UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ
UNIOESTE - CAMPUS DE FOZ DO IGUAÇU
CENTRO DE ENGENHARIAS E CIÊNCIAS EXATAS
CURSO DE CIÊNCIA DA COMPUTAÇÃO

TCC - TRABALHO DE CONCLUSÃO DE CURSO

Proposta de Trabalho de Conclusão de Curso

Interface de Voz da Operação (IVO)

Vinicius de Oliveira Jimenez
Orientador(a): William Francisco da Silva
Co-orientador(es): Filipe Ventura Muggiati e Hilton Carlos Miranda Lessa

Foz do Iguaçu, 22 de maio de 2024

1 Identificação

1.1 Área e Linha de Pesquisa

Grande Área: Ciências Exatas e da Terra
Código: 1.00.00.00-3

Linha de Pesquisa: Ciência da Computação
Código: 1.03.00.00-7

Especialidade: Metodologia e Técnicas da Computação
Código: 1.03.03.00-6

1.2 Palavras-chave

1. Transcrição de fala;
2. Síntese de fala;
3. *Speech-To-Text (STT)*;
4. *Text-To-Speech (TTS)*;
5. Indústria 4.0;
6. Aplicação web;
7. API.

2 Introdução e Justificativa

A Indústria 4.0 surge a partir de uma nova fase caracterizada pela transformação digital ou informatização, e conectividade. Nesse sentido, a indústria se consolida com umas das principais protagonistas nos avanços em tecnologias, como Internet das Coisas (*Internet of Things* — IoT), computação em nuvem e Inteligência Artificial (IA) (BASCO, A. I. et al., 2018). Ainda, de acordo com Basco (2018), a IA é um dos 10 (dez) pilares fundamentais da Indústria 4.0, fundamentada na criação de algoritmos capazes de possibilitar que os computadores processem informações com uma velocidade excepcional.

No contexto industrial, a aplicação de algoritmos de IA possibilita o desenvolvimento de Redes Neurais — *Neural Networks* (NNs), voltadas à automação de

processos operacionais complexos e morosos, análise de grande volume de dados, dando apoio à tomada de decisões estratégicas, e resultando em uma gestão mais inteligente e otimizada das estações de trabalho (BASCO, A. I. et al., 2018). Com essa finalidade, em 2019, o Departamento de Operação do Sistema (OPS.DT) da Itaipu Binacional (IB), criou um grupo de trabalho para analisar e indicar ações para modernizar os processos da divisão com base nos fundamentos da Industria 4.0.

Este grupo de trabalho realizou uma análise de diversas tecnologias, com destaque para os assistentes virtuais acionados por comandos de voz. Tais sistemas são constituídos por um conjunto de módulos interdependentes, que operam de maneira integrada para viabilizar seu funcionamento. Dentro esses módulos, destaca-se aquele responsável pela conversão de fala em texto — *Speech-To-Text* (STT), e o módulo responsável pela síntese de texto em fala — *Text-To-Speech* (TTS). Diante disso, foi solicitado ao Centro de Gestão Energética (GE.DT) do Itaipu Parquetec o desenvolvimento de um sistema capaz de atender a essas demandas.

A tecnologia de transcrição de fala em texto, também conhecida como reconhecimento automático de fala — *Automatic Speech Recognition* (ASR), consiste na identificação de palavras pronunciadas e suas subsequente conversão em texto escrito. Esse campo tem apresentado avanços notáveis em termos de precisão e eficiência, impulsionados principalmente pelo progresso da IA. Um exemplo expressivo desse avanço é a introdução da arquitetura *transformer*, desenvolvida pela Google (VASWANI et al., 2017), representando um marco significativo ao permitir processamento de sentenças completas, em vez de palavras isoladas, por meio de mecanismos de atenção capazes de direcionar o foco às partes mais relevantes da entrada (áudio) (ANKIT, 2024).

Para Tokuda et al. (2013), a síntese de texto para fala é uma técnica utilizada para gerar fala artificial que seja inteligível e com características naturais a partir de um texto de entrada. Analogamente aos modelos STT, os modelos TTS também evoluíram significativamente ao longo do tempo, resultando em sistemas capazes de gerar vozes quase indistinguíveis da fala humana. De acordo com Barakat (2024), mesmo com técnicas avançadas, como modelos ocultos de Markov (HMMs) e modelos guassianos (GMMs), a fala gerada por esses métodos ainda era visivelmente artificial. O cenário alterou a partir do surgimento de aprendizado profundo — *Deep Learning* (DL), dando início a aplicação de NNs na síntese de fala (BARAKAT, H.; TURK, O.; DEMIROGLU, C., 2024).

Diante desse contexto, percebe-se a importância da adoção de um sistema para a realização de conversões *Speech-to-Text* (STT) e *Text-to-Speech* (TTS) no ambiente de supervisão de uma usina, podendo contribuir significativamente para o aumento da agilidade e eficiência operacional nas salas de controle. Em determinadas situações, o esforço cognitivo necessário para solicitar e receber informações

por meio da fala é inferior ao exigido por textos escritos, o que possibilita a realização simultânea de múltiplas atividades por um mesmo indivíduo de forma mais intuitiva e natural.

Adicionalmente, ao sistema proposto, especificamente, o módulo de STT, pode ser utilizado para realizar transcrições de ligações telefônicas entre o despachante e o operador nos centros de controle. Essas transcrições podem contribuir para facilitar a passagem de turno das equipes em tempo real, estabelecer conexões entre as ligações, eventos e registros para análises pós-operacionais, além de fornecer uma base de dados consultável para que as equipes revisem conversas anteriores durante o turno.

3 Objetivos

3.1 Objetivo Geral

Este projeto tem como objetivo geral o desenvolvimento do sistema IVO, voltado à conversão de fala em texto (*Speech-to-Text* – STT) e de texto em fala (*Text-to-Speech* – TTS), bem como ao aprimoramento contínuo dos modelos de reconhecimento de fala. O sistema é composto por quatro aplicações integradas: uma aplicação web piloto, uma aplicação web destinada ao treinamento, uma API[1] principal e uma API dedicada à IA.

3.2 Objetivos Específicos

Dentre os principais objetivos específicos destacam-se:

- Realizar levantamento e análise crítica de bibliotecas voltados à implementação de modelos STT e TTS;
- Projetar e implementar uma API dedicada à integração e execução dos modelos STT e TTS;
- Estruturar uma API para viabilizar o consumo unificado dos serviços desenvolvidos neste projeto e que internos da IB;
- Desenvolver uma aplicação *web* que permita ao usuário final realizar transcrições de áudio e sínteses de fala consumindo os *endpoints*[2] da API principal;
- Criar uma aplicação voltada à gestão e ao treinamento de modelos STT.

[1] Interface de programação de aplicações – *Application Programming Interface* (API)

[2] Um endpoint de API é um endereço onde a API recebe pedidos para acessar informações ou serviços disponíveis no servidor. (NOSOWITZ; GOODWIN, 2024)

4 Plano de Trabalho e Cronograma de Execução

Nessa seção pretende-se a apresentar o plano de trabalho e cronograma de execução, contemplando as atividades a serem desenvolvidas ao longo do trabalho.

1. Atividade 1: conduzir uma Revisão Sistemática da Literatura (RSL) com base em critérios bem definidos, a fim de identificar, avaliar e sintetizar pesquisas relevantes sobre tecnologias de transcrição de fala e síntese de fala, bem como arquiteturas utilizadas para a API, e aplicações *web*;

2. Atividade 2: elaborar o documentação do projeto e os protótipos de telas das aplicações piloto e de treinamento. Os artefatos resultantes desta atividade devem ser aprovados pelos responsáveis pelo projeto pelo lado da IB;

3. Atividade 3: desenvolver a API responsável pela integração com os modelos STT e TTS, além de realizar o processo de ajuste fino – *fine-tuning*[3], para gerar novos modelos STT treinados;

4. Atividade 4: implementar a API principal que atue como orquestradora do sistema, responsável pelo controle de fluxo entre os serviços, acesso ao banco de dados, autenticação, autorização, roteamento de requisições e demais funcionalidades;

5. Atividade 5: desenvolver a aplicação piloto que permita ao usuário realizar transcrições STT e sínteses TTS, utilizando os serviços da API principal;

6. Atividade 6: implementar a aplicação de treinamento dos modelos STT, com funcionalidades que incluam a listagem e visualização dos modelos disponíveis e dos áudios transcritos;

7. Atividade 7: executar testes que validem o funcionamento conjunto de todos os componentes do sistemas — APIs, modelos, aplicações *web* e banco de dados — assegurando a comunicação correta entre os módulos da solução;

8. Atividade 8: redigir o texto da monografia.

Na Tabela 1 é apresentado o cronograma de execução das atividades apresentadas anteriormente nesta mesma seção.

[3]O ajuste fino ou *fine-tuning* em ML é o processo de adaptar um modelo previamente treinado para tarefas ou casos de uso específicos (BERGMANN, 2024).

Atividades	Período								
	Jun	Jul	Ago	Set	Out	Nov	Dez	Jan	Fev
1 - Atividade 1	•	•	•	•					
2 - Atividade 2	•	•							
3 - Atividade 3		•	•	•	•	•			
4 - Atividade 4		•	•	•	•	•			
5 - Atividade 5		•	•	•	•				
6 - Atividade 6			•	•	•	•			
7 - Atividade 7					•	•	•		
8 - Atividade 8	•	•	•	•	•	•	•	•	•

Tabela 1: Cronograma de atividades

5 Material e Método

5.1 Material

Serão utilizados os seguintes materiais e recursos para a implementação do sistema IVO:

- Notebook: será utilizado um notebook do modelo Lenovo Legion Slim 5i com as seguintes configurações:
  - Processador: Intel® Core™ i7-13700H de 13ª geração;
  - Sistema Operacional: Microsoft Windows 11 Pro;
  - Placa de vídeo: GPU para laptop NVIDIA® GeForce RTX™ 4050 6GB GDDR6;
  - Memória: 16 GB DDR5-5.200MHz (SODIMM)(2 x 8 GB);
  - Armazenamento: 512 GB SSD M.2 2280 PCIe Gen4 TLC.
- Subsistema do Windows Para Linux (WSL[4]): será utilizado um subsistema dentro do sistema operacional Windows devido a possibilidade de utilizar comandos Docker[5] via linha de comando. O subsistema possui as seguintes configurações:
  - Distribuição: Ubuntu;
  - Versão: 22.04.3 LTS.
- Máquinas virtuais: serão utilizados duas máquinas virtuais — *Virtual Machine* (VM) para realizar o *deploy*[6] das aplicações necessárias. As VMs contam com as seguintes configurações:

[4] *Windows Subsystem Linux* (WSL).

[5] Docker é uma plataforma *open source* que possibilita aos desenvolvedores criar, executar, distribuir, atualizar e administrar contêineres de forma eficiente (SUSNJARA; SMALLEY, 2024).

[6] O *deploy* de uma aplicação consiste em torná-la acessível para que possa ser utilizada por usuários finais ou integrada a outros sistemas (IBM, 2025)

  - VM IVO API
    - Sistema Operacional: Ubuntu 22.04 LTS;
    - Processadores: 8;
    - Memória: 17 GB;
    - Armazenamento: 80 GB.
  - VM IVO IA
    - Sistema Operacional: Ubuntu 22.04 LTS;
    - Processadores: 20;
    - Memória: 46 GB;
    - Armazenamento: 80 GB.

- Gerenciamento de contêineres: será utilizado a ferramenta Podman que se destaca por ser *open-source* não depender de um *daemon*, tornando-o um alternativa segura e acessível (HAT, 2024). Além disso, é a tecnologia adotada pela IB;

5.2 Método

Os métodos adotados para a realização das atividades serão organizados de acordo com as fases do plano de atividades (Tabela 1). Algumas tarefas, por serem semelhantes, adotaram a mesma metodologia de execução.

- **Revisão Sistemática da Literatura (RSL)**

  Realizar busca por referências bibliográficas utilizando a abordagem RSL. De acordo com Galvão (2019), a ênfase desta técnica de revisão recai sobre a reprodutibilidade por parte de outros pesquisadores, detalhando de maneira explícita as bases de dados bibliográficas utilizadas, as estratégias de busca adotadas, em cada uma delas, o processo de seleção de artigos e textos científicos, bem como os critérios estabelecidos para inclusão e exclusão.

- **Documentação e prototipagem**

  Escrever a documentação do projeto empregando os padrões do centro GE.DT. Entende-se como documentação: documento de requisitos, visão geral do sistema, diagramas de classe e de casos de uso e manual do usuário. No que tange à prototipagem das telas das aplicações *web*, a ferramenta utilizada para tal será o Figma[7]. Deve-se seguir o guia de estilos do centro GE.DT.

[7]Site oficial do Figma: <https://www.figma.com/>

- **Desenvolvimento das APIs**

  Inicialmente, criar os repositórios no GitLab[8] interno do GE.DT. Após essa etapa inicial, iniciar a implementação das arquiteturas escolhidas para cada API, definindo suas devidas bibliotecas, o Sistema Gerenciador de Banco de Dados (SGBD) a ser utilizado. Para o caso da API de STT e TTS, definir modelos *base* para testes iniciais. Ja à API principal, estudar métodos de comunicação entre as APIs. A autenticação e autorização será realizada utilizando o servidor de autenticação Keycloak[9], utilizando a técnica de *Single Sign-On* (SSO)[10].

- **Desenvolvimento das aplicações web**

  A metodologia de desenvolvimento das aplicações web abrange, de forma similar a do desenvolvimento de APIs, a criação dos repositórios no GitLab, bem como a definição das bibliotecas a serem utilizadas. Ademais, é imprescindível o estudo e a aplicação de técnicas de responsividade e acessibilidade, a fim de garantir que as aplicações sejam utilizáveis por todos os usuários e possam ser acessadas de maneira intuitiva e eficiente em diferentes dispositivos, incluindo *smartphones* e *tablets*, além de computadores e *notebooks*.

6 Critérios de Avaliação

Os resultados obtidos serão avaliados com base no desempenho das aplicações que compõem o sistema IVO, bem como no funcionamento integrado do sistema como um todo. Espera-se, entre outras funcionalidades, que o sistema seja capaz de transcrever fala em texto, sintetizar texto em fala, armazenar os áudios recebidos para posterior transcrição, permitir a avaliação das transcrições pelos usuários finais e, sobretudo, viabilizar o treinamento contínuo dos modelos de reconhecimento automático de fala STT. Ademais, o *feedback* dos usuários finais da IB será um elemento relevante e primordial para aferir se a proposta atendeu aos objetivos estabelecidos.

[8] Site oficial do GitLab: <https://about.gitlab.com/>

[9] Site oficial do Keycloak: <https://www.keycloak.org/>

[10] Single sign-on (SSO) é um método de login que permite ao usuário acessar vários sistemas usando um somente as credenciais, sem precisar se autenticar novamente em cada aplicação (SCAPICCHIO; FORREST, 2024).

7 Referências

ANKIT, U. *Transformer Neural Networks: A Step-by-Step Breakdown*. 2024. Disponível em: <https://builtin.com/artificial-intelligence/transformer-neural-network>. Citado na página 3.

BARAKAT, H.; TURK, O.; DEMIROGLU, C. Deep Learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. 2024. Citado na página 3.

BASCO, A. I. et al. *Industria 4.0: fabricando el futuro*. Buenos Aires: Inter-American Development Bank, 2018. Citado 2 vezes nas páginas 2 e 3.

BERGMANN, D. *What is fine-tuning?* 2024. Disponível em: <https://www.ibm.com/think/topics/fine-tuning>. Acessado em: 21 maio 2025. Citado na página 5.

GALVãO, M. C. B.; RICARTE, I. L. M. Revisão sistemática da literatura: Conceituação, produção e publicação. *Logeion: Filosofia da Informação*, Rio de Janeiro, RJ, v. 6, n. 1, p. 57–73, 2019. Citado na página 7.

HAT, R. *What is Podman?* 2024. Disponível em: <https://www.redhat.com/en/topics/containers/what-is-podman>. Acessado em: 18 maio 2025. Citado na página 7.

IBM. *Deploying software*. 2025. Disponível em: <https://www.ibm.com/docs/en/zos/3.1.0?topic=task-deploying-software>. Acessado em: 18 maio 2025. Citado na página 6.

NOSOWITZ, D.; GOODWIN, M. *What is an API endpoint?* 2024. Disponível em: <https://www.ibm.com/think/topics/api-endpoint>. Acessado em: 21 maio 2025. Citado na página 4.

SCAPICCHIO, M.; FORREST, A. *What is single sign-on (SSO)?* 2024. Disponível em: <https://www.ibm.com/think/topics/single-sign-on>. Acessado em: 21 maio 2025. Citado na página 8.

SUSNJARA, S.; SMALLEY, I. *What is Docker?* 2024. Disponível em: <https://www.ibm.com/think/topics/docker>. Acessado em: 21 maio 2025. Citado na página 6.

TOKUDA, K. et al. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, v. 101, n. 5, p. 1234–1252, 2013. Citado na página 3.

VASWANI, A. et al. Attention is all you need. In: NIPS. *Advances in Neural Information Processing Systems 30 (NIPS 2017*. Long Beach, Estados Unidos, 2017. Citado na página 3.

8 Síntese Bibliográfica

ANKIT, U. *Transformer Neural Networks: A Step-by-Step Breakdown*. 2024. Disponível em: <https://builtin.com/artificial-intelligence/transformer-neural-network>. Nenhuma citação no texto.

BARAKAT, H.; TURK, O.; DEMIROGLU, C. Deep Learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. 2024. Citado na página 3.

BASCO, A. I. et al. *Industria 4.0: fabricando el futuro*. Buenos Aires: Inter-American Development Bank, 2018. Citado 2 vezes nas páginas 2 e 3.

GALVÃO, M. C. B.; RICARTE, I. L. M. Revisão sistemática da literatura: Conceituação, produção e publicação. *Logeion: Filosofia da Informação*, Rio de Janeiro, RJ, v. 6, n. 1, p. 57–73, 2019. Citado na página 7.