Transcrição e Correção Automática de Conversas em Diversos Idiomas

Matheus Henrique de Macedo Barreto
Luciene Cavalcanti Rodrigues (Orientadora)

e-mail:
matheus.barreto@fatec.sp.gov.br

Resumo: Este trabalho apresenta o desenvolvimento de um sistema de transcrição e correção automáticas de conversas em diversos idiomas, visando aprimorar a habilidade de conversação na respectiva língua. O sistema transcreve o áudio da conversa em texto e, em seguida, uma inteligência artificial corrige o texto produzido, fornecendo correções aos usuários. A metodologia envolve o uso de algoritmos de reconhecimento de fala e processamento de linguagem natural. Os resultados alcançados demonstram a viabilidade e eficácia do sistema na correção de erros gramaticais e na melhoria da fluência na conversação em outros idiomas.

Palavras-chave: Transcrição Automática, Correção por IA, Conversação em outros idiomas, Reconhecimento de Fala, Processamento de Linguagem Natural.

Abstract: *This work presents the development of a system for automatic transcription and correction of conversations in other languages, aiming to enhance conversational skills in the respective language. The system transcribes the conversation audio into text, and then artificial intelligence corrects the produced text, providing feedback to users. The methodology involves the use of speech recognition algorithms and natural language processing. The results demonstrate the system's feasibility and effectiveness in correcting grammatical errors and improving fluency in conversations in other languages.*

Keywords: *Automatic Transcription, AI Correction, Multilingual Conversation, Speech Recognition, Natural Language Processing.*

1 Introdução

Neste estudo, busca-se criar um sistema para ajudar no aprendizado de diversos idiomas, aproveitando tecnologias modernas. Nota-se a falta de ferramentas acessíveis e eficazes para melhorar a habilidade de conversação em outros idiomas. Nosso objetivo é desenvolver um sistema que transcreva automaticamente conversas e as corrija usando inteligência artificial. Com isso, facilitando o aprimoramento da fluência e precisão na conversação do idioma escolhido, preenchendo uma lacuna no mercado de recursos educacionais.

2 Justificativa

Este projeto é motivado pela necessidade de oferecer uma solução inovadora para melhorar as habilidades de conversação em outros idiomas. Diante da crescente importância desse idioma em um mundo globalizado, a falta de recursos acessíveis e eficazes para o aprendizado linguístico representa um obstáculo significativo. Ao aproveitar tecnologias modernas, como transcrição automática e correção por inteligência artificial, a aplicação busca preencher essa lacuna, fornecendo aos usuários um sistema que ofereça correções imediatas e personalizadas, visando aprimorar a fluência e precisão na conversação em inglês.

3 Objetivo(s)

- Desenvolver um sistema de transcrição automática de conversas, capaz de converter áudio em texto em 20 idiomas.
- Implementar um mecanismo de correção por inteligência artificial que analise o texto transcrito e forneça correções imediatas sobre erros gramaticais e de pronúncia.
- oferecer uma ferramenta acessível e eficaz para o aprendizado de idiomas de forma a aprimorar a fluência e a precisão na conversação dos usuários,
- Contribuir para a pesquisa e o desenvolvimento de tecnologias educacionais inovadoras que explorem o potencial da inteligência artificial no ensino e aprendizado de idiomas.

4 Fundamentação Teórica

A base teórica deste projeto se concentra no reconhecimento de fala e na aplicação da inteligência artificial no contexto do ensino de idiomas. O reconhecimento de fala engloba técnicas de processamento de sinais e linguística computacional para converter áudio em texto de forma precisa. Por outro lado, a inteligência artificial é fundamental para corrigir automaticamente erros gramaticais e de pronúncia, oferecendo correções personalizadas aos usuários. Combinando essas áreas, o sistema proposto visa melhorar a fluência e a precisão na conversação, proporcionando uma abordagem inovadora para o aprendizado de idiomas.

De acordo com a publicação de MULTILINGUAL MASTERY (2023), que buscou estudar e analisar aplicativos de idiomas, o Duolingo (uma das aplicações estudadas), que é uma plataforma de aprendizado de idiomas online que oferece uma abordagem interativa e gamificada para o ensino de inglês e outros idiomas, busca oferecer aos usuários o poder de praticar habilidades de conversação, vocabulário, gramática e compreensão auditiva por meio de lições curtas e jogos divertidos. Ele também oferece correções imediatas e personalizadas para ajudar os alunos a progredirem em seu aprendizado.

Ao analisar o aplicativo, percebe-se seu ponto forte na abordagem lúdica e interativa, mantendo os usuários engajados. Ao desenvolver a aplicação de correção de conversação, pode-

se considerar a inclusão de elementos semelhantes de gamificação para motivar a prática regular. No entanto, é importante superar a crítica de foco excessivo na gramática e vocabulário em detrimento da fluência e pronúncia autêntica.

Outra aplicação analisada foi o Babbel, um aplicativo de aprendizado de idiomas que se concentra em fornecer lições estruturadas e conteúdo prático para ajudar os usuários a aprender inglês e outros idiomas. Ele oferece uma variedade de exercícios de compreensão auditiva, leitura, escrita e conversação, adaptados ao nível de proficiência de cada aluno. Babbel também se destaca por sua abordagem baseada em situações do cotidiano, tornando o aprendizado mais relevante e aplicável.

Esta aplicação se destaca por sua estrutura organizada e foco em situações do cotidiano, o que pode inspirar a inclusão de exercícios práticos e contextualizados na aplicação de correção de conversação. No entanto, é fundamental evitar que a abordagem estruturada sacrifique a sensação de envolvimento e motivação dos usuários, buscando equilibrar a formalidade com a interatividade.

Por fim, o último aplicativo estudado, foi o Rosetta Stone, uma plataforma de aprendizado de idiomas que utiliza o método de imersão total para ensinar inglês e outros idiomas. Ele enfatiza o aprendizado por meio de imagens, associação visual e repetição para ajudar os alunos a desenvolverem habilidades de conversação, compreensão auditiva, leitura e escrita de forma natural e intuitiva. Rosetta Stone também oferece correções em tempo real e rastreamento de progresso para incentivar os alunos a continuarem aprendendo de forma consistente.

A plataforma tem ênfase na imersão total e associação visual, sugere a inclusão de elementos visuais e auditivos na aplicação de correção de conversação. Isso pode facilitar a associação entre palavras e frases em inglês com imagens e sons correspondentes. No entanto, é essencial evitar uma abordagem excessivamente visual que possa alienar usuários que prefiram aprender de outras maneiras, como por meio da prática direta de conversação.

5 Metodologia

Foi utilizada a pesquisa de campo na área de educação e tecnologia e a pesquisa aplicada ao desenvolvimento do sistema de transcrição e correção automáticas de conversas em outros idiomas será detalhado, incluindo a implementação de tecnologias específicas para cada etapa do processo.

Após estudos sobre o uso de inteligência artificial e as diversas ferramentas que podem ser utilizadas para processamento de linguagem natural optou-se pelo GEMINI, que, segundo Haas (2024) é um modelo fundacional desenvolvido pelo Google, com diferencial em sua construção multimodal, capaz de integrar informações de texto, imagens, áudio, vídeos e código para gerar conteúdo.

- Desenvolvimento da Interface do Usuário
    - A interface foi desenvolvida utilizando React[1] para criar uma interface interativa e responsiva
    - Utilizou-se da biblioteca Three.js com o GLTFJSX[2] para integrar e animar um avatar 3D criado no ReadyPlayerMe[3] e animado no Mixamo.
    - É exibido um microfone na interface para envio de áudio ao servidor.
    - O histórico de mensagens exibe a transcrição, correções e mensagens de áudio com sincronização labial do avatar.

- Desenvolvimento do Servidor
    - Implementado com JavaScript e Express[4], o servidor processa mensagens de áudio capturadas na interface do usuário.
    - Integração com GEMINI para transcrição, análise gramatical e sugestões de melhoria.
    - Utilização da biblioteca node-edge-tts[5] para gerar áudio das respostas e do Rhubarb Lip Sync [6]para sincronizar os movimentos labiais do avatar.
- Integração com o GEMINI para Correções e Dicas
    - O servidor utiliza a API do GEMINI para transcrição, análise gramatical e correção automática.
    - Fluxo de Funcionamento:
        1. O áudio capturado pelo microfone é enviado ao servidor.
        2. O GEMINI realiza a transcrição, correção gramatical e fornece correções e sugestões de melhoria.
        3. O sistema exibe as correções no front-end, destacando erros e melhorias, além de retornar uma resposta em áudio para proporcionar uma experiência de conversação natural.

5.1 Testes e Avaliação

Para validar a eficácia do sistema, foi realizado testes com a professora de inglês Maura, utilizando a aplicação nas variantes da língua inglesa disponíveis. Os testes demonstraram que o sistema foi eficiente ao reconhecer diferentes sotaques de inglês australiano, britânico, canadense e americano.

---

1 Disponível em: https://react.dev/
2 Disponível em: https://github.com/pmndrs/gltfjsx
3 Disponível em: https://docs.readyplayer.me/ready-player-me/api-reference/rest-api/avatars/get-3d-avatars
4 Disponível em: https://github.com/expressjs/express
5 Disponível em: https://github.com/SchneeHertz/node-edge-tts
6 Disponível em: https://github.com/DanielSWolf/rhubarb-lip-sync

Além disso, o sistema adaptou corretamente o sotaque nas respostas de acordo com a variante selecionada, proporcionando uma experiência mais natural e personalizada para os usuários. Foi observado que o sistema não apresentou insights ou correções errôneas, reforçando sua confiabilidade.

No entanto, os testes também evidenciaram algumas limitações do sistema. Em algumas ocasiões, o microfone não era reconhecido, dificultando o envio de áudio pelo usuário. Além disso, a resposta da API GEMINI, em certos momentos, demorava mais do que o esperado, resultando em atrasos significativos e, em alguns casos, ocasionando erros no sistema.

Os problemas encontrados destacam áreas importantes de melhoria, especialmente no que se refere à robustez da integração com o microfone e à otimização do tempo de resposta da API. Apesar dessas limitações, os resultados gerais validam o potencial do sistema, demonstrando sua eficácia na transcrição, correção gramatical e adaptação a diferentes sotaques.

6 Desenvolvimento

O desenvolvimento do sistema foi dividido em duas partes principais: front-end (interface do usuário) e back-end (servidor e processamento de áudio). O tópico a seguir, detalha as etapas realizadas para atingir os objetivos propostos.

- Desenvolvimento da interface do usuário com React
  - A interface foi desenvolvida utilizando React, permitindo maior flexibilidade e integração com recursos modernos.
  - Um avatar 3D, criado no ReadyPlayerMe e animado no Mixamo, foi integrado a interface usando a biblioteca Three.js com GLTFJSX.
  - É exibido um microfone na interface, permitindo que o usuário envie áudios ao servidor.
  - O histórico de mensagens exibe a transcrição, a correção e as respostas de áudio, sincronizando os lábios do avatar para oferecer uma experiência mais imersiva.
  - Os usuários podem selecionar entre 20 idiomas suportados(desses 20, todas as variações do inglês disponíveis foram testadas na interação inicial por uma professora de inglês) e diferentes perfis de personalidade para o avatar (professora, amigo ou familiar), configurando a experiência de acordo com suas preferências e nível de fluência.
- Desenvolvimento do servidor com JavaScript
  - O servidor foi implementado com JavaScript usando a biblioteca Express.
  - O áudio enviado é processado e enviado para a API do GEMINI, que retorna transcrição, correção gramatical e sugestões sobre a conversação.
  - Utilizou-se a biblioteca node-edge-tts para converter a transcrição em áudio e o Rhubarb Lip Sync para sincronizar os movimentos labiais do avatar com a fala.

- As respostas processadas são enviadas de volta, onde são exibidas na interface do usuário e reproduzidas pelo avatar para proporcionar uma interação fluida e natural.

O código-fonte está disponível em: https://github.com/tauk7/teacher

7 Resultados e Discussões

Descrever o funcionamento do sistema e colocar pelo menos 3 ou 4 telas

A interface inicial do sistema solicita que o usuário escola o idioma a ser trabalhado (Figura 1),

*Figura SEQ Figura * ARABIC 1: Interface onde você deve escolher 1 entre 20 idiomas disponíveis e em seguida, clicar no botão próximo.*

Interface onde se é possível escolher o nível de fluência (Figura 2),

*Figura SEQ Figura * ARABIC 2: Interface onde se deve escolher uma entre as 3 opções do nível de fluência na língua selecionada.*

Interface onde se é possível escolher a personalidade do avatar com quem você irá conversar (Figura 3),

*Figura SEQ Figura * ARABIC 3: Interface onde você deve escolher uma entre 3 opções da personalidade do avatar, sendo a Maura focada em correções, Sasha em perguntas triviais e Rita em conversas citando sua família e coisas do seu dia a dia.*

Interfaces onde poderá conversar com um dos seguintes avatares (Figura 4, 5 e 6),

*Figura SEQ Figura * ARABIC 4: Tela onde se é possível conversar com a personagem Maura, onde você pode enviar a mensagem de voz pressionando o ícone de microfone e ver seu histórico de mensagens da sessão. Ao lado direito na parte inferior, é exibido sugestões a respeito de sua conversação.*

*Figura SEQ Figura * ARABIC 5: Tela onde se é possível conversar com a personagem Sasha, com as mesmas funcionalidades da Figura 4.*

Os

*Figura SEQ Figura * ARABIC 6: Tela onde se é possível conversar com a personagem Rita, com as mesmas funcionalidades da Figura 4.*

resultados obtidos com os testes demonstraram que o sistema transcreve e corrige conversas em inglês com alta precisão, reconhecendo diferentes sotaques (australiano, britânico, canadense e americano) e ajustando as respostas conforme a variante selecionada. A integração de avatares 3D e perfis de personalidade personalizáveis foi elogiada, proporcionando uma experiência imersiva e motivadora.

Entretanto, foram observados problemas como o não reconhecimento ocasional do microfone e atrasos na resposta da API GEMINI, que em alguns casos causaram erros no sistema. Esses pontos indicam áreas para melhorias, especialmente na integração com o microfone e na otimização do tempo de resposta.

Apesar disso, os resultados indicam o potencial do sistema para auxiliar no aprendizado de idiomas. A migração da interface de Angular para React foi essencial para a implementação de funcionalidades avançadas, como animações sincronizadas no avatar e suporte a múltiplos idiomas.

8 Conclusões

Este trabalho apresentou o desenvolvimento de um sistema inovador para transcrição e correção automáticas de conversas, com foco na melhoria das habilidades de conversação em inglês e outros idiomas. A metodologia combinou tecnologias modernas, como avatares 3D, inteligência artificial e processamento de áudio, para criar uma experiência interativa e personalizada. Os resultados mostraram que os objetivos foram alcançados: o sistema não apenas demonstrou alta precisão na transcrição e correção gramatical, mas também ajudou os usuários a aprimorar sua fluência. Apesar do sucesso, limitações foram identificadas, como a necessidade de uma maior variedade de personalidades e perfis, além de melhorias na latência do sistema. Futuras pesquisas podem explorar a expansão dessas funcionalidades e a adaptação para outros contextos educacionais.

Agradecimentos

Agradeço à minha orientadora, professora Luciene Cavalcanti Rodrigues, por sua orientação e apoio durante o desenvolvimento deste trabalho. Expresso também minha gratidão ao professor Henrique Dezani, que me ajudou a contornar o desafio de transcrição de áudio, ensinando-me sobre as capacidades da API GEMINI. Por fim, agradeço aos meus colegas e familiares pelo suporte contínuo ao longo deste projeto.

Referências

ANGULAR. **What is Angular?**. 2024. Disponível em: https://angular.dev/overview. Acesso em: 23/06/2024.

EXPRESS. **Fast, unopinionated, minimalist web framework for Node.js**. 2024. Disponível em: https://github.com/expressjs/express. Acesso em: 23/06/2024.

GEMINI API. **Começar a usar a API Gemini**. 2024. Disponível em: https://ai.google.dev/gemini-api/docs. Acesso em: 03/10/2024.

GLTFJSX. **Usage**. 2024. Disponível em: https://github.com/pmndrs/gltfjsx. Acesso em: 03/10/2024.

MIXAMO. **Get animated.** 2024. Disponível em: https://www.mixamo.com. Acesso em: 03/10/2024.

MULTILINGUAL MASTERY. **Babbel vs. Rosetta Stone vs. Duolingo: Qual é o melhor para você?**. 2023. Disponível em: https://multilingualmastery.com/babbel-vs-rosetta-stone-vs-duolingo/. Acesso em: 21/04/2024.

NODE EDGE TTS. **node-edge-tts**. 2024. Disponível em: https://github.com/SchneeHertz/node-edge-tts. Acesso em: 03/10/2024.

OPENAI. **Introducing APIs for GPT-3.5 Turbo and Whisper**. 2024. Disponível em: https://openai.com/index/introducing-chatgpt-and-whisper-apis/. Acesso em: 23/06/2024.

REACT. **Create user interfaces from components**. 2024. Disponível em: https://react.dev/. Acesso em: 03/10/2024.

READYPLAYERME. **GET - 3D avatar**. 2024. Disponível em: https://docs.readyplayer.me/ready-player-me/api-reference/rest-api/avatars/get-3d-avatars. Acesso em: 03/10/2024.

RHUBARD LIP SYNC. **Rhubard Lip Sync**. 2024. Disponível em: https://github.com/DanielSWolf/rhubarb-lip-sync. Acesso em: 03/10/2024.

SANTANDER OPEN ACADEMY. **A importância da língua inglesa**. 2023. Disponível em: https://www.santanderopenacademy.com/pt_br/blog/a-importancia-da-lingua-inglesa.html. Acesso em: 21/04/2024.

THE PYTHON CODE. **How to Convert Speech to Text in Python**. 2024. Disponível em: https://thepythoncode.com/article/using-speech-recognition-to-convert-speech-to-text-python. Acesso em: 23/06/2024.

YOUTUBE. **How to Build a 3D Chatbot with ChatGPT & ElevenLabs**. 2024. Disponível em: https://youtu.be/EzzcEL_1o9o?si=b76TxyFo-Gu4VKU-. Acesso em: 03/10/2024.