Sistema de Conversão de Voz com GPT

Christian Junji Litzinger State
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
christian.state@edu.ufes.br

Filipe Gomes Arante de Souza
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
filipe.ga.souza@edu.ufes.br

***Resumo*—Este documento apresenta um projeto que permite uma interação fluida entre um humano e um sistema de inteligência artificial (IA) através de comandos de voz. O sistema recebe uma entrada de voz, a converte em texto e a envia para a API do Gemini. A IA processa o pedido e gera uma resposta textual, que é posteriormente convertida em voz, permitindo uma comunicação contínua. A solução proposta integra o reconhecimento de voz, o processamento de linguagem natural e a conversão de texto em voz para criar uma experiência de conversação intuitiva.**

***Palavras-chave*—Reconhecimento de Voz, Processamento de Linguagem Natural, Inteligência Artificial**

I. Introdução

O desenvolvimento de sistemas de IA com interfaces baseadas em voz tem ganhado relevância em virtude de suas amplas aplicações, que vão desde assistentes virtuais até ferramentas de acessibilidade. A capacidade de interagir com um sistema de forma natural por meio da fala melhora a experiência do usuário e facilita a comunicação com tecnologias digitais. Este trabalho propõe um sistema de conversão de voz que integra reconhecimento de voz, processamento de linguagem natural por meio da API do Gemini e síntese de fala, proporcionando uma comunicação bidirecional. Além da implementação técnica, realizamos um experimento prático de conversação para avaliar a performance do sistema.

II. Trabalhos Correlatos

Diversos sistemas de interação por voz já estão consolidados no mercado, como o Siri da *Apple* [1], a Alexa da Amazon [2] e o Google Assistant [3]. Tais soluções geralmente dependem de motores proprietários de processamento de linguagem natural e estão integradas em ecossistemas específicos. Em contraste, o sistema proposto utiliza a API do Gemini da Google [4], proporcionando uma abordagem mais adaptável e personalizável, que pode ser aplicada em diferentes contextos e dispositivos.

Na literatura acadêmica, trabalhos relevantes no reconhecimento de voz e síntese de fala, como o Deep Speech [8] e o Tacotron [9], demonstram avanços significativos com o uso de redes neurais profundas e abordagens end-to-end. A integração dos conceitos desses estudos com a capacidade de diálogo do Gemini resulta em um sistema que combina robustez e flexibilidade, diferenciando-se dos métodos convencionais.

III. Metodologia

O sistema proposto segue um fluxo modular, conforme ilustrado na Figura 1, onde cada componente desempenha um papel específico na comunicação entre o usuário e a IA:

1) **Reconhecimento de Voz:**
   - Captura de áudio: O sistema utiliza o microfone do usuário para capturar o sinal de voz.
   - Pré-processamento: Aplicação de técnicas de normalização e redução de ruídos, utilizando algoritmos de filtragem para melhorar a qualidade do áudio.
   - Conversão de áudio para texto: O áudio é processado pela biblioteca Speech Recognition, que utiliza a API de reconhecimento de voz do Google para converter o sinal em texto. Nota-se que essa etapa depende de uma conexão com a internet.
2) **Processamento de Linguagem Natural:**
   - Envio do texto para a API do Gemini: O texto gerado é encaminhado para a API, que utiliza o modelo Gemini-2.0-Flash para compreender o contexto e prever a sequência de *tokens* que compõem a resposta.
   - Gerenciamento de tokens e latência: São implementadas estratégias para controle de requisições e otimização da comunicação, visando minimizar a latência e garantir a coerência dos diálogos.
3) **Síntese de Fala:**
   - Conversão de texto em áudio: A resposta textual é convertida em voz utilizando a biblioteca gTTS, que se apoia no serviço de síntese de fala do Google.
   - Entrega da resposta: O áudio sintetizado é reproduzido para o usuário, completando o ciclo de conversação.

A implementação foi realizada em Python [5], utilizando as bibliotecas Speech Recognition [6] para o reconhecimento de voz e gTTS [7] para a síntese de fala.

IV. Experimentos

Foi conduzido um experimento prático de conversação utilizando a API do Gemini, demonstrando as habilidades do software em reconhecer voz, converter o áudio em texto e gerar fala. A interação completa pode ser visualizada através do

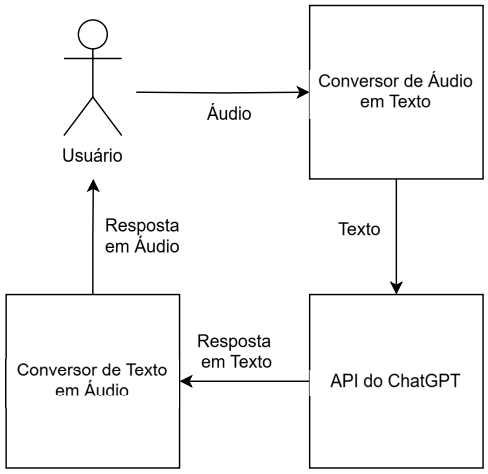

Fig. 1. Fluxo de comunicação do sistema: da captura de voz ao retorno da resposta sintetizada.

seguinte link no YouTube: https://www.youtube.com/watch?v=Wejg9pDkcE4.

V. Resultados

O experimento demonstrou que, embora o sistema apresente uma comunicação eficaz, ocorreram falhas pontuais, tais como:

- Dificuldade em reconhecer a primeira palavra da frase em algumas ocasiões.
- Incapacidade de captar corretamente a entonação em perguntas, o que pode afetar a compreensão do contexto.

Esses resultados evidenciam a viabilidade do sistema para aplicações práticas, apontando, contudo, desafios relacionados à robustez do reconhecimento de voz que devem ser abordados em trabalhos futuros.

Bibliografia

[1] Apple Inc., "Siri: Intelligent Personal Assistant," Cupertino, CA, USA, 2011. Disponível em: https://www.apple.com/siri. Acesso em: 18/02/2025.

[2] Amazon.com, Inc., "Alexa: Voice Service and AI Assistant," Seattle, WA, USA, 2014. Disponível em: https://developer.amazon.com/alexa. Acesso em: 18/02/2025.

[3] Google LLC, "Google Assistant: AI-Powered Virtual Assistant," Mountain View, CA, USA, 2016. Disponível em: https://assistant.google.com. Acesso em: 18/02/2025.

[4] Google LLC, "Gemini: A Next-Generation Foundation Model," Mountain View, CA, USA, 2023. Disponível em: https://gemini.google.com/. Acesso em: 13/03/2025.

[5] Python Software Foundation, "Python Language Reference," versão 3.x, 2023. Disponível em: https://www.python.org. Acesso em: 19/02/2025.

[6] A. Zhang, "SpeechRecognition: Library for performing speech recognition with support for several engines and APIs," 2023. Disponível em: https://pypi.org/project/SpeechRecognition/. Acesso em: 19/02/2025.

[7] P. G. e colaboradores, "gTTS: Google Text-to-Speech API wrapper," 2023. Disponível em: https://pypi.org/project/gTTS. Acesso em: 19/02/2025.

[8] A. Hannun et al., "Deep Speech: Scaling up end-to-end speech recognition," 2014. Disponível em: https://www.researchgate.net/publication/269722411_DeepSpeech_Scaling_up_end-to-end_speech_recognition. Acesso em: 24/02/2025.

[9] Y. Wang et al., "Tacotron: Towards End-to-End Speech Synthesis," 2017. Disponível em: https://www.researchgate.net/publication/315696313_Tacotron_A_Fully_End-to-End_Text-To-Speech_Synthesis_Model. Acesso em: 24/02/2025.