Integração de ferramentas de Inteligência Artificial para assistentes virtuais

Alexandre Costa Ferro Filho
alexandre_ferro@discente.ufg.br
Decivan Damasceno Araujo Filho
decivan.damasceno@discente.ufg.br
Evellyn Nicole Machado Rosa
nicole@discente.ufg.br
Guilherme Henrique dos Reis
guilherme_reis@discente.ufg.br
Marcelo Henrique Lopes Ferreira
marcelomarcelo2@discente.ufg.br
Murilo de Oliveira Guimarães
muriloguimaraes@discente.ufg.br

UFG – Universidade Federal de Goiás
INF – Instituto de Informática

Resumo:

Este relatório propõe a elaboração de uma assistente pessoal que integre o ChatGPT com voz, áudio e Stable Diffusion, permitindo aos usuários interagir com o modelo de linguagem de maneira mais intuitiva e natural. Esta integração aumenta a eficiência e conveniência do uso da tecnologia, tornando-a mais acessível e ampliando suas aplicações em vários setores da sociedade.

Palavras Chave:

ChatGPT, Stable Diffusion, Processamento de Áudio e Voz, STT, TTS .

1- Introdução

Ferramentas usando Inteligência Artificial (IA) estão cada vez mais populares e presentes no cotidiano, como por exemplo o ChatGPT, que conseguiu a marca de 5 milhões de usuários em apenas 1 semana [1]. Tal sucesso acontece não somente devido a fascinante tecnologia que envolve o funcionamento dessas ferramentas, como também a utilidade e auxílio proporcionado por tais. Nessa conjuntura, há de valer diferentes maneiras possíveis de utilizar a IA no cotidiano humano.

Em paralelo, outra ferramenta tecnológica cada vez mais utilizada são as assistente virtuais, que podem auxiliar o usuário nas tarefas diárias como fazer uma pesquisa, tocar uma música, realizar um cálculo, e até mesmo conversar . Portanto, a fim de melhorar o uso de assistente virtuais, é de suma necessidade o desenvolvimento de metodologias envolvendo IA que promovam seu uso mais efetivo. Nessa conjuntura, o relatório a seguir propõe a elaboração de um assistente virtual que integra diferentes tecnologias de IA para promover novas funcionalidades e melhor auxiliar seus usuários.

Tais funcionalidades que serão abordadas a seguir envolvem, principalmente, a geração de imagens e o processamento de linguagem natural por parte do assistente virtual. Isso será possível através da utilização de métodos Speech-to-Text (STT) , que converte a fala do usuário para um texto, posteriormente interpretado pelas tecnologias de IA do ChatGPT, que gera um prompt para o Stable Diffusion, gerando uma imagem, se desejada. Além desse prompt, será utilizado o método de Text-to-Speech (TTS), que converte a resposta do ChatGPT em forma de texto para áudio, facilitando a interação com o usuário.
Para atingir esse objetivo, ressalta-se a seguir os fundamentos teóricos das técnicas utilizadas pelas tecnologias IA assim como outras técnicas usadas ao longo da metodologia.

2 - Fundamentação teórica

A fim de compreender a metodologia a ser apresentada futuramente, vale entender conceitos

prévios, técnicas e algoritmos a serem utilizados para a resolução do problema.

2.1 - Pré Processamento de Áudio

A captura de áudio para a utilização no projeto deve ser realizada através de um dispositivo de gravação, como microfone, o qual obtém o sinal analógico do áudio. Esse sinal deve ser amostrado adequadamente para ser armazenado e processado por computadores, a fim de evitar a perda de informações importantes no sinal, logo, deve-se seguir o Teorema de Amostragem ou regra da amostragem de Nyquist, que estabelece que a taxa de amostragem- velocidade em que o sinal é amostrado deve ser pelo menos duas vezes maior que a frequência máxima do sinal.

A etapa de pré-processamento de dados de áudio é fundamental antes de serem processados por algum tipo de algoritmo ou modelo, para que esse sinal seja analisado com mais precisão e eficiência. Dentre as tarefas do pré-processamento, que possivelmente serão utilizadas dentro do projeto, podem-se destacar a remoção de ruído, advindo do ambiente em que foi feita a gravação da fala ou do aparelho que obteve o áudio, aplicação de filtros, utilizados muitas vezes para remover frequências indesejadas, ou suavizar o áudio para corrigir alguma distorção existente no sinal. Além dessas, outras possíveis técnicas que podem ser utilizadas existe a normalização de volume, usada para garantir que o áudio tenha um nível de volume padrão, ou a segmentação de palavras que se trata de identificar os limites das palavras individuais dentro de um sinal de áudio contínuo, permitindo que o sistema de reconhecimento de fala as trate como unidades separadas.

Por fim, a última realização de um pré-processamento do sinal de áudio seria dependente do tipo de entrada do modelo adotado, podendo ser necessária a extração do espectrograma do sinal, o qual o representa graficamente a distribuição de frequência do sinal de áudio. Esse tipo de representação pode ser obtido através da aplicação da Transformada de Fourier, uma transformação matemática que permite a análise da frequência.

Para se obter um espectrograma, o sinal de áudio é primeiramente dividido em pequenos segmentos de tempo conhecidos como janelas. Em seguida, a Transformada Discreta de Fourier é aplicada em cada janela, permitindo a obtenção da representação de frequência. Ao visualizar os resultados da DFT ao longo do tempo, com a frequência no eixo y e o tempo no eixo x, é possível obter uma imagem que representam os espectrogramas, possivelmente utilizados.

2.2 - STT e TTS

STT é uma tecnologia de processamento de fala que permite a conversão da fala em texto. Ele é usado dentro do projeto para capturar a voz do indivíduo e convertê-la em texto para compreensão do ChatGPT. O funcionamento dessa ferramenta pode ser dividida em etapas: aquisição do áudio, processamento de fala e reconhecimento de fala.[2]

O algoritmo de STT usa técnicas de inteligência artificial e aprendizado de máquina para analisar o sinal de áudio e produzir a transcrição da fala. Isso envolve a identificação das palavras e frases faladas, bem como a avaliação de probabilidades para determinar a transcrição mais provável.

Text-to-Speech (TTS) se trata do processo inverso do STT, sendo o processo pelo qual um texto é convertido em fala artificial, usado dentro do projeto para transmitir o texto gerado pelo ChatGPT, em forma de voz. Esse processo envolve diversas etapas, como: a análise do texto, em que é analisado o texto de entradas- palavras, frases e símbolos-; a geração de fonemas, etapa em que as palavras do texto são convertidas em fonemas- unidade básica de som utilizada para formar palavras-; modelagem da fala, o qual envolve a seleção de parâmetros como velocidade, entonação e inflexão; e por fim a síntese da voz, técnica para combinar os fonemas em um áudio, incorporando as informações sobre a modelagem da fala.[3]

2.3 - ChatGPT

O ChatGPT foi desenvolvido pela OpenAI, uma organização de pesquisa de inteligência artificial sem fins lucrativos, fundada em 2015. A ferramenta é

um sistema de chatbot que permite fazer perguntas e, muitas vezes, receber respostas rápidas e úteis.

O ChatGPT é baseado em uma arquitetura de transformers, que é uma abordagem inovadora para o processamento de linguagem natural. A arquitetura de transformers foi desenvolvida para superar algumas limitações dos modelos anteriores, como a dificuldade em processar informações em sequência. O seu diferencial está na capacidade de atender a relações de dependência entre elementos em uma sequência, como palavras em uma frase. Essa arquitetura é baseada em redes neurais, onde as informações de entrada são codificadas em uma representação densa de vetor, que é passada por várias camadas para realizar tarefas como classificação de texto, geração de texto e tradução. Além disso, o modelo transformers é altamente escalável e pode ser facilmente treinado com grandes quantidades de dados, o que o torna uma opção atraente para aplicações em processamento de linguagem natural.

No entanto, a ferramenta disponibilizada pela OpenAI ainda sofre limitações, uma delas é relacionada a sua dependência textual, já que só é possível realizar perguntas em texto e obter respostas em texto. Essa desvantagem torna o ChatGPT menos conveniente para alguns usuários, principalmente aqueles que preferem usar comandos de voz ou têm dificuldades para digitar. Outra desvantagem enfrentada pelo ChatGPT é em relação a não citação das fontes utilizadas pelo modelo para gerar respostas, o que gera uma desconfiança em relação às informações dadas pela ferramenta. Apesar dessas dificuldades, o chat é de excelente auxílio individual, assim, para torná-lo ainda mais utilizável, seria interessante a criação de uma assistente pessoal integrada com o ChatGPT e que funcione por comandos de voz e retorne respostas em áudio.

2.4 - Stable Diffusion

O modelo de Aprendizagem Profunda (Deep Learning) de texto para imagem, conhecido como Stable Diffusion, é uma tecnologia de ponta lançada em 2022. Ele permite a geração de imagens detalhadas a partir de descrições em texto, além de ser aplicável a outras tarefas [4]. Com o código disponível publicamente e sua compatibilidade com a maioria das máquinas comuns, a tecnologia tem atingido uma ampla base de usuários.

O Stable Diffusion possui a habilidade única de gerar novas imagens completamente a partir do zero, bastando uma descrição textual dos elementos desejados na imagem. Imagens existentes também podem ser editadas para incluir ou excluir elementos. No entanto, o modelo ainda apresenta limitações em seu treinamento para entender membros e faces humanas, o que pode resultar em imagens confusas.

Além disso, a Stable Diffusion também se destaca por ser uma tecnologia aberta e colaborativa, o que significa que qualquer pessoa interessada pode contribuir para o seu desenvolvimento e aprimoramento. Ao disponibilizar o código e dar liberdade de uso das imagens geradas, a Stable Diffusion incentiva a criatividade e a imaginação dos usuários.

Portanto, devido às grandes capacidades do Stable Diffusion e do ChatGPT, o projeto engloba ambas tecnologias. Enquanto o ChatGPT é capaz de descrever cenários complexos com extrema precisão, o Stable Diffusion consegue transformar essas descrições em imagens realistas e impactantes. O projeto, então, permite que o usuário descreva o que deseja na imagem, logo, o ChatGPT o transforma em uma descrição detalhada e o Stable Diffusion gera a imagem final e a retorna para o usuário.

2.5 - Interface e Grad.io

O avanço do aprendizado de máquina possibilitou o desenvolvimento de novas Inteligências Artificiais (IA) e modelos de Processamento de Linguagem Natural (NLP), o que resultou em uma ampla utilização desses modelos por grande parte da população, como é o exemplo do ChatGPT e algumas IAs que geram imagens. Contudo, mesmo que o conhecimento desses chatbots sejam muito amplos, não é o suficiente para que o usuário consiga interagir com esse chatbot, por isso é muito importante uma interface simples, de fácil interação e intuitiva.

Uma interface é uma forma de comunicação entre um usuário e um sistema, o design de uma interface deve pensar nas formas que o usuário deve interagir com o sistema de forma que o usuário termine a utilização sem reclamações e satisfeito. Por isso uma boa interface deve conter botões, menus, ícones, caixas de texto, entre outros. Outro aspecto

importante para uma boa interface é que ela seja atraente e bem projetada, pois quanto mais claro e atraente for a interface, mais ela vai cativar e chamar a intenção do usuário[5]

Com isso, para criar uma interface atraente e cativante para o chatbot, foi utilizado o Grad.io, que é um pacote do Python open-source que auxilia a fazer interfaces para o usuário. O Grad.io consegue disponibilizar para o modelo uma página web, dessa forma, facilitando a utilização de outras pessoas que possam estar interessadas no chatbot. A implementação é feita em python e conforme a documentação para a criar a interface para o chatbot com todos os botões, ícones e caixas necessárias.

O Grad.io utiliza como input diversos tipos de dados, como áudio, texto, imagem, entre outros[6]. Dessa forma, é possível integrar todos os algoritmos utilizados no chatbot e facilitar a interação do usuário com todos esses sistemas.

3 - Metodologia

Foi optado pela utilização do modelo Whisper para a implementação do Speech-to-Text, devido à sua transcrição satisfatória sem a necessidade de um pré-processamento mais complexo, à disponibilidade do modelo para o idioma português (já que o modelo é multilinguístico) e à sua capacidade de realizar transcrições precisas sem a necessidade de treinamento adicional, o que economiza tempo.

O processo de pré-processamento começa com a amostragem do sinal analógico, seguida pelo recorte do áudio em segmentos de 30 segundos. Esses segmentos de áudio são então convertidos em imagens de log-mel-espectrograma, que representam a energia de frequência do sinal de áudio em uma escala logarítmica. O modelo Whisper é configurado para transcrever apenas em português, é o tamanho do modelo adotado foi o “base”, pois pareceu o mais adequado[7]. O resultado do STT é então passa por uma análise de palavras-chave, que substitui palavras que possam prejudicar a geração de prompts de qualidade no ChatGPT. Somente após esse processo, o resultado é utilizado como entrada no chatbot inteligente.

O prompt resultante é utilizado como entrada no TTS, o qual teve como modelo optado o GTTS, devido à sua facilidade de implementação e disponibilidade em português. O TTS, então, gera uma saída de áudio contendo a resposta do chatbot.

Por fim, para a geração de imagens, a mesma entrada utilizada para gerar o prompt no ChatGPT é analisada em busca de palavras-chave que indicam se o usuário deseja que seja gerada uma imagem ou se basta a resposta gerada pelo chatbot. Dependendo da análise, o prompt resultante é utilizado como entrada no modelo Stable Diffusion, a fim de gerar uma imagem mais detalhada, que gera uma imagem correspondente, ou é retornada uma imagem padrão quando a geração de imagem não é necessária.

Para a integração com a interface, utilizamos os métodos do Grad.io. Para esses métodos precisamos passar o input e ele retorna um output na saída diretamente para o usuário. Além disso, temos que passar como parâmetro a função que a interface deve utilizar, que no nosso caso foi toda a comunicação entre os algoritmos, citado anteriormente. Para os inputs foram utilizados o audio do usuário, que é feito a partir da gravação com o microfone, e assim, após passar pela função ele nos retorna 4 outputs, o primeiro sendo o STT, mostrando a conversão do áudio do usuário para texto, a segunda sendo a resposta do ChatGPT, após a pergunta ou comando do usuário passar pela API, o terceiro, caso o usuário requisite a geração de uma imagem, será a descrição do ChatGPT para o Stable Diffusion, caso ele não requisite uma imagem, aparecerá uma imagem pré-definida, e por último o TTS, com a transcrição da resposta do ChatGPT de texto para áudio. Com essa interface também é possível gerar um link que permite o compartilhamento para outros usuários. Consulte o Apêndice B para acessar o código completo.

4 - Resultados e Discussão

O resultado obtido foi relativamente satisfatório, isso é perceptível pelo produto final através de testes realizados.

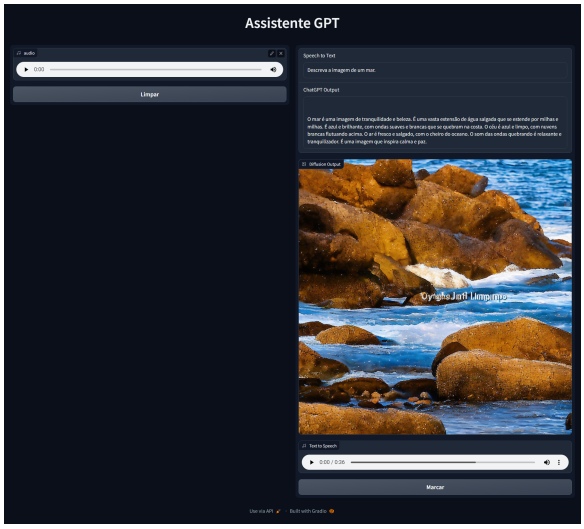


***Figura 1**: Exemplo de uma resposta fornecida pelo nosso chatbot em que o usuário pede uma imagem*

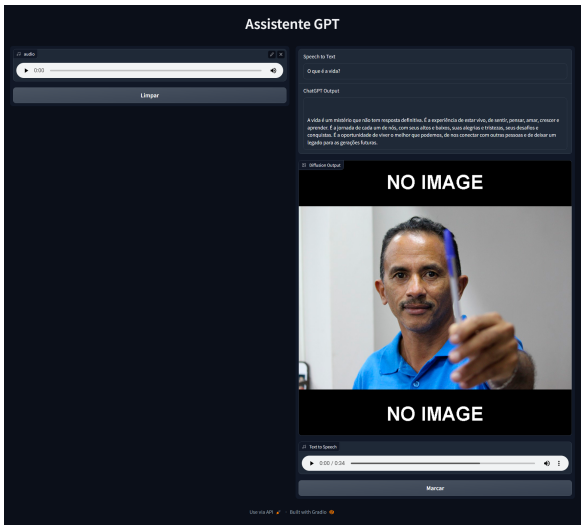


***Figura 2**: Exemplo de uma resposta fornecida pelo nosso chatbot em que o usuário não pede uma imagem*

Contudo, existem duas inconsistências principais no produto obtido que devem ser abordadas.
O primeiro é referente ao ChatGPT em si. Como mencionado anteriormente, o chatbot de inteligência artificial não fornece fontes a suas respostas ou provê respostas necessariamente corretas. Desse modo, o assistente criado pode eventualmente disponibilizar informações falsas, já que o ChatGPT é especializado em emular linguagem natural e não fornecer resultados de pesquisa sempre precisos.

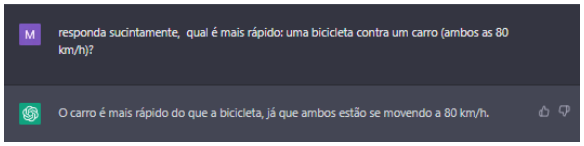


***Figura 3**: Exemplo de uma resposta errônea fornecida pelo ChatGPT*

Já o segundo problema de performance do assistente ocorre durante a fase de STT. Nosso palpite é que talvez fosse necessário um fine-tuning do modelo para realizar a transcrição com maior precisão em português ou analisar melhor como o ruído ou outros fatores estava afetando o desempenho, por mais que o modelo lide bem com ruídos. Dessa maneira, o assistente muitas vezes mal interpreta o que foi lhe referido, comprometendo a performance do produto como todo.

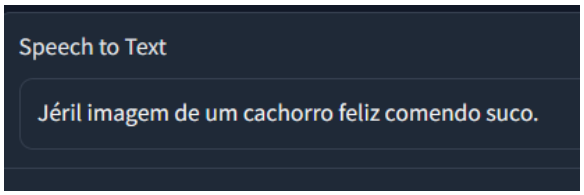


***Figura 4**: Exemplo de um erro de funcionamento do STT. A saída gerada é correspondente a fala: "Gere uma imagem de um cachorro feliz tomando suco".*

5 - Conclusão

Foi possível compreender através do que foi desenvolvido neste trabalho tópicos muito importantes abordados pelo bacharelado de Inteligência Artificial, como a integração do

ChatGPT com ferramentas de inteligência artificial, o processamento digital de sinais , TTS, STT e o funcionamento do Stable Diffusion. Eles, quando integrados, representam um avanço significativo na capacidade de conversação e interação do modelo de linguagem GPT.

Além disso, foi revelado que o produto -antes criado pela simples motivação de integrar o ChatGPT a uma assistente- ataca problemas maiores do que fora antecipado.

A utilização de TTS e SST permite a interação mais natural com os usuários, proporcionando uma experiência de conversação mais humana, além de ser uma maneira de permitir uma maior acessibilidade ao chat, já que pessoas mais idosas ou que possuem alguma deficiência visual terão uma facilidade maior para utilizar o assistente virtual. Por fim, com a implementação do ChatGPT para criar prompts para o Stable Diffusion, obtém-se imagens de alta qualidade devido a descrição bem detalhada.

Logo, com o uso dessas tecnologias avançadas, o ChatGPT pode se tornar uma ferramenta poderosa e eficiente para comunicação e interação com os usuários, abrindo novas possibilidades em áreas como atendimento ao cliente, assistentes virtuais, e outras aplicações que exigem interação humano-máquina.

6 - Referências

[1] “ChatGPT hit 1 million users in 5 days: Here’s how long it took others to reach that milestone”, IndianExpress. Disponível em<https://indianexpress.com/article/technology/chatgpt-hit-1-million-users-5-days-vs-netflix-facebook-instagram-spotify-mark-8394119/>. Acesso em: 09 de fev. de 2023.

[2] "O que é conversão de fala em texto?". Amazon Web Services, 2023. Disponível em: <https://aws.amazon.com/pt/what-is/speech-to-text/#:~:text=Speech%20to%20text%20is%20a,recognition%20or%20computer%20speech%20recognition.>. Acesso em: 09 de fev. de 2023

[3] PORTASIO, Luiza. "Você sabe o que é TTS?". Medium, 2021. Disponível em: <https://medium.com/dialograma/voc%C3%AA-sabe-o-que-%C3%A9-tts-dfa2e3835c6b.>. Acesso em: 09 de fev. de 2023

[4] ALAMMAR, Jay. "The Illustrated Stable Diffusion". *jalammar.github.io*. Acesso em: 09 de fev. de 2023.

[5] SMESTAD, Tuva Lunde, and Frode Volden. "Chatbot personalities matters: improving the user experience of chatbot interfaces." *Internet Science: INSCI 2018 International Workshops, St. Petersburg, Russia, October 24–26, 2018, Revised Selected Papers 5*. Springer International Publishing, 2019.

[6] GRAD.IO. Quickstart. Disponível em: <.https://gradio.app/qNextuickstart/>. Acesso em: 09 de fev. de 2023

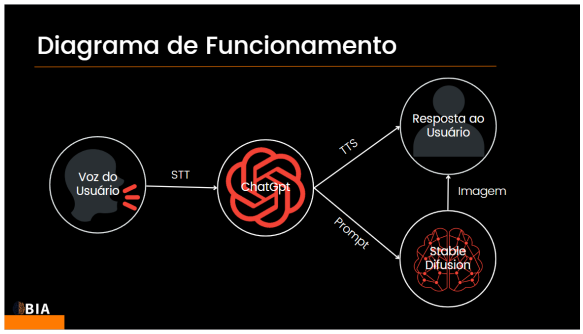
[7]OPENAI. Whisper. GitHub, 2018. Disponível em: <https://github.com/openai/whisper>. Acesso em: 14 de fev. de 2023

[8]GTTS. Documentação do módulo GTTS. Read the Docs, 2023. Disponível em: <https://gtts.readthedocs.io/en/latest/module.html#languages-gtts-lang>. Acesso em: 14 de fev. de 2023

7 - Apêndices

APÊNDICE A -

DIAGRAMA DE FLUXO DO FUNCIONAMENTO DO NOSSO CHATBOT



APÊNDICE B - CÓDIGO FUNCIONAL DO NOSSO CHATBOT

https://colab.research.google.com/drive/1dGk1oYfbe83KEwJkR-xmGkuJEdioL6Uv?usp=sharing

APÊNDICE C - CANVAS DE APRESENTAÇÃO DO PROJETO

https://www.canva.com/design/DAFarYYC1b4/kcGhU1MNeQd6nA3tlA0zUw/edit?utm_content=DAFarYYC1b4&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton