UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

Processamento de Áudio e Voz

202005471 - Állan Christoffer Pereira Silva

202005475 - Gabriel da Mata Marques

202005484 - Heinz Felipe Cavalcante Rahmig

202005494 - Luís Augusto do Prado Assunção

Relatório sobre a Geração de Conteúdo Acessível para Deficientes Visuais Utilizando TTS (Text-To-Speech) para Audiodescrição

1. Introdução

A acessibilidade é um pilar fundamental na construção de uma sociedade inclusiva. Para deficientes visuais, a falta de audiodescrição em vídeos representa uma barreira significativa para o acesso total ao conteúdo multimídia. A audiodescrição é um recurso que traduz imagens em palavras, permitindo que pessoas com deficiência visual compreendam melhor o conteúdo visual. No entanto, a criação manual de audiodescrição pode ser demorada e cara. Neste contexto, a síntese de voz a partir de textos (TTS) surge como uma solução promissora para automatizar o processo de audiodescrição. Este trabalho aborda a criação de um algoritmo de TTS focado na audiodescrição, visando uma voz natural, ritmo adequado e conteúdo inteligível.

2. Dataset Utilizado

O dataset desempenha um papel vital no treinamento e avaliação de qualquer modelo de aprendizado de máquina. Neste projeto, o dataset utilizado foi extraído do Kaggle e consiste em scripts de texto do Jornal Nacional, bem como gravações de um locutor em ambiente controlado. Este dataset específico foi escolhido devido à sua riqueza em variações de entonação e conteúdo, representando um desafio real para o treinamento de um modelo TTS. As gravações totalizam 20 horas, com amostras variando de 5 a 31 segundos, média de 13 segundos, e uma taxa de amostragem de 44100 Hz.

3. Metodologia

A metodologia adotada neste projeto é um processo estruturado, dividido em três tarefas principais, cada uma focada em uma etapa crítica da geração de audiodescrição através da síntese de fala. A seguir, são detalhadas as tarefas e o pipeline adotado:

3.1 Task 1: Geração de Text-To-Speech a partir dos Roteiros Já Anotados

Esta tarefa envolve a conversão de roteiros de texto em fala sintetizada. Utilizando modelos de TTS, como os descritos nos experimentos, o texto é transformado em áudio, mantendo a naturalidade e a inteligibilidade da voz. A anotação prévia dos roteiros facilita este processo, permitindo uma maior consistência na geração de voz.

3.2 Task 2: Conversão da Voz para um Locutor Específico

A personalização da voz é um aspecto vital para tornar a audiodescrição mais envolvente e menos robótica. Esta tarefa envolve o ajuste da voz sintetizada para corresponder às características de um locutor específico. Pode ser feito através de técnicas de transferência de estilo de voz ou treinamento com amostras de voz do locutor desejado.

3.3 Task 3: Inserção Automatizada dos Trechos de Audiodescrição no Vídeo

Esta etapa finaliza o processo, inserindo automaticamente os trechos de audiodescrição no vídeo original. A inserção precisa ser feita de forma a não interferir na trilha sonora original, mantendo a coesão do conteúdo. A automação deste processo garante eficiência e precisão na integração da audiodescrição com o vídeo.

3.4 Pipeline

O pipeline adotado para realizar essas tarefas é composto por várias etapas, cada uma contribuindo para o produto final:

- Leitura dos Arquivos de Metadados: Importa as informações necessárias, como roteiros e anotações.
- Síntese de Fala a partir de Texto: Utiliza modelos de TTS para converter texto em áudio.

- Conversão de MP4 para WAV: Converte os arquivos de vídeo em áudio, permitindo a manipulação e a mixagem do som.
- Carregamento do Áudio Base: Importa o áudio original para ser mesclado com a audiodescrição.
- Equalização e Mixagem: Ajusta os níveis de som e mescla o áudio sintetizado com o áudio original, mantendo a qualidade e o equilíbrio do som.

A combinação dessas etapas forma um processo coeso que transforma o texto em uma audiodescrição integrada, contribuindo para a acessibilidade e a inclusão de deficientes visuais.

4. Pré-processamento

O pré-processamento é um estágio crucial para garantir a qualidade e a eficácia do treinamento de modelos de TTS. Neste trabalho, várias técnicas foram aplicadas:

- Remoção de Silêncio: Eliminar os silêncios no início e no final das amostras garante que o modelo não seja treinado em partes irrelevantes, melhorando a eficiência do treinamento.
- Normalização: A normalização garante que todas as amostras tenham uma escala uniforme, facilitando o processo de treinamento e levando a um modelo mais robusto.
- Limiar de Silêncio: Definir um limiar de 60 dB para determinar o silêncio ajuda na identificação de segmentos úteis, evitando o treinamento em partes desnecessárias.
- Faixa de Frequência para Transformação Mel: A limitação da faixa de frequência a 0 a 8000 Hz concentra o modelo nas frequências mais relevantes para a percepção humana, otimizando o desempenho.

5. Experimentos

5.1 Experimento Preliminar: CoquiAI/GlowTTS

O CoquiAI/GlowTTS é uma arquitetura de Text-to-Speech baseada no modelo Glow. Utiliza um mecanismo de normalização de fluxo invertível que permite um treinamento eficiente e uma síntese de alta qualidade. Neste experimento, o modelo foi treinado por 150 épocas, com uma divisão de 80% para treino e 20% para validação. A redução de 60% das amostras foi necessária devido a limitações de hardware. Uma dificuldade encontrada foi a obtenção das saídas após a inferência, destacando desafios na implementação prática.

5.2 Segundo Experimento: Tacotron 2 e MultiBand MelGAN

Este experimento explorou a combinação de dois modelos:

- Tacotron 2: Um modelo de síntese de fala que recebe um texto e infere um Mel espectrograma. É composto por uma rede de codificação de texto e uma rede de decodificação de espectrograma, com um módulo de atenção entre eles.

- MultiBand MelGAN: Um vocoder modelo convolucional que mapeia um espectrograma para áudio. Utiliza várias bandas de frequência para acelerar o treinamento e melhorar a qualidade do áudio.

A integração desses dois modelos permite uma geração de voz eficiente e de alta qualidade, combinando a interpretação semântica do Tacotron 2 com a síntese de áudio precisa do MultiBand MelGAN.

5.3 Terceiro Experimento: FastSpeech 2

O FastSpeech 2 é uma evolução do modelo FastSpeech original, que é um modelo de Text-to-Speech não autoregressivo. Ao contrário dos modelos autoregressivos, o FastSpeech 2 pode gerar áudio em paralelo, tornando a síntese mais rápida. Ele utiliza duração e tom previamente extraídos, juntamente com uma codificação de texto, para gerar um espectrograma Mel, que é então convertido em áudio. É um modelo End-to-End, o que significa que ele pode receber um texto como entrada e fornecer um waveform como saída, sem a necessidade de estágios intermediários.

6. Mixagem Automatizada

A mixagem automatizada desempenha um papel crucial na solução proposta, servindo como a ponte entre a síntese de fala e a integração harmoniosa da audiodescrição no conteúdo original do vídeo. O algoritmo de mixagem adotado neste projeto realiza várias funções essenciais:

- **Carregamento do Áudio Sintetizado:** O arquivo de áudio sintetizado é carregado, preparando-o para a manipulação e integração.
- **Aplicação de Ganho a um Clipe de Áudio:** Ajusta-se o ganho do clipe de áudio para corresponder ao nível desejado de dBFS (Decibels Full Scale), assegurando que o áudio sintetizado esteja em conformidade com os padrões de volume.
- **Aplicação de Aceleração a um Clipe de Áudio:** Modifica-se a velocidade do clipe de áudio com base no conteúdo do texto, permitindo um controle fino sobre o ritmo e a cadência da fala.
- **Mistura de Áudio Sintetizado com Áudio Base:** O áudio sintetizado é mesclado com o áudio original do vídeo, aplicando técnicas de equalização e outros efeitos para criar uma trilha sonora integrada.

A mixagem automatizada é fundamental para a escalabilidade da solução. Ela permite que grandes volumes de conteúdo sejam processados com eficiência, mantendo a qualidade e a coesão da saída de áudio. Ao eliminar a necessidade de ajustes manuais demorados, a mixagem automatizada garante uma consistência que pode ser difícil de alcançar de outra forma. Em um contexto de audiodescrição, isso significa que a trilha sonora original e a fala sintetizada podem ser combinadas de maneira suave e natural, enriquecendo a experiência para o público-alvo sem comprometer a integridade do conteúdo original.

7. Conclusão

Este projeto apresentou uma solução inovadora para a geração de conteúdo acessível para deficientes visuais, utilizando tecnologias de síntese de fala (TTS) para a audiodescrição. Através de um processo metodológico bem definido, composto por geração de Text-To-Speech, personalização de voz, e inserção automatizada de audiodescrição, foi desenvolvido um sistema capaz de transformar textos em audiodescrições naturais e envolventes.

Uma das tarefas desenvolvidas neste trabalho foi o algoritmo de mixagem automatizada, que não apenas facilitou a integração da audiodescrição com o conteúdo original, mas também demonstrou o potencial para escalabilidade. A sua implementação bem-sucedida abre caminhos para a produção em massa de conteúdo acessível, mantendo a qualidade e a coesão do áudio.

O trabalho também enfrentou desafios, incluindo a escassez de documentação, incompatibilidades entre bibliotecas, e a falta de modelos pré-treinados em português brasileiro. Essas dificuldades refletem os obstáculos ainda presentes no campo emergente da TTS e ressaltam a necessidade de colaboração e pesquisa contínua.

Em suma, a Prova de Conceito (POC) realizada neste projeto validou a ideia central da solução, fornecendo uma base sólida para futuras expansões e melhorias. A importância da audiodescrição na inclusão de deficientes visuais no mundo do conteúdo multimídia não pode ser subestimada, e este trabalho contribui para tornar essa inclusão uma realidade mais tangível e acessível.

Referências

1. Van Den Oord, A., et al. "WaveNet: A Generative Model for Raw Audio." 2016.
2. Shen, J., et al. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions." 2017.
3. Vasquez, D. and Lewis, M. "Multi-Band MelGAN: Faster Waveform Generation for High-Quality Text-to-Speech." 2020.
4. Ren, Y., et al. "FastSpeech 2: Fast, High-Quality, and Robust Text-to-Speech System." 2020.
5. Coqui, The AI Voice Research Foundation. "Coqui TTS." GitHub.