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, interação vista apenas em filmes de ficção científica. (SILVA, 2010) Para esses estudos virarem realidade, os computadores terão de possuir total entendimento da fala humana, capacidades como: falar, ouvir, ler, escrever, alem do reconhecimento de pessoas pela voz, devem ser estabelecidas. Essas capacidades são os objetivos dos sistemas de RAV, permitindo que o computador "entenda"o que está sendo dito.(SILVA, 2008) Os sistemas de reconhecimento automático de voz *(RAV)* evoluiram consideravelmente 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) Com a melhoria do hardware, os jogos de computadores, se tornaram cada vez mais parecidos com a realidade em gráficos e na interatividade, a tendencia sugere que os famosos joysticks poderão ser aposentados em pouco tempo.

O primeiro jogo, foi desenvolvido em 30 de julho de 1961, por Steve Russel, não tinha objetivos comerciais, apenas acadêmicos. O principal objetivo de Steve Russel era poder mostrar todo o poder de processamento do computador DEC PDP-1, Para isso criou SpaceWar. Inicialmente a ideia de Russel era fazer um filme interativo, mas acabou se tornando o pai dos jogos eletrônicos. [Henrique Moraes Ramos , 2007(Historia dos jogos)]

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 Motivações

Aumento de desempenho individual, pois sendo o meio de comunicação mais natural para o ser humano, os comandos por voz seriam mais rápidos que por joystick, além de permitir utilizar as mãos para fazer outras coisas em quanto estivesse jogando.

1.3 Aplicações do reconhecimento automático de fala

As aplicações para sistemas com reconhecimento de voz, podem ser aplicados em qualquer interação homem-máquina, e nas mais diversas áreas. As áreas mais comuns são: (MARTINS, 1997a)

- Sistemas de controle e comando: Estes sistemas utilizam a fala para realizar determinadas funções;
- Sistemas de telefonia: O usuário pode utilizar a voz para fazer uma chamada, ao invés de discar o número;
- Sistemas de transcrição: Textos falados pelo usuário podem ser transcritos automaticamente por estes sistemas;
- Acesso à informação: O usuário recebe algum tipo de informação, que se encontra armazenada em um banco de dados. Exemplo: notícias, previsão do tempo, hora certa, etc.
- Centrais de atendimento ao cliente: Uma atendente virtual pode ser utilizada a fim de realizar o atendimento ao cliente;

- Operações bancárias: O usuário efetua operações bancárias, como informações do seu saldo, trasnferências de dinheiro.
- Preenchimento de formulários: O usuário entra com os dados via fala.
- Robótica: Robôs podem se comunicar pela fala com seus donos.

1.4 Visão Geral do Trabalho

2 Referencial teórico

Este capítulo descreve os principais elementos teóricos utilizados no desenvolvimento desta pesquisa. As seções 2.3 e **??** resumem os principais conceitos sobre redes neurais, assim como as principais técnicas utilizadas nesta área. A seção 2.2.1 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 4.1. E por fim, alguns conceitos chave sobre imagens digitais são apresentados na seção 4.2.

2.1 Fala

A fala é a forma de comunicação mais utilizada pelos seres humanos.(SILVA, 2008) Através da fala, o cérebro humano consegue interpretar informações extremamente complexas, tais como identificar a pessoa que está falando, sua posição no espaço físico, seu estado emocional e outros dados como a ironia, seriedade ou tristeza. Os computadores, apesar de fazerem cálculos mais rápidos que o homem, não conseguem reconhecer através da fala informações como os seres humanos.

2.1.1 Vantagens da comunicação pela fala em sistemas homém-máquina

Segundo (FURUI, 1989) podemos citar:

- Natural: Não precisa de treinamento especial e nem de habilidades especiais;
- Rapidez: A informação é transmitida mais rapidamente que pelas outras formas de comunicação.
- Flexível: Deixa as mãos, olhos livres;

- Eficiente: Tem uma elevada taxa de informação;

2.1.2 Desvantagens no uso da fala em sistemas homem-máquina

Mesmo possuindo vantagens significativas, a comunicação por fala também possui desvantagens, como (FURUI, 1989) descreveu:

- Ruidos: O sistema fica suscetível a interferência do ambiente, necessitando de um removedor de ruídos para ambientes com alto índice de ruídos.
- Diversidade da língua: Características que variam de pessoa para pessoa, como sotaque, velocidade da fala, condições físicas e emocionais do locutor.

2.2 O Sistema de Reconhecimento de Voz

Sistemas de reconhecimento automático de voz, tem como objetivo, transformar um sinal analógico(fala) obtido através de um transdutor, mapeando-o a fim de produzir como saída a palavra, uma sequencia de fonemas ou uma sentenças correspondentes ao sinal de entrada. Com o resultado da tradução, pode-se tomar decisões, traduzir para outra língua, etc. Reconhecedores de voz, podem ser divididos em três grandes classes: reconhecimento por comparação de padrões, reconhecimento baseado na análise acústico-fonética e reconhecimento empregando inteligência artificial.(RABINER, 1993) No reconhecimento por comparação de padrões, existem duas formas distintas: treinamento e reconhecimento. Na fase de treinamento, são apresentados padrões ao sistema para criação de representantes, para cada um dos padrões. A fase de reconhecimento compara um padrão ainda desconhecido, com os padrões existentes no sistema, o que mais se aproximar do padrão existente, é escolhido como o padrão reconhecido. A fase de treinamento é fundamental para o sucesso do sistema, portanto uma quantidade considerável de material será necessário para a fase de treinamento. Sistemas com Modelos Ocultos de Markov (HMM) utilizam essa classe de reconhecimento.(MARTINS, 1997a) Nos sistemas com reconhecimento baseado na análise acústico-fonema, o sinal de fala é decodificado baseado em suas características acústicas e nas relações entre essas características. (INCE, 1992) É identificada as unidades fonéticas da fala a ser reconhecida, e concatenando essas unidades é reconhecida a palavra. Nessa análise é necessário considerar as propriedades invariantes da fala. Segundo (MARTINS, 1997a) Um analisador acústico-fonética apresenta as

sequintes fases: análise espectral, detecção das características que descrevem as unidades fonéticas, a fase mais importante de todo o processo que é: segmentação do sinal de fala e identificação das unidades fonéticas e escolha da palavra que melhor corresponde a sequência de unidades. Reconhecimento empregando inteligência artificial explora os conceitos tanto do reconhecimento por padrões quanto o baseado em análise acústico-fonema. (RABINER, 1993) Utilizando redes neurais, cria-se uma matriz de ponderações que representa os nós das redes, e suas saídas, estão relacionadas as unidades a serem reconhecidas. (MARTINS, 1997a) O processo para o reconhecimento de voz pode ser dividio em quatro fases: aquisição do sinal de voz, pré-processamento, extração de informações e geração dos padrões de voz. (SILVA, 2009)

2.2.1 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) As redes neurais também surgiram nos anos 50, mas não houve prosseguimento nos estudos, devido a problemas práticos. Muitos reconhecedores de voz, foram criados nas décadas de 50 e 60.(FURUI, 1995) No início dos anos 70, surgiram os algoritmos para sistemas de fala contínua, graças as técnicas de *Linear Predictive Coding* (LPC) e *Dynamic Time Warping* (DTW). (RABINER, 1993) E os anos 80 foram marcados pela disseminação dos metodos estáticos, como *M*odelos Ocultos de Markov (HMM). (RABINER, 1993) Esse período foi de grande evolução para os sistemas de reconhecimento de voz, as redes neurais passaram a ser usadas no desenvolvimento dos sistemas, sendo possível implementar sistemas mais robustos, com vocabulários grandes e com taxas de acerto de mais de 90%.(MARTINS, 1997a)

2.2.2 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, 1997b) Essas categorias que definem a precisão do sistema de reconhecimento.

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.

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)

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.

Sistemas RAV com suporte a grandes vocabulários são chamados de Large Vocabulary Continuos Speech Recognition (LVCSR). Existem muitas dificuldades encontradas na criação de sistemas LVCSR, como: a disponibilidade de um corpus de voz digitalizada e transcrita grande o suficiente para treinamento do sistema, recursos como bases de textos de tamanho elevado e um dicionario fonético de amplo vocabulário.(SILVA, 2008)

Perplexidade

Relação sinal – ruído

2.3 Reconhecimento de voz baseado em padrões

De acordo com (MARTINS, 1997a), o reconhecimento baseado em padrões, é a técnica que oferece melhor resultado nos sistemas de reconhecimento de fala, então a implementação do sistema será usando essa técnica. Um sistema de reconhecimento de voz usando reconhecimento de padrões poder ser representado na figura X (RABINER, 1993):

2.3.1 Fatores que Interferem no Desempenho

3 Modelos Ocultos de Markov

4 Resultado

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.

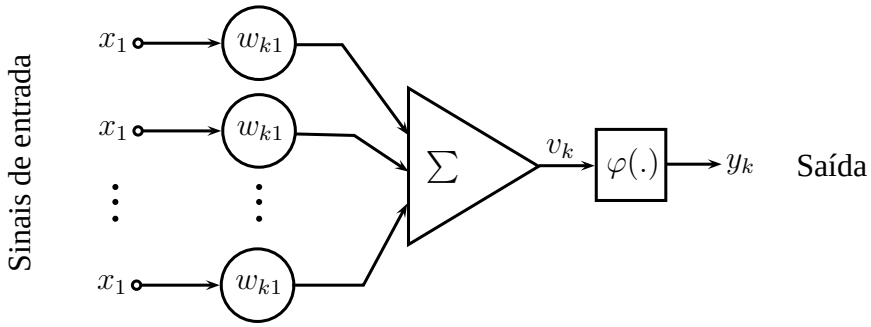

Figura 1: Texto da figura

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 4.1:

$$d_i(t) = \sqrt{\sum_{j=1}^{N} (x_j(t) - w_{ij}(t))^2} \tag{4.1}$$

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

4.1 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 4.2:

$$m_{pq} = \iint x^p y^q f(x, y) \mathrm{d}x\mathrm{d}y, p, q \in \tag{4.2}$$

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

$$m_{pq} = \sum_{x,y} x^p y^q f(x,y), p, q \in \tag{4.3}$$

A massa total da função $f(x,y)$ é determinado pelo momento m_{00}, conforme a equação 4.4:

$$m_{pq} = \sum_{x,y} f(x,y), p, q \in \tag{4.4}$$

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 4.5:

$$\bar{x} = \frac{1}{m_{00}} \sum x f(x,y) = \frac{m_{10}}{m_{00}} \tag{4.5a}$$

$$\bar{y} = \frac{1}{m_{00}} \sum y f(x,y) = \frac{m_{01}}{m_{00}} \tag{4.5b}$$

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

$$\mu_{pq} = \sum_{x,y} (x - \bar{x})^p (y - \bar{y})^q f(x,y) \tag{4.6}$$

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 4.7:

$$\gamma = 1 + \frac{p+q}{2} \tag{4.7a}$$

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{4.7b}$$

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:

$$\varphi_1 = \eta_{20} + \eta_{02} \tag{4.8a}$$

$$\varphi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{4.8b}$$

$$\varphi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{4.8c}$$

$$\varphi_4 = (\eta_{30} + \eta_{12})^2 + (3\eta_{21} + \eta_{03})^2 \tag{4.8d}$$

$$\varphi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] \tag{4.8e}$$

$$+(3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{4.8f}$$

$$\varphi_6 = (\eta_{20} - \eta_{02}) \left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{4.8g}$$

$$+4\eta_{11}(\eta_{30} - \eta_{12})(\eta_{21} + \eta_{03}) \tag{4.8h}$$

$$\varphi_7 = (3\eta_{21} - \eta_{30})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] \tag{4.8i}$$

$$+(3\eta_{12} - \eta_{03})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{4.8j}$$

4.2 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.

4.2.1 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*).

4.2.2 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.

FURUI, S. *Digital Speech Processing, Synthesis, and Recognition. Marcel Dekker*. Monografia, 1989.

FURUI, S. *Speech Recognition - Past, Present and Future. NTT Review*. Monografia, 1995.

INCE, A. N. *Digital speech processing: speech coding, synthesis, and recognition*. [S.l.]: Kluwer Academic Publishers, 1992. ISBN 0-7923-9220-5.

MARTINS, J. A. *Avaliação de diferentes técnicas para reconhecimento de fala*. Monografia (Doutorado), Campinas, 1997.

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

RABINER, L. R. *Fundamentals of speech recognition*. [S.l.]: PTR Prentice Hall, 1993.

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

SILVA, C. P. A. da. *Sistemas de Reconhecimento de Voz para o Português brasileiro utilizando os Corpora Spoltech e OGI-22*. Monografia (Graduação), Belém, 2008.

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