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 2.1 e 2.2 resumem os principais conceitos sobre redes neurais, assim como as principais técnicas utilizadas nesta área. A seção 2.3 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 2.4. E por fim, alguns conceitos chave sobre imagens digitais são apresentados na seção 2.5.

2.1 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.

2.1.1 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.

2.1.2 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)

2.1.3 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.

2.1.4 Perplexidade

2.1.5 Relação sinal – ruído

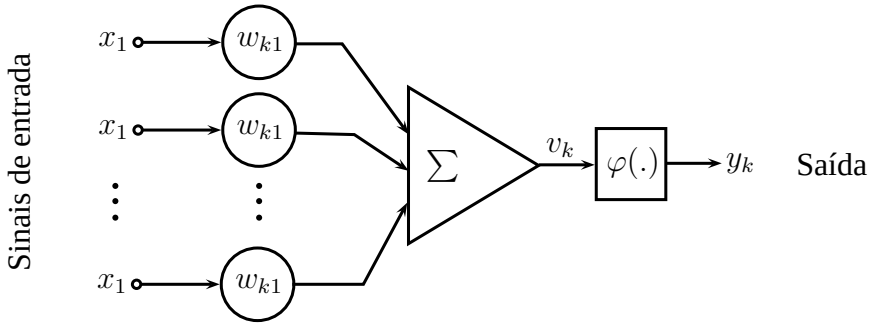

Figura 1: Texto da figura

2.2 Principais tipos de redes neurais

Dentre a grande variedade de tipos de redes neurais artificiais, as mais importantes, seja por sua contribuição teórica ou pela praticidade, são:

2.2.1 Rede perceptron multicamada

A perceptron multicamada é (a) uma rede direta, isto é, o sinal passa da entrada até a saída sem ciclos; (b) possui camadas intermediárias de neurônios entre as camadas de entrada e saída, ditas camadas ocultas; (c) utiliza funções de ativação não lineares, comumente a função sigmoide e (d) os neurônios são altamente conectados, em geral cada neurônio é conectado a todos os neurônios da camada anterior e da camada seguinte.

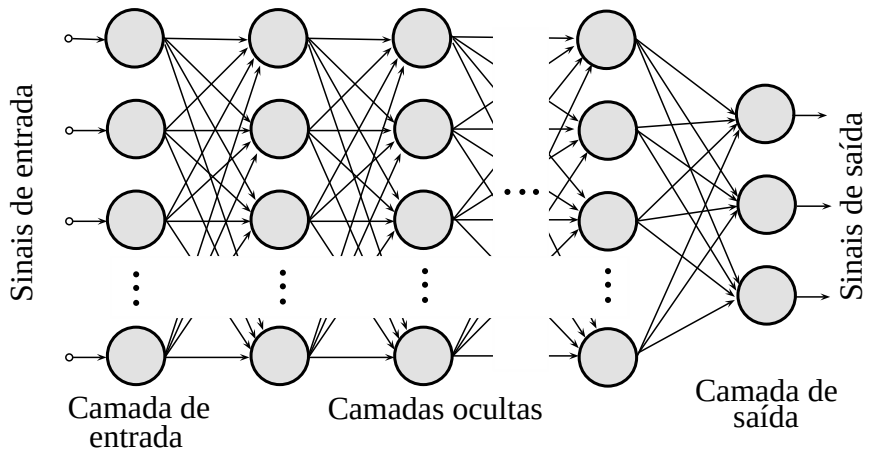

Figura 2: Texto da figura

A presença de múltiplas camadas permite a este tipo de rede resolver uma enorme variedade de problemas, ou em outros termos, reconhecer uma vasta variedade de padrões. As redes multicamadas são computacionalmente completas, ou seja, são equivalentes à classe das máquinas de Turing.

A retropropagação é o algoritmo de treinamento mais utilizado para esta variante de rede neural. Cada iteração deste algoritmo é dividido em dois passos, no primeiro a rede é alimentada com um dos exemplos, o resultado é capturado e comparado com o valor esperado, com isso o erro geral da rede é calculado; segue-se então ao segundo passo, que atualiza os pesos sinápticos penalizando cada neurônio segundo sua influência no erro geral, essa etapa é feita da camada de saída para a de entrada, retroativamente.

2.2.2 Rede de Hopfield

A rede de Hofield é (a) uma forma de rede neural recorrente, isto é, determinadas conexões realimentam alguns neurônios, formando ciclos na rede; (b) apresenta um atraso temporal, ou seja, a propagação dos sinais não é instantânea e (c) a saída é um estado de convergência, isto é, após se apresentar uma entrada a rede opera em ciclos até que a saída não mude mais, situação onde se diz que a rede alcançou o equilíbrio.

ESQUEMA DE UMA REDE DE HOPFIELD AQUI

A rede de Hopfield funciona como uma memória endereçada por conteúdo, ou memória associativa, por exemplo, muitas vezes lembramos de fatos inteiros apenas com uma pequena lembrança do acontecimento. Uma memória associativa é, deste modo, um conjunto de padrões armazenados de tal modo que, quando se apresenta um novo padrão, a resposta será o padrão armazenado que seja mais parecido a este que foi apresentado.

Geralmente as redes de Hopfield não possuiem um método de aprendizado associado, os pesos sinápticos são calculados por métodos matemáticos provenientes de sua definição formal. A definição formal garante que a rede sempre irá convergir, contudo, em algumas situações esta convergência pode não ocorrer para o padrão mais próximo da entrada, e ainda não existe um método conhecido que resolva este problema.

2.2.3 Redes de Kohonen

As redes de Kohonen apresentam apenas duas camadas de neurônios, a camada de entrada e a de saída. A camada de saída é uma espécie de malha de neurônios não conectados

entre si, mas amplamente conectados com os neurônios da camada de entrada. Esta malha funciona como um mapa, onde para cada padrão de entrada apenas um neurônio é ativado, padrões semelhantes ativam neurônios dentro de uma mesma região da malha.

ESQUEMA DE UMA REDE DE KOHONEN AQUI

As redes de Kohonen possuem um algoritmo próprio de treinamento, dividido em três etapas; na primeira, chamado de processo competitivo, uma determinada entrada ativa apenas um neurônio da malha; na segunda, chamado de processo cooperativo, o neurônio escolhido estabelece uma vizinhança de neurônios que serão ajustados para, junto com ele, identificar padrões semelhantes ao que foi apresentado; e por fim, na terceira etapa, chamada de processo adaptativo, os pesos são atualizados com base no neurônio vencedor e na vizinhança topológica. Este algoritmo de treinamento é dito não supervisionado, pois não depende de um par *(entrada, saída esperada)*, já que a própria rede estabelece como será a configuração dos resultados.

2.3 Redes neurais de Kohonen

Esta seção irá apresentar mais detalhadamente como é a configuração de uma rede de Kohonen, seu algoritmo de treinamento e os usos comuns deste tipo de rede.

2.3.1 Topologia de uma rede de Kohonen

Como dito anteriormente, a rede de Kohonen apresenta apenas duas camadas de neurônios, a camada de entrada e a camada de saída. A camada de entrada deve possuir tantos neurônios quanto forem à quantidade de elementos do padrão de entrada. A camada de saída é uma grade de geometria livre, geralmente retangular, de neurônios que não estão ligados entre si, mas estão, cada um, ligados a todos os neurônios da camada de entrada. As conexões apresentam pesos para escalar o sinal enviado.

ESQUEMA DE UMA CONEXÃO DA REDE DE KOHONEN AQUI

2.3.2 Treinamento da rede

O treinamento requer que os pesos sinápticos sejam iniciados com valores bem pequenos, para que a rede não apresente inicialmente nenhuma configuração. Três processos são executados para cada entrada do conjunto de treinamento, o processo competitivo, o

processo cooperativo e o processo adaptativo.

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.

Processo cooperativo

Estudos biológicos indicam que ao ser excitado, um neurônio estimula seus vizinhos topológicos, de forma que quanto mais próximo um neurônio está do neurônio ativo, mais excitado pelo estímulo do neurônio ativo ele é. O processo cooperativo busca simular este mecanismo biológico.

Em termos matemáticos, o que se deseja é um parâmetro h_{ik} , dito *vizinhança topológica*, que indica o gral de cooperação entre o neurônio vencedor i e o seu vizinho k,

que deve ser simétrico em relação ao neurônio k e deve decrescer constantemente com o aumento da distância l_{ik} , até que $\lim_{l_{ik}\to\infty} h_{ik} = 0$. A função gaussiana 2.2 atende a estas duas exigências:

$$h_{ik} = e^{\left(\frac{l_{ik}^2}{2\sigma^2}\right)} \tag{2.2}$$

O parâmetro σ é denominado *largura efetiva da vizinhança*, e deve diminuir a cada iteração, indicando uma tendência de especialização da rede. Neste trabalho o parâmetro σ é a equação 2.3:

$$\sigma(t) = \sigma_0 e^{t/\tau_l} \tag{2.3}$$

Onde:

- σ_0 é o valor inicial de σ;
- t é a iteração atual;
- τ_l é uma constante de tempo.

Processo adaptativo

O processo adaptativo atualiza os pesos sinápticos a cada iteração, levando em consideração o neurônio vencedor e a vizinhança topológica. O ajuste dos pesos deve decrescer com o tempo, para evitar que novos dados comprometam seriamente o conhecimento já adquirido, substituindo padrões já estabelecidos por novos. Algo semelhante ocorre com o cérebro humano, ao decorrer do envelhecimento o aprendizado vai se tornando mais difícil.

O ajuste Δw_{ij} que a sinapse entre o neurônio de entrada i e um neurônio da malha j deve sofrer é expresso pela equação 2.4:

$$\Delta w_{ij} = \eta(t) h_{ik}(t)(x_j - w_{ij}) \tag{2.4}$$

Onde $h_{ik}(t)$ é o parâmetro vizinhança topológica na iteração t , referente ao neurônio

vencedor k . O parâmetro *taxa de aprendizagem* $\eta(t)$ é definido pela expressão 2.5:

$$\eta(t) = \eta_0 e^{t/\tau_l}, \eta_0 \in [0, 1] \tag{2.5}$$

Onde τ_l é uma constante de tempo.

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

2.4 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 2.6:

$$m_{pq} = \iint x^p y^q f(x, y) \mathrm{d}x\mathrm{d}y, p, q \in \tag{2.6}$$

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

$$m_{pq} = \sum_{x,y} x^p y^q f(x,y), p, q \in \tag{2.7}$$

A massa total da função $f(x,y)$ é determinado pelo momento m_{00}, conforme a equação 2.8:

$$m_{pq} = \sum_{x,y} f(x,y), p, q \in \tag{2.8}$$

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 2.9:

$$\bar{x} = \frac{1}{m_{00}} \sum x f(x,y) = \frac{m_{10}}{m_{00}} \tag{2.9a}$$

$$\bar{y} = \frac{1}{m_{00}} \sum y f(x,y) = \frac{m_{01}}{m_{00}} \tag{2.9b}$$

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

$$\mu_{pq} = \sum_{x,y} (x - \bar{x})^p (y - \bar{y})^q f(x,y) \tag{2.10}$$

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 2.11:

$$\gamma = 1 + \frac{p+q}{2} \tag{2.11a}$$

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{2.11b}$$

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:

$$\varphi_1 = \eta_{20} + \eta_{02} \tag{2.12a}$$

$$\varphi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{2.12b}$$

$$\varphi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{2.12c}$$

$$\varphi_4 = (\eta_{30} + \eta_{12})^2 + (3\eta_{21} + \eta_{03})^2 \tag{2.12d}$$

$$\varphi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] \tag{2.12e}$$

$$+(3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{2.12f}$$

$$\varphi_6 = (\eta_{20} - \eta_{02}) \left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{2.12g}$$

$$+4\eta_{11}(\eta_{30} - \eta_{12})(\eta_{21} + \eta_{03}) \tag{2.12h}$$

$$\varphi_7 = (3\eta_{21} - \eta_{30})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] \tag{2.12i}$$

$$+(3\eta_{12} - \eta_{03})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{2.12j}$$

2.5 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.

2.5.1 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*).

2.5.2 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

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.