Universidade Federal da Paraíba

Desenvolvimento de jogo com reconhecimento de voz em tempo real

Artur Dartagnan de Oliveira Vasconcelos (20210026643)

Jose Augusto da Silva Barbosa (20210094705)

Thiago Rodrigues Cruz Justino (20220007276)

2024

Sumário

Apresentação do problema

Na realidade contemporânea, a utilização de tecnologias de reconhecimento de fala tem se tornado cada vez mais proeminente, integrando-se a diversas ferramentas do cotidiano das pessoas. A popularização de assistentes virtuais, como Google Assistant, Alexa e Siri, exemplifica como o reconhecimento de voz está moldando a interação humana com a tecnologia. Essa tendência também se reflete em funções comuns, como a transcrição de áudios no WhatsApp, que facilitam a comunicação e promovem uma experiência mais fluida e acessível no dia a dia.

Entretanto, quando olhamos para o setor de jogos, notamos uma lacuna significativa. Embora as tecnologias de reconhecimento de fala estejam amplamente presentes em outros contextos, sua aplicação nos jogos ainda é limitada e pouco explorada. A falta de jogos que incorporam essa tecnologia impede que muitos jogadores possam desfrutar de experiências interativas e dinâmicas que poderiam ser enriquecidas por comandos de voz.

Nesse contexto, decidimos desenvolver um jogo que integra o reconhecimento de fala em tempo real. Este projeto não apenas busca inovar na forma como os jogos são jogados, mas também se relaciona com questões importantes na área da fonoaudiologia. A necessidade de uma pronúncia clara e precisa das palavras se torna essencial para o sucesso dentro do jogo, o que pode auxiliar na reabilitação da fala e na prática de habilidades linguísticas para jogadores que buscam aprimorar sua dicção.

Além disso, a criação de modelos especializados para a interpretação de comandos de voz com alta eficiência é uma parte crucial do nosso projeto. Esse aspecto é vital, pois o jogo requer que os jogadores emitam comandos verbais para controlar um navio, realizando ações como navegar, manobrar e executar manobras táticas. A precisão e a resposta imediata do reconhecimento de fala são fundamentais para garantir uma experiência de jogo envolvente e satisfatória.

Outro aspecto relevante que nossa iniciativa aborda é a acessibilidade. Muitos jogos populares de hoje em dia não consideram as necessidades de jogadores com deficiências, especialmente aqueles que têm dificuldades motoras ou de interação. Ao integrar o reconhecimento de fala, nosso jogo busca criar um espaço inclusivo, permitindo que pessoas com diferentes capacidades possam participar e se divertir em um ambiente de jogo que valoriza suas habilidades e promove a igualdade de oportunidades.

Objetivos

- **Explorar a utilização do reconhecimento de fala em jogos**: o relatório visa investigar como a tecnologia de reconhecimento de fala pode ser integrada aos jogos, destacando a escassez de jogos existentes que utilizam essa tecnologia. A pesquisa busca compreender o potencial dessa integração para enriquecer a experiência de jogo e ampliar as possibilidades de interação.
- **Desenvolver um jogo com reconhecimento de fala em tempo real**: um dos principais objetivos do projeto é a criação de um jogo que utilize reconhecimento de fala em tempo real, permitindo que os jogadores emitam comandos verbais para controlar as ações dentro do jogo. Este objetivo inclui a definição das mecânicas do jogo, a implementação do sistema de reconhecimento de voz e a realização de testes para garantir sua eficácia.
- **Criar modelos especializados para comandos de voz**: outro objetivo importante é a criação de modelos de reconhecimento de fala especializados que possam interpretar comandos verbais com alta eficiência. O relatório abordará como esses modelos são treinados e ajustados para atender às necessidades específicas do jogo e dos jogadores.
- **Contribuir para a discussão sobre inovações em jogos**: Por fim, o relatório pretende contribuir para a discussão mais ampla sobre inovações tecnológicas no setor de jogos. Ao apresentar as descobertas e as lições aprendidas durante o desenvolvimento do projeto, espera-se inspirar outros desenvolvedores e pesquisadores a explorar o potencial do reconhecimento de fala em suas próprias criações.

Dados utilizados e pré-processamento dos dados

Para este projeto, foi criado um conjunto de dados contendo comandos de movimentação e ações relacionados a um jogo utilizando a biblioteca Pygame. As teclas de movimentação incluem as direções de esquerda, direita, frente e trás, e também foram adicionados comandos relacionados a duas ações principais: turbo e pause.

Os dados foram gerados com base em combinações de diferentes prefixos para as movimentações e ações, resultando em uma variedade de frases que simulam comandos possíveis durante o jogo. Abaixo, detalhamos como cada conjunto de dados foi gerado:

Para cada movimento, utilizamos uma lista extensa de prefixos e os movimentos como esquerda, direita, cima e baixo.

Para as ações de turbo e pause, além de utilizarmos prefixos similares aos das movimentações, também incluímos variações mais detalhadas para cada ação, como "acelerar", "ligar o turbo", "pausar o jogo", entre outras. Essa abordagem expandiu a diversidade dos comandos para essas duas ações, gerando frases mais naturais e variadas.

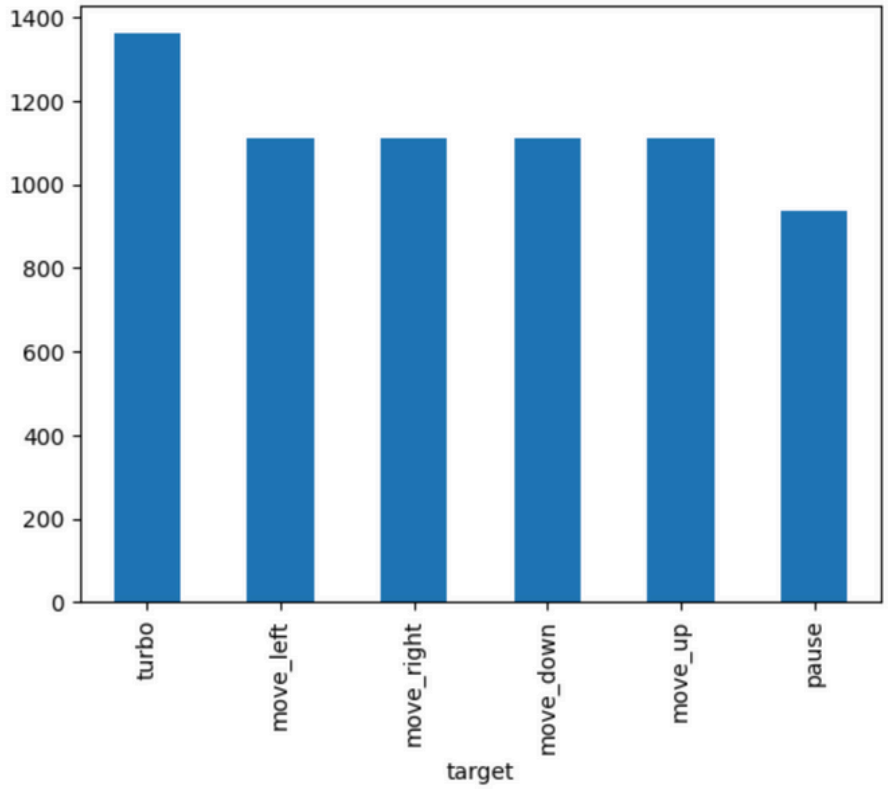
Com base nas teclas de movimentação e ações, foi criada uma função de geração de frases que combina os prefixos com os termos de movimentação e as ações do jogo. A função foi executada diversas vezes para garantir uma diversidade nos exemplos gerados. Além disso, para as ações, utilizamos uma lista mais extensa de sinônimos e variações para aumentar ainda mais o vocabulário dos comandos.

Esses comandos foram associados a seus respectivos movimentos e ações, formando o conjunto de dados final, no qual cada frase foi atribuída a um "target" (ou seja, o movimento ou ação correspondente).

O conjunto final de dados foi armazenado em um arquivo CSV contendo todas as frases geradas e seus respectivos targets. Ao todo, foram geradas 6.735 frases, cada uma mapeada para um movimento ou ação. Esse arquivo CSV foi utilizado para treinar modelos de machine learning, permitindo que eles aprendessem a associar comandos de linguagem natural a ações específicas dentro do jogo.

Dados utilizados e pré-processamento dos dados

Antes de iniciar o treinamento dos modelos, os dados passaram por um pré-processamento, onde foram removidas duplicatas e inconsistências nas frases. Além disso, as frases foram transformadas em uma representação adequada para o modelo RandomForest, utilizando técnicas de tokenização e vetorização para converter as frases de texto em vetores numéricos.



1-Gráfico com o número de frases por Target

	comando	target
0	Vá para esquerda	move_left
1	Vá para direita	move_right
2	Vá para frente	move_down
3	Vá para trás	move_up
4	Ande para esquerda	move_left

2-Tabela de exemplo

Metodologia - Técnica utilizada

Para reconhecer comandos de voz em tempo real, utilizamos a biblioteca Vosk, que oferece um sistema de reconhecimento de fala offline baseado no Kaldi. O Vosk foi escolhido por sua precisão, suporte a vários idiomas, incluindo o português, e pela capacidade de funcionar sem a necessidade de uma conexão com a internet.
A configuração do sistema envolve:
Vosk Model: Utilizamos o modelo pré-treinado em português, vosk-model-small-pt-0.3. Esse modelo foi carregado localmente e usado para processar o áudio em tempo real.
PyAudio: Foi utilizado para capturar áudio do microfone, com uma taxa de amostragem de 16 kHz, que é ideal para reconhecimento de fala.O fluxo de execução do reconhecimento de voz ocorre da seguinte forma:
Inicializa o sistema de captura de áudio via PyAudio.
1.O áudio capturado é processado pelo Vosk em tempo real, e os resultados parciais e finais da transcrição são gerados.
2.Cada transcrição é processada e mapeada para os comandos correspondentes no jogo.
Para associar os comandos transcritos às ações e movimentos do jogo, foi utilizado o algoritmo RandomForest. Esse algoritmo foi escolhido por sua robustez e capacidade de generalização em problemas de classificação, além de ser menos suscetível a overfitting em relação a outros métodos.
O fluxo de trabalho para a classificação dos comandos envolve os seguintes passos:

- Pré-processamento: As frases transcritas pelo SpeechRecognition são vetorizadas e transformadas em representações numéricas adequadas.
- Treinamento: Um modelo RandomForest foi treinado com um conjunto de dados sintéticos gerados a partir de frases de comando, conforme descrito na seção anterior.
- Predição em Tempo Real: Para cada comando transcrito em tempo real, o modelo RandomForest classifica se o comando é relacionado a um movimento (esquerda, direita, frente, trás) ou uma ação (turbo, pause).

Metodologia - Experimento para avaliar a técnica utilizada

Avaliação da Acurácia da Classificação de Comandos

- Critério: A acurácia foi avaliada utilizando validação cruzada nos dados de treinamento.
- Métricas: Foram analisadas as seguintes métricas: acurácia, precisão, recall e F1 score.

A avaliação do modelo de transcrição em tempo real foi realizado de forma prática: realizou-se testes em tempo real durante a execução do jogo, onde o jogador emitia comandos de voz e o sistema respondia em tempo real. A avaliação considerou a latência entre o comando falado e a resposta do jogo, bem como a precisão na interpretação dos comandos.

Resultados

O sistema de transcrição de voz e classificação de comandos para o jogo se mostrou altamente eficaz. Utilizando o SpeechRecognition para reconhecimento de fala em português, o sistema capturou os comandos com precisão, mesmo diante de variações na fala ou interferências de ruído. A transcrição foi quase instantânea, garantindo que os comandos do jogador fossem executados de maneira fluida e sem atrasos perceptíveis durante a jogabilidade.

O modelo Random Forest utilizado para classificar os comandos apresentou uma acurácia de 100% nos testes, mostrando-se capaz de associar corretamente as transcrições de voz aos comandos esperados, mesmo com diferentes formas de expressar os comandos. Isso garantiu uma experiência de jogo natural, onde o jogador podia controlar ações como "ativar turbo" ou "mover-se para a direita" sem interrupções.

Nos testes práticos, o sistema respondeu de forma rápida e precisa aos comandos de voz, proporcionando uma jogabilidade imersiva e interativa. A robustez do modelo Random Forest e a eficiência do SpeechRecognition em transcrever a fala em tempo real garantiram um desempenho consistente, abrindo caminho para a ampliação da tecnologia para outros jogos ou aplicações.

3-Imagem do jogo