Raphael Ferreira Ramos

Jogos fisicamente interativos baseados em reconhecimento de voz

Campos dos Goytacazes/RJ

2012

Raphael Ferreira Ramos

Jogos fisicamente interativos baseados em reconhecimento de voz

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 do Prof^{o}. Luis Antônio Rivera Escriba, DrSc.

Tutor: Luis Antônio Rivera 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

Muitas pessoas me apoiaram em minha graduação e neste trabalho, então deixo aqui minhas homenagens.

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 (RAV)* 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 *RAV*, permitindo que o computador "entenda" o que está sendo dito (SILVA, 2008).

De acordo com Silva (2009) e Silva (2010) os sistemas *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, segurança etc. 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.

Com a tendência da melhora dos processadores, memórias e placas de vídeo, esta linha de processamento de voz está em bastante evidência. Os jogos de computadores, são uma área que também está acompanhando essa evolução, e também demandam reconhecimento de voz para controle de comandos. 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, que 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 foi criado o SpaceWar. Inicialmente a ideia de Russel era fazer um filme interativo,

mas acabou se tornando o pai dos jogos eletrônicos (RAMOS, 2007). Milhares de jogos foram desenvolvidos nas décadas seguintes, passando por tetrix do russo Alexey Pajitnov, super Mário, que foi o jogo mais vendido da época, até chegar nos games atuais, que surpreendem pelo realismo. Videogames com as mais modernas tecnologias vem sendo lançados ultimamente, um exemplo é o Xbox 360, fabricado pela Microsoft Corporation, que surpreendeu ao fazer o joystick que possui um sistema inteligente de profundidade de seus botões traseiros, similares a um gatilho. Com isso, os comandos são interpretados de acordo com a intensidade em que estes são pressionados. Em um game de corrida, por exemplo, faz uma enorme diferença na hora de acelerar mais suavemente com o seu carro, ou simplesmente "afundar" o pé no acelerador (BORGES, 2010). Mas a grande revolução ainda estava por vir, em novembro de 2010, a Microsoft lançou o kinect, um sensor de movimento que veio para revolucionar o mundo dos games, promovendo uma integração total com o jogador, e acabando com a mística de que jogar videogame é sinal de sedentarimo (BORGES, 2011).

1.1 Aplicações do reconhecimento automático de fala

Segundo Martins (1997a) os sistemas com reconhecimento de voz podem ser aplicados em qualquer atividade que demande interação homem-máquina, e nas mais diversas áreas. Há mais de uma década ele já mostrava a importância do uso de voz em diversas aplicações, algumas dessas áreas são:

- 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.

  Nos últimos anos, jogos de computador estão sendo produzidos para responderem a comandos de voz, como o *Tom Clancy's EndWar*, um jogo de estratégia que simula uma terceira guerra mundial, e sua equipe recebe seus comandos falados, simulando as ordens de um oficial.

1.2 Objetivos

O objetivo geral deste trabalho é desenvolver um jogo interativo guiado por comandos de voz ditados pelo usuário, o método abordado será testado em um clássico do mundo dos games, pacman [1]. 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.

1.3 Motivações

A maior motivação seria o 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, permitindo utilizar as mãos para fazer outras coisas em quanto estivesse jogando, além das diversas aplicações que são cada vez maiores nas atividades humanas.

1.4 Metodologia

Para o propósito desde trabalho se estabeleceu um estudo de técnicas de reconhecimento de voz de comandos, técnicas de análise de sons para extração de atributos que serão alimentados como treinamento para um modelo de grafos de *Cadeia de Markov Oculto*,

[1] O jogador é uma cabeça redonda com uma boca que se abre e fecha, posicionado em um labirinto simples repleto de pastilhas e 4 fantasmas que o perseguem. O objetivo era comer todas as pastilhas sem ser alcançado pelos fantasmas, em ritmo progressivo de dificuldade.

implementação dos diferentes modelos, acoplação no jogo clássico *pacman* e verificação dos resultados.

A implementação será desenvolvida para dispositivos móveis com sistema operacional Android [2], que utiliza como linguagem de programação a linguagem Java.

1.5 Visão Geral do Trabalho

Neste trabalhou buscou-se desenvolver um sistema *RAV* com baixa taxa de erros, um dicionário pequeno de palavras e reconhecimento de palavras isoladas, que são a melhor forma de introdução nos estudos de sistemas com reconhecimento de fala, possibilitando estudos futuros em aplicações com reconhecimento de palavras contínuas e vocabulários grandes. O sistema desenvolvido é baseado nos modelos ocultos de Markov *(HMM)*.

Este trabalho está dividido em capítulos, que são descritos a seguir:

O capítulo 2 tem como objetivo fazer as referências teóricas sobre o processamento do sinal de fala, sistemas *RAV*, como suas características, histórico, reconhecimento de padrões e alguns trabalhos relacionados.

O capítulo 3 visa mostrar toda estrutura empregada no desenvolvimento do sistema RAV, como aquisição da fala, que são as formas para captura do som, pré-processamento que é a filtragem do sinal capturado, extração do parêmtros necessários e o treinamento dos padrões para comparações posteriores.

[2] Sistema operacional móvel que roda sobre o núcleo Linux. Foi inicialmente desenvolvido pelo Google e posteriormente pela Open Handset Alliance.

2 Reconhecimento Automático de Voz

Sistemas de reconhecimento automático de voz, tem como objetivo, transformar um sinal analógico(fala) obtido através de um transdutor [1], 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. Geralmente sistemas de reconhecimento de fala são divididos em 4 fases, que serão detalhadas mais a frente neste trabalho, fases essas de: aquisição de voz, pré-processamento, extração de informações e geração de padrões.

A primeira etapa é a *aquisição de voz*, processo pelo qual ondas sonoras são convertidas em sinais elétricos, e depois é feita uma conversão analógico-digital do mesmo. Algumas características do ambiente de gravação podem atrapalhar no processo de reconhecimento, como ruídos, distância do microfone, etc. Assim é preciso passar por uma fase que é feita uma filtragem afim de tornar o sinal o mais próximo possível da fala pura, o nome desta etapa é *pré-processamento*. Logo depois, é feita a *extração de informações* do sinal, que consiste em representar segmentos, fonemas ou qualquer outra unidade de fala com o menor número possível de parâmetros, de forma que estes contenham informações suficientes para caracterizar o sinal de fala, já que um sinal digital possui uma grande quatidade de parâmetros, então a exigência por tempo e processamento seriam muito altas. Outra etapa, e uma das mais importantes é a *geração de padrões*, também chamada de padrões de referência, que é realizada durante o treinamento do sistema, nesta fase é gravado o maior número possível de padrões, para cada palavra cadastrada no dicionário, para garantir o melhor funcionamento e tempo de resposta do reconhecedor. (SILVA, 2009), (LOUZADA, 2010).

[1] Dispositivo que transforma um tipo de energia em outro, utilizando para isso um elemento sensor

2.1 Fala

Segundo Silva (2008) a fala é a forma de comunicação mais utilizada pelos seres humanos. 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.

Vantagens da comunicação pela fala em sistemas homém-máquina

Segundo Furui (1989) podemos citar:

- Naturalidade: Não precisa de treinamento especial e nem de habilidades especiais;
- Rapidez: A informação é transmitida mais rapidamente que pelas outras formas de comunicação.
- Flexibilidade: Deixa as mãos, olhos livres;
- Eficiência: Tem uma elevada taxa de informação;

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.

2.1.1 Aquisição do sinal de fala

Aquisição do sinal de fala é a primeira etapa de um sistema *RAV*, ele é responsável por capturar e converter um sinal analógico em um sinal elétrico, esse processo pode ser feito através de um microfone ou telefone. Todas as etapas de aquisição de voz, podem ser vistas na figura 1 (SILVA, 2009):

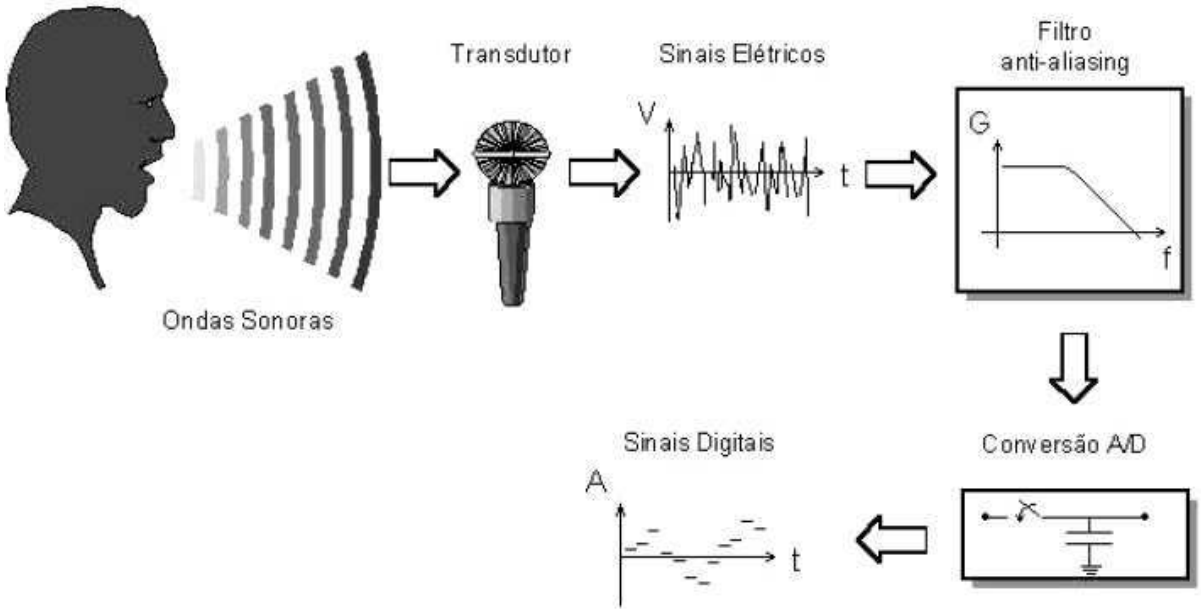

Figura 1: Processo de aquisição do sinal da fala

2.2 O Sistema de Reconhecimento de Fala

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). Rabiner (1993) mostra que 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). E os anos 80 foram marcados pela disseminação dos metodos estáticos, como *M*odelos Ocultos de Markov (HMM). 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).

Rabiner (1993) classifica os reconhecedores de voz em três grandes classes: *reconhecimento por comparação de padrões*, *reconhecimento baseado na análise acústico-fonética* e *reconhecimento empregando inteligência artificial*. 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. 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). Como dito anteriormente o processo para o reconhecimento de fala pode ser dividido em quatro fases: aquisição do sinal de voz, pré-processamento, extração de informações e geração dos padrões de voz, que podem ser vistas na figura 2, que além de mostrar essas etapas, também é ilustrado a fase de reconhecimento (SILVA, 2009).

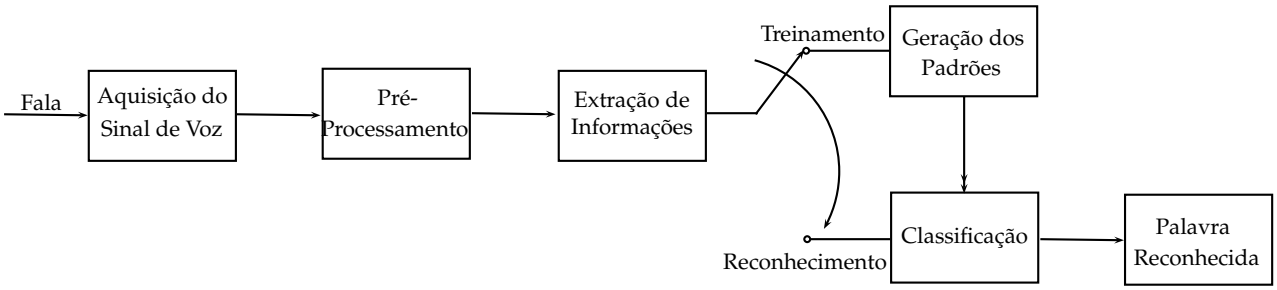

Figura 2: Diagrama de blocos de um sistema de reconhecimento de voz

2.2.1 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, o tamanho do vocabulário, a dependência ou independencia do locutor, perplexidade e relação sinal-ruído (MARTINS,

1997b). As categorias que serão detalhadas a seguir definem a precisão do sistema de reconhecimento fala:

A) 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.

B) 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 segundo Silva (2010) alcançam taxa de menos de 2% de erro para dígitos de 0 à 10. 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”.

Neste trabalho é considerado que o reconhecedor de fala conectada é similar a um reconhecedor de fala contínua, diferente do que propõe Martins (1997a), onde ele define fala conectada como um padrão em que as palavras ditas fazem parte de um vocabulário restrito e faladas de forma contínua, além do reconhecimento ser feito usando padrões de referência para cada palavra. Já no reconhecedor de fala contínua

os padrões a serem reconhecidos são sentenças e frases, envolvendo o reconhecimento de unidades básicas como fones, difones e outros, implicando em uma segmentação do sinal de fala.

C) 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. 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.

D) Perplexidade:

Uma medida popular que mede a dificuldade da tarefa, combinando o tamanho do vocabulário e o modelo de linguagem, que pode ser basicamente definida como a média do número de palavras que pode seguir uma palavra depois que o modelo de linguagem for aplicado. Pode ser pequena sendo menor que 10 ou grande sendo maior que 100 (YNOGUTI, 1999). A perplexidade de um modelo de linguagem depende do domínio de discurso. Na tabela 1 tem-se um quadro comparativo para diversas aplicações (COLE, 1997) :

Tabela 1: Perplexidades típicas para vários domínios.

Domínio	Perplexidade
Radiologia	20
Medicina de emergência	60
Jornalismo	105
Fala geral	247

E) Relação sinal - ruído: Também chamado de *(SNR)*, do inglês Signal Noise Ratio, são problemas que podem prejudicar o desempenho do sistema, como: ruídos, ambiente, distorção acústica, diferentes microfones e outros.

2.3 Reconhecimento da fala 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 foi representando por Rabiner (1993) e pode ser visto na figura 3:

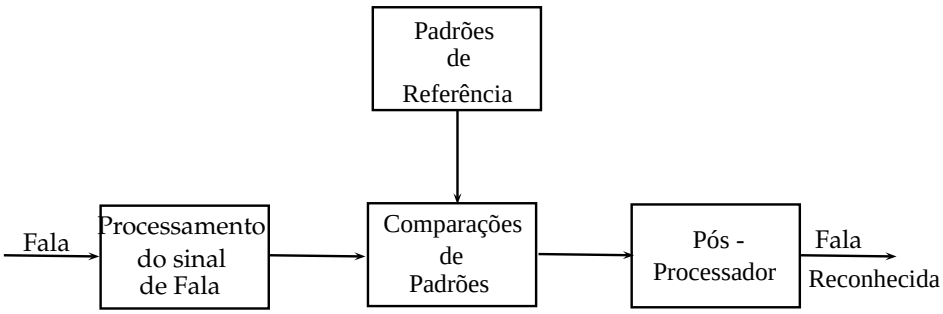

Figura 3: Sistema básico de reconhecimento de fala baseado na comparação de padrões

2.3.1 Processamento do Sinal de Fala

Nessa fase, o sinal analógico é digitalizado para ser comparado com os diferentes tipos de padrões, para essa comparação o sinal digital é convertido em um conjunto de parâmetros espectrais e temporáis. As comparações entre formas de ondas da fala são muito complicadas, e isso justifica o uso de parâmetros, como exemplo, podemos citar uma distorção de fase que é imperciptível ao ouvido humano, mas altera a forma da onda, dificultando as comparações de padrões (MARTINS, 1997a). Um grande número de parâmetros tem sido propostos, segundo Martins (1997a) os parâmetros mais usados são os: derivados dos coeficientes Linear Predictive Coding *(LPC)* e os *derivados diretamente do espectro do sinal* Como já citado, os reconhecedores de palavras isoladas, necessitam de capturar os pontos limitantes de cada palavra. Existem vários algoritmos de detecção desse início e fim das palavras, usando parâmetros como: energia e taxa de cruzamento de zero para separar o sinal de fala do ruído.

2.3.2 Padrões de Referência

Padrão de referência é o processo conhecido como treinamento, pois é nessa fase, que são criados os exemplares das unidades a serem reconhecidas. Como a maioria dos sistemas de reconhecimento de voz são reconhecedores independentes de locutor, são necessários a apresentação de vários exemplos de cada unidade, com a maior variedade de diferentes locutores possível, para criação de um sistema robusto.

Ince (1992) sugere dois tipos de padrão: Um tipo chamado modelo estático que faz um modelamento estático das características exemplares do padrão, *Modelos Ocultos de Markov* (HMM) são exemplos desse método. Outro tipo é conhecido como padrão de referência não paramêtrico, podendo ser um exemplo do padrão a ser reconhecido ou um padrão médio do padrão a ser reconhecido. Nos *Nos Modelos Ocultos de Markov* (HMM) cada padrão é representado por uma rede com N estados, que são caracterizados por uma função de probabilidade de transição entre estados e um conjunto de funções de probabilidade de símbolos de saída (MARTINS, 1997a).

2.3.3 Comparação de Padrões

A comparação de padrões, é a fase em que os dados são cruzados, o conjunto de parâmetros que representa o padrão desconhecido é comparado com os diversos padrões de referência, esses parâmetros são da mesma natureza que os padrões já referenciados. Nos padrões de referência gerados por *Modelos de Markov Ocultos*(HMM), a comparação resulta na probabilidade de que cada modelo de referência tenha gerado o conjunto de paramêtros de entrada (MARTINS, 1997a).

2.3.4 Pós-Processador

A última fase seria a escolha do melhor padrão referencial, resultado da comparação de padrões, para o padrão desconhecido. Martins (1997a) mostra que como auxílio na escolha do melhor padrão, pode-se usar restrições sintáticas e semânticas, eliminando os candidatos não razoáveis.

2.4 Avaliação de Desempenho de um Reconhecedor

Vários fatores interferem no desempenho de um reconhecedor de voz, segundo Martins (1997a) um reconhecedor de palavras isoladas pode ser avaliado com essas medidas:

- Porcentagem de acerto: Porcentagem de palavras que foram reconhecidas corretamente;
- Porcentagem de rejeição: Porcentagem de palavras que pertencem ao vocabulário e foram rejeitadas erradamente;
- Porcentagem de erro: Porcentagem de palavras que foram reconhecidas erradamente.

Já no caso de reconhecedores de fala contínua, as medidas seriam (MARTINS, 1997a):

- Porcentagem de inserção: Porcentagem de palavras extras inseridas na sentença reconhecida;
- Porcentagem de omissão: Porcentagem de palavras corretas omitidas na sentença reconhecida;
- Porcentagem de substituição: Porcentagem de corretas substituidas por palavras incorretas na sentença reconhecida.

2.5 Modelos Ocultos de Markov

Segundo Waghabi (2009) a idéia principal dos *Modelos Ocultos de Markov* (MOM) é representar uma variável aleatória como uma *Cadeia de Markov*, com a idéia que essa variável não pode ser medida diretamente, mas sim com observações possíveis a partir de cada estado da variável. Em Rabiner (1989) uma cadeia oculta de Markov é definida como uma máquina de estados finita que modifica seu estado a cada unidade de tempo. Podendo ser mais especificado como um modelo matemático formado por uma cadeia de estados conectados entre si, onde cada transição entre os estados possui uma probabilidade de ocorrência, além de cada estado está vinculado a um processo estocástico, que pode ser discreto ou contínuo, conhecido como processo de observação. Jurafsky e Martin (2006) em seu trabalho apresenta uma cadeia com 3 estados, mostradas na figura 4, que detalha a probabilidade de ocorrência de um estado de tempo tendo-se por base o estado atual.

Como exemplo é mostrado um dia de sol, as chances de um próximo dia ensolarado é de 50%, do próximo dia ser chuvoso as chances são de 10% e de um dia nublado 40%. As probabilidades para os outros estados são diferentes mas seguem a mesma idéia.

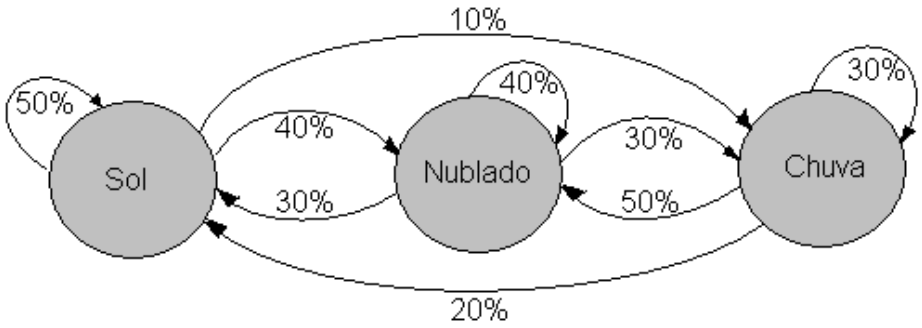

Figura 4: Cadeia de Markov para previsão do tempo

Waghabi (2009) explica o exemplo: "Imaginemos que um profissional que prevê o tempo não possa observar diretamente o clima. Trabalhando em um ambiente confinado, sua única dica de como está o tempo baseia-se nas vestimentas dos outros companheiros de trabalho. Como nem sempre seus companheiros conseguem prever corretamente o tempo, uma pessoa portando um guarda-chuva não significa certeza de chuva, mas tão somente uma probabilidade alta. Da mesma forma alguém de camiseta poderia ser surpreendido por um mau tempo, mas não tão frequentemente. Neste cenário ligeiramente estranho, a cadeia de estados do tempo está oculta ao meteorologista, e suas aferições têm que se basear em observações relacionadas com o tempo através de uma matriz de probabilidades". Como na figura 5:

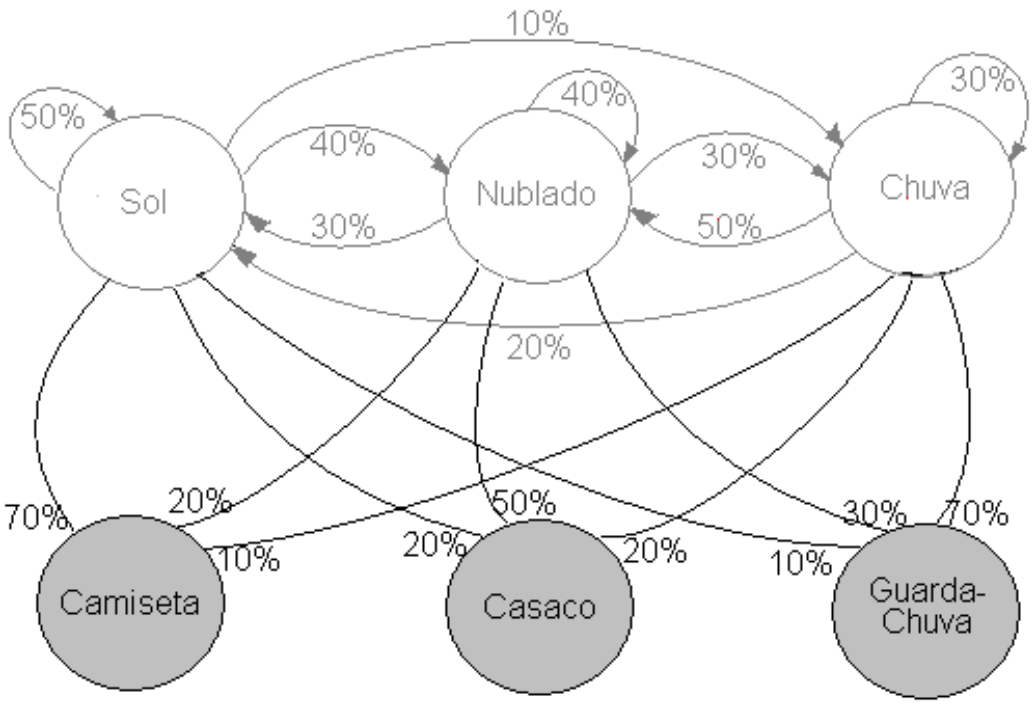

Figura 5: MOM para previsão do tempo

Com este modelo acima o meteorologista pode deduzir, observando o comportamento de seus companheiros durante um determinado tempo, qual foi a seqüência de estados ocultos(tempo) mais provável de produzir a seqüência de observações, obtendo assim uma expectativa de em qual estado o sistema está no momento atual, consequentemente permitindo a previsão do dia seguinte.

Um Modelo Oculto de Markov (MOM), é composto por (WAGHABI, 2009):

- Um conjunto $Q = \{q_i\}$ de estados ocultos do modelo, ou apenas estados.
- Uma matriz de probabilidades de transições $A = \{a_{ij}\}$ entre os estados q_i e q_j, onde $a_{ij} = [0, 1]$ com $i, j \in [1, |Q|]$, e $\sum a_{ij} = 1$ para um mesmo estado q_i.
- Um conjunto $O = o_n$ de estados observáveis, ou observações.
- Uma matriz de probabilidades $B = b_{in}$ indicando a chance do estado q_i produzir a observação $O = o_n$, onde $b_{in} = [0, 1]$ com $i \in [1, |Q|]$, $n \in [1, |O|]$.
- Uma distribuição $\Pi = \pi_i$ de probabilidades da modelo iniciar no estado q_i, onde $\pi_i = [0, 1]$ com $i \in [1, |Q|]$.

Segundo Rabiner (1989) existem 3 problemas que são básicos no desenvolvimento de sistemas RAV modelados por HMMs, o problema da avaliação, problema da decodificação

e o problema do treinamento. Dentre os algoritmos que podem resolver esses problemas encontrados: o algoritmo Forward ou o algoritmo Backward para o problema de seleção caso existam vários modelos, o algoritmo de Viterbi, que é próximo ao ótimo para o problema da decodificação ou reconhecimento e o algoritmo Baum-Welch para o problema de treinamento (SILVA, 2009).

2.6 Trabalhos Relacionados

Como essas áreas são muito importantes para os seres humanos, muitos estudosforam feitos no sentido de utilizar a fala pra realizações de ações. Como exemplo temos:

Barcelos (2007) desenvolve uma aplicação de reconhecimento de voz para aplicações em cadeira de rodas, onde o cadeirante se movimenta através de comandos de voz, como facilitador da implementação, foi utilizado o software IBM Via Voice, que segundo Damasceno (2005) obteve um melhor desempenho e aplicabilidade quando comparado a outros softwares, considerando a língua falada, a robustez do reconhecimento e a interface de trabalho com outros programas devido à aplicação deste desenvolvimento ser no Brasil.

Já em Rodrigues (2009), para efetuar o reconhecimento de voz foi utilizado redes neurais artificiais, também chamadas de *(RNA)*. Usando como base RNA foi criado uma rede para identificar comandos básicos de voz, e assim, efetuar o acionamento de um robô móvel. Outra característica importante no projeto é o identificador neural, que foi desenvolvido como dependente do locutor, onde um sistema é desenvolvido com base nas características vocais de um locutor. Para novos locutores seria necessário um novo treinamento da rede com as características vocais dos novos locutores.

3 Estruturas de um Sistema RAV

Os conceitos e definições necessárias para o desenvolvimento deste trabalho, foram apresentados nos capítulos anteriores, neste capítulo será apresentando toda a estrutura para implementação do sistema de reconhecimento de voz.

O sistema RAV proposto, foi implementado para um sistema independente do locutor, visando um jogo divertido para o maior número de pessoas possíveis, com o modo de pronúncia de palavras isoladas, e um vocabulário pequeno, que faz parte de um diciónario pré-definido.

Este sistema foi desenvolvido em Java, com suas Apis e direcionada para sistemas operacionais Android, tendo como base a teoria dos Modelos Ocultos de Markov para o modelamento de sequencias de frames. Então cada elocução é dividida em quadros de tempos com iguais durações, extraindo seus parâmetros de cada um deles para se criar os modelos Hmms para cada palavra do dicionário, como sistemas de reconhecimento de palavras isoladas necessitam da captura do início e fim das palavras pronunciadas, o usuário deve pronunciar um comando, e depois de um breve intervalo, pronunciar o próximo, os comandos disponíveis no dicionário da aplicação são: “Direita”, “Esquerda”, “Acima” e “Abaixo”. O personagem do jogo só responderá ao comando dito, depois de reconhecer qual é o comando falado, em caso de sucesso, o jogo continua normalmente, até a vitória ou derrota do jogador, em caso do não reconhecimento da palavra, o sistema ignora a palavra dita. Outra característica importante do reconhecedor é a tentativa de capturar o humor do jogador com palavras ofensivas gravadas no dicionário, o sistema apresenta uma penalidade em caso dessas palavras serem pronunciadas.

Este trabalho pode ser dividido em 6 etapas: Aquisição da fala, Pré-Processamento, Extração de Parâmetros, Criação de referências, Classificação e Execução dos comandos.

3.1 Arquitetura de um sistema RAV

O sistema RAV pode ser definido em 4 processos que são mostrados na figura 6, esses processos são compostos por outras etapas. O primeiro processo da figura 6 é chamado de *Front-End*, que engloba as etapas de captura do sinal da fala, conversão do sinal elétrico em sinal analógico-digital e o processo de filtragem. Outro processo ilustrado na figura 6 é o *Modelo Acústico* que é a fase responsável pelo treinamento das unidades a serem reconhecidas. O *Reconhecedor* é a parte que une todas os processos, é onde é feito as comparações entre o sinal falado e filtrado com os padrões de referência, que fazem parte da gramática adotada no sistema, o processo *Gramática* contém o dicionário de modelos da aplicação.

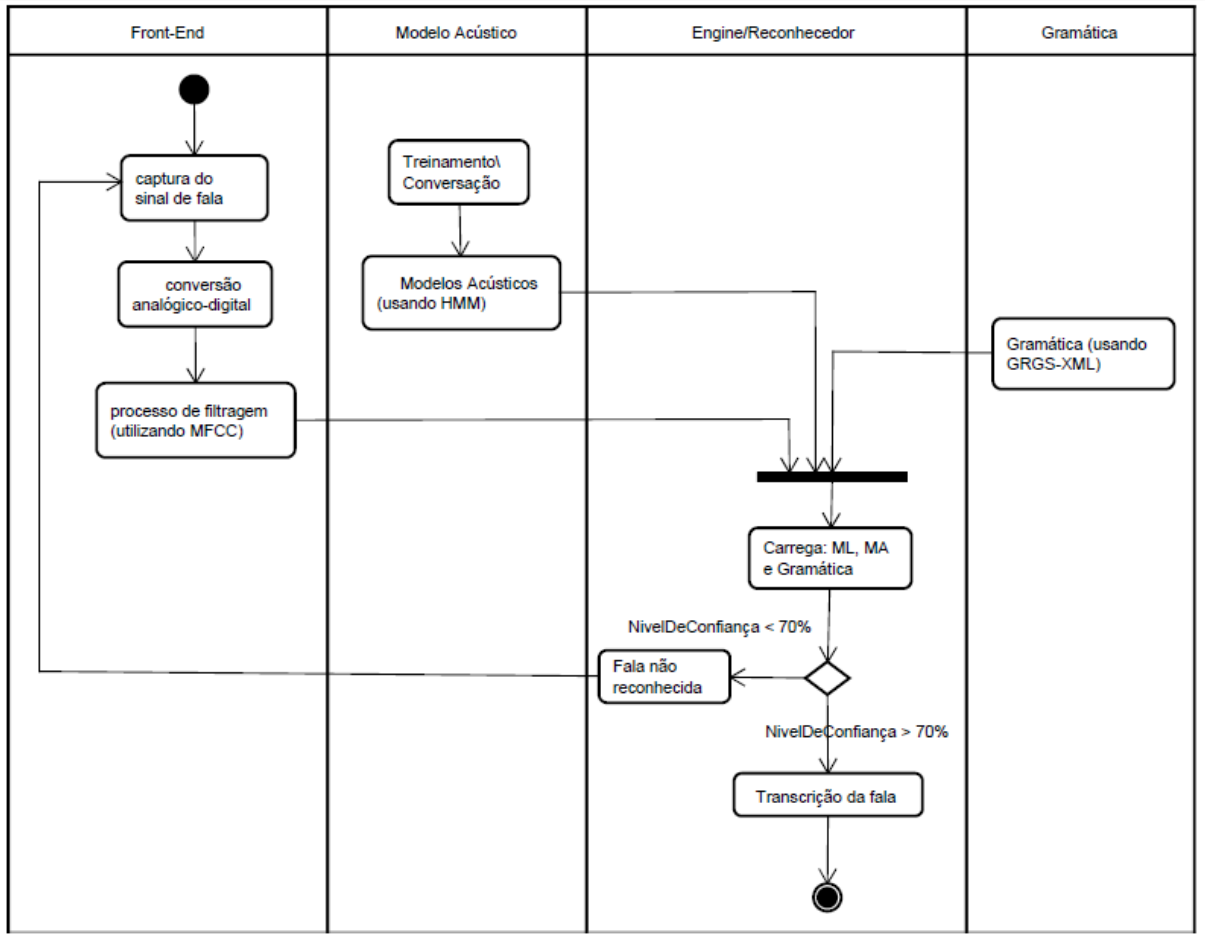

Figura 6: Diagrama de Atividades do Sistema

Com resultado da transcrição, são executados os comandos do jogo, no dispositivo móvel.

3.1.1 Front-End

Nesta seção serão mostrados os processos necessários para etapa *Front-End*, que são: *Aquisição de fala* junto com a *conversão analógico-digital* e o *Processo de filtragem* que também é chamado de pré-processo.

A) Aquisição da fala:

A interação do usuário com a aplicação é feita apenas com a voz, ao iniciar o sistema, a interface de voz com usuário é habilitada, permitindo ao usuário interagir com os comandos definidos no vocabulário. Como a aplicação é destinada a dispositivos móveis, a captura do som, é feita pelo microfone do celular ou tablet, fornecendo um sinal elétrico, sendo nessário uma filtragem do sinal analógico resultante por um filtro passa-baixas, chamado de anti-aliasing, para depois ser feita a conversão analógio-digital (RABINER, 1978). Esse filtro tem o intuito de suprimir componentes de frequência superiores à metade da frequência de amostragem, sendo chamado de Nyquist (PROAKIS, 1995).

A última etapa da aquisição da fala é a conversão do sinal de fala analógico em digital através de um amostrador, possibilitando o processamento digital. Segundo Chou (2003) é nesta fase que são escolhidas a taxa de amostragem, impossibilitando a ocorrência do efeito de aliasing e a precisão usada para a gravação do sinal, a partir do número de níveis que esse sinal poderá assumir. Todas etapas podem ser vistas na subseção 2.1.1 deste trabalho.

B) Pré-Processamento:

Sistemas RAV sofrem com características do ambiente de gravação e o canal de comunicação, como ruídos de alta frequência, distância do microfone, períodos de silêncio, etc. Uma forma de amenizar esse proplemas é fazer o sinal passar por um processo chamado de pré-processamento, deixando o sinal mais próximo da fala pura. As etapas desse processo podem ser mostradas na figura 7 (SILVA, 2009).

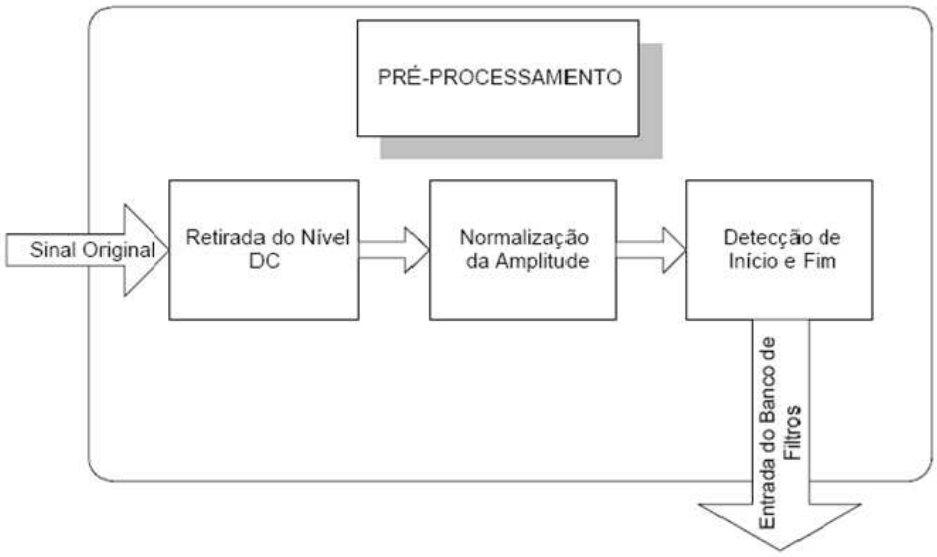

Figura 7: Diagrama de blocos da fase de pré-processamento

Calculando a média aritmétrica das amplitudes do sinal digital e subtraindo de cada aplitude esta média, consegue-se retirar o nível DC, que é uma componente contínua que atrapalha a comparação em valores absolutos. Outro problema encontrado é o diferencial entre sons mais baixos e sons mais altos, que é resolvido com a normalização da amplitude, esse pré-processamento do sinal faz com que todos os valores de amplitude de todos os sinais estejam na faixa de -1 e 1, garantindo que esses sinais sejam processados igualmente no algoritmo de reconhecimento. Esse processo é possível dividindo o valor de cada amostra do sinal pelo maior valor de aplitude do mesmo.

A última fase do pré-processamento no sistema RAV de palavras isoladas é a detecção do início e fim da locução, a fim de remover de forma precisa períodos de silêncio, que podem conter ruídos, sinais indesejados e a duração do sinal falado (SILVA, 2009). De acordo com Chu (2003) esse processo também tem como objetivo diminuir a carga computacional e economizar tempo, já que o sistema poderá processar apenas trechos que fazem parte da fala. O extremo inicial é determinado pelo primeiro quadro onde realmente se inicia a fala e o extremo final é determinado pelo último quadro que ainda há fala.

3.1.2 Modelo Acústico

Processo que define as unidades a serem modeladas, extração de parâmetros e modelos ocultos de markov se encaixam nessa fase do processo.

A) Extração de Parâmetros:

A extração de parâmetros é uma etapa de grande importância em um sistema RAV, pois o sinal digital possui uma grande quantidade de dados e uma análise direta necessitaria tempo e processamento consideráveis e ainda sim, não apresentariam um resultado expressivo. Muitas informações existentes no sinal digital puro não possuem significância alguma para a distinção fonética, assim o classificador empregado dificilmente conseguirá diferenciar amostras de palavras distintas.

No trabalho de Silva (2009) é mostrado a idéia básica da extração de parâmetros, que é representar segmentos, fonemas ou qualquer outra unidade de fala com o menor número possível de parâmetros, com informações necessárias para caracterizar o sinal de fala. Por melhor que seja o classificador, este só apresentará bons resultados se os parâmetros utilizados durante o treinamento ou reconhecimento contiverem informações relevantes. Uma redução no volume de dados mantendo informações suficientes para a caracterização do sinal viabilizará uma classificação robusta e confiável.

Algumas técnicas de análise espectral são mostradas por Rabiner (1978) e são usadas para obter os parâmetros do sinal digital, elas são: a transformada rápida de Fourier (Fast Fourier Transform ou FFT), os métodos de banco de filtros (Filter Bank), os de análise homomórfica ou análise cepstral (mel-cepstrum) e os de codificação por predição linear (Linear Predictive Coding ou LPC).

A técnica FFT, os métodos de banco de filtros e o LPC foram muito utilizados para a análise espectral da fala, no entanto, elas possuem algumas restrições, por isso Deller (1993) propõe o uso da técnica *mel-cepstrum*, cujo os coeficientes mel-cepstrais (Mel-Frequency Cepstral Coefficients ou MFCC), são obtidos pela representação em freqüência na escala *Mel*, a que considera a técnica mais apropriada para ser utilizada no processo de reconhecimento de voz. Com vantagens no uso dessa técnica, atualmente os coeficientes MFCC são os mais populares (BOUROUBA, 2007).

B) Treinamento dos Modelos Ocultos de Markov:

A fase de treinamento é uma das etapas de maior importancia em um sistema de reconhecimento de voz independente do locutor e ser o fator determinante na obtenção de um sistema com bons resultados ou não. É o momento em que são definidos os modelos HMMs para cada palavra do vocabulário utilizado.

A definição da quantidade de estados necessários para modelar uma palavra e o número de misturas por estado não são definidos por uma regra, mas sim, dependente

da familiaridade com os modelos HMMs ou por intuição, além de serem necessários muitos testes para obtenção do melhor resultado.

3.1.3 Reconhecedor

Pode ser considerado a central do sistema RAV, é a etapa que une todas os outros processos e define o resultado de sucesso ou fracasso do reconhecimento.

A) Reconhecimento:

Como citado em Silva (2009) a fase de reconhecimento consiste em dada uma elocução, descobrir qual o modelo que tem a maior probabilidade de gerá-la. Nesta etapa é necessário apenas o pré-processamento e da extração de características que gera a sequência de observações a ser utilizada no algoritmo de reconhecimento. O procedimento de reconhecimento pode ser visualizado na figura baseada em (RABINER, 1989).

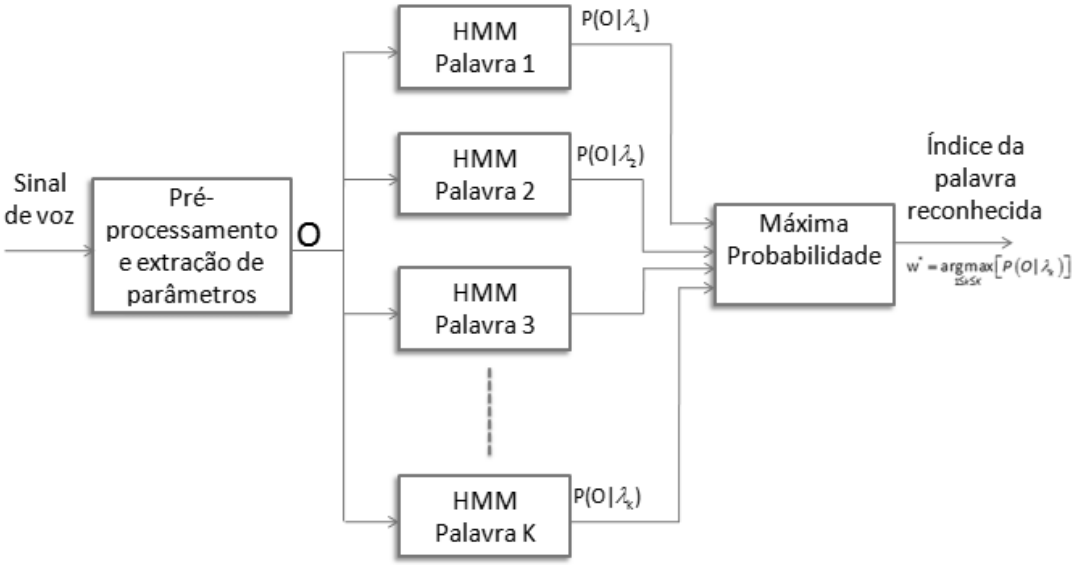

Figura 8: Procedimento de reconhecimento

O algoritmo *forward* pode ser usado para o reconhecimento a fim de se determinar a probabilidade de cada modelo de palavra gerar uma dada elocução. O modelo com maior probabilidade é o escolhido como correspondente à palavra falada, assumindo que todas as palavras têm uma mesma probabilidade de ocorrência. O algoritmo de *Viterbi* também pode ser utilizado para esta classificação, embora este resulte apenas em uma aproximação para a probabilidade de uma dada sequência de observações.

Os resultados utilizando-se um ou outro algoritmo são praticamente idênticos (SILVA, 2009).

3.1.4 Gramática

As unidades treinadas que serão modelos do reconhecimento são consideradas a gramática do sistema.

A) Construção do dicionário de códigos:

Também chamado de codeblock é gerado a partir da base de dados de treinamento seguindo um critério de otimização. Y. (1980) propos um algoritmo muito eficiente para o treinamento, conhecido como algoritmo LBG. No caso do reconhecimento de voz para palavras isoladas, cada elocução é dividida em vetores (frames) com os parâmetros obtidos. Para cada um dos modelos HMM, todos os frames de cada elocução da palavra que ele representa são distribuídos entre todos os estados do mesmo de maneira uniforme. Após esta divisão calcula-se para cada estado um vetor centróide a partir de todos os vetores pertencentes a este estado (SILVA, 2009).

Referências Bibliográficas

BARCELOS, A. *Reconhecimento de voz para aplicação em cadeira de rodas*. 2007. http://www.aedb.br/seget/artigos08/44_Reconhecimentodevozaplicadoemcadeiraderodas.pdf. Acesso em: 14/03/2012.

BORGES, D. *XBox 360 review*. TechTudo, 2010. Disponível em: <http://www.techtudo.com.br/review/xbox-360/um-dos-melhores-consoles-da-atual-geracao.html> Acesso em: 13/03/2012.

_____. *Kinect review*. TechTudo, 2011. Disponível em: <http://www.techtudo.com.br/review/kinect/o-acessorio-revolucionario-da-microsoft.html>. Acesso em: 13/03/2012.

BOUROUBA, E.-H. Isolated words recognition system based on hybrid approach dtw/ghmm. MacMillan Publishing, 2007.

CHOU, B.-H. J. W. *Pattern Recognition in Speech and Language Processing*. [S.l.]: CRC Press, 2003. ISBN 0849312329.

CHU, W. C. *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. [S.l.]: Wiley-Interscience, 2003. ISBN 0471373125.

COLE, R. *Survey of the state of the art in human language technology*. [S.l.]: Cambridge University Press, 1997. ISBN 88-427-0018-5.

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

DAMASCENO, E. F. *Implementação de Serviços de Voz em Ambientes Virtuais*. February 2005. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos% -/v4.3/art09.pdf>. Acesso em: 14/03/2012.

DELLER, J. R. Discrete-time processing of speech signals. Macmillan Publishing Company, 1993.

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.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing*. [S.l.]: Prentice Hall, 2006.

LOUZADA, J. A. *Reconhecimento automático de fala por computador*. Monografia (Graduação) — PUC-GO, Goiás, 2010.

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.

PROAKIS, D. K. M. J. G. *Digital Signal Processing: Principles, Algorithms and Applications*. [S.l.]: Prentice Hall, 1995. ISBN 0133737624.

RABINER, L. R. *A tutorial on hidden Markov models and selected applications in speech recognition*. [S.l.]: Proceedings of the IEEE, 1989.

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

RABINER, R. W. S. L. R. *Digital processing of speech signals*. [S.l.]: Prentice Hall; US edition, 1978. ISBN 0132136031.

RAMOS, H. M. Disciplina: Computadores e sociedade, *A história dos jogos de computadores*. 2007. Disponível em: <http://www-usr.inf.ufsm.br/~hramos/elc1020/historia_jogos.pdf>. Acesso em: 13/03/2012.

RODRIGUES, F. F. *Acionamento de um robô lego mindstorms por comandos vocais utilizando redes neurais artificiais*. Monografia (Graduação), Ouro Preto, 2009.

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.

WAGHABI, M. R. F. B. E. R. *Aplicação de Modelos Ocultos de Markov na Teoria dos Jogos*. 2009.

Y., L. An algorithm for vector quantizer design. Communications, IEEE Transactions on, 1980.

YNOGUTI, C. A. *Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov*. Monografia (Doutorado) — Unicamp, Campinas, 1999.