Projeto de IA Geradora de Áudio

O objetivo deste projeto é desenvolver uma Inteligência Artificial (IA) capaz de gerar áudios de voz humana de alta qualidade com base em textos ou comandos específicos(como outro áudio). Esta IA será projetada para síntese da fala. O projeto explorará técnicas avançadas de processamento de linguagem natural (NLP) e aprendizado profundo para criar um sistema robusto e eficiente.

Objetivos

- Desenvolver um modelo de IA que possa gerar voz humana a partir de textos e outros comandos.
- Avaliar a qualidade e a naturalidade do áudio gerado.
- Testar e validar o modelo em diferentes cenários de aplicação e diferentes entradas, podendo ser possível espeificar até intonação da voz.

Metodologia

1. Preparação dos Dados

- **Coleta de Dados**: Reunião de um grande conjunto de dados de voz emparelhado com transcrições textuais.
- **Pré-processamento**: Validar transcrições e qualidade dos áudios coletados.

2. Arquitetura do Modelo

- **Tacotron-2**: Tacotron-2 é o modelo estado-da-arte para síntese da fala. Trata-se da segunda versão da Tacotron, um modelo desenvolvido pelo Google e que possui uma arquitetura seq2seq, o qual recebe como entrada o texto e produz como saída um espectrograma em escala mel.
    - **Encoder**: Recebe tokens de input.
    - **Attention Mechanism**: Alinha os tokens de input com os espectrogramas Mel.

- **Decoder**: Gera espectrogramas de Mel, que são convertidos para áudio.

3. Treinamento

- **Configuração do Treinamento**: Definição de hiperparâmetros, como taxa de aprendizado, número de épocas e tamanho do batch.
- **Algoritmo de Otimização**: Uso de otimizadores como Adam para ajustar os pesos da rede.
- **Loss Function**: Utilização de funções de perda adequadas para medir a diferença entre os espectrogramas preditos e os espectrogramas de referência.

4. Pós-processamento

- **Conversão de Espectrogramas**: Conversão dos espectrogramas Mel gerados, para o áudio final.

5. Avaliação e Ajustes

- **Avaliação da Qualidade**: A avaliação consiste em escutar os áudios gerados e comparar com uma fala natural humana, sendo o modelo ótimo, o que se aproximar mais a isso.
- **Refinamentos**: Em suma, aumento no conjunto de dados, tendo, segundo análises feitas, uma melhora quase que linear na relação qualidade do áudio x quantia de dados utilizadas.

Dados

O conjunto de dados consiste em :

1. **Áudios de pessoas falando**: os áudios foram retirados em sua maioria de vídeos do youtube e de sites como o Common Voice, os áudios utilizados foram escolhidos baseando se em sua uniformidade, buscamo áudios mais limpos e com menos ruídos.
2. **Transcrição dos áudios usados**: A maior parte das transcrições foi feita a partir da Whisper Transcriber, que é uma IA que converte áudios para textos, a ferramenta comete alguns erros então foi necessário tomar alguns cuidados para que o texto transcrito fosse fiel ao áudio.

3. **Espectrogramas Mel**: Esses espectrogramas eram o que a IA geradora de voz de fato utilizava para gerar o áudio em si.

Resultados

A aplicação do modelo Tacotron-2 para a síntese de voz humana mostrou resultados satisfatórios. O modelo foi treinado usando cerca 30 min de áudios o que foi uma quantidade suficiente para um resultado razoável contendo apenas alguns erros de fala ,intonação e tonalidade das silábas.