Raphael Ferreira Ramos

Jogos fisicamente interativos baseados em reconhecimento de voz usando Modelos Ocultos de Markov

Campos dos Goytacazes/RJ

2012

Raphael Ferreira Ramos

Jogos fisicamente interativos baseados em reconhecimento de voz usando Modelos Ocultos de Markov

Monografia apresentada ao Curso de Graduação em Ciência da Computação da Universidade Estadual do Norte Fluminense Darcy Ribeiro como requisito para obtenção do título de Bacharel em Ciência da Computação, sob orientação da Profº. Rivera Antônio Escriba, DrSc.

Tutor: Rivera Antônio Escriba, DrSc.

Universidade Estadual do Norte Fluminense Darcy Ribeiro

Campos dos Goytacazes/RJ

2012

"Somos quem podemos ser, sonhos que podemos ter."

Humberto Gessinger

AGRADECIMENTOS

À meu pai, minha mãe e especialmente à você.

Lista de Figuras

Lista de Códigos

Resumo

Aqui entra o resumo do meu trabalho que será a última coisa a ser feita.

Sumário

1 Introdução

A fala é a principal forma de comunicação dos seres humanos, desde o início dos computadores, a busca por computadores mais inteligentes, levam cientistas ao estudo de sistemas de *reconhecimento automático de voz*, visando uma comunicação natural entre o homem e a máquina. (SILVA, 2010) Os sistemas de reconhecimento automático de voz *(RAV)* evoluiram considerávelmente com o passar dos anos, e sua aplicação se encontra em diversas áreas, como: sistemas para atendimento automático, ditado, interfaces para computadores pessoais, controle de equipamentos, robôs domésticos, indústrias totalmente à base de robôs inteligentes, etc. (SILVA, 2010) Mas mesmo com toda evolução do hardware dos computadores e otimização dos algoritmos e métodos, os sistemas *(RAV)* estão longe de compreender um discurso sobre qualquer assunto, falado de forma natural, por qualquer pessoa, em qualquer ambiente.(SILVA, 2009)

1.1 Objetivos

O objetivo geral deste trabalho é desenvolver um jogo interativo guiado por comandos voz ditados pelo usuário, o jogo é baseado em um clássico do mundo dos games, pacman, onde o objetivo do personagem principal é comer todas as pastilhas, e não ser devorado pelos 4 fanstamas que o perseguem por um labirinto. A interação é feita usando comandos de fala pré-definidos em sua gramática, que são: DIREITA, ESQUERDA, SUBIR, DESCER. Além de ser guiado por esses comandos, o jogo também reconhece determinadas palavras que podem caracterizar o humor do usuário, como: BURRO, DROGA, MERDA, pronunciadas essas ofensas, o usuário recebe uma penalidade, até perder a partida.

1.2 Visão Geral do Trabalho

2 *Referencial teórico*

Este capítulo descreve os principais elementos teóricos utilizados no desenvolvimento desta pesquisa. As seções 2.2 e **??** resumem os principais conceitos sobre redes neurais, assim como as principais técnicas utilizadas nesta área. A seção 2.1 é dedicada especificamente as redes de Kohonen, descrevendo sua estrutura conceitual e seu algorítimo de treinamento, esta categoria de rede neural é o núcleo da técnica de *clustering* de imagens proposta neste trabalho, assunto abordado no próximo capítulo. Uma breve formalização dos descritores de Hu é feita na seção 2.3. E por fim, alguns conceitos chave sobre imagens digitais são apresentados na seção 2.4.

2.1 Histórico dos Sistemas de Reconhecimento de Voz

Sistemas de reconhecimento automático de voz vem sendo estudados desde os anos 50 nos laboratórios Bell, quando foi criado, o primeiro reconhecedor de dígitos isolados com suporte a um locutor.(CUNHA, 2003)

2.2 Características de Sistemas RAV

Existem várias maneiras de categorizar um sistema de reconhecimento de voz, os mais importantes são: o estilo de pronuncia que é aceito, ao tamanho do vocabulário e à dependência ou independencia do locutor. (MARTINS, 1997) Essas categorias que definem a precisão do sistema de reconhecimento.

2.2.1 Dependência do locutor

Podemos classificar sistemas de reconhecimento como dependentes e independentes do locutor. Um sistema dependente de locutor reconhece a fala das pessoas cujas vozes foram utilizadas para treinar o sistema, apresentando uma pequena taxa de erros, para o locutor

para qual foi treinado o sistema, implementação mais simples que sistemas independentes do locutor, que reconhecem a fala de qualquer pessoa com uma taxa de acerto aceitável. Neste caso é necessário realizar o treino do sistema com uma base que inclua diferentes pessoas com diferentes idades, sexo, sotaques, etc. O que dificulta a construção desses sistemas.

2.2.2 Modo de pronúncia

Sistemas RAV podem ser classificados quanto ao modo de pronúncia de duas formas, sistemas de palavras isoladas e os de fala conectadas(contínua). Reconhecedor de palavras isoladas são sistemas que reconhecem palavras faladas isoladamente, isto é, entre cada palavra deve existir uma pausa mínima, para que seja detectado o início e o fim da mesma. Isso proporciona um resultado muito superior aos de fala contínua, estes sistemas são os mais simples de serem implementados. Um exemplo clássico de reconhecedores de palavras isoladas são os reconhecedores de dígitos, que alcançam taxa de menos de 2% de erro para dígitos de 0 à 10.(SILVA, 2010) Já o reconhecedor de palavras conectadas são sistemas mais complexos que os de palavras isoladas e utilizam palavras como unidade fonética padrão. São capazes de reconhecer sentenças completas, pronunciadas sem pausa entre as palavras, e por isso não se tem informação de onde começam e terminam determinadas palavras, muitas palavras são mascaradas, encurtadas e as vezes não pronunciadas. Esses sistemas precisam lidar com todas as características e vícios da linguagem natural, como o sotaque, a duração das palavras, a pronúncia descuidada, etc. Tornando ainda mais difíceis as tarefas do reconhecedor em casos como “ele vai morrer em dois dias” que muitas vezes é dito como “ele vai morrerem dois dias”.(SILVA, 2010)

2.2.3 Tamanho do vocabulário

Um fator muito importante na precisão de um RAV, é o tamanho do vocabulário, quanto maior seu tamanho, maior a quantidade de palavras ambíguas, com realizações sonoras semelhantes, ocasionando maior chance de erros por parte do decodificador responsável pelo reconhecimento.(SILVA, 2010) Segundo (SILVA, 2009) vocabulários podem ser definidos como:

- Vocabulário pequeno: reconhecem até 20 palavras.
- Vocabulário médio: reconhecem entre 20 e 100 palavras.

- Vocabulário grande: reconhecem entre 100 e 1000 palavras.
- Vocabulário muito grande: reconhecem mais de 1000 palavras.

2.2.4 Perplexidade

2.2.5 Relação sinal – ruído

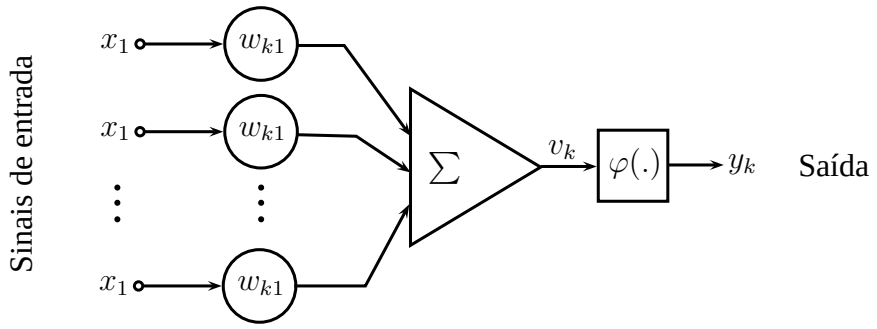

Figura 1: Texto da figura

Processo competitivo

Quando uma entrada $x = [x_1, x_2, ..., x_n]^T$ é apresentado à rede, o neurônio da grade que melhor responder a este padrão será ativado, este neurônio é dito vencedor, e será recompensado ajustando-se seus componentes para mais próximo do vetor de entrada.

O critério escolhido para determinar o neurônio vencedor é a distância euclidiana entre o vetor de entradas e o vetor de pesos das sinapses do neurônio, como indicado na equação 2.1:

$$d_i(t) = \sqrt{\sum_{j=1}^{N} (x_j(t) - w_{ij}(t))^2} \tag{2.1}$$

Onde:

- $d_i(t)$ é a distância euclidiana entre o vetor de pesos do neurônio i e o vetor de entradas na iteração t;
- i é o índice do neurônio da grade;
- j é o índice do neurônio de entrada;
- N é o número de entradas;

- $x_j(t)$ é o sinal de entrada na entrada j na iteração t;
- $w_{ij}(t)$ é o valor do peso sináptico entre o neurônio de entrada j e o neurônio da grade i na iteração t.

Algoritmo geral de treinamento

O algoritmo 1 resume as três etapas anteriores e descreve todo o processo de treinamento de uma rede de Kohonen:

Algorithm 1: Treinamento de uma rede de Kohonen

Entrada: σ_0 , τ_l , η_0 e o valor do *erro*
início
  repita
    Calcular a *largura efetiva* $\sigma(t)$;
    Calcular a *vizinhança topológica* h;
    Calcular a *taxa de aprendizado* $\eta(t)$;
    para cada *conexão* **faça**
      Calcular Δw;
      Ajustar o arco;
    fim para cada
  até *distâncias auclidianas* $\leq$ *erro*;
fim

2.3 Descritores de Hu

Os descritores de Hu são um conjunto de sete momentos invariantes a rotação, translação e escala.

O momento bidimensional de ordem $(p+q)$ é dado pela equação 2.2:

$$m_{pq} = \iint x^p y^q f(x,y) \mathrm{d}x\mathrm{d}y, p, q \in \tag{2.2}$$

A equação num domínio discreto, pode ser reescrita na forma:

$$m_{pq} = \sum_{x,y} x^p y^q f(x,y), p, q \in \tag{2.3}$$

A massa total da função $f(x, y)$ é determinado pelo momento m_{00}, conforme a equação 2.4:

$$m_{pq} = \sum_{x,y} f(x, y), p, q \in \tag{2.4}$$

Existe um ponto no qual a aplicação pontual da massa total gera o mesmo momento que a massa distribuída, este ponto é dito centroide de $f(x, y)$ e suas coordenadas x e y são dadas pela equação 2.5:

$$\bar{x} = \frac{1}{m_{00}} \sum x f(x, y) = \frac{m_{10}}{m_{00}} \tag{2.5a}$$

$$\bar{y} = \frac{1}{m_{00}} \sum y f(x, y) = \frac{m_{01}}{m_{00}} \tag{2.5b}$$

O momento central é obtido se deslocando a imagem para o centroide, da seguinte forma:

$$\mu_{pq} = \sum_{x,y} (x - \bar{x})^p (y - \bar{y})^q f(x, y) \tag{2.6}$$

Ainda é necessário normalizar o momento para que os valores resultantes não sejam extremos a ponto de serem ignorados pelo sistema de reconhecimento de padrões. O momento central de ordem $(p + q)$ normalizado é obtido dividindo o momento central de y mesma ordem por um fator definido por μ_{00}^{γ} , conforme indicado pela equação 2.7:

$$\gamma = 1 + \frac{p + q}{2} \tag{2.7a}$$

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{2.7b}$$

A partir dessas equações são estabelecidos sete momentos invariantes à translação, rotação e escala, chamados de momentos de Hu, ou descritores de Hu. São eles:

$$\varphi_1 = \eta_{20} + \eta_{02} \tag{2.8a}$$

$$\varphi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{2.8b}$$

$$\varphi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{2.8c}$$

$$\varphi_4 = (\eta_{30} + \eta_{12})^2 + (3\eta_{21} + \eta_{03})^2 \tag{2.8d}$$

$$\varphi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] \tag{2.8e}$$

$$+(3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{2.8f}$$

$$\varphi_6 = (\eta_{20} - \eta_{02}) \left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{2.8g}$$

$$+4\eta_{11}(\eta_{30} - \eta_{12})(\eta_{21} + \eta_{03}) \tag{2.8h}$$

$$\varphi_7 = (3\eta_{21} - \eta_{30})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] \tag{2.8i}$$

$$+(3\eta_{12} - \eta_{03})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{2.8j}$$

2.4 Imagens digitais

Imagens digitais são representações computacionais de imagens bidimensionais, codificadas de modo a permitir seu armazenamento, exibição e manipulação por dispositivos eletrônicos. Há dois tipos fundamentais de imagens digitais, os mapas de bits (*bitmaps*) e as imagens vetoriais.

2.4.1 Mapas de bits

Mapa de bists é a representação matricial de uma imagem, onde cada posição, chamada de *pixel*, armazena uma cor. Normalmente os *pixels* são codificados no padrão RGB (*Red*, *Green*,*Blue*), que utiliza três *bytes* para armazenar um inteiro para as cores vermelha, verde e azul, respectivamente. Em mídias impressas é comum que as imagens *bitmaps* utilizem o padrão CMYK (*Cian*, *Magenta*, *Yelow*, *Black*) ao invés do RGB.

Embora uma imagem bitmap seja armazenada na RAM com todos os *pixels*, é comum, por uma questão de economia de memória e tempo de transmissão, a compressão destes arquivos. Entre os principais formatos de compressão estão o GIF (*Graphics Interchange Format*), o JPEG (*Joint Photographic Experts Group*) e o PNG (*Portable Network Graphics*).

2.4.2 Imagens vetoriais

As imagens vetoriais são formadas pela descrição geométrica de objetos. Por serem compostas de vetores, este tipo de imagem ocupa menos espaço na memória comparado com as bitmaps, e não perdem a qualidade quando aplicado transformações de escala e rotação sobre elas.

Referências Bibliográficas

CUNHA, A. M. da. *Métodos probabilísticos para reconhecimento de voz*. Monografia (Graduação), Rio de Janeiro, 2003.

MARTINS, J. A. *Reconhecimento de voz para palavras isoladas*. Monografia (Doutorado), Campinas, 1997.

SILVA, A. G. da. *Reconhecimento de voz para palavras isoladas*. Monografia (Graduação), Recife, 2009.

SILVA, C. P. A. da. *Um Software de Reconhecimento de Voz para Português Brasileiro*. Monografia (Pós-Graduação), Belém, 2010.