Sistema de Conversão de Voz-Texto com GPT

João Vitor Roriz da Silva
Universidade Federal do Espírito Santo
Vitoria, Brasil
joaovitor.roriz@gmail.com

***Abstract*—O avanço nas tecnologias de inteligência artificial (IA) e processamento de linguagem natural (PLN) tem impulsionado o desenvolvimento de sistemas interativos e intuitivos para conversão de voz em texto. Este estudo apresenta uma abordagem mais simples que combina a capacidade de reconhecimento de voz com a complexidade do modelo de linguagem GPT da OpenAI, criando uma plataforma dinâmica para interação voz-texto. O sistema utiliza bibliotecas como PyAudio e Speech Recognition para a captação e transcrição da voz do usuário. Paralelamente, emprega-se a API da OpenAI em conjunto com a Google Text-to-Speech para processar consultas e gerar respostas coerentes e contextualizadas em formato de áudio. Além disso, o trabalho incorpora o uso do VITS (Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech), um modelo vanguardista na síntese de voz, para aprimorar a qualidade e a naturalidade da voz sintetizada utilizada nas respostas do sistema.**

***Index Terms*—Autoencoder, PLN, OpenAI, Voz-Texto, VITS**

I. Introdução

Recentemente, temos testemunhado uma transformação significativa no campo da tecnologia, especialmente com o avanço da Inteligência Artificial (IA) e do Processamento de Linguagem Natural (PLN). Esses desenvolvimentos estão redefinindo nossa interação com máquinas e dispositivos eletrônicos. Dentro desse panorama evolutivo, destaca-se o progresso no desenvolvimento de sistemas de conversão de voz em texto. Este trabalho apresenta uma abordagem simples nessa área, combinando a precisão do reconhecimento de voz com a sofisticação na compreensão e geração de linguagem natural, características marcantes do modelo GPT da OpenAI. O foco deste estudo é fazer um sistema baseado em GPT que receba voz, converte em texto, gerando uma demanda para GPT e que a GPT retorne de volta o texto no formato de in-streaming que seria convertido de volta em voz. O sistema que desenvolvemos une a praticidade da captura de voz, empregando ferramentas como PyAudio e Speech Recognition, à robusta capacidade de processamento de linguagem do GPT. Além de transcrever com eficácia a voz humana, nosso sistema processa consultas em linguagem natural e gera respostas pertinentes, que são convertidas em fala. Esta integração entre a API da OpenAI e o Google Text-to-Speech facilita uma interação fluida e natural. Inovamos também ao incorporar o modelo VITS (Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech), que representa um avanço significativo na qualidade da síntese de voz. O VITS é essencial para melhorar a naturalidade e qualidade da voz sintetizada nas respostas do sistema, um fator crucial para a usabilidade e aceitação do sistema pelos usuários. Assim, este trabalho se insere na confluência de campos como IA, PLN e interação humano-computador, visando apresentar uma solução técnica de vanguarda e contribuir para o entendimento de como as interfaces de voz podem aprimorar a acessibilidade e eficiência em sistemas baseados em IA. A análise dos resultados busca oferecer insights relevantes para o avanço dos sistemas de conversão de voz em texto, tornando-os mais eficientes e humanizados."

II. Trabalhos Correlatos

Este estudo se baseia em avanços significativos no campo do processamento de linguagem natural e síntese de voz. Um componente central de nossa abordagem é o modelo VITS, desenvolvido por Walnut [1], que representa uma inovação significativa na síntese de voz end-to-end. Uma adaptação desse modelo para o português [2] foi fundamental para a aplicabilidade do sistema em ambientes lusófonos. Além disso, o tutorial disponível no YouTube [3] foi uma referência valiosa para a compreensão e implementação prática do modelo VITS em nosso trabalho. Paralelamente, a integração com a API da OpenAI [4] permitiu o acesso a um dos mais avançados modelos de linguagem baseados em IA, essencial para a conversão eficaz de texto em linguagem natural e para o processamento de consultas complexas.

III. Metodologia

O processo de gravação e transcrição de voz no sistema desenvolvido é descrito a seguir:

A. Processo de Gravação e Transcrição

1) **Iniciação da Gravação de Áudio**: A gravação é iniciada pelo usuário através da interação com o sistema. Após a execução do arquivo `main.py` no terminal, o usuário tem a opção de digitar uma pergunta ou pressionar 'Enter' para iniciar a gravação de áudio. Uma vez concluída a fala, o usuário deve pressionar 'Enter' novamente para finalizar a gravação e enviar o áudio para processamento.

Imagem ilustra o processo de iniciação da gravação de áudio.

2) **Captura de Áudio com PyAudio**: O áudio é capturado em tempo real utilizando a biblioteca PyAudio. Esta biblioteca é configurada para gravar áudio em formato WAV, armazenando o arquivo de áudio temporariamente no sistema. A configuração do PyAudio inclui a definição da taxa de amostragem de 44100 Hz, utilizando um único canal (mono) e no formato da amostra de áudio (pyaudio.paInt16) para garantir uma gravação de alta qualidade.

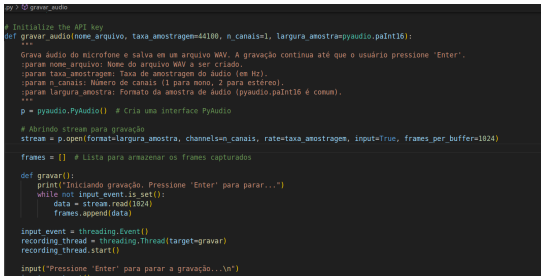

Imagem ilustra o trecho de código do gravar audio.

3) **Transcrição de Áudio com Speech Recognition**: Após a gravação, o arquivo de áudio WAV é processado pela biblioteca Speech Recognition para converter o áudio em texto. Esta etapa envolve a análise do arquivo de áudio e a identificação precisa das palavras faladas, convertendo-as em um formato textual que possa ser processado posteriormente pelo sistema.

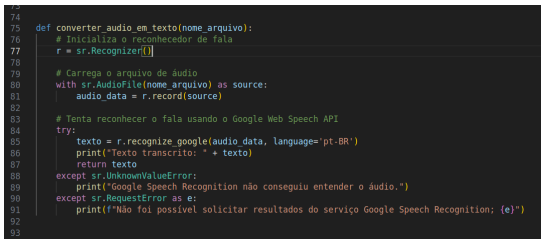

Imagem ilustra o trecho de código de converter audio em texto.

B. Integração com OpenAI GPT e gTTS

A integração com a API da OpenAI e o Google Text-to-Speech (gTTS) é um componente essencial do nosso sistema, permitindo a transformação eficiente do texto transcrito em respostas faladas. O processo é detalhado abaixo:

1) **Processamento do Texto com OpenAI GPT:** Após a transcrição do áudio, o texto é enviado para a API da OpenAI para processamento. Utilizamos especificamente o modelo ”gpt-3.5-turbo” para este fim. A API é configurada para receber o texto transcrito como parte de uma série de mensagens, onde cada mensagem é um input ou resposta anterior. O código para esta integração é:

```
response = openai.chat.completions.
    create(
    model="gpt-3.5-turbo",
    messages=messages,
    temperature=0.5
)
```

Este trecho de código configura o modelo GPT-3.5 para processar as mensagens com uma ’temperatura’ de 0.5, equilibrando entre previsibilidade e criatividade nas respostas.

2) **Conversão de Texto em Fala com gTTS:** As respostas geradas pelo modelo GPT são então convertidas em fala usando o Google Text-to-Speech (gTTS). O gTTS é uma ferramenta que permite converter texto em um arquivo de áudio MP3 com uma voz sintetizada. O idioma e a velocidade da fala podem ser configurados. Por exemplo:

```
tts = gTTS(text=answer[0], lang=
    idioma, slow=False)
arquivo_audio = "audio.mp3"
tts.save(arquivo_audio)
```

Aqui, ’answer[0]’ contém o texto gerado pelo GPT, que é convertido em fala no idioma especificado e salvo como um arquivo MP3.

3) **Uso Opcional do Cliente de Áudio OpenAI:** Adicionalmente, oferecemos a opção de usar o cliente de áudio da própria OpenAI para a síntese de voz. Isso proporciona uma alternativa ao gTTS, possibilitando uma variedade de vozes e estilos. O código correspondente é:

```
response = client.audio.speech.
    create(
    model="tts-1",
    voice="nova",
    input=answer[0],
)
```

Este código utiliza o modelo ”tts-1” com a voz ”nova” para criar um arquivo de áudio a partir do texto gerado pelo GPT.

Este processo de integração assegura que nosso sistema forneça respostas rápidas e naturais, enriquecendo a interação do usuário com a máquina.

C. Uso Opcional do VITS

Além de usar o cliente de áudio da própria OpenAI,temos a opção de utilizar o modelo VITS para a síntese de voz. O VITS, especialmente em sua versão adaptada para o português, oferece uma qualidade superior de voz sintetizada, com maior naturalidade e expressividade. Para a implementação do VITS, especialmente sua versão em português, é necessário seguir os seguintes passos:

1) Instalar as dependências necessárias para o VITS, conforme documentado em [1].
2) Configurar o ambiente de execução para suportar a síntese de voz em português, seguindo as orientações de [2].
3) Integrar o VITS com nosso sistema para permitir a conversão do texto gerado pelo GPT em fala sintetizada.

Para esse trabalho , optamos por utilizar o TTS-Portuguese Corpus, um conjunto de dados extensivo especificamente voltado para a síntese de fala em português. Este corpus é

referenciado pelo artigo *TTS-Portuguese Corpus*, que oferece uma análise detalhada do mesmo. Os passos para a integração e treinamento incluem:

1) Aquisição do TTS-Portuguese Corpus, assegurando que todas as diretrizes éticas e de uso de dados sejam seguidas.
2) Preparação dos dados, o que envolve a limpeza, a segmentação e a normalização do texto e áudio contidos no corpus.
3) Treinamento do modelo de síntese de voz com o corpus preparado, utilizando técnicas avançadas de aprendizado de máquina para garantir a máxima eficácia e naturalidade da voz sintetizada.
4) Avaliação e ajuste contínuo do modelo, com base nos resultados obtidos e no feedback dos usuários, para aprimorar a qualidade e a precisão da síntese de voz.

O uso do TTS-Portuguese Corpus é fundamental para garantir que o sistema desenvolvido seja capaz de produzir uma voz sintetizada de alta qualidade e que soe natural para falantes nativos de português. Além disso, a flexibilidade do VITS permite que ele seja adaptado e treinado com outras bases de dados, o que o torna uma ferramenta valiosa para a síntese de voz em diversos idiomas e contextos. Esta capacidade de treinar o modelo com diferentes conjuntos de dados é crucial para expandir a aplicabilidade do sistema em ambientes multilíngues e para diferentes variantes linguísticas, proporcionando uma gama mais ampla de opções de voz e maior personalização na experiência do usuário.

IV. Experimentos e Resultados

A. Configuração do Experimento

O computador utilizado é um notebook equipado com 16 GB de memória RAM e uma placa de vídeo NVIDIA GeForce GTX 1060. As instruções detalhadas para a implementação do sistema estão disponíveis no arquivo README no GitHub: https://github.com/joaovitorroriz/VoztextocomGPT.git. Este arquivo README oferece um guia para a instalação do sistema, bem como informações sobre as versões das bibliotecas utilizadas, como PyAudio, Speech Recognition e OpenAI API. A implementação do VITS está referenciada na bibliografia como [2]. Os testes realizados no sistema são demonstrados através do vídeo disponível em https://youtu.be/QBWT2b3iVIA. Incluiremos uma imagem do teste que ilustra a mensagem falada, a resposta em texto gerada pelo GPT, assim como o áudio da conversa.

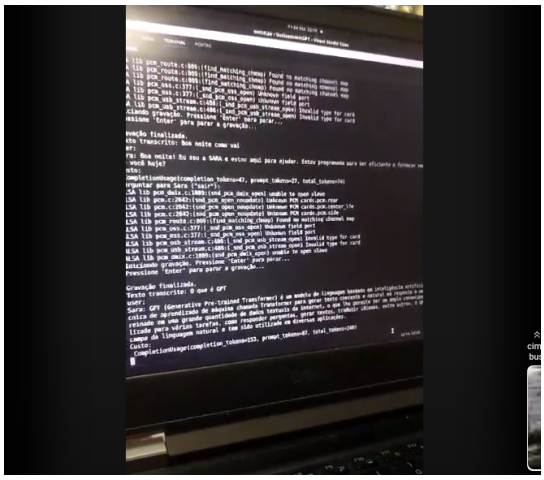

Imagem ilustra a tela do video treinamento.

B. Resultados Esperados

- **Precisão do Reconhecimento de Voz:** Em geral, o sistema demonstra uma boa capacidade de reconhecimento de voz, e a API do GPT interpreta adequadamente as perguntas. No entanto, ruídos ambientais, como conversas paralelas, trânsito ou ventiladores, podem afetar negativamente o reconhecimento. Testes iniciais foram realizados usando o microfone do notebook e um microfone de fone de ouvido. Futuros testes poderão explorar o uso de equipamentos diferentes e técnicas de redução de ruído.
- **Naturalidade da Voz Sintetizada pelo VITS:** Para a síntese de voz, utilizou-se a base de dados TTS-portuguese. Conforme descrito no artigo relacionado, esta base contém um total de 71.358 palavras, com 13.311 palavras distintas. O conjunto de dados desenvolvido abrange aproximadamente 10 horas e 28 minutos de gravação de um único locutor, com qualidade de 48 kHz, totalizando 3.632 arquivos de áudio em formato Wave. A duração dos arquivos varia entre 0,67 e 50,08 segundos. Com a utilização do VITS, observou-se uma execução satisfatória na síntese de voz. Há potencial para melhorias, como a expansão da base de dados ou a experimentação com outras bases. Durante o treinamento da rede do VITS, foram utilizados os arquivos de áudios com duração de até 15 segundos, chegando aproximadamente a 2700 arquivos de teste. Atualmente, estão sendo conduzidos testes com uma nova base de dados, que inclui cerca de 6 horas de gravações de áudios de um mesmo locutor.
- **Ideias para Melhorias Futuras e Métodos de Avaliação:** Há uma série de ideias em consideração para aprimorar o sistema, incluindo a possibilidade de introduzir métodos de avaliação mais avançados. Essas ideias visam principalmente aprimorar a precisão do reconhecimento de voz e a qualidade das respostas geradas pelo GPT. Uma das propostas é o desenvolvimento de uma interface web, que poderia tornar o sistema mais acessível e intuitivo para os usuários. Essas ideias, ainda em fase inicial de concepção, sugerem caminhos promis-

sores para futuras melhorias, expandindo a funcionalidade e a usabilidade do sistema.

REFERENCES

[1] J. Walnut, *VITS: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech*,GitHub Repository, https://github.com/jaywalnut310/vits,2020.

[2] ProgramadorArtificial, *VITS Portuguese*, GitHub Repository, https://github.com/ProgramadorArtificial/vits-portuguese, 2021.

[3] Como Treinar VITS para Transforma Texto em Áudio, YouTube Video, https://www.youtube.com/watch?v=m8UNUtA0Imk, 2021.

[4] OpenAI, *API Reference*, https://platform.openai.com/docs/api-reference, 2022.

[5] TTS-Portuguese Corpus, *TTS-Portuguese Corpus*,GitHub Repository, https://github.com/Edresson/TTS-Portuguese-Corpus.git,2023.