Universidade de São Paulo
Instituto de Matemática e Estatística
Bachalerado em Ciência da Computação

Leonardo Pereira Macedo

Desenvolvimento de um módulo de reconhecimento de voz para a *game engine* Godot

São Paulo
9 de outubro de 2017

Desenvolvimento de um módulo de reconhecimento de voz para a *game engine* Godot

Monografia final da disciplina
MAC0499 – Trabalho de Formatura Supervisionado

Supervisor: Prof. Dr. Marco Dimas Gubitoso

São Paulo
9 de outubro de 2017

Agradecimentos

Resumo

A área de *games* evoluiu muito desde o início da década da 70, quando começaram a ser comercializados. As principais causas estão relacionadas aos avanços em diferentes áreas da Computação.

Com o passar do tempo, surgiram as *game engines*: *frameworks* voltados especificamente para a criação de jogos, visando a facilitar o desenvolvimento e/ou algumas de suas etapas.

Focaremos em uma *game engine* em particular, *Godot*. Por possuir código aberto, este *software* permite a extensão de suas funcionalidades através da criação de novos módulos.

Este projeto busca implementar um módulo de reconhecimento de voz para *Godot*, depois demonstrando a nova capacidade em um jogo simples desenvolvido na própria plataforma.

Palavras-chave: *software*, *game engine*, *Godot*, desenvolvimento de módulo, extensão de funcionalidade.

Abstract

Video games have evolved considerably since the beginning of the 70's, when they started to be commercialized. The main reasons are related to several advances in different fields of Computer Science.

Over time, *game engines* started appearing: *frameworks* designed specifically to assist on game creation, simplifying the process and/or some of its steps.

We will focus on a specific game engine, *Godot*. Since it is an open source project, it is possible to extend its funcionalities by creating new modules.

This project's goal is to implement a speech recognition module for *Godot*, then showing the new feature in a simple game developed on the engine itself.

Keywords: software, game engine, *Godot*, module development, functionality extension.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Reconhecimento de voz

Neste capítulo, abordaremos a parte teórica do reconhecimento de voz, sem nos preocuparmos com sua aplicação no contexto deste trabalho. Em particular, analisaremos brevemente os diferentes parâmetros que influenciam seu uso.

1.1 Definição

Reconhecimento automático de voz (ou da fala), ou *speech to text* (STT), é um campo multidisciplinar que envolve as áreas de Inteligência Artificial, Estatística e Linguística. Busca-se desenvolver metodologias e tecnologias para que computadores sejam capazes de captar, reconhecer e traduzir a linguagem falada para texto.

A figura 1.1 apresenta os três componentes de um programa genérico STT:

- O **usuário**, que codifica um comando através de sua voz;
- O **dispositivo**, que converte a mensagem falada para um formato interpretável;
- O **software de aplicação**, que recebe a saída do dispositivo e realiza uma ação apropriada.

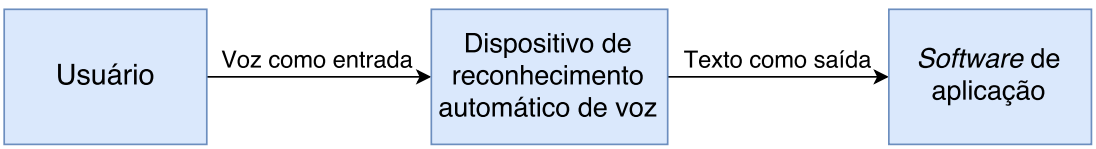

Figura 1.1: *Sistema genérico de reconhecimento automático de voz*

1.2 História

O primeiro sistema de reconhecimento de voz conhecido foi o *Audrey*, construído em 1952 por três pesquisadores do *Bell Labs* para reconhecer dígitos falados por um único usuário.

10 anos depois, a IBM apresentou o *Shoebox*, que reconhecia 16 palavras em inglês, entre elas os dígitos de 0 a 9. Quando captava palavras como *plus*, *minus* ou *total*, *Shoebox* instruía outra máquina de adições a realizar cálculos ou imprimir o resultado. A entrada era feita por um microfone (figura 1.2), que convertia a voz do usuário em impulsos elétricos, classificados internamente por um circuito de medição.

Figura 1.2: *Máquina Shoebox sendo operada*

Sistemas de reconhecimento de voz só tiveram um avanço realmente significativo na década de 80, devido a um método estatístico denominado *modelo oculto de Markov* (ou **HMM**, sigla para *Hidden Markov Model*). Ao invés de procurar por modelos de palavras em padrões de som, HMM considerava a probabilidade de um som desconhecido possuir palavras, o que acelerou o processo e tornou possível usar um vocabulário maior nos computadores. Outro modelo que ganhou bastante popularidade na época foi o de redes neurais, que é efetivo para classificar palavras isoladas e fonemas individuais mas encontra problemas em tarefas envolvendo reconhecimento contínuo. Ao contrário do HMM, este método não consegue modelar bem dependências temporais.

A evolução na tecnologia de reconhecimento de voz foi tamanha que, atualmente, é inegável seu impacto em nosso dia a dia. Um celular moderno consegue captar palavras ou pequenas frases de seu usuário dentre um enorme vocabulário para fazer buscas na Internet, tocar uma música ou fazer uma ligação. Muitas empresas utilizam máquinas para receber ligações de seus clientes; de acordo com o que interpretam, a chamada é redirecionada para um funcionário mais adequado. Alguns países chegam até a usar reconhecimento de voz para autenticar a identidade de alguém por telefone, com o objetivo de evitar fornecer dados pessoais pelo mesmo. Também há usos em transportes, na área médica e para fins educativos.

1.3 Parâmetros principais

Há diversos tipos de parâmetros que caracterizam as capacidades de um sistema de reconhecimento de voz. Eles se subdividem nos três tipos a seguir.

1.3.1 Específicos ao aplicativo

Estão relacionados à *forma* com que o aplicativo em si realiza o reconhecimento de voz. Inclui dois parâmetros:

- A **forma de falar**, que pode ser através de *palavras isoladas*, com pausas entre elas; *palavras conectadas*, que são concatenadas sem pausas; ou *fala contínua*, onde o fluxo de palavras é semelhante a uma fala natural.
- **Existência de treinamento**, que subdivide aplicativos em dois grupos:
  - Os sistemas **dependentes** (*speaker-dependent*), caracterizados pelo *treinamento* feito pelo usuário. Isto é, são computadores que analisam e se adaptam aos padrões particulares da fala captada, resultando em uma maior acurácia. Geralmente, o usuário deve ler algumas páginas de texto para a máquina antes de começar a usar o sistema. Esta variante é comumente usada em casos particulares, onde um número limitado de palavras deve ser reconhecido com bastante precisão.
  - Os sistemas **independentes** (*speaker-independent*), que são desenvolvidos para reconhecer a voz de qualquer pessoa e não requerem treinamento. É a melhor opção para aplicações interativas que usam voz, já que não é viável fazer com que os usuários leiam páginas de texto antes do uso. Sua desvantagem é a acurácia menor se comparado ao reconhecimento dependente; para contornar isso, costuma-se limitar o vocabulário reconhecido pelo sistema.

1.3.2 Específicos à tarefa

Dependendo do objetivo a ser alcançado com o reconhecimento de voz, alguns parâmetros podem ser melhor ajustados para obter maior velocidade ou acurácia. São eles:

- O **vocabulário**, referente a quantas palavras são reconhecidas pelo sistema. O tamanho pode ser pequeno (menor que 20 palavras) até muito grande (mais de 20 mil palavras), sendo diretamente proporcional à velocidade do reconhecimento. Além disso, a similaridade entre a pronúncia de algumas palavras pode afetar a acurácia, uma vez que a distinção entre elas torna-se mais complicada.

- A **sintaxe**, isto é, a gramática artificial que o sistema aceita para uma determinada tarefa. O exemplo mais simples seria uma máquina de estados finita, onde as palavras permitidas após um estado ou nó são definidas explicitamente.
- O **fator de ramificação**, que é uma forma de medir a complexidade da sintaxe. É definido como o número médio de palavras permitidas em cada nó da gramática, e possui grande impacto sobre o desempenho do sistema.

1.3.3 Ambientais

Dentre os vários parâmetros externos ao sistema e que podem interferir no reconhecimento de voz, destacam-se:

- A taxa **sinal-para-ruído**, que avalia a intensidade média do sinal recebido em relação ao ruído de fundo, tipicamente medido em decibéis (dB). Quanto menor a taxa, maior a dificuldade no reconhecimento de voz.
- O **próprio usuário**, o que inclui o volume de sua voz, a velocidade com que fala e até mesmo sua condição psicológica: o nível de estresse de um piloto sob ataque em uma aeronave é diferente de alguém simplesmente querendo ouvir uma música, por exemplo.