição, pulando assim a etapa de transformação de um vídeo para áudio.

A Figura 1 apresenta um diagrama conceitual de cada uma das etapas de transcrição automática descritas.

Figura 1: *Etapas da transcrição automática de vídeos e áudios*

3.2 Implementação

O código-fonte do projeto capaz de realizar o processo de transcrição foi construído aliando-se a concepção das etapas planejadas na Seção 3.1 com as tecnologias e fundamentos estudados e descritos ao longo da Seção 2. O desenvolvimento desta implementação pode ser observado na Figura 2.

```python
import whisper
from pytube import YouTube

def VideoTranscription(url, file_name):
    youtube_video_url = url
    youtube_video = YouTube(youtube_video_url)

    print(youtube_video)

    streams = youtube_video.streams.filter(only_audio="true")
    stream = streams.first()

    stream.download(filename=file_name)

    model = whisper.load_model("base")

    result = model.transcribe(file_name)
    print(result["text"])

video_url = input()
file_name = input()

VideoTranscription(video_url, file_name)
```

Figura 2: *Código-fonte da transcrição de vídeos*

Neste contexto, as etapas seguem o mesmo fluxo da concepção do projeto, com o usuário inserindo um *link* para um vídeo do *YouTube*®. A biblioteca *Python Pytube* é utilizada para a criação de uma classe do tipo YouTube, que contém como parâmetro a URL do vídeo a ser analisado. A partir disso, são selecionadas as *streams* dentro da classe criada, e é feita uma filtragem para que somente o áudio dos arquivos seja armazenado, por meio da propriedade only_audio="true". Com isso, a biblioteca é capaz de fazer o download da *stream* originada, armazenando um arquivo do tipo .mp4 na máquina.

A partir disso, já com o arquivo de áudio gerado, o modelo de reconhecimento automático de fala *Whisper* é utilizado, carregando o modelo "base"para este projeto em específico. Com o modelo já selecionado, é iniciada a transcrição do áudio por meio do método *transcribe*, já gerando o resultado final em formato de texto. Um diagrama completo desta implementação com suas tecnologias é mostrado na Figura 3.

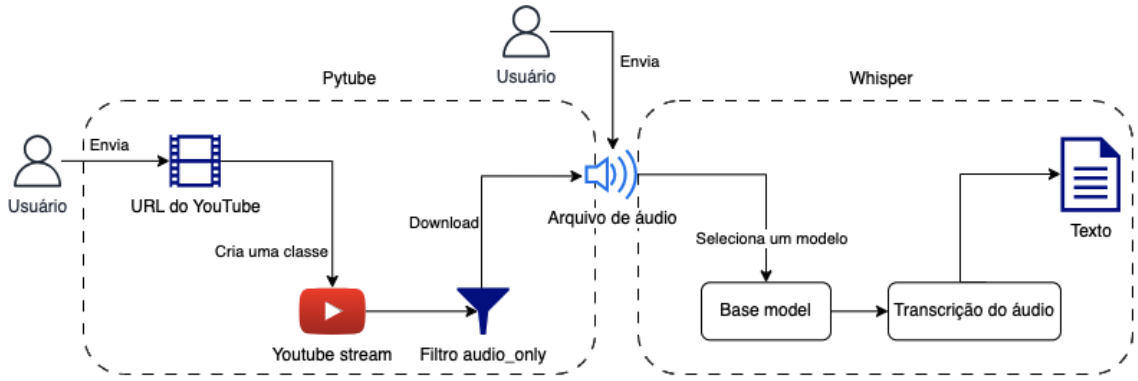

Figura 3: *Solução de transcrição automática de vídeos e áudios por meio da biblioteca Python Pytube e do ASR Whisper*

Durante o processo de transcrição do áudio utilizando o *Whisper*, a ferramenta inicia suas etapas verificando, para cada *token*, se há alguma fala sendo expressa. Em caso negativo, o modelo já passa para o próximo *token* do áudio e recomeça o procedimento. Em caso positivo, é feita uma identificação da linguagem falada dentro daquele intervalo. A partir disso, é feita a transcrição daquele *token* para o arquivo que irá gerar o texto completo, e essa ação é repetida até que se atinja o último *token*. Esta descrição pode ser visualizada em detalhes por meio da arquitetura apresentada na Figura 4. O diagrama também representa a etapa de "Transcrição do áudio"sinalizada na Figura 3.

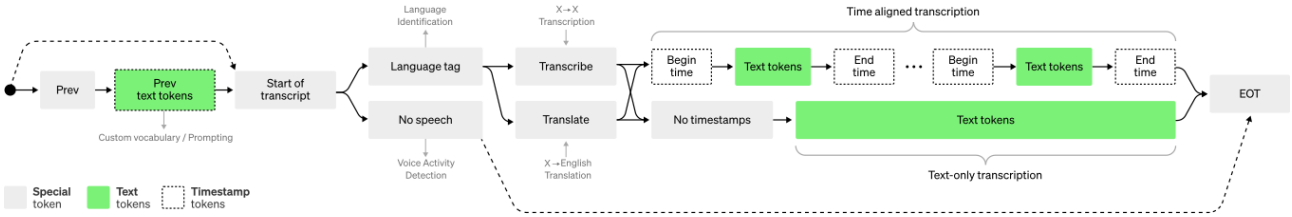

Figura 4: *Diagrama da arquitetura do modelo de reconhecimento automático de fala Whisper*[3]

4 Avaliação Experimental

Esta seção visa apresentar a definição e os resultados de avaliações experimentais conduzidas para avaliar a solução. A subseção 4.1 apresenta o protocolo dos experimentos realizados, incluindo materiais; subseção 4.2 apresenta as métricas exploradas no experimento; e a seção 4.3 apresenta os resultados dos experimentos realizados.

4.1 Protocolo experimental e materiais

Ao longo deste trabalho, realizamos experimentos para verificar a efetividade e taxa de precisão do modelo de transcrição de vídeos abordado, de forma a usufruir de diversas situações e temas e medir sua efetividade. Conduzimos no total seis experimentos, que se diferenciam tanto pela língua do vídeo, quanto pelo sotaque do criador de conteúdo, teor técnico e tipo de vocabulário do material em questão. Todos os vídeos utilizados como base possuem uma duração menor que 4 minutos, de modo a poder realizar a comparação manual da saída esperada com o texto obtido. A solução, no entanto, é capaz de realizar transcrições para qualquer duração de vídeo inserida.

Em todos os experimentos os mesmos passos foram realizados, utilizando a biblioteca *Pytube* para a conversão de vídeos para áudios em quaisquer qualidades, e fazendo uso do modelo *Whisper OpenAI* para a conversão do áudio obtido em texto. Esse passo a passo foi conduzido em notebooks Python. Por se tratarem de uma biblioteca e uma ferramenta *open-source*, a utilização deste modelo para a realização das transcrições automáticas não apresentou qualquer tipo de custo ou limitação.

Ao final do experimento, realizamos uma comparação manual de cada arquivo de saída com o texto esperado com todas as frases corretas, de forma a entender e averiguar a qualidade dos resultados pelo modelo estudado. Esta análise manual contou com os seguintes passos, repetidos em cada um dos experimentos:

- Transcrição do texto original com base na visualização do vídeo, realizada de forma manual. Para casos em que os vídeos dispunham de legendas, estas foram

utilizadas de forma a auxiliar essa transcrição.

- Verificação da transcrição realizada com falantes nativos da língua a ser analisada. Esta etapa foi realizada para a transcrição de vídeos em espanhol.

- Comparação entre a transcrição realizada manualmente e a transcrição obtida pelo modelo de reconhecimento automático de fala. As palavras que se encontravam na transcrição manual e não na transcrição do modelo foram demarcadas com a cor verde, e o oposto obteve uma marcação em vermelho. As palavras que eram congruentes em ambos os textos não receberam nenhuma marcação.

- Contabilização do número de palavras modificadas, adicionadas e removidas ao comparar as duas transcrições para entender a efetividade de cada experimento.

O detalhamento de cada um dos seis experimentos realizados, juntamente com seus propósitos e motivações, pode ser observado nesta Seção.

- **Experimento #1**[1]

  - Conteúdo em inglês / Autor estadunidense / Diálogo entre pessoas;

  Realizamos um primeiro experimento tendo como base um vídeo na língua Inglesa estadunidense, no qual duas pessoas interagem entre si e introduzem o tema *Design of Everyday Things* de uma forma didática e sem termos muito técnicos. Este primeiro experimento teve como objetivo entender a efetividade do modelo de transcrição sob condições mais simples, como a presença de sotaques que receberam um maior volume de treinamento, a boa qualidade de som e ausência de ruídos de fundo, e a ausência de uma linguagem técnica.

- **Experimento #2**[2]

  - Conteúdo em inglês / Autor alemão / Linguagem técnica;

  O segundo experimento foi realizado utilizando como base um depoimento de Albert Einstein acerca da Teoria da Relatividade. O foco principal desse experimento foi validar o entendimento de fala do modelo sob diversas condições adversas, como a presença da língua inglesa com um forte sotaque alemão, o uso de termos técnicos na fala e a baixa qualidade do áudio de entrada, por se tratar de um depoimento do século passado.

[1] O vídeo base para o Experimento #1 pode ser visualizado em `https://www.youtube.com/watch?v=pAOyWFOFhsg`

[2] O vídeo base para o Experimento #2 pode ser visualizado em `https://www.youtube.com/watch?v=aNuuYKieHRY`

- **Experimento #3**[3]

  - Conteúdo em português / Linguagem técnica / Fala rápida;

  No terceiro experimento, decidimos testar como entrada vídeos na língua Portuguesa. De modo a iniciar o processo, utilizamos como amostra um guia básico e introdutório ao *ReactJS*. No vídeo, o interlocutor faz uso de alguns termos técnicos para a área e se pronuncia de maneira rápida e com poucas pausas entre cada sentença. Este experimento teve como principais focos comparar a efetividade do modelo sob diferentes linguagens, e entender o impacto da velocidade da fala para o resultado final.

- **Experimento #4**[4]

  - Conteúdo em português / Linguagem menos técnica / Fala pausada;

  No quarto experimento tivemos como objetivo novamente testar a efetividade do modelo de transcrição para a língua portuguesa. Nesta iteração, no entanto, optamos por utilizar como base um vídeo da professora do Instituto de Computação da UNICAMP, Cláudia Bauzer Medeiros, apresentando suas pesquisas e trabalhos. O discurso em questão possui uma fala mais pausada do que no experimento anterior, e uma menor quantidade de terminologias técnicas.

- **Experimento #5**[5]

  - Conteúdo em espanhol / Linguagem técnica / Fala pausada;

  Para o quinto experimento, optamos por validar o funcionamento do modelo de transcrição a partir de uma terceira linguagem. Escolhemos um vídeo em espanhol que descreve de forma introdutória alguns conceitos de *Interação Humano-Computador*, fazendo uso de algumas palavras técnicas e utilizando uma fala ainda mais pausada e calma que no Experimento #4.

- **Experimento #6**[6]

  - Conteúdo em espanhol / Fala rápida;

  Por fim, realizamos um sexto e último experimento, também explorando conteúdo na língua Espanhol, com o intuito de verificar a eficácia do projeto nesta linguagem. Por termos testado no Experimento #5 uma entrada de vídeo com falas

[3] O vídeo base para o Experimento #3 pode ser visualizado em `https://www.youtube.com/shorts/o5LJyOUUqpk`

[4] O vídeo base para o Experimento #4 pode ser visualizado em `https://www.youtube.com/watch?v=ue8ylamxGKQ`

[5] O vídeo base para o Experimento #5 pode ser visualizado em `https://www.youtube.com/watch?v=p1ghnX7hRbA`

[6] O vídeo base para o Experimento #6 pode ser visualizado em `https://www.youtube.com/watch?v=hH_C9KPvWe4`

mais lentas, decidimos por inserir um segundo vídeo em espanhol no qual o interlocutor se comunica de maneira propositalmente mais rápida que o comum, enquanto recita versos da obra "Hamlet", de William Shakespeare.

4.2 Métricas

Em nossos experimentos, usamos a *Word Error Rate* (WER)[11] como ferramenta para verificar a precisão das transcrições obtidas. A WER (Taxa de Erro de Palavras), é uma métrica fundamental na avaliação da qualidade de sistemas de reconhecimento automático de fala e transcrição de voz para texto. Essa métrica quantifica a discrepância entre a transcrição gerada automaticamente e a transcrição de referência, expressando o número total de substituições, inserções e deleções de palavras necessárias para alinhar ambas as transcrições.

A WER é calculada dividindo o número total de operações de edição (substituições, inserções e deleções) pelo número total de palavras na transcrição de referência. Quanto menor a taxa de erro de palavras, mais precisa é a transcrição. A fórmula pode ser representada matematicamente como:

$$WER = (S + I + D)/N,$$

em que S representa o número de palavras substituídas, I o número de palavras inseridas, D o número de palavras removidas e N o número total de palavras no texto.

Esta ferramenta é essencial na avaliação da efetividade e precisão de algoritmos de transcrição automática, fornecendo uma medida quantitativa para entender e aprimorar a qualidade do processamento de fala em aplicações diversas. A principal limitação da ferramenta, no entanto, consiste no fato de que ela não é capaz de analisar, para cada caso, a natureza e a gravidade do erro na transcrição. Assim, toda modificação no texto original é contabilizada com o mesmo peso.

Para suprir essa ausência e complementar o estudo do projeto, realizaremos uma análise humana manual de cada caso selecionado experimentalmente. Para isso, a comparação manual entre a transcrição obtida e o texto de saída gerado será feita pelos integrantes deste trabalho, incluindo a validação através do auxílio de pessoas que têm o espanhol como língua nativa. Em suma, a comparação manual será realizada através da leitura do texto gerado ao mesmo tempo em que o vídeo era reproduzido, para os casos em que não há legenda disponível. Porém, quando existente, a legenda será comparada diretamente ao texto gerado, efetuando assim as correções necessárias destacadas neste documento. Aliada a essa comparação textual, é necessária também uma análise mais abstrata da gravidade de cada erro de transcrição, ou seja, o impacto que aquele erro proporciona para o entendimento geral do conteúdo e para a futura geração de quizzes.

4.3 Resultados

Apresentamos os principais resultados para cada um dos experimentos citados. Para cada um dos resultados, apresentamos as transcrições de saída, juntamente com os termos inseridos de forma indevida grifados em vermelho; e os textos que deveriam ter sido inseridos no lugar grifados em verde.

4.4 Experimento #1: Conteúdo em inglês / Autor estadunidense / Diálogo entre pessoas

Fornecendo o vídeo "*Introduction to Conceptual Models - Intro to the Design of Everyday Things*"como primeira entrada, a Figura 5 apresenta o resultado da transcrição obtida ao final do Experimento #1, juntamente com a correção do texto recebido.

Hey Christian, want some water? Hey Don, yes please. Sure. That's a fancy tea pot. Yeah, it's that's one of my favorites. I'm sorry, doesn't have that water. I'd like a little bit more than that no. Well, yeah, why don't you fill it and I want to explain this coffee pot? Sure. So, this is a joke. It's one of my favorite jokes. It's sometimes called a coffee pot for masochists. It was done originally by a French artist. And this is a copy made just for me. And I've used to the cover of recover my books. I love it so much. And you know, it's what's nice about it. It's obvious you grab the handle. It's obvious that's the spout. But it's also obvious that it's the wrong way. It won't work. The conceptual model is clear and is clear that it's well, that it's a joke. It's impossible on purpose. What's the matter? What are you doing? Well, it doesn't have a place to fill. So, I'm not sure I recommend that. Okay. Then the other choice is at the bottom. Which I'm going to try. Good. Seems work. So far. That's a funny way to use a tea pot. Yeah, that doesn't seem to work. That's not the way I use it before. Hey. Why doesn't the water drip out? Well, look at that. So, how do you think it works? What's your conceptual model with the way it works? It's a mystery. It's supposed to be a mystery. It's called a puzzle pot. The Chinese invented it to go about 400 years ago. This is a copy. And you're right. There's no obvious way to put the water in. You don't really want to put it in there. And you turn it upside down. And you put it in there. And you put it in there. And you put it in there. And you put it in there. And you put it in there. And you turn it upside down. And there's a hole, sure. But come on. You pull the water there. You turn it upside down. The water will flow out. So, yet if I do that, what? How does it work? You're not supposed to be able to figure it out. That's the whole point. That's why it's called a puzzle pot. So this was designed by a trickster. A good designer. And this is the challenge for design. We'll design things such that somebody can have an effective conceptual model and understand how it works. Or go out of the way so you don't have a model. If in fact that's the goal, you know, to fool you. And the only way a designer can communicate is through the objects that they design. And there is the challenge. And what this lesson is about.

Figura 5: *Resultado da transcrição do vídeo Introduction to Conceptual Models - Intro to the Design of Everyday Things*

A partir do resultado, observamos que a maior parte do texto obtido se manteve exato às falas ao longo do vídeo, com exceção de alguns termos que não impactam o entendimento geral do conteúdo. O maior problema se apresenta pela repetição da mesma frase do vídeo algumas vezes. Tentamos realizar o reprocessamento do vídeo outras três vezes, e em todos os casos a transcrição gerada se dava da mesma exata maneira, e o motivo para esta repetição no trecho exposto ainda é desconhecido.

4.5 Experimento #2: Conteúdo em inglês / Autor alemão / Linguagem técnica

Utilizando o vídeo "*Albert Einstein Explains Theory of Relativity — Albert Einstein Real Video — Colour Footage*"como modelo, a Figura 6 apresenta a transcrição obtida ao final do Experimento #2.

```
It followed from the special theory of relativity that mass and
   energy are both but put at both different manifestations of
   the same thing, a somewhat unfamiliar conception for the
   average mind of the revered mind. Furthermore, the equation E
    is equal m C square in which energy is put third equal to
   mass multiplied by square of the velocity of light showed that
    very small amount of mass may be converted into a very large
   amount of energy and vice-versa we see well. The mass and
   energy were in fact equivalent work in fact exceedal this.
   According A coroutine to the formula mentioned before, this
   was demonstrated by Cockroft and Walton Kocras and Bryson in
    1932 experimentally.
```

Figura 6: *Resultado da transcrição do vídeo "Albert Einstein Explains Theory of Relativity — Albert Einstein Real Video — Colour Footage"*

Apesar do texto apresentado na Figura 6 possuir um menor tamanho que na figura 5, observamos que houve uma maior quantidade de erros na transcrição do modelo adotado. A presença de erros na transcrição pode ser atribuída à influência do sotaque alemão do falante, o que, por vezes, dificultou a compreensão precisa do conteúdo. Ademais, a inclusão de nomes de pessoas reais acrescentou uma camada adicional de complexidade, contribuindo para eventuais imprecisões na transcrição.

4.6 Experimento #3: Conteúdo em português / Linguagem técnica / Fala rápida

A Figura 7 apresenta o resultado da transcrição do vídeo "*ReactJS: Guia básico para começar - Parte 01*", já com as devidas correções analisadas manualmente.

> React JS DS, essa tecnologia e uma das mais populares do mundo do front-end Front Change e nesse video eu vou te trazer aqui um guia dir que voce pode usar para conseguir trabalhar ter a ver com ela eu estou aberta aqui com o Tech Guide e qual que e a qualquer proposta eu vou te fazer um tour pelas coisas iniciais que ele tem aqui para voce conseguir trabalhar uma coisa que e super eu sou por importante a gente ter nivel um de profundidade aqui e a gente ter tem uma base boa de JavaScript mas nao so de JavaScript como de CSS, Dom, que sao coisas padroes do navegador que vao te ajudar bastante no dia a dia se voce quiser se aprofundar em qualquer um desses topicos que tem sopossos ter aqui dentro da Tech Guide e so voce clicar em um dos cards que tem aqui e quando ele abrir voce vai ter dentro dele varias coisas que voce pode usar para conseguir estudar, então isso da entro por exemplo tem tanto um link de documentacao quanto video do YouTube como do meu canal quanto conta ate mesmo os cursos da Alura Lura que e o pessoal que fez o Tech Guide gostou? manda aqui nos comentarios se voce tem outra dica para dar pro pessoal a pessoa tambem.

Figura 7: *Resultado da transcrição e anotação do vídeo "ReactJS: Guia básico para começar - Parte 01"*

A tarefa de transcrição do vídeo foi significativamente desafiadora devido à fala acelerada e à supressão de algumas sílabas pelo locutor do vídeo em questão, o que impactou diretamente na precisão do processo. A presença de cortes abruptos entre as frases dificultou a identificação adequada dos pontos de pontuação, como vírgulas e pontos finais, ao longo do discurso. Assim, a ausência de pausas naturais entre as ideias apresentadas resultou em uma transcrição que não conseguiu discernir de maneira adequada os limites de cada frase, impactando assim a compreensão coerente do texto transcrito. Ainda assim, no entanto, é possível entender a ideia geral do texto e o que o conteúdo quer transmitir.

4.7 Experimento #4: Conteúdo em português / Linguagem menos técnica / Fala pausada

O resultado da transcrição do vídeo "*IC - Nossos Professores: Claudia Bauzer Medeiros*", disponível no *YouTube*®, pode ser visualizado na figura 8.

```
Meu nome e Claudia Bauzer Medeiros Bausei-Medero, sou
professora do de Instituto de Computacao da Unicamp, e minha
 presquisa visa permitir que pesquisadores de varias
deviam as areas espalhados por todo mundo, possam vao a
interligar seus dados para obter resultados cientificos em que
 todos aproveitem o que cada um homem faz. Isso se chama
pesquisa colaborativa mediada por dados. Em epoca de pandemia,
 imaginem imagina em pesquisadores da saude colaborando por
meio de dados com pesquisadores que desenvolve equipamentos,
vacinas, ecologos, sociologos, economistas e tantos outros,
para que ao final aparecam solucoes apoiadas em ciencia para a
 saude fisica, mental e o bem-estar de todos. Fantastico, ne?
Sou da Soda Unicamp e marcho hoje machuujo virtualmente
pela ciencia.
```

Figura 8: *Resultado da transcrição do vídeo "IC - Nossos Professores: Claudia Bauzer Medeiros"*

Comparando com os resultados apresentados na Figura 7, também na língua portuguesa, observamos que a utilização de um vídeo composto por diálogos mais lentos e pausados contribui positivamente para uma transcrição mais coesa, com mais pontuações ao longo do conteúdo e com uma menor taxa de erros na transcrição de palavras. Assim, a ideia geral do texto pode ser entendida de forma mais clara, e grande parte dos erros cometidos durante a transcrição não possui um impacto tão relevante para o entendimento do conteúdo como um todo.

4.8 Experimento #5: Conteúdo em espanhol / Linguagem técnica / Fala pausada

Utilizando como base o vídeo "*¿Qué es la interacción Humano-Computadora?*", a Figura 9 apresenta o resultado da transcrição realizada, com as devidas correções realizadas manualmente.

Que es interaccion humano computadora? La interaccion humano computadora tambien se conoce como IHC y se refiere a la disciplina que estudia como las personas interactuan con las computadoras, es decir, la relacion, interaccion y convivencia entre humano y computadora. Para la IHC, lo mas importante es la usabilidad, la seguridad y la funcionalidad. La IHC o interaccion humano computadora surgio a partir de la necesidad que se creo con la introduccion de computadoras en la vida humana. Al notar que se afectaba tanto la cotidianidad del ser humano, llegando incluso a lograr cambios de conductas significativos en el dia a dia. En el articulo interaccion ser humano computadora, usabilidad y universalidad en la era de la informacion, se define la IHC como una ciencia multidisciplinaria y emergente, que se situa en una interseccion entre la psicologia cognitiva, la ingenieria de aplicaciones ergonomicas, las ciencias sociales y la informatica aplicada. Y ademas nos menciona que una forma simple de definir la IHC seria proveer un entendimiento de la forma en que los usuarios trabajan y la forma en que los sistemas computacionales y sus interfaces necesitan ser estructuradas para facilitar el logro de sus tareas. Si analizamos todo nuestro entorno, notaremos que las pantallas estan presentes en casi todas nuestras actividades. Y que interactuar con una o varias al dia es una conducta comun del ser humano, ya sea que se utilicen para trabajo y productividad, entretenimiento entre tenimiento y ocio o simplemente para realizar un proceso. Estudiar y determinar la forma en la que un usuario interactua con una interface o pantalla, se ha convertido en algo vital si queremos encontrar la forma de brindar la mejor experiencia al usuario. Y asi poder facilitar todas las tareas diarias para que el usuario final se beneficie de la efectividad que brindan las computadoras para ayudarnos a resolver problemas o realizar gestiones de forma rapida, segura y sencilla. Finalmente, debemos saber que la IHC no se enfoca en definir si la interaccion con las computadoras es positiva o negativa. Sino Si no en entender la forma en la que se interactua para generar beneficios al usuario, en una era en el que el uso de las computadoras es parte de la conducta del ser humano.

Figura 9: *Resultado da transcrição do vídeo "¿Qué es la interacción Humano-Computadora?"*

Com exceção de dois termos transcritos de forma disjunta, todo o texto transcrito se mostrou completamente coincidente com a fala do autor do vídeo. Como os dois erros apontados não possuem impacto algum sobre o entendimento geral do texto, podemos dizer que a transcrição se mostrou exata ao conteúdo e serviu perfeitamente ao seu propósito. Acreditamos que o principal ponto para que houvesse uma quantidade tão minoritária de erros se deve à interlocução clara, pausada e concisa do autor do vídeo, dizendo cada palavra em um passo inferior a de uma conversa habitual.

4.9 Experimento #6: Conteúdo em espanhol / Fala rápida

Por fim, o resultado do vídeo "*El que habla rapido*", no qual um sujeito recita um trecho de uma obra de Hamlet, pode pode ser visualizado na Figura 10.

```
    y lo voy luego acelerar el ritmo un poco mas. Si, por supuesto.
      Ser o no ser... E aqui el dilema. Que es mejor para el alma?
      Sufrir insulto de fortuna golpes dardos o levantarse en armas
      contra el oce a no del mal o , y o ponerse a el y que asi
      cesen se se morir, dormir nada mas y decir asi que con un
      sueno damos fin a las llagas del corazon y a todos los males,
      herencia evidencia de la carne y decir, ven consumacion te
      con su maciote deseo morir, dormir, dormir, sonar acaso
      sonir a caso que dificil, pues en el sueno de la muerte que
      suenos sobrevendran sobrevengaran cuando despojados de
      ataduras mortales encontremos encontramos la paz pasa .
      Aqui la razon por la que tan longeva ojeva llega a ser la
      desgracia de gracia . Pero quien podra soportar los azotes y
      las burlas laburas del mundo la le injustiza del tirano,
      la afrenta del soberbio frente al sovervio , la angustia del
      amor , despreciado la espera del juicio, la ar rogancia del
      poderoso y la humillacion que la virtud recibe de quien es
      quienes indigno. Eso fue increible. Ya, aun fue absolutamente
       sorprendente en lo disfrute, gracias.
      Credes que hablar rapido Queres que arrastrase? Es
      contagioso? Bien, despues de ese trabalenguas es
      trabal, lenguaces hora de una breve pala.
```

Figura 10: *Resultado da transcrição do vídeo "El que habla rapido"*

Como era esperado, a transcrição deste vídeo se mostrou menos fiel ao aúdio original se comparado com o experimento #5. Claramente, observa-se que a diferença

na velocidade de fala contribui diretamente na taxa de compreensão do modelo de transcrições adotado. As falas rápidas e palavras se mesclando entre si ao longo do vídeo fez com que alguns termos fossem suprimidos ou transformados em uma única palavra ao longo da transcrição, gerando novos vocábulos que não fazem tanto sentido para o contexto da obra. No entanto, é importante ressaltar que, apesar de este experimento apresentar um resultado com mais erros que o anterior, a saída obtida ainda é satisfatória, dado que o vídeo por si só já oferece uma grande dificuldade de compreensão para qualquer pessoa que tentar ouvi-lo naturalmente.

4.10 Taxa de erros de transcrições para cada experimento

A Figura 1 apresenta um compilado dos resultados da métrica de WER para cada um dos experimentos realizados, juntamente com um panorama do número de palavras modificadas (adições, remoções ou substituições) e o número de palavras totais para cada texto.

Tabela 1: Palavras totais, modificações feitas e axa de erros das palavras transcritas pela métrica WER para cada experimento realizado

Experimento	Modificações	Palavras totais	WER
#1	39	469	0,083
#2	20	102	0,196
#3	21	204	0,102
#4	12	109	0,110
#5	4	372	0,010
#6	34	190	0,178

A partir dos dados obtidos, pudemos observar uma média de taxa de erros de palavras de 0,113 dentre todos os experimentos. De forma geral, todos os experimentos possuíram valores relativamente baixos de WER, com nenhum atingindo um valor superior a 0,2, o que demonstra uma boa efetividade do modelo utilizado.

5 Aplicando a transcrição em uma plataforma Web

Com o propósito de avaliar a aplicabilidade prática da funcionalidade de transcrição, integramos essa ferramenta ao contexto da plataforma Web do Quizzing[7]. Na referida plataforma, cada usuário tem a capacidade de criar ou acessar um "Espaço", o qual representa uma temática específica ou um foco de aprendizado. Dentro de cada

[7]A implementação de todas as seções da plataforma pode ser visualizada em `https://gitlab.ic.unicamp.br/quizzing/quizapp`

Espaço, os usuários têm a possibilidade de criar um ou mais "Estudos", que constituem seções delineadas relacionadas à temática do Espaço em questão. A Figura 11 apresenta um esboço da tela de Espaços, com cada *card* representando um diferente Estudo dentro daquela temática específica.

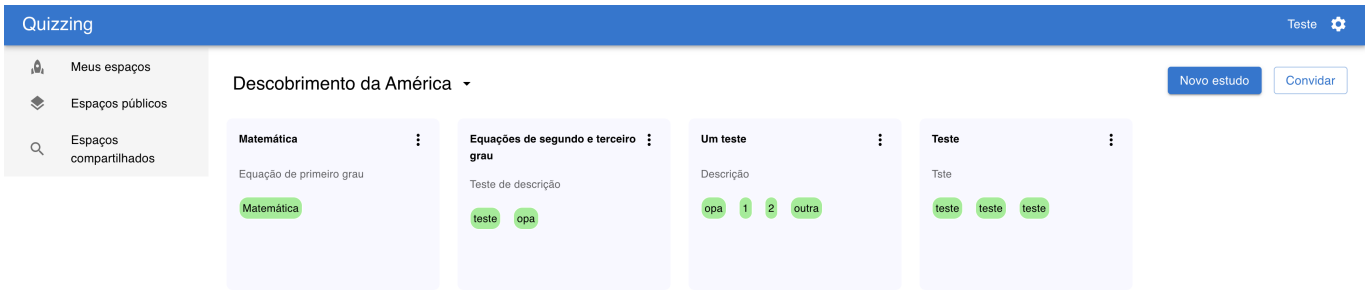

Figura 11: *Versão inicial da tela de Espaços da plataforma Web Quizzing*

No âmbito de cada "Estudo", os usuários têm a liberdade de incorporar novos conteúdos. Para isso, podem optar por inserir diretamente um documento em formato PDF, redigir um texto livre ou incorporar um link de um vídeo do *YouTube*®, que, por sua vez, utiliza todo o procedimento de transcrição detalhado neste trabalho.

A partir desses documentos, os usuários podem gerar novos quizzes, utilizando como base de dados os materiais e conteúdos previamente inseridos dentro da seção de Estudos específica. Essa abordagem prática, incorporando a transcrição de vídeos diretamente na plataforma, não apenas amplia a variedade de recursos educativos, mas também proporciona uma experiência de aprendizado interativa e personalizada para os usuários da plataforma Quizzing.

6 Discussão

Ao realizar diversos experimentos com o modelo de transcrição de vídeos gerado e ao analisar os resultados para cada um dos experimentos, encontramos que a biblioteca PyTube e a ferramenta OpenAI Whisper cumprem com o seu propósito na transcrição de vídeos de forma multilíngue e gratuita. A sua taxa de precisão acaba sendo impactada por uma série de fatores únicos e próprios de cada vídeo, mas em todos é possível compreender a ideia geral que o conteúdo passa.

Dentre os experimentos realizados, alguns dos fatores que impactaram a precisão das transcrições apresentadas foram:

- **Falas rápidas e pouco pausadas** - Observada principalmente nos Experimento #3 e #6, a presença de diálogos mais dinâmicos, com a omissão eventual

de algumas sílabas durante o discurso, faz com que o áudio não seja completamente compreensível e algumas palavras sejam misturadas com outras durante o processo de transcrição, ou até palavras que não existem sejam criadas. Ademais, a presença de cortes no vídeo para trazer essa velocidade impactou no entendimento da transcrição do momento correto de se inserir uma vírgula ou ponto final, gerando textos corridos e com poucas pontuações e dificultando sua assimilação de modo geral.

- **Uso de sotaques** - Como o modelo de transcrição de áudios do *OpenAI Whisper* foi treinado na língua inglesa principalmente por meio de conteúdos provenientes dos Estados Unidos e Reino Unido, os áudios que apresentaram sotaques distintos daqueles com maior treinamento na base possuíram como saída algumas palavras entendidas de maneira errônea. É o caso do Experimento #2, no qual o sotaque alemão de Albert Einstein fez com que a palavra "put"fosse entendida como "third", por exemplo.

- **Ruídos e outros sons de fundo** - Este tópico, de forma geral, apresentou um resultado muito positivo, dado que a transcrição foi capaz de se ater sempre aos sons de falas humanas, e não chegou em momento algum a transcrever de forma errônea outros sons provenientes do áudio. O único ponto claro de impacto sobre o texto de saída é a qualidade do áudio, que pode fazer com que o texto seja cada vez menos confiável à medida que o áudio se torna mais difícil de compreender.

- **Diferentes linguagens** - De forma geral, as transcrições apresentadas mostraram resultados semelhantes independente da língua utilizada, inglês, português ou espanhol. A única distinção marcante quando comparado o inglês com o espanhol e português reside na ausência, nas transcrições em inglês, da criação de palavras que não pertencem ao vocabulário original. Nas outras duas línguas, houve uma propensão para a geração de termos inexistentes.

Analisando todos os experimentos realizados, podemos afirmar que o Experimento #5 apresentou os resultados mais próximos do ideal, com uma média de 1 erro de transcrição a cada 100 palavras. De forma geral, ele foi o que que apresentou um menor número de modificações e o que menos alterou o sentido de alguma frase ou contexto em relação ao conteúdo original.

O Experimento #1 também teve um ótimo desempenho na transcrição do áudio, mas a presença de um erro com a repetição de uma frase, que fez com que seu WER aumentasse. De qualquer modo, acreditamos que isso traria um menor impacto na geração de um quiz por meio do texto obtido do que se comparado com outras modificações textuais realizadas nos outros experimentos. Os Experimentos #2 e #6 desempenharam abaixo da média, com valores de 0,196 e 0,178, respectivamente.

Podemos observar, nesses casos, que os sotaques fortes e a velocidade de interlocução desempenharam um papel relevante para que a transcrição fosse menos exata.

Para o futuro, temos como principais objetivos a conexão do modelo de transcrição de vídeos para texto criado com a plataforma Web do Quizzing. Assim, será possibilitado que esta criação de documentos de texto por meio da submissão de links de vídeos no *YouTube*® ocorra em uma plataforma mais amigável e familiar para que qualquer pessoa usufrua, por meio de uma API que conecte o sistema de transcrição feito em Python ao React da plataforma. Esperamos que esses textos gerados possam ser utilizados como base na geração de novos quizzes, fazendo com que haja um novo meio de captação de conteúdos e dados que não apenas documentos em PDF, livros e apostilas.

7 Conclusão

Neste trabalho, realizamos uma exploração abrangente da transcrição automática de vídeos para textos, visando a geração de conteúdos textuais em linguagem natural para servir como conteúdo base para a criação de quizzes educativos de forma automatizada. Em nossa solução, exploramos o uso da biblioteca *Python Pytube* em conjunto com a ferramenta *OpenAI Whisper* para realizar essa transcrição. Ao longo do processo, submetemos a transcrição a uma série de desafios, incluindo a variação multilinguística, a presença de sons e ruídos de fundo, sotaques, e a velocidade da fala do interlocutor, a fim de mensurar sua efetividade em diferentes cenários e perspectivas. Contribuímos para a criação de um ecossistema educativo ao desenvolver uma versão inicial do *front-end* e do *back-end* da plataforma Web do Quizzing. Esta iniciativa proporciona um ambiente interativo no qual indivíduos podem compartilhar, participar e acessar quizzes sobre uma ampla gama de temas, promovendo assim a disseminação do conhecimento de maneira dinâmica e acessível.

Referências

[1] Chunliang Yang, Liang Luo, Miguel Vadillo, Rongjun Yu, David Shanks. Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review, 2021. 10.1037/bul0000309. Acesso em: 18 dez. 2023.

[2] OpenAI. Whisper: Speech Recognition with Transformer. Disponível em: `https://openai.com/research/whisper`. Acesso em: 29 nov. 2023.

[3] Fonte: OpenAI. Disponível em: `https://images.openai.com/blob/18ff9c06-7853-4e3b-946f-508f0cd7ed13/asr-details-desktop.svg?width=10&height=10&quality=50`. Acesso em: 29 nov. 2023.

[4] John Levis, Ruslan Suvorov. Automatic Speech Recognition, 2012. 10.1002/9781405198431.wbeal0066. Acesso em: 30 nov. 2023.

[5] Nick Ficano. pytube Documentation, Release 15.0.0, 2023. Acesso em: 01 dez. 2023.

[6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, 2022. Acesso em: 03 dez. 2023.

[7] Lawrence Rabiner, Biing-Hwang Juang. Fundamentals of Speech Recognition, 1993. Acesso em: 03 dez. 2023.

[8] Peidong Wang, Tara N. Sainath, Ron J. Weiss. Multitask Training with Text Data for End-to-End Speech Recognition, 2020. Acesso em: 04 dez. 2023.

[9] Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, Alexis Conneau. Massively multilingual joint pre-training for speech and text, 2022. Acesso em: 04 dez. 2023.

[10] Tomás Baviera. Ser o no ser: La cuestión sobre Hamlet, 2019. Acesso em: 09 dez. 2023.

[11] AssemblyAI. Dylan Fox. Is Word Error Rate Useful?, 2023. Disponível em `https://www.assemblyai.com/blog/word-error-rate`. Acesso em: 13 dez. 2023.

[12] Rohit Prabhavalkar, Tara N. Sainath, Yonghui Wu, Patrick Nguyen, Zhifeng Chen, Chung-Cheng Chiu, Anjuli Kannan. Minimum Word Error Rate Training for Attention-based Sequence-to-Sequence Models, 2017. Acesso em: 13 dez. 2023.