Raphael Ferreira Ramos

Reconhecimento de fala isolada em dispositivos móveis

2014

Raphael Ferreira Ramos

Reconhecimento de fala isolada em dispositivos móveis

Orientador: Luis Antônio Rivera Escriba, DSc

Universidade Estadual do Norte Fluminense

2014

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

Humberto Gessinger

Agradecimentos

Agradeço a Deus, por ter me ajudado em todas as escolhas, aos meus pais, Ronaldo e Lina que me incentivaram desde muito cedo a continuar nos estudos, à minha família que sempre esteve ao meu lado me apoiando e cobrando, a todos amigos que de forma direta ou indireta me ajudaram nesses anos de graduação, a minha namorada que soube compreender meus momentos de dificuldades, ao meu orientador Rivera que me ajudou e cobrou quando precisei. Faltaria espaço neste projeto se eu escrevesse o nome de todos que me ajudaram nesses anos, então deixo aqui o meu muito obrigado a todos vocês.

Resumo

A fala é a principal forma de comunicação dos seres humanos, quando falamos permitimos ao outro que conheça nossos pensamentos, sentimentos, necessidades e passamos a conhecer os sentimentos, pensamentos e necessidades do outro. Essa forma de comunicação humana ainda é um sonho se tratando de comunicação homem-máquina. O fator motivador para pesquisa nesta área foi a possibilidade de desenvolvimento de um sistema de reconhecimento de fala que não necessitasse de algum tipo de aprendizado para sua utilização, possibilitando que qualquer pessoa através da fala pudesse manipular o sistema, já que a fala é o meio de comunicação mais natural para o ser humano. Esse trabalho visa construir um jogo para descoberta de cores utilizando a voz para o seu reconhecimento, este sistema servirá como base para trabalhos futuros sobre reconhecimento de fala já que as técnicas mais tradicionais nesses sistemas foram aplicadas. O reconhecedor desenvolvido é aplicado ao reconhecimento de palavras isoladas com dependência de locutor que roda em dispositivos móveis com sistema operacional *Android*.

Abstract

The speech is the main form of communication between human beings ,when we speak we are allowing others know our thoughts, feelings, needs and so we come to know the feelings, thoughts and needs of others. This human form of communication is still a dream in case of communication between man-machine. The factor motivating for research in this area was the possibility of developing a speech recognition system that did not require any type of learning for their use, allowing anyone through speech could manipulate the system, since speech is the most natural humans means of communication. This work aims to build a game to discovery colors using speech for recognition. This system will serve as a basis for future work on speech recognition since as the more traditional techniques were applied in these systems. The recognizer developed is applied to the recognition of isolated words with dependency speaker that runs on mobile devices with operational system Android.

Lista de Figuras

Lista de Tabelas

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 Fala (RAF)*, visando uma comunicação natural entre o homem e a máquina, interação vista apenas em filmes de ficção científica. Para esses estudos virarem realidade, os computadores terão de possuir total entendimento da fala humana, capacidades como: falar, ouvir, ler, escrever, além do reconhecimento de pessoas pela voz, devem ser estabelecidas. Essas capacidades são os objetivos dos sistemas de reconhecimento de fala, permitindo que o computador "entenda" o que está sendo dito (SILVA, 2008).

De acordo com Silva (2009) e Silva (2010) os sistemas de *RAF* evoluíram 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, navegação em *smartphones*, etc. Mas mesmo com toda evolução do hardware dos computadores e otimização dos algoritmos e métodos, esses sistemas estão longe de compreender um discurso sobre qualquer assunto, falado de forma natural, por qualquer pessoa, em qualquer ambiente. Em aplicações presentes em celulares, *tablets* e *smartphones* o problema se agrava devido a falta de recursos computacionais e fatores ambientais.

Neste trabalho o foco são os dispositivos móveis já que o mundo converge para que todos nós tenhamos um destes dispositivos no bolso sendo um computador portátil com acesso a internet e diversas funções. Esses dispositivos já possuem o reconhecimento de voz como realidade, empresas como *Google* e *Apple* estão investindo em pesquisas nessas áreas e oferecem sistemas que controlam o dispositivo por voz para tarefas como: iniciar navegação *GPS*, buscar no *Google*, obter a previsão do tempo e a temperatura do momento, etc. Com a popularização desses dispositivos, comandos por voz viraram tendência para uma comunicação mais natural com o aparelho permitindo uma interação

sem restringir mãos, olhos e necessitando menos atenção que os outros meios de interação. Como forma de inicialização nessa área o trabalho propõe um reconhecedor de palavras isoladas pois o desenvolvimento é mais simples que um reconhecedor de palavras contínuas ou conectadas, que capturam palavras de comando em meio a frases. Um jogo que utiliza interação por comandos de voz é proposto para o sistema operacional *Android*, sistema presente na maioria dos dispositivos móveis atuais no mercado.

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 controladores de jogos tradicionais poderão ser aposentados em pouco tempo. Segundo Ramos (2007) 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. 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 um controle para os jogos 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 jogo de corrida por exemplo, faz uma enorme diferença na hora de acelerar mais suavemente o seu carro ou acelerar mais forte (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 sedentarismo (BORGES, 2011).

1.1 Aplicações do reconhecimento automático de fala

Segundo Martins (1997) 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, transferências de dinheiro;
- Preenchimento de formulários: O usuário entra com os dados via fala.
- Robótica: Robôs podem se comunicar pela fala com seus donos.

Atualmente existem novas aplicações, como:

- Jogos: 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. A *EA Sports* também utiliza comandos de voz no *fifa 2013* e *fifa 2014*, onde você pode substituir jogadores com comandos de voz sem pausar o jogo;
- *Smart Tvs*: Em substituição ao controle remoto é possível mudar de canal, aumentar o volume, desligar a televisão por comandos de voz;
- *Smartphones*: Envio de *SMS*, chamadas, navegação na internet, tirar fotos, aplicativos, etc.

A evolução nas pesquisas e desenvolvimento de sistemas de reconhecimento de voz é evidente, o que a pouco tempo parecia distante já está sendo utilizado. Filmes como *her*, onde um sistema interage com um homem solitário reconhecendo seus sentimentos

através da sua fala (GRACA, 2014), *Oblivion*, onde computadores capturam situações e obedecem a comandos de voz (OBLIVIONMOVIE2013, 2013), retratam o que é esperado em um futuro próximo.

O que nos interessa, no enfoque deste trabalho é a aplicação relacionada com sistemas de controle, onde o usuário determina um comando e o sistema realiza a tarefa.

1.2 Objetivos

O objetivo geral deste trabalho é desenvolver um sistema de reconhecimento de fala que reconheça um pequeno conjunto de palavras isoladas da língua portuguesa para controle e comando de determinadas funções. Para testar o funcionamento do sistema será desenvolvido um jogo interativo guiado por comandos de voz ditados pelo usuário utilizando comandos de fala pré-definidos em sua gramática, que são: **Amarelo**, **Azul**, **Branco**, **Verde**, **Vermelho**, **Preto**.

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 qualquer controle físico, permitindo ao usuário utilizar as mãos para fazer outras coisas enquanto utiliza o sistema, além das diversas aplicações que são cada vez maiores nas atividades humanas. Em dispositivos móveis permite a criação de aplicações com interfaces naturais de comunicação homem-máquina.

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 cálculo de distâncias chamado Dynamic time warping *(DTW)*, desenvolvimento do jogo de cores e verificação dos resultados.

A implementação será desenvolvida para dispositivos móveis com sistema operacional Android [1], que utiliza como linguagem de programação a linguagem Java. O sistema

[1] Sistema operacional móvel que roda sobre o núcleo Linux. Desenvolvido pela *Open Handset Alliance*, liderada pelo *Google* e outras empresas.

funciona utilizando o microfone do dispositivo para coleta do sinal de entrada e geração dos padrões de treino. Com a captura do sinal de voz é possível treinar o sistema ou tentar descobrir as cores apresentadas, a fase de treinamento consiste nas etapas de pré-processamento onde são aplicadas filtros para retirada dos momentos de silêncio, normalização, retirada do nível descontínuo, pré-enfase e janelamento e a etapa de extração de características onde são aplicados as transformadas de *Fourier*, e as técnicas de extração dos coeficientes mel-cepstrais (do inglês *Mel-frequency cepstral coefficients (MFCC))*. Esses coeficientes são usados como padrões para comparação com as elocuções executadas no momento do jogo. A elocução teste que será pronunciada no momento do jogo também passa pelos processos de pré-processamento e extração de características, com a diferença de que os coeficientes resultantes serão comparados com os padrões criados no momento do treinamento, essa comparação é realizada através da técnica *Dynamic Time Warping (DTW)* que utiliza a métrica chamada de distância *euclidiana* encontrando a menor distância entre os coeficientes, se a menor distância referenciada for referente ao padrão de uma cor do dicionário da aplicação a cor é mostrada na tela, em caso do padrão com a menor distância não possuir algum padrão referente, uma mensagem é exibida.

1.5 Visão Geral do Trabalho

Neste trabalhou buscou-se desenvolver um sistema de *RAF* 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 utiliza vários algoritmos para preparação do sinal de voz, como: “Pré-Enfase, janela de Hamming, coeficientes mel-cepstrais, DTW”.

Este trabalho está dividido em 5 capítulos, que são descritos a seguir:

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

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

O capítulo 4 mostra as técnicas utilizadas na implementação do sistema, a aplicação

desenvolvida e resultados.

O capítulo 5 mostra as conclusões do trabalho.

2 Reconhecimento Automático de Fala

Sistemas de reconhecimento automático de fala *(RAF)*, 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. Geralmente sistemas de reconhecimento de fala são divididos em 4 fases: **aquisição** da fala onde é capturado o sinal sonoro, **pré-processamento** onde o sinal é purificado, **extração de informações** onde é extraído as informações mais relevantes para próxima etapa e **reconhecimento** onde são feitas as comparações com os padrões e elocução teste.

A primeira etapa *aquisição de fala*, é o processo pelo qual ondas sonoras são convertidas em sinais elétricos, sendo representados a partir desse momento por números (flutuantes). Algumas características do ambiente de captura podem atrapalhar no processo de reconhecimento, como ruídos, distância do transdutor, etc. Assim é preciso passar por uma fase que é feita uma purificação afim de tornar o sinal o mais próximo possível da fala pura, removendo os períodos de silêncio, normalizando o volume da elocução e janelando o sinal já que na fase seguinte de extração de características se obtêm melhores resultados com sinais estacionários, fenômeno que não ocorre com o sinal sonoro, então é feito a divisão do sinal em quadros de poucos milisegundos, 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 quantidade de parâmetros a exigência por tempo e processamento seriam muito altas. Com os coeficientes gerados da extração de características é possível aplicar várias técnicas de operação para o reconhecimento da fala como: Cadeias de Markov que hoje é o método mais utilizado em sistemas de reconhecimento de fala ou também Redes Neurais que são treinadas para executar a

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

função de reconhecimento de padrões com amostras dos comandos de fala pré-definidos. Neste trabalho foi adotado a classificação por padrões, utilizando *(DTW)*.

2.1 Fala e aquisição

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, existem limitações como de processamento, métodos para identificação, armazenamento, ambiguidades de informações, complexidade dos algoritmos etc. Neste sentido a fala reconhecida é limitada.

Existem vantagens e desvantagens de se usar a fala como meio de comunicação com a máquina tal como explica Furui (1989):

Vantagens

- 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 os olhos e as mãos livres que podem complementar com gestos;
- Eficiência: Tem uma elevada taxa de informação para estabelecer contexto de fala.

Desvantagens

- Ruídos: 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 de *RAF*, ele é responsável por capturar e converter um sinal analógico em um sinal elétrico, esse processo pode ser feito através do aparelho transdutor. Todas as etapas de aquisição de voz, podem ser vistas abaixo:

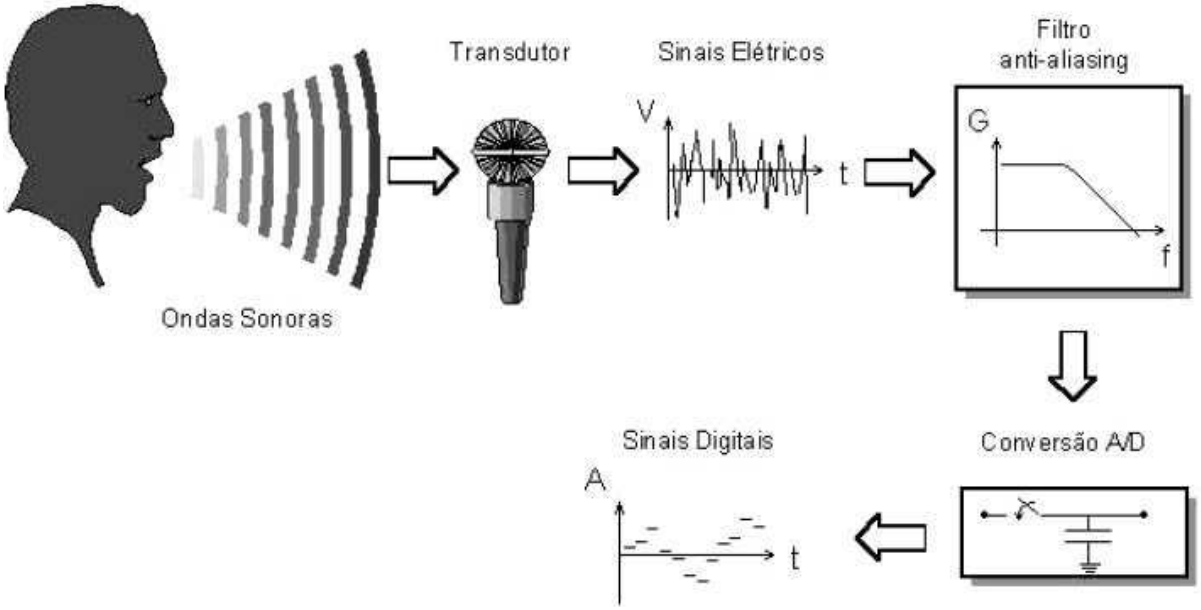

Figura 1: Processo de aquisição do sinal da fala (SILVA, 2009).

A Figura 1 ilustra todo o processo de captura da fala, onde um transdutor captura as ondas sonoras ditas pelo usuário e seus sensores convertem o sinal sonoro em sinais elétricos, o resultado dessa conversão passa por um filtro passa-baixa que garante que a frequência de amostragem seja igual ao dobro da frequência máxima do sinal, como exigido pelo teorema de *Nyquist* (FARIAS, 2011). Então é realizada a conversão do sinal de fala analógico em digital através de um amostrador possibilitando o processamento digital. Nesta etapa são escolhidas a taxa de amostragem de forma a assegurar a não ocorrência do efeito de *aliasing* e a precisão usada para a gravação do sinal. A precisão é baseada em níveis que são representados por uma cadeia de bits e deve ser escolhido de forma a conseguir uma boa precisão. Quanto maior o número de níveis maior será a precisão.

2.2 Sistemas de RAF

Sistemas de reconhecimento automático de fala 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). Nos anos 80 foram marcados pela disseminação dos métodos estáticos, como Modelos Ocultos de Markov do inglês Hidden Markov Model *(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, 1997).

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 aproxima 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 (MOM) dos inglês *HMM* utilizam essa classe de reconhecimento. 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. É identificada as unidades fonéticas da fala a serem reconhecidas, e concatenando essas unidades é reconhecida a palavra. Nessa análise é necessário considerar as propriedades invariantes da fala. Segundo Martins (1997) um analisador acústico-fonética apresenta as seguintes 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. 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 fala, que podem ser vistas na Figura 2, que além de mostrar essas etapas, também é ilustrado a fase de reconhecimento (SILVA, 2009).

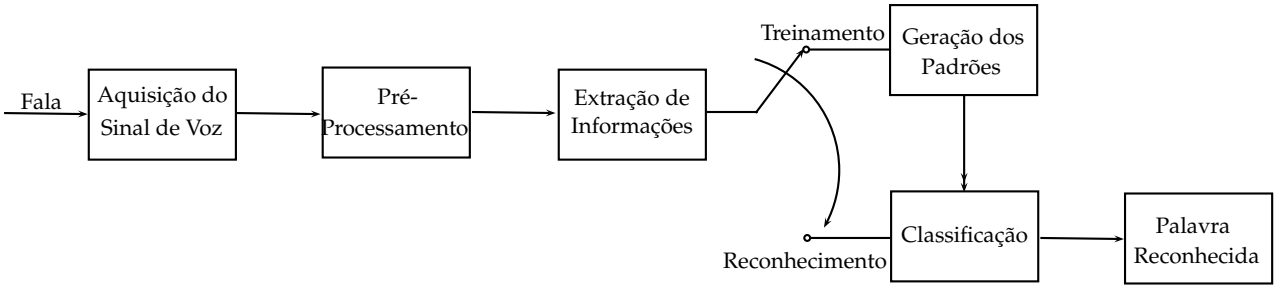

Figura 2: Diagrama de blocos de um sistema de reconhecimento de voz.

A Figura 2 são mostradas todas as etapas do *RAF*, iniciando com a captura da fala, onde um sinal digital é enviado para fase de pré-processamento, nesta fase o sinal é purificado e transformado em um sinal estacionário pelo janelamento do sinal, com o sinal janelado é feita a extração de informações desse sinal, retirando toda informação irrelevante para o reconhecimento, essas etapas apresentadas são necessárias tanto para a fase de treinamento quanto a fase de classificação. No treinamento é processada a etapa para geração de padrões, que serão as amostras usadas no reconhecimento como padrões referentes as palavras pré-definidas no vocabulário do sistema, na etapa de classificação é feita a comparação entre a elocução teste e as amostras criadas no treinamento, essa classificação é realizada pelo método *DTW* que calcula as distâncias entre os vetores e encontra a menor distância entre as amostras e apresenta a palavra reconhecida.

2.2.1 Características de Sistemas de RAF

Existem várias maneiras de categorizar um sistema de reconhecimento de fala, os mais importantes são: o estilo de pronuncia que é aceito, o tamanho do vocabulário, a dependência ou independência do locutor, perplexidade e relação sinal-ruído (MARTINS, 1997). 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 de *RAF* 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".

Alguns trabalhos apresentam reconhecedores de fala conectada como um sistema diferente dos reconhecedores de fala contínua, neste trabalho é considerado que o reconhecedor de fala conectada é similar a um reconhecedor de fala contínua, diferente do que propõe Martins (1997), 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 *RAF*, 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 *RAF* 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:

É muito importante poder quantificar a dificuldade que as comparações dos modelos de linguagem impõe aos sistemas de *RAF*, de acordo com (YNOGUTI, 1999) a melhor maneira de avaliar um modelo de linguagem é utiliza-lo em um sistema de reconhecimento e determinar qual obtém a menor taxa de erro.

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:

Tabela 1: Perplexidades típicas para vários domínios (COLE, 1997).

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

Na Tabela 1 é mostrada alguns valores típicos de perplexidades, onde é possível observar que ambientes com um vocabulário mais restrito possuem uma taxa de perplexidade menor que ambientes que possuem vocabulários maiores. O sistema produzido neste trabalho é considerado com uma perplexidade baixa já que pode reconhecer até 6 palavras.

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 (1997), 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:

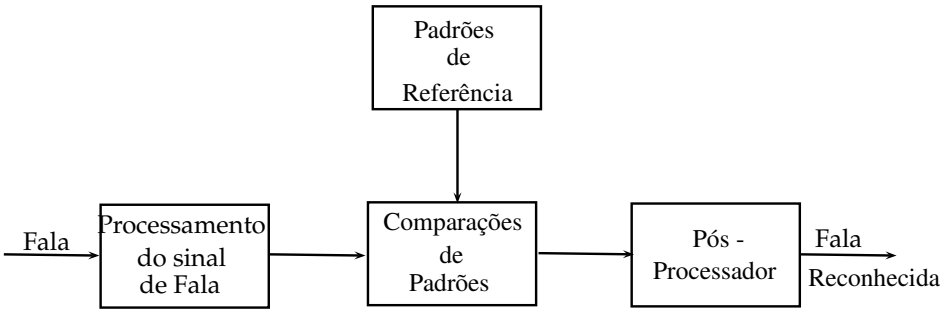

Figura 3: Sistema de RAV baseado na comparação de padrões (RABINER, 1993).

Na Figura 3 é apresentado a estrutura de um sistema de fala baseado em padrões, onde é possível observar os padrões de referência que são criados no treinamento e as

comparações entre essas amostras criadas no treinamento e elocução de teste, o resultado dessa comparação é processado no pós-processador apresentando a palavra reconhecida.

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 temporais. 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 imperceptível ao ouvido humano, mas altera a forma da onda, dificultando as comparações de padrões (MARTINS, 1997). Um grande número de parâmetros tem sido propostos e 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* (MOM) são exemplos desse método. Outro tipo é conhecido como **padrão de referência não paramétrico**, podendo ser um exemplo do padrão a ser reconhecido ou um padrão médio do padrão a ser reconhecido. Nos *Modelos Ocultos de Markov* (MOM) 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.

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 pelo *MOM* do inglês Hidden Markov Model *HMM*, a comparação resulta na probabilidade de que cada modelo de referência tenha gerado o conjunto de parâmetros de entrada (MARTINS, 1997). No modelo *DTW* é calculado a menor distância entre os coeficientes de características do sinal de voz dito pelo usuário e os padrões definidos no treinamento.

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 (1997) 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 (1997) 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:

- 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 substituídas por palavras incorretas na sentença reconhecida.

2.5 Trabalhos Relacionados

Como essas áreas são muito importantes para os seres humanos, muitos estudos foram 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.

Ruaro (2010b) que foi o trabalho de maior inspiração para este trabalho apresenta um *RAF* para dispositivos móveis em java, que reconhece os dígitos de 0 a 10, além das técnica *MFCC* e *DTW*.

RNA também são usadas em Paula (2000) onde é criado um sistema de *RAF* para palavras faladas na língua portuguesa como: “um”, “dois”, “três”.

Uma comparação entre a técnica de extração de características utilizada neste projeto é comparada com uma técnica proposta para trabalhos futuros no projeto de Cuadros (2007), onde são abordados e implementados, tais como: dígitos isolados, dígitos concatenados e frases completas, com e sem ruído.

Um sistema de *RAF* para comandar um equipamento elétrico qualquer foi apresentado por Bresolin (2003), utilizando o *MATLAB* para criação do sistema.

Em Ramiro (2010), é criado um sistema de *RAF* para palavras isoladas utilizando *HMM*, onde é possível "ligar" e "desligar" uma "televisão" ou uma "lâmpada".

Louzada (2010) apresenta um sistema de *RAF* que só realiza os comandos ditos com mais de 70% de acerto no reconhecimento utilizando *HMM* para um sistema independente do locutor.

Outro sistema de *RAF* independente de locutor foi proposto por Alvarenga (2012), sendo o objetivo específico do trabalho conceber e desenvolver um sistema de reconhecimento de voz capaz de identificar comandos de voz. A finalidade precípua do sistema era controlar movimentos de robôs, com aplicações na indústria e no auxílio de deficientes físicos.

3 Sistema de RAF proposto

O sistema de *RAF* deste trabalho tinha como proposta inicial um sistema independente de locutor, porém, devido a implementação ser genérica em *smartphones* com sistema operacional *Android* que possuem limitações de hardware e a escolha das técnicas propostas consumirem muito processamento e armazenamento, o sistema implementado se tornou dependente do locutor, então cada usuário precisa treinar o sistema para conseguir uma melhor eficiência no reconhecimento, com o modo de pronúncia de palavras isoladas onde cada amostra a ser reconhecida é salva no dicionário do sistema e o reconhecimento é realizado comparando a elocução teste com esses padrões, e um vocabulário pequeno que faz parte de um dicionário pré-definido com as palavras: **Amarelo, Azul, Branco, Preto, Verde, Vermelho**. Este sistema foi desenvolvido em Java e *APIs* direcionadas para o sistema operacional *Android*.

O sinal sonoro capturado é convertido em sinal digital para as operações de reconhecimento. Esse sinal digital precisa passar por um pré-processamento antes de ser extraídas as características que serão usadas no reconhecimento. Nessa etapa de pré-processamento são realizados vários filtros no sinal digital, como: detecção de extremos, normalização, retirada do nível *DC*, pré-enfase e janelamento. Com o sinal pré-processado ele está pronto para ir para etapa de extração de características, onde são removidas as informações redundantes através do método Mel-frequency cepstrum *(MFCC)* que aplica a transformada de Fourier do inglês Fast Fourier transform *(FFT)*, escala mel e logaritmo no sinal processado. Com o resultado dos coeficientes *(MFCC)* é possível selecionar os padrões de treinamento, esses padrões serão usados para comparação com o sinal gerado pelo usuário do jogo. O sistema de *RAF* desenvolvido nesse trabalho quando está em execução é chamado de jogo, pois é o momento onde o sistema apresenta as cores na tela e o usuário escolhe determinadas cores para serem mostradas. Com os padrões definidos é aplicado o método para cálculo das distâncias entre os sinais, neste trabalho o método escolhido para comparação das distâncias foi o *Dynamic Time Warping (DTW)*. Com um dicionário de cores definido o jogador precisa acertar as cores que são mostradas na tela,

sendo criado um quadro de acertos e erros.

Sistemas de reconhecimento de voz podem ser dividido em 6 etapas: Aquisição da fala, Pré-Processamento, Extração de Parâmetros, Criação de referências, Classificação e Execução dos comandos.

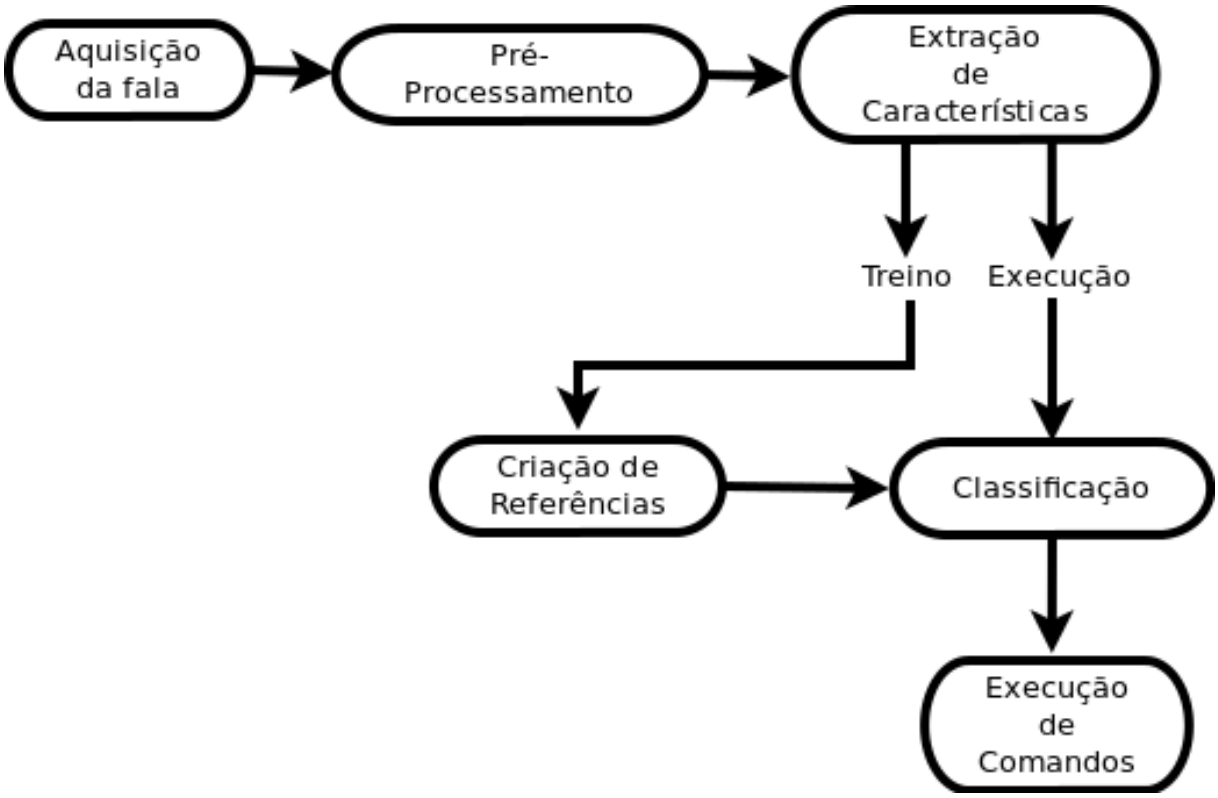

Figura 4: Etapas de um RAF.

3.1 Arquitetura do sistema de RAF

O sistema de *RAF* pode ser definido em 4 estágios que são mostradas na Figura 5, *Conversor*, *Modelo Acústico*, *Reconhecedor* e *Gramática* esses estágios são compostos por outras etapas. *Conversor* engloba as etapas de captura do sinal da fala, conversão do sinal elétrico em sinal analógico-digital e o processo de filtragem que termina gerando coeficientes *(MFCC)*. *Modelo Acústico* é a fase responsável pelo treinamento das unidades a serem reconhecidas. O *Reconhecedor* é a parte que une o resultado de todos os estágios, 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.

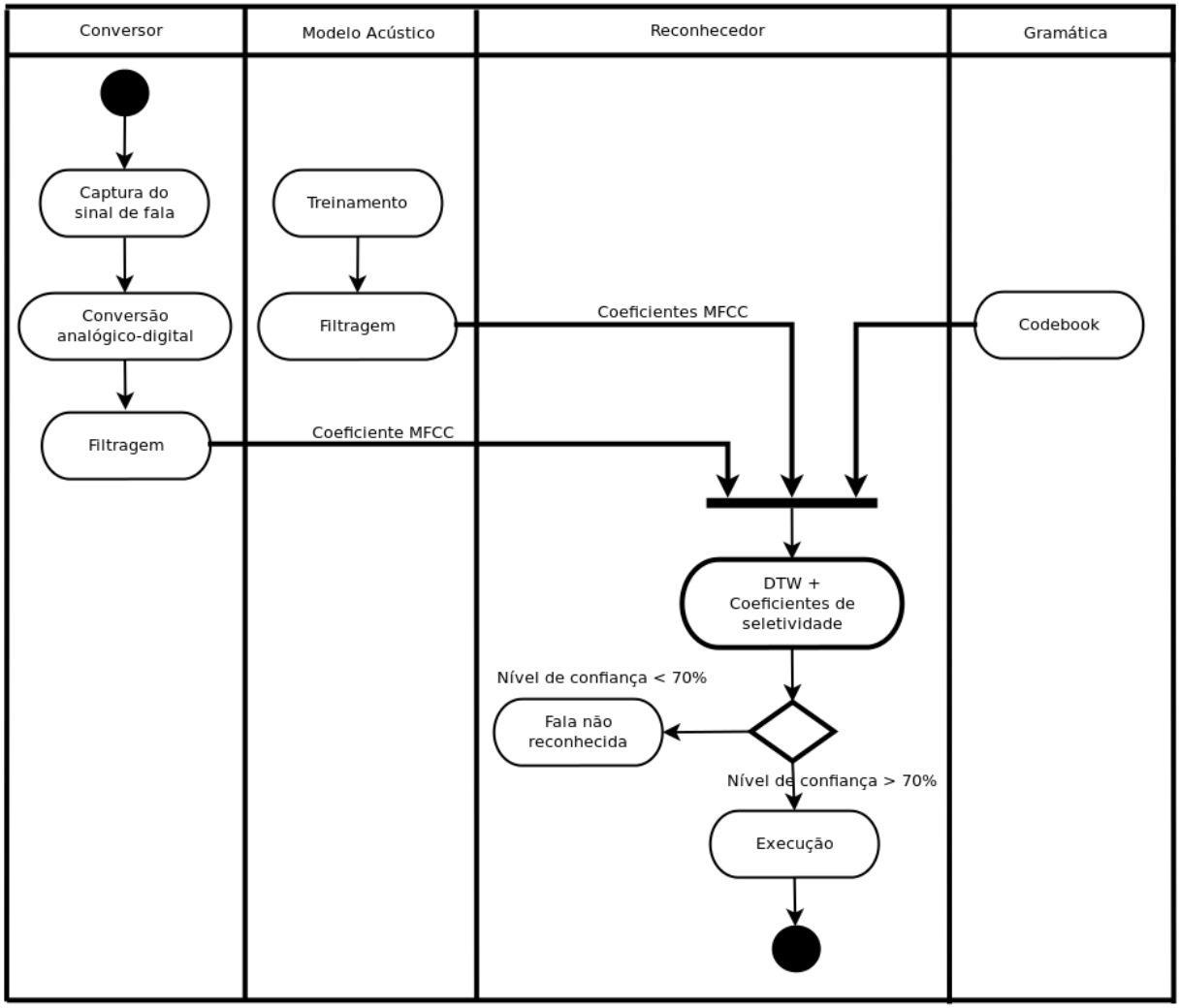

Figura 5: Arquitetura de sistemas de RAF.

Transcrição é o nome dado ao fim do processo de reconhecimento, onde é verificado se o sinal pronunciado foi reconhecido ou não. Com resultado da transcrição são executadas as decisões no dispositivo móvel.

3.1.1 Conversor

Os processos que compõem são: **Captura da fala** junto com a **conversão analógico-digital** e o **Processo de filtragem** que realiza o pré-processamento e extração de características. Nesta etapa o sinal sonoro pronunciado pelo usuário é convertido para um sinal elétrico para realização dos filtros necessários para sua digitalização na conversão analógico-digital. Esse sinal digital então está pronto para receber a purificação no pré-processamento e então podem ser retirados as informações relevantes na extração de características.

A) Captura do sinal de fala:

Como a aplicação é destinada a dispositivos móveis, a captura do som, é feita pelo microfone do celular ou tablet sendo necessário uma filtragem do sinal analógico resultante por um filtro passa-baixas, chamado de *anti-aliasing* [1] essa função pode ser encontrada dentro da função de treinamento ou jogando. Esse filtro tem o intuito de suprimir componentes de frequência superiores à metade da frequência de amostragem, sendo chamado de Nyquist (PROAKIS, 1995). Um figura demonstrando todo processo de captura do sinal pode ser encontrado na Figura 1 do Capítulo 2, subseção 2.1.1.

B) Conversão analógico-digital:

A etapa de conversão do sinal de fala analógico em digital é realizada 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.

C) Filtragem:

O processo de filtragem engloba outros processos que são: Pré-processamento e extração de características.

Pré-Processamento:

Sistemas de RAF 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 problemas 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 abaixo:

[1] Filtragem analógica cujo objetivo é eliminar frequências altas, evitando a interferência no espectro relevante durante a análise espectral. Tal interferência produz um ruído chamado alias.

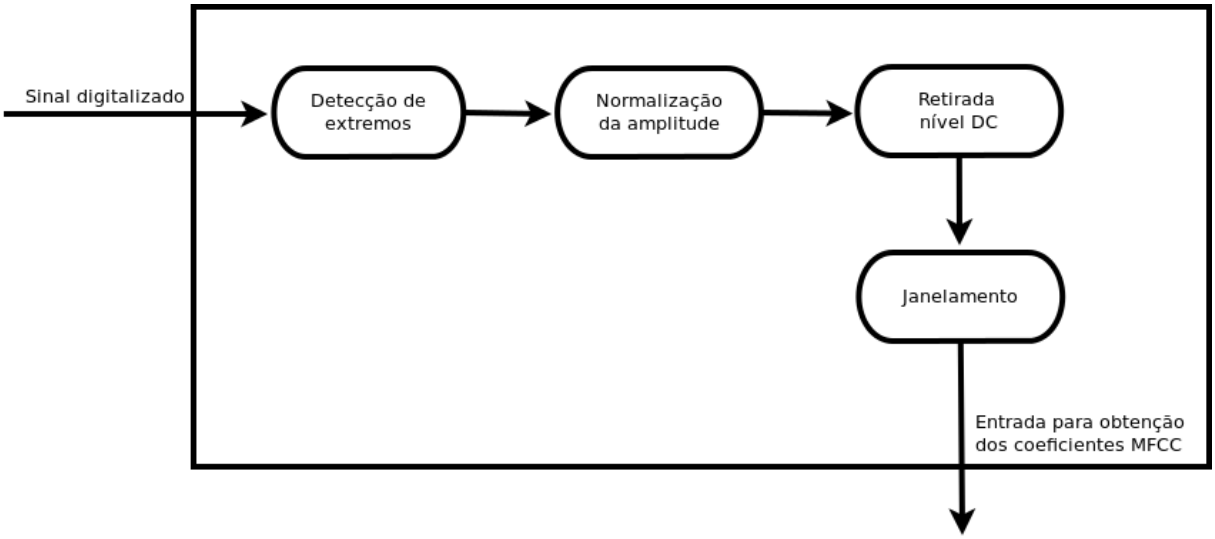

Figura 6: Diagrama de blocos da fase de pré-processamento.

Com o sinal digitalizado é necessário remover o cabeçalho do arquivo *PCM* para melhora de desempenho nos processos seguintes.

(a) Detecção de extremos:

Com o sinal digital sem cabeçalho é feita 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. 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.

(b) Normalização:

Para acabar com o problema de sons mais baixos e sons mais altos, é realizado 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 amplitude do mesmo.

(c) Nível DC:

Calculando a média-aritmética das amplitudes do sinal digital e subtraindo de cada amplitude esta média, consegue-se retirar o *nível DC*, que é uma componente contínua que atrapalha a comparação em valores absolutos.

(d) Janelamento:

Com o sinal digital filtrado e sem partes com silêncio é necessário dividir o sinal em intervalos de tempo mínimos para uma análise estacionária do sinal de fala, esse processo se chama janelamento, essa etapa é extramente importante para a etapa de extração de características.

Extração de Parâmetros:

A extração de parâmetros é uma etapa de grande importância em um sistema de *RAF*, 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 ideia 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 frequência 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).

3.1.2 Modelo Acústico

Algumas amostras precisam ser pré-gravadas para serem usadas como padrões das palavras incluídas na gramática e no modelo acústico são realizados o treinamento e uma filtragem dessas amostras.

A) Treinamento:

A fase de treinamento é uma das etapas de maior importância em um sistema de reconhecimento de fala e é o fator determinante na obtenção de um sistema com bons resultados ou não. É o momento em que são definidos os coeficientes *MFCC* para cada palavra do vocabulário utilizado, são gerados sinais de fala para serem usados como padrões, quanto maior a quantidade de padrões, maior a probabilidade de acerto do reconhecedor. Esses padrões são criados com várias formas de pronúncia, as palavras são faladas de forma: rápida, lenta, vozes diferentes como homem e mulher, etc.

B) Filtragem:

A filtragem que ocorre no **modelo acústico** é o mesmo que ocorre no modelo **conversor**, porém as amostras não são filtradas em tempo de execução, elas precisam ser gravadas anteriormente no modo de treinamento.

3.1.3 Reconhecedor

Essa fase pode ser considerada a central do sistema de *RAF*, é a etapa que une todos os outros processos e define o resultado de sucesso ou fracasso do reconhecimento. Com o sinal pré-processado é feita a extração das características e uma comparação com os coeficientes padrões. O padrão que possuir a menor distância e relação linear em relação ao sinal sonoro dito pelo usuário define o resultado do reconhecimento.

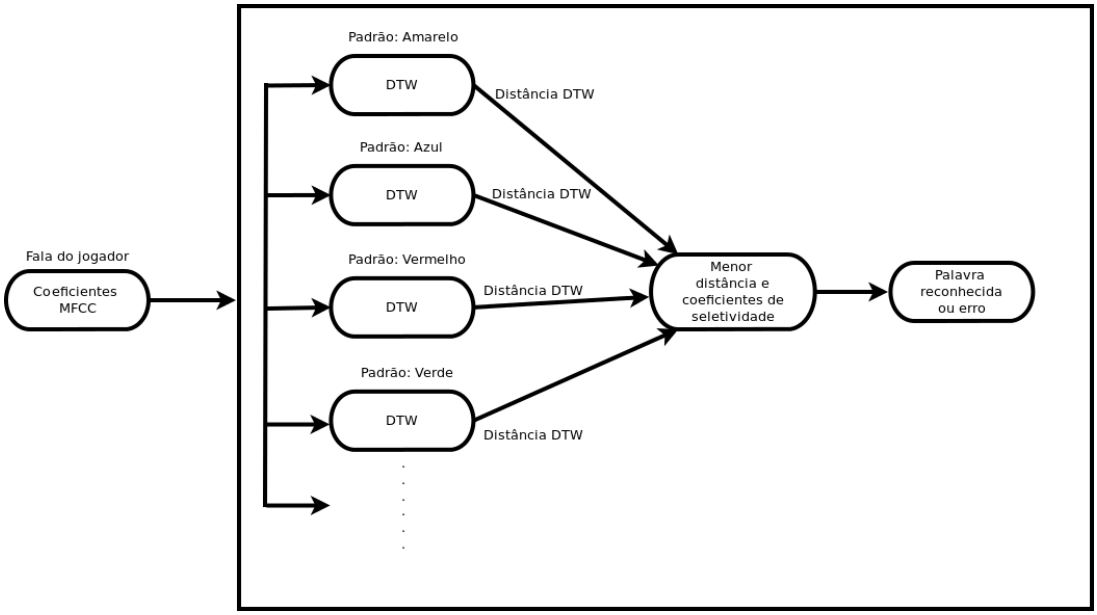

Figura 7: Procedimento de reconhecimento.

A Figura 7 mostra o funcionamento da etapa de reconhecimento, onde um sinal digital já tratado pelos processos de pré-processamento e extração de características é enviado para o reconhecedor, na Figura 8 é possível observar um pequeno trecho dos coeficientes *MFCC* de uma elocução da palavra **Amarelo**. Então é calculado as distâncias entre os vetores das amostras de treino e a elocução de teste, com a menor distância é encontrada a palavra do vocabulário que faz referência essa distância e é apresentada a palavra reconhecida ou não.

```
[-9.954634,3.924831,0.09159435,3.7373376,1.2764467,1.6818748,-1.8694402,-0.40146604,0.059863012,2.9115093,1.1654348,-4.2948227,
6.0429187,-5.5116105,7.8866925,-2.446282,0.05742232,2.4454744,-1.5085335,0.92411816,1.8841515,-2.5831501,-5.4172783,4.900813,
-6.0175095,6.244034,-3.2063339,-1.7793913,0.66172284,0.5955804,0.8727893,1.9654167,0.38268334,-1.928737,8.882959,-4.58244,8.359941
-2.2044387,0.9677965,-0.4917093,-0.58421826,2.787202,4.568597,-2.9135232,-5.694151,5.538808,-5.496592,6.135084,-1.5738153,1.711737
0.3681012,0.59990567,3.0794773,4.3359528,-1.1278439,-7.3747773,5.7290673,0.34593102,2.8988848,-1.4917376,0.18919547,-2.2361424,
-0.17264338,-0.13710678,2.8519642,0.017018111,-8.241397,5.8858905,2.3524783,5.0694737,-1.9207405,-0.88898206,0.6918082,0.8438,
1.0517799,3.2456987,-1.3473012,-10.724303,2.5525663,-1.2185276,0.5109417,-2.4223988,0.23792024,0.41653675,-1.1325183,-2.1177375,
1.6876777,1.8947194,-10.962196,1.8502649,-3.2905343,0.794774,-0.99146,1.8191034,0.3143207,0.84158,0.2853882,2.3661768,0.39915186,
```

Figura 8: Trecho dos coeficientes MFCC da palavra Amarelo.

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 que o sinal já tenha sido pré-processado e os coeficientes de características para serem a sequência de observações a ser utilizada no algoritmo de reconhecimento.

A) Dynamic Time Warping (DTW):

O método *DTW* ou Modelagem Determinística é um algoritmo para calcular o caminho ótimo entre duas séries temporais (RUARO, 2010b). O algoritmo também é uma

solução para os problemas causados pela variação de tempo no pronunciamento das palavras. Nesta modelagem utiliza-se uma métrica entre os padrões de referência e a palavra teste, resultando em distâncias entre os vetores, distância que pode ser obtida através de cálculos vetoriais tais como: *Malahanobis*, a distância de *Bhattacharyya* e a *distância Euclidiana*, esta última, segundo Ruaro (2010b) é a mais comumente usada no reconhecimento de voz devido ao seu baixo custo computacional e simplicidade de implementação.

- Distância Euclidiana:

  A distância euclidiana é uma forma de computar a semelhança através da distância entre duas distribuições vetoriais. Quanto menor for a distância calculada, mais próximos são os padrões comparados. A distância euclidiana pode ser obtida para duas sequencias através da seguinte equação:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} |x[i] - y[i]|^2} \tag{3.1}$$

Chamando o coeficiente *MFCC* do padrão de referência de R e o coeficiente do padrão de teste de T com N e M quadros de coeficientes MFCC cada, onde quadros são o tamanho dos vetores de cada elocução, podem ser representados por sequencias espectrais mostrados abaixo:

$$R = r_1, ..., r_n \tag{3.2}$$

$$T = t_1, ..., t_m \tag{3.3}$$

onde r_n e , t_m n = 1,...,N e m = 1,...M, são os vetores de parâmetros de características acústicas. Aplicando o DTW as sequencias R e T são comprimidas ao longo do eixo temporal resultante para se obter um alinhamento não-linear eliminando a diferença temporal, essa características permite comparar elocuções de tamanhos diferentes.

Em FURTUNĂ (2008) é mostrado o objetivo do DTW, que é encontrar o menor caminho ou caminho ótimo através da menor distância da matriz de distâncias mínimas. As distâncias então são somadas e obtém-se uma distância mínima geral. Com as distâncias calculadas é necessário encontrar os coeficientes de seletividade de reconhecimento, para verificar se o reconhecimento foi bem sucedido ou não.

B) Coeficientes de Seletividade de Reconhecimento:

Estes coeficientes são utilizados para quantificar e qualificar uma determinada distância em relação às outras distâncias obtidas e assim classificar se o reconhecimento pode ser considerado correto ou não. No reconhecimento a menor distância encontrada geralmente é a palavra reconhecida, mas palavras fora do banco de padrões podem ser ditas e seriam referenciadas a alguma menor distância, causando um erro.

Um exemplo do reconhecimento que pode causar erro é mostrado na Figura 9, que o jogador pronuncia a palavra *Rosa*, uma palavra não presente no banco de padrões, a menor distância calculada foi do padrão *Azul* então o reconhecimento seria impreciso.

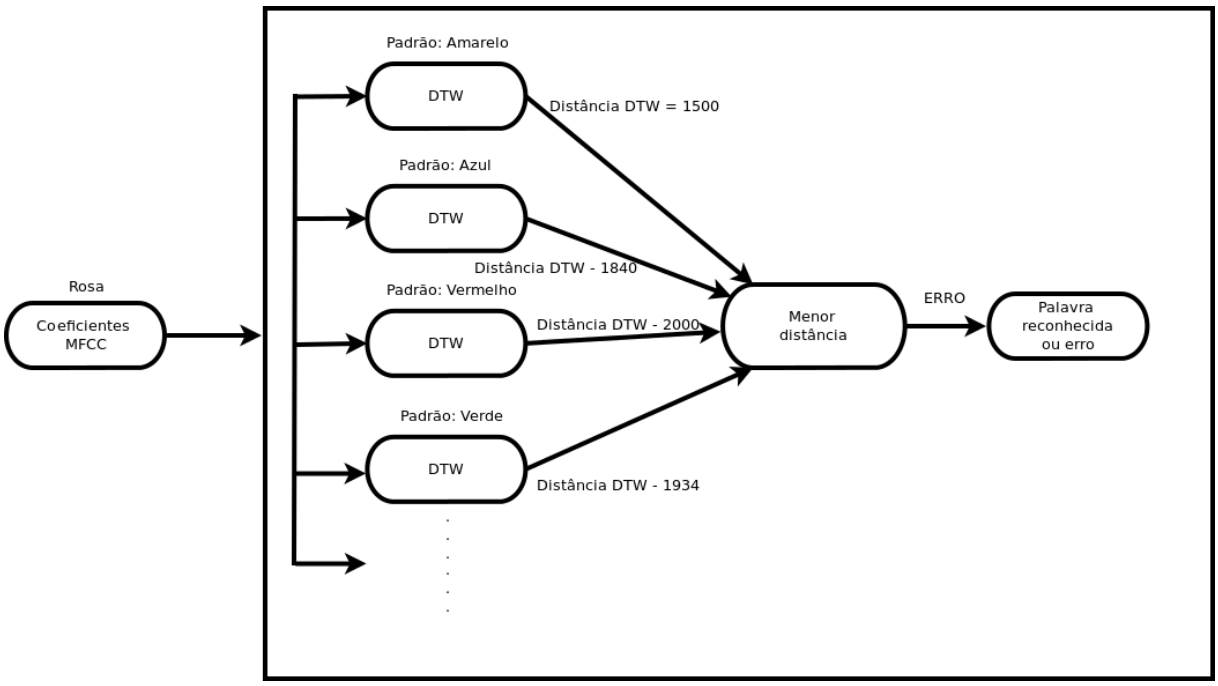

Figura 9: Erro com a menor distância no *DTW*.

A Figura 9 mostra o mesmo procedimento de reconhecedor, porém enfatiza o problema da sempre existir uma menor distância no cálculo com os padrões, problema resolvido com auxílio dos coeficientes de seletividade.

Baseado em Ruaro (2010b) este trabalho propõe analisar a qualidade do reconhecimento através dos coeficientes de seletividade definidos por:

- C_1 = menor distância entre a elocução de teste e as elocuções de referência;
- C_2 = diferença entre a segunda menor distância e C_1;
- C_3 = diferença entre as médias das distâncias e a menor distância entre a elocução de teste e as de referência que não pertencem a mesma palavra da elocução de teste ou que não pertença ao padrão que resultou na menor distância geralmente sendo associada a segunda menor distância.

Assim, quanto menor for C_1 melhor a qualidade do reconhecimento e quanto maior forem C_2 e C_3 maior a qualidade do reconhecimento, sendo uma boa métrica de qualidade das distâncias. Os coeficientes de seletividade Sel_1 e Sel_2 também são usados nesse trabalho. Dado que a distância final da DTW era a soma das distâncias entre os quadros ao longo do caminho ótimo de comparação, as parametrizações geradas geraram coeficientes numericamente diferentes. As distâncias resultantes da aplicação da DTW em vez de caracterizarem um número maior que 0 e menor que 9, são resultado da soma das distâncias do caminho ótimo, resultando em valores fora deste padrão. Então usa-se dois novos fatores.

$$Sel_1 = \frac{C_2}{C_1} \quad Sel_2 = \frac{C_3}{C_1} \tag{3.4}$$

Esses coeficientes são uma boa mediada da qualidade de parametrização. Isto porque a normalização C_2 e C_3 em relação a C_1 na determinação de Sel_1 e Sel_2 elimina as diferenças numéricas entre as diferentes parametrizações e permite a comparação direta dos diferentes procedimentos de extração de parâmetros. Da maneira em que Sel_1 e Sel_2 foram definidos quanto maiores eles forem melhor será a qualidade ou seletividade do reconhecimento.

3.1.4 Gramática

As unidades treinadas que serão modelos do reconhecimento são consideradas a gramática do sistema, no sistema desenvolvido neste trabalho 6 palavras foram escolhidas para fazerem parte da gramática do sistema de *RAF*, as palavras são: **Amarelo, Azul, Branco, Preto, Verde, Vermelho**

A) Codebook:

Também chamado de *dicionário de códigos*, o *codebook* é gerado a partir da base de dados de treinamento seguindo um critério de otimização. O *codebook* criado nesse trabalho foi definido pelos sinais de fala que foram ditos de maneira mais clara e de diferentes formas sendo salvos seus respectivos coeficientes resultados do método *MFCC*. Quando se decide treinar o sistema, é feita uma busca pelo cartão de memória do dispositivo móvel na pasta da aplicação e todos os arquivos do tipo *WAV* com os nomes das palavras salvas no dicionário são submetidos as fases de pré-processamento e extração de características, obtendo os coeficientes. Esses coeficientes são salvos

em arquivos de texto no cartão de memória do dispositivo. Quando o usuário entrar no modo de **Jogo**, esses coeficientes salvos no cartão de memória são recuperados e comparados com a elocução pronunciada pelo usuário.

4 Implementação

Como abordados nos capítulos anteriores existem inúmeras combinações de técnicas para um sistema de reconhecimento de fala, neste capítulo são apresentados os detalhes técnicos das técnicas usadas para o propósito de implementação deste trabalho.

4.1 Captura da fala

Para captura da fala foram utilizadas funcionalidades nativas do sistema operacional *Android*, que possui algumas classes responsáveis por esse tipo de manipulação de dados e o microfone nativo do dispositivo móvel que será o responsável por converter ondas sonoras em sinais elétricos de tensão analógica. A classe *AudioRecorder* foi escolhida por ser mais flexível no processo de captura, podendo ser editado os valores de entrada que são importantes para o desenvolvimento do reconhecedor de voz. Esse processo de conversão de um sinal analógico em sinal digital é necessário para o tratamento computacional. Segundo Ruaro (2010b) o objetivo da digitalização é representar um sinal analógico em níveis de representação discretos no tempo que correspondem aproximadamente às variações contínuas no tempo presentes no sinal a ser digitalizado, além de citar as quatro etapas mínimas para digitalização, que são:

1. Filtragem anti-aliasing;
2. Amostragem;
3. Quantização;
4. Codificação.

Antes de detalhar os processos da digitalização, é importante citar o Teorema de Amostragem de *Nyquist*, segundo Carvalho (2008):

> "A frequência de amostragem de um sinal analógico, para que possa posteriormente ser reconstituído com o mínimo de perda de informação, deve ser igual ou maior a duas vezes a maior frequência do espectro desse sinal."

A metade da frequência de amostragem é chamada frequência de *Nyquist* e corresponde ao limite máximo de frequência do sinal que pode ser reproduzido.

4.1.1 Filtro anti-aliasing

O sinal sonoro pode possuir componentes de frequência em uma faixa de amplitude muito larga e também não possuir limitação constante em frequência, então como não é possível garantir que o sinal não contenha frequências com distorções, ruídos, interferências é realizado um filtro-passa baixas chamado *anti-aliasing*. Esse filtro é realizado pela placa de som do dispositivo. O efeito de aliasing nada mais é do que a superposição dos espectros de cada pulso modelado em amplitude por falta de espaço. Na restituição do sinal pelo filtro passa baixa com frequência de corte do tamanho da metade da taxa de amostragem escolhida que pode ser chamado de **fn**, a parte do espectro original acima de **fn** aparece como se tivesse sido dobrada em torno de **fn** e invertida espectralmente, ou seja, frequências mais altas passam a ser menores. O sinal indesejável de aliasing que aparece na reprodução é uma réplica do sinal original, porém com frequência errada.

4.1.2 Amostragem

O sinal sonoro é propagado no formato de ondas no espaço, por ser analógico o sinal possui infinitas variações através do tempo. A medição desses valores só é possível com auxílio de técnicas que captam amostras regulares do sinal analógico e os codifique em um conjunto sequencial sem perca de informações relevantes.

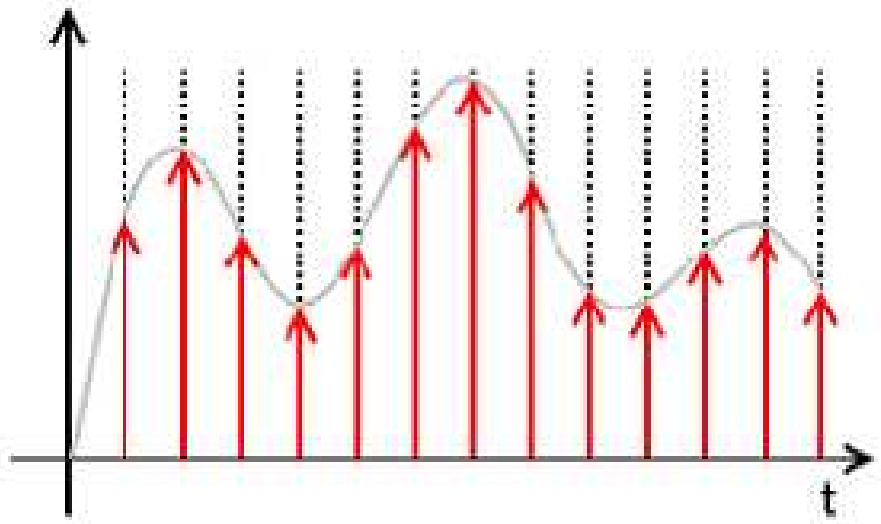

Figura 10: Amostragem de um sinal contínuo (FILHO, 2006).

A Figura 10 representa um exemplo da aplicação da técnica de amostragem em um sinal contínuo, dividindo o sinal em amostras regulares no tempo.

O teorema da amostragem proporciona condições para que o sinal analógico possa ser recuperado através de suas amostras, obtendo amostras do sinal em instantes de tempo discretos e é definida uma frequência de amostragem que representa o número de amostras registradas por segundo, os valores mais comuns para gravação de áudio são de 8 KHz, 11 KHz, 16 KHz e 22 KHz, neste trabalho o taxa escolhida foi de 22 Khz.

Aumentando a taxa de amostragem é possível diminuir o erro no reconhecimento, porém essa melhoria ocorre até o limite de 22 Khz pois a maioria das características marcantes da fala estão na faixa de 8 Khz, em (OLIVEIRA, 2002) foi apresentado uma tabela mostrando as melhorias com o aumento das taxas de amostragem, partindo do valor inicial de 8Khz.

Tabela 2: Redução da taxa de erros com o aumento da taxa de amostragem.

Taxa de amostragem	Redução relativa da taxa de erros
8KHz	Limite mínimo
11KHz	+10%
16KHz	+10%
22KHz	+0%

Esta tabela parte do limite mínimo de frequência para o reconhecimento, aumentando 10% na melhoria do reconhecimento a cada acréscimo na frequência definida até o limite de 22KHz, que a partir desse valor não há melhora no reconhecimento.

4.1.3 Quantização

Quantização é o processo que transforma um sinal discreto nos sinais digitais que são filtrados pelo processador, nesse processo é feita uma aproximação de uma faixa contínua de valores para um conjunto relativamente pequeno de símbolos discretos ou valores inteiros.

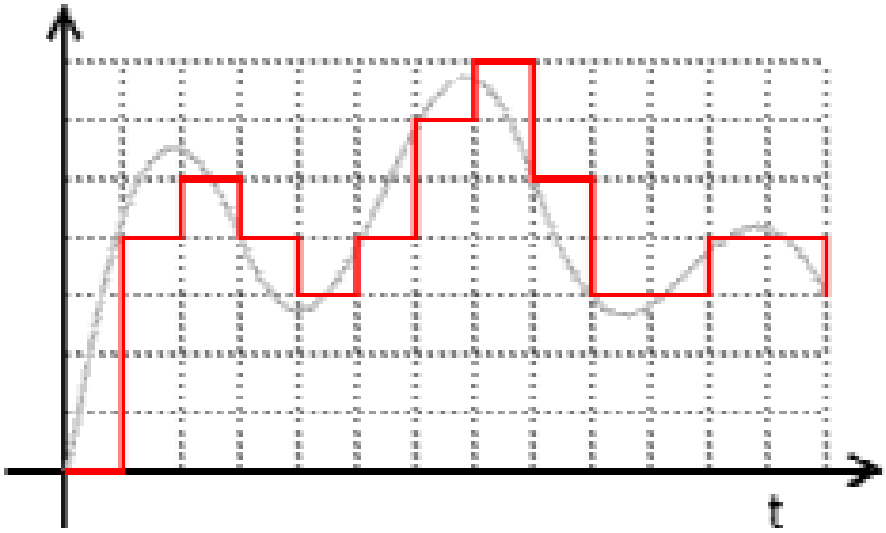

Figura 11: Sinal digital (FILHO, 2006).

Na Figura 11 a técnica de amostragem é aplicado em um sinal contínuo.

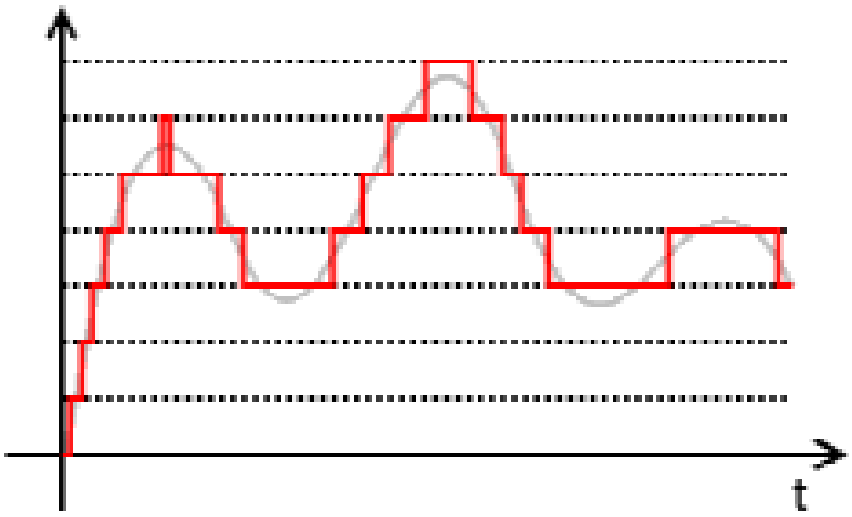

Figura 12: Sinal quantizado (FILHO, 2006).

Já na Figura 12 o mesmo sinal amostrado é quantizado para se obter uma aproximação maior do sinal original.

Com os sinais amostrados inicia-se a quantização em uma determinada resolução. Neste trabalho, cada amostra do sinal recebe a resolução de 16 bits, mas em muitos outros trabalhos 8 bits também são muito utilizados. Em (RUARO, 2010b) pode-se encontrar a seguinte afirmação sobre o que é quantização:

> "O processo através do qual um sinal contínuo em amplitude é transformado em um sinal discreto em amplitude."

4.1.4 Codificação

Este é o último processo da conversão analógico digital, na codificação é utilizado os resultados obtidos e refinados nas etapas de amostragem e quantização os transformando em um formato sequencial binário pois um sinal digital binário só pode ter dois valores diferentes "0" ou "1". 8 níveis de quantização podem ser encontrados: 0,1,2,3,4,5,6,7 *Volts* e cada nível é representado por uma sequência de 3 bits.

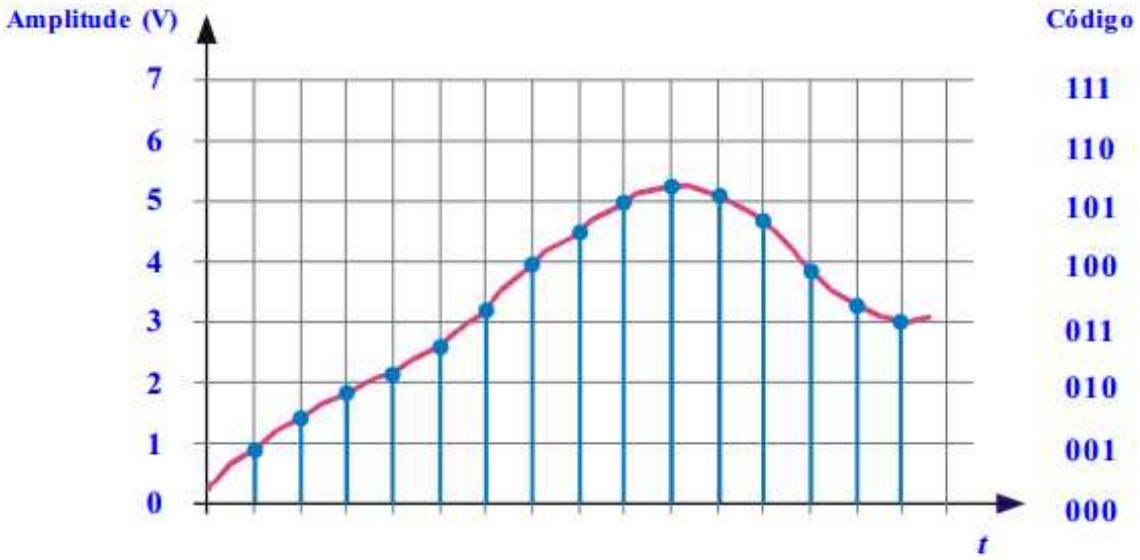

Figura 13: Codificação (FILHO, 2013).

Na Figura 13 a sequência resultante seria: **001 001 010 010 011 011 100 101 101 101 101 101 100 011 011**.

O resultado da captura é um vetor de bytes cujo os níveis representam a intensidade do sinal de voz. Os parâmetros utilizados na gravação foram:

- Taxa de amostragem: 22 KHz
- Número de bits por amostra: 16 bits
- Número de canais: 1 canal

Onde *taxa de amostragem* é a quantidade de amostras de um sinal analógico coletadas em uma determinada unidade de tempo para conversão em um sinal digital. Sendo uma frequência é comumente medida em *Hertz (Hz)*.

Número de bits por amostra podemos ter 8 ou 16 bits. Com 8 bits é possível gravar 256 valores diferentes de amplitude do sinal. Com 16 bits é possível gravar 65536 diferentes valores. A escolha depende da capacidade de processamento, sendo 8 bits suficiente, mas 16 bits é melhor.

Número de canais pode ser 1 para *mono* e 2 para *estéreo*. Em muitos casos 1 canal é suficiente. Quando tivermos dois canais, se usarmos 8 bits para representar uma amostra, teremos 8 bits para cada canal por amostra, resultando num total de 16 bits por amostra. Se usarmos 16 bits para representar cada amostra, com 2 canais, teremos 16 bits para cada canal por amostra, resultando num total de 32 bits por amostra (OLIVEIRA, 2002).

A partir dessa captura é feita a digitalização do sinal convertendo o arquivo de som em um arquivo de áudio Pulse-code modulation *PCM* no formato *WAVE*. Esse arquivo *Wave* fica salvo no cartão de memória do dispositivo.

4.2 Pré-Processamento

Como resultado da captura do sinal de fala temos um vetor de bytes com as amplitudes relativas ao sinal sonoro, esse vetor herda todos os dados do arquivo *wav* criado no cartão de memória. Esse formato de arquivo possui um cabeçalho que não possui utilidade na extração de características e se não for removido pode causar problemas no reconhecimento. Então é feita a remoção do cabeçalho do vetor retirando os índices de valor menor que 56 antes de iniciar a etapa de pré-processamento. O pré-processamento visa preparar o sinal para a extração de parâmetros, normalizando, detectando o início e fim da palavra, retirando o nível *DC*, realizando a pré-enfase do sinal e janelamento as cinco etapas são descritas a seguir:

4.2.1 Normalização

Inicia-se o processo de pré-processamento com a normalização. A normalização da amplitude se refere à altura do som, esse filtro faz com que todos os valores de amplitude estejam na mesma faixa de valores, tornando que tanto os sons mais altos quanto os mais baixos possam ser processados igualmente pelo algoritmo de reconhecimento. Para ser realizado esse processo, é necessário converter o vetor de bytes do arquivo do tipo *wav* gravado no cartão de memória para um vetor de valores flutuantes, cada índice desse vetor do tipo *float* representa a amplitude do sinal. Esse processo é realizado encontrando o maior valor de amplitude do sinal e subtraindo cada valor de amplitude do sinal por esse valor.

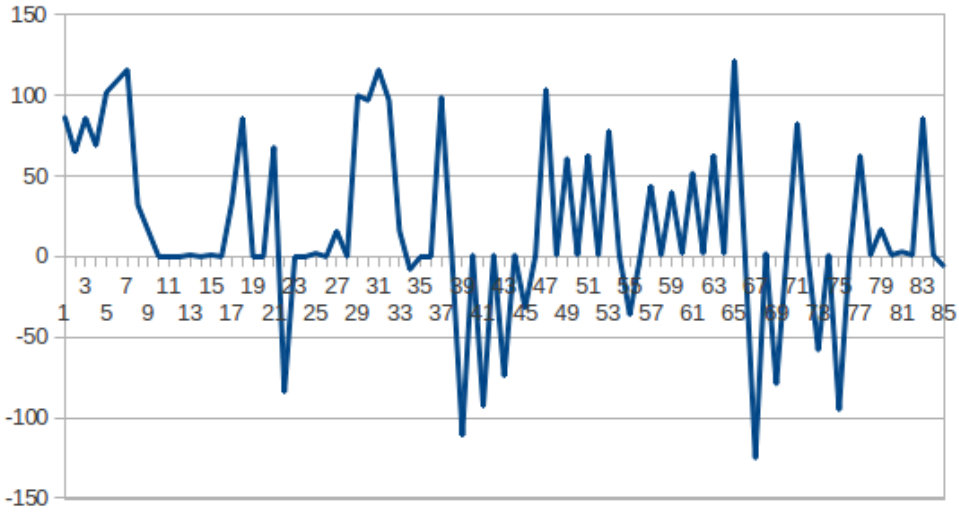

Figura 14: Sinal sem normalização.

Sinal sem normalização é mostrado na Figura 14.

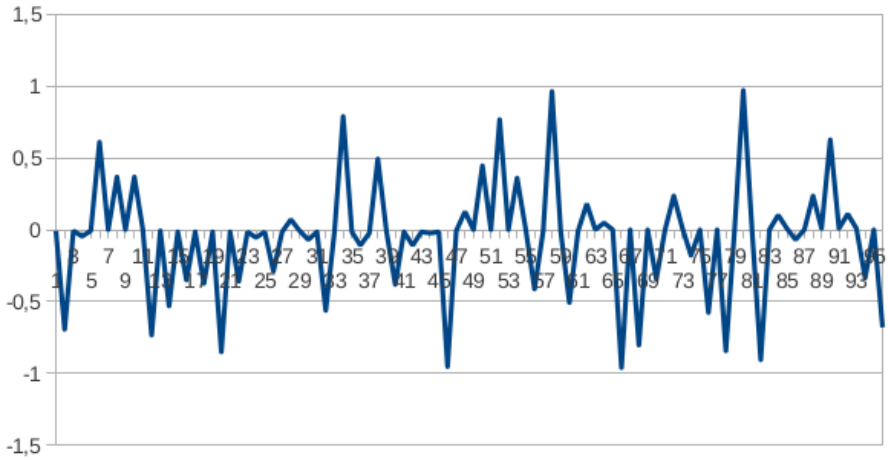

Figura 15: Sinal normalizado.

Na Figura 15 é mostrado um trecho da palavra “Amarelo” com o sinal normalizado, como se pode observar todos os valores se encontram na faixa de -1 a 1.

4.2.2 Detecção de extremos

A etapa da detecção de extremos consiste em detectar o início e fim do sinal de voz, reduzindo o custo computacional na fase de extração de parâmetros, esta fase é muito importante para o reconhecimento da voz, se feita de maneira errada pode mudar consideravelmente os valores das características e impossibilitar o reconhecimento. Essa etapa é de tamanha importância que já existe uma ramo de estudos voltados apenas para algoritmos de detecção de extremos. O algoritmo utilizado neste trabalho é o apresentando em (SAHA; CHAKROBORTY; SENAPATI, 2005) que propõe um novo algoritmo superior aos modelos convencionais de cruzamentos por zero ou por energia que apesar da fácil implementação apresentam algumas limitações.

O algoritmo de Saha, Chakroborty e Senapati (2005) para detecção de extremos e remoção do silêncio:

1. Calcule a média e desvio padrão das primeiras 1.600 amostras do vetor já normalizado. Se **M** e **D** são a média e o desvio padrão, respectivamente, em seguida,

analiticamente, podemos escrever:

$$M = \frac{1}{1600} \sum_{i=1}^{1600} x(i) \tag{4.1}$$

$$D = \sqrt{\frac{1}{1600} \sum_{i=1}^{1600} (x(i) - M)^2} \tag{4.2}$$

Onde i é o índice do vetor com valores flutuantes. Nota-se que o ruído de fundo é caracterizado por **M** e **D**.

2. Percorrer o vetor de amostras e em cada amostra verificar se a função distância de *Mahalanobis* é maior do que 3 ou não, a amostra deve ser tratada como amostra sonora caso contrário, é um silêncio.

3. Marque a amostra sonora como 1 e o silêncio como 0. Divida o sinal de fala todo em 10 ms. Agora um novo vetor é representado por apenas zeros e uns.

4. Considere que exista M número de zeros e N número de uns em uma janela. Se M $\geq$ N, é necessário converter cada um dos zeros e vice-versa. Este método adotado aqui mantendo em mente que um sistema de produção da fala que consiste em cordas vocais, língua, etc trato vocal não pode mudar abruptamente em um curto período de janela de tempo tomado aqui como 10ms.

5. Colete a parte sonora só de acordo com o rotulado amostras "1" a partir da matriz de janela e defina um novo vetor. Recupere a parte sonora do vetor original a partir do novo vetor gerado.

As Figura 16 ilustra um sinal antes da passagem pelo filtro de detecção de extremos e a Figura 17 apresenta o sinal com os extremos detectados e o silêncio removido.

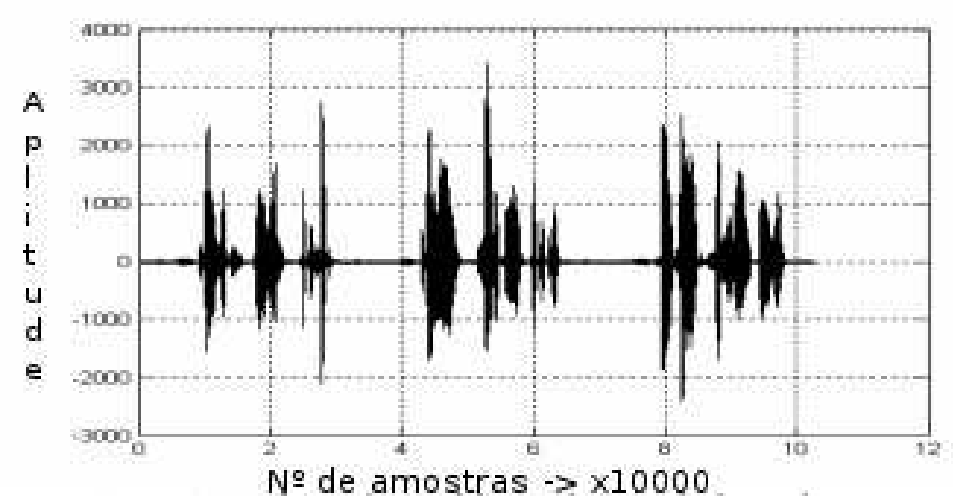

Figura 16: Sinal original (SAHA; CHAKROBORTY; SENAPATI, 2005).

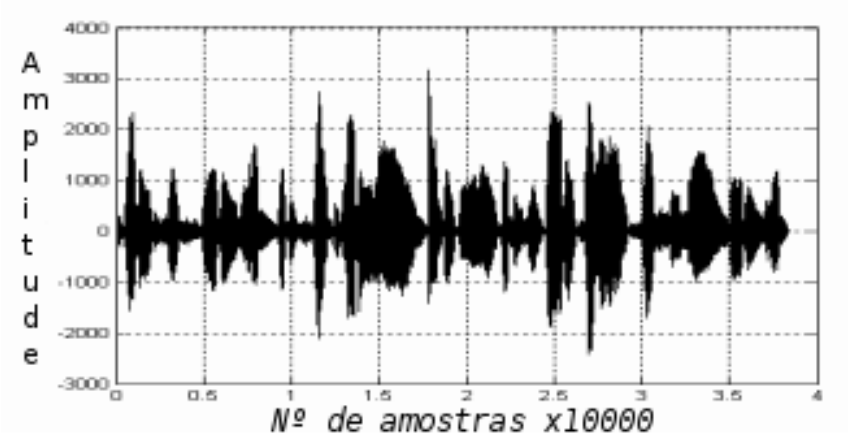

Figura 17: Sinal com os silêncios removidos (SAHA; CHAKROBORTY; SENAPATI, 2005).

4.2.3 Retirada da componente contínua

Geralmente os sinais de voz apresentam uma componente contínua que atrapalha a comparação em valores absolutos, então é necessário remover esse nível DC afim de deixar todas as amostras oscilando em torno do valor 0. Depois de normalizado e sem os períodos de silêncio, é aplicado o filtro para retirada do nível DC(contínua) no vetor com valores flutuantes, para isso calcula-se a média aritmética das amplitudes do sinal e subtrai-se cada amplitude por essa média. A eliminação deste nível coloca todos os sinais em relação a mesma referência.

Na Figura 18 é representada a palavra “desce” com a componente contínua.

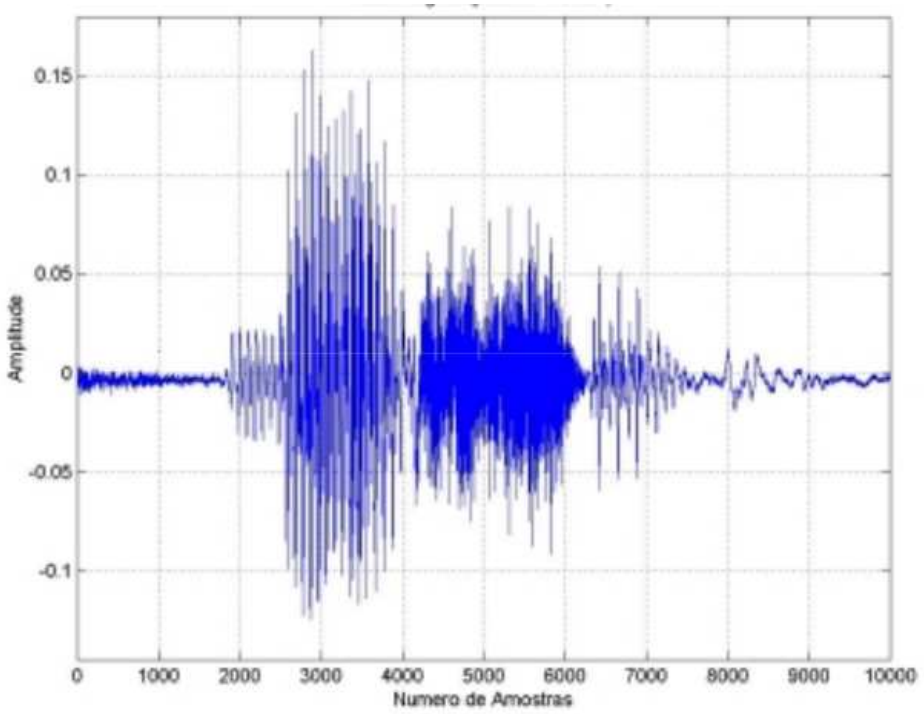

Figura 18: Sinal de voz com a componente DC (LEMOS; RODRIGUES; HERNANDEZ, 2004).

Já na Figura 19 o filtro de remoção da componente contínua foi aplicado.

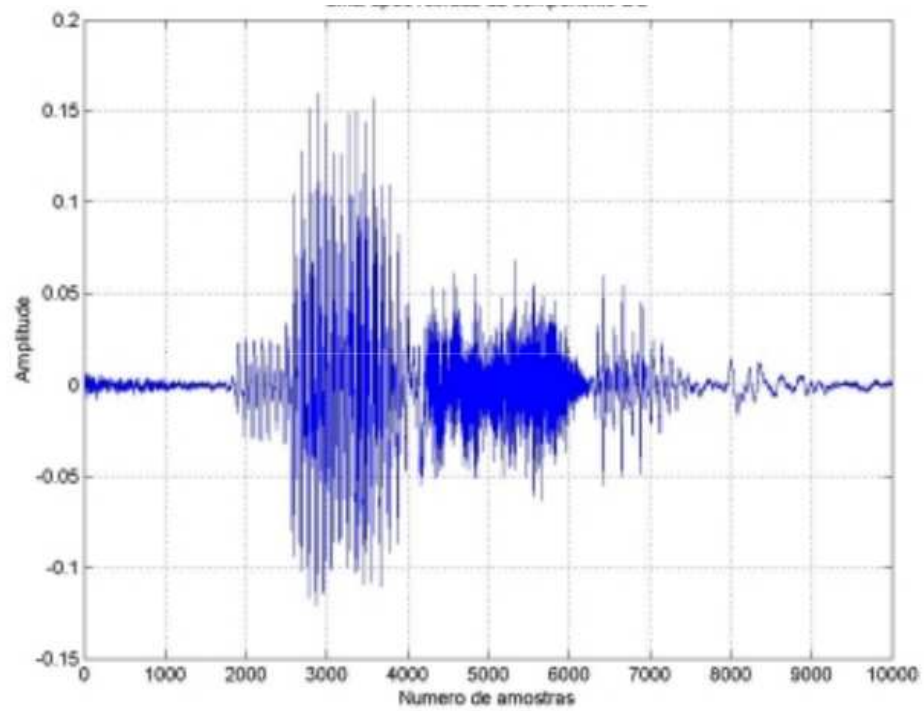

Figura 19: Sinal de voz sem a componente DC (LEMOS; RODRIGUES; HERNANDEZ, 2004).

4.2.4 Pré-Enfase

A filtragem de pré-enfase é necessária para eliminar uma tendência espectral de aproximadamente -6dB/oitava na radiação dos lábios. Essa distorção espectral não traz informação adicional e é eliminada pelo filtro de resposta aproximadamente +6dB/oitava, que ocasionaria um nivelamento no espectro (SILVA, 2009).

Basicamente o filtro de pré-enfase realça as frequências que o sistema auditivo humano é mais sensível convertendo o sinal tratado em um vetor de características.

É possível observar em sinais de voz que a energia presente nas frequências altas é menor se comparada as baixas frequências.

No domínio do tempo, o sinal de saída $\mathbf{y(n)}$ relaciona-se com o sinal de entrada $\mathbf{x(n)}$ onde $\mathbf{x}$ é o vetor resultante dos filtros anteriores, $\mathbf{n}$ o índice do vetor e $\mathbf{y}$ o novo valor para amplitude no índice $\mathbf{n}$, pela fórmula abaixo:

$$y(n) = x(n) - \alpha x(n-1) \tag{4.3}$$

O coeficiente α neste trabalho utilizou-se o coeficiente $\alpha = 0.95$, que é o mais utilizado em reconhecedores de voz, podendo ser encontrado por (RUARO, 2010b), (SILVA, 2009) em trabalhos onde foram encontrados bons índices de reconhecimento.

Um sinal sem aplicação do filtro de pré-enfase pode ser visto na figura 20

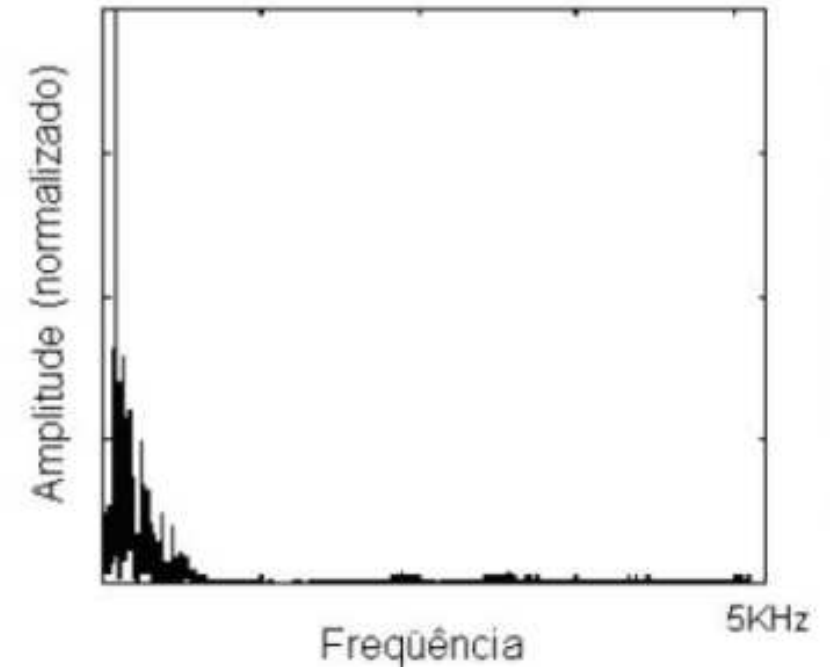

Figura 20: Sinal sem filtro de pré-enfase (SILVA, 2009).

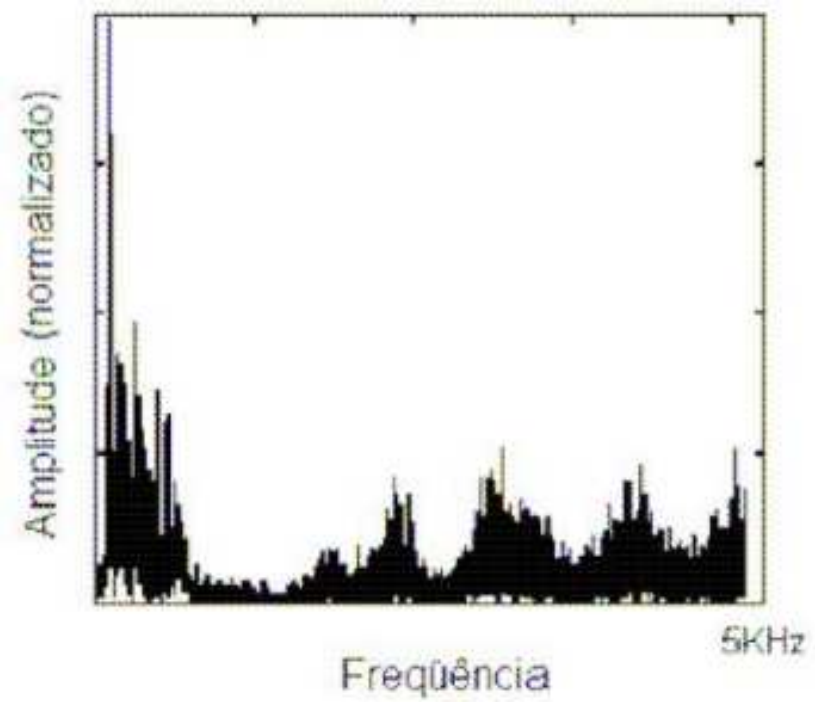

Figura 21: Sinal com pré-enfase (SILVA, 2009).

Existe uma grande variedade nos trabalhos utilizados como referência das técnicas de pré-processamento, neste trabalho o sinal original é normalizado, removido o silêncio do sinal, retirado a contínua e aplicado o filtro de pré-enfase antes de janelar o sinal.

4.2.5 Divisão do sinal em quadros e Janelamento

Depois do sinal pré-enfatizado é necessário segmentar este sinal em janelas, já que as técnicas de extração de parâmetros conseguem bom resultados para sinais estacionários, o que não ocorre na voz. Essa segmentação ocorre dividindo o sinal em N quadros*(frames)* de amostras. Para o sinal ser considerado quase estacionário a divisão do sinal varia no intervalo de 10 ms a 30 ms. Essa divisão é possível, devido a se assumir que o sinal de fala é invariante no tempo sobre um intervalo menor que 30 ms (RUARO, 2010b). Com a minimização das margens de transição em forma de ondas truncadas e de uma melhor separação do sinal de pequena amplitude de um sinal de grande amplitude com frequências muito próximas uma da outra consegue-se um aumento de informações espectrais, efeito que é possível devido ao janelamento do sinal (BRAGA, 2006).

Ruaro (2010b) compara o janelamento do sinal de voz com os *frames* em um vídeo, os mesmos representam estaticamente algum momento de imagem e seu conjunto corresponde a um vídeo em si. Existem várias técnicas existentes de janelamento como a *Retangular, Bartlett, Blackman, Hanning, Welch.* A mais utilizada em sistemas de reconhecimento de voz é a janela de *Hamming* do inglês *Hamming Window* que é uma versão modificada da janela de *Hanning.* Sua função é suavizar as bordas de cada segmento que por consequência da segmentação podem conter variações abruptas do sinal. Essa suavização é realizada pelo próprio aspecto da janela e também pela sobreposição entre janelas. Sua função de aplicação é dada pela seguinte equação, dada por Oliveira (2002):

$$Hw[n] = \begin{cases} 0,54 - 0,46\cos\left(\frac{2\pi n}{n-1}\right), & 0 \leq n \leq N-1 \\ 0, & n > N-1 \end{cases} \tag{4.4}$$

Janelas de Hamming sobrepostas podem ser vistas na Figura 22.

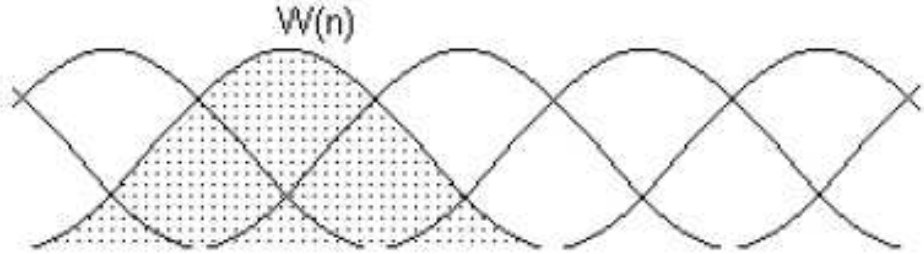

Figura 22: Janelas de Hamming (quadros) sobrepostas (MULATINHO; MESQUITA, 2011).

Basicamente a aplicação da janela a um sinal do domínio de tempo corresponde a multiplicação do sinal pela função da janela representada. A sobreposição das janelas podem variar entre 0% a 70%. Quanto mais alta a sobreposição mais suave a transição dos parâmetros extraídos, porém, estimativas amplamente suavizadas podem ocultar variações reais do sinal e caso a última janela ultrapassar os limites do sinal, deve-se completar com zeros até o final da janela. O tamanho da janela utilizada neste trabalho foi de 256 amostras por janela, com 50% de sobreposição.

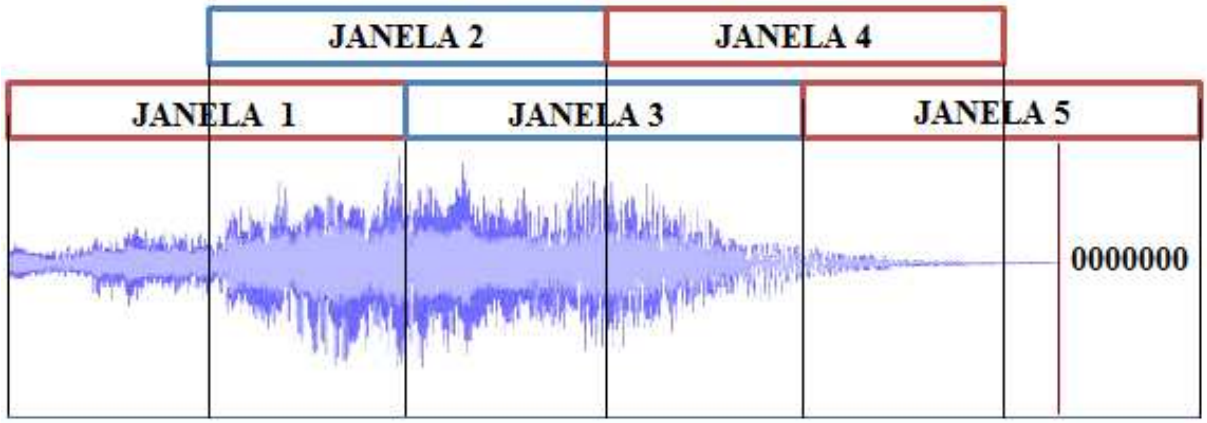

Figura 23: Janelas de Hamming com sobreposição de 50% (RUARO, 2010b).

4.2.6 Transformada Rápida de Fourrier

Segundo (BRESOLIN, 2008) a análise de *Fourier* é uma técnica com enorme número de aplicações. Além de ser muito utilizada no cálculo numérico, ciências aplicadas, engenharias a análise de *Fourier* se tornou a base do processamento de sinais. Transformada rápida de *Fourier* do inglês *Fast Fourier Transform*(FFT) são vários algoritmos combinados para otimizar o cálculo da Transformada Discreta de *Fourier* do inglês *Discrete Fourier Transform (DFT)* e sua inversa. A *DFT* é calculada no sinal depois da etapa de janelamento e é utilizada para extrair o conteúdo da frequência(espectro) do quadro corrente, em cada quadro é aplicado uma FFT *(Fast Fourier Transform)*, obtendo assim seu espectro, que é passado por um conjunto de filtros triangulares na escala Mel. Em cada janela é estabelecido um vetor onde cada índice representa uma escala de frequências baseada na faixa de frequências de amostragem definida na captura do sinal de voz, desse vetor procura-se a maior amplitude e com o índice desse valor é possível identificar a frequência da janela analisada (RUARO, 2010a).

4.3 Extração de parâmetro

Na etapa da extração de parâmetros busca-se extrair as informações mais importantes do sinal acústico, como informações que possam ser utilizadas para identificar diferenças fonéticas, além de diminuir o custo computacional na fase de reconhecimento. Devido a essa extração de informações normalmente não é possível recuperar o sinal original. Neste caso usa-se o *MFCC* a partir das informações da transformada de *Fourier*.

4.3.1 Parâmetros MFCC

O método escolhido para gerar os coeficientes com as características foi o *Mel-Cepstrais (MFCC do inglês Mel-Frequency Cepstral Coefficients)* que é muito utilizado em sistemas de *RAF*.

Em Siqueira e Alcaim (2011) é mostrado que um sinal senoidal de 880 Hz não soa duas vezes mais agudo que um de 440 Hz e nem quatro vezes mais agudo que um de 220 Hz. Ou seja, a escala em Hertz não reflete bem a percepção auditiva humana. Para representá-la melhor, foi criada a escala mel. Com os experimentos de diversos ouvintes foi criada a convenção de que f (em Hertz) para m (em mel):

$$m = M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) (1)$$

$$f = M^{-1}(m) = 700\left(\exp^{\frac{m}{1125}} - 1\right) (2)$$

A nova medida se mostrou muito eficaz para extrair dados para o reconhecimento de voz e então a técnica de extração de parâmetros *MFCC* baseia-se no uso do espectro de som alterado segundo a escala Mel.

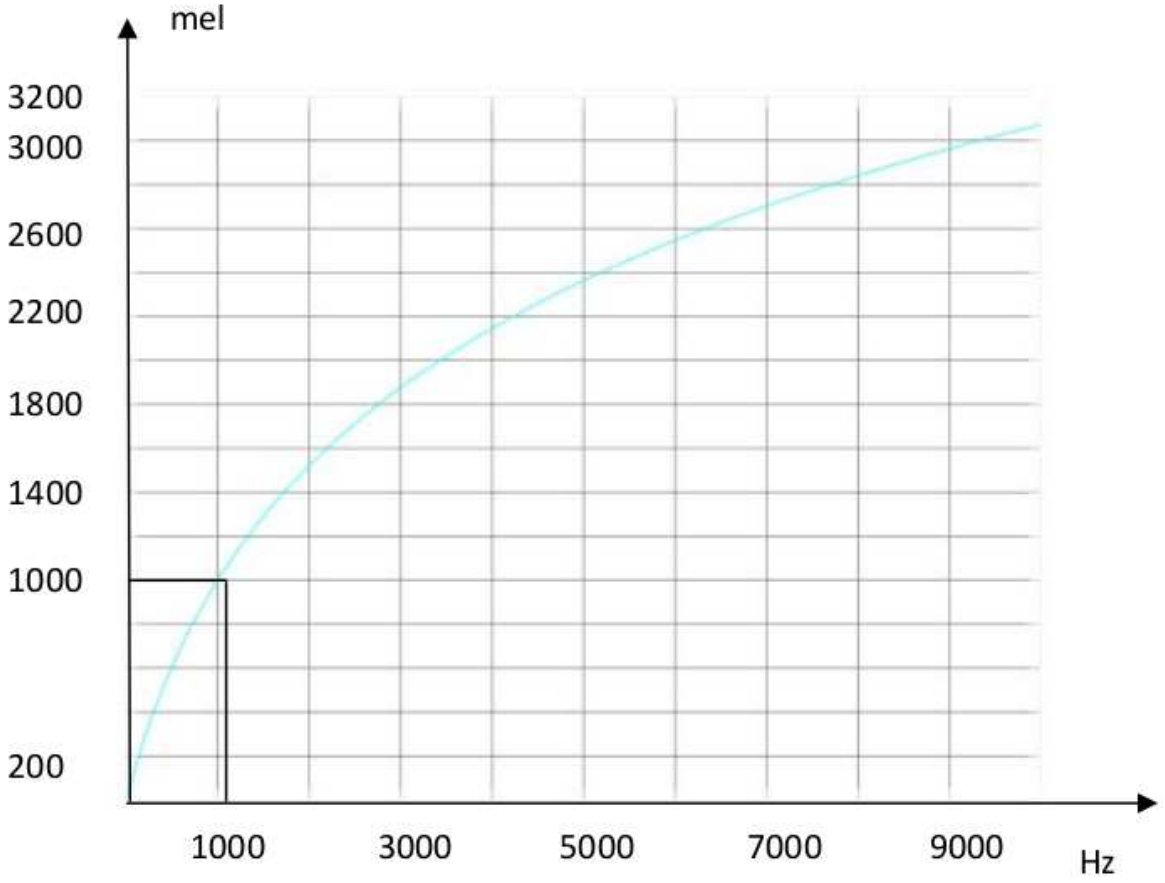

Figura 24: Escala de Frequência Mel (ALVARENGA, 2012).

Em Rabiner (1993) é mostrado que para obtenção dos coeficientes *MFCC*, filtra-se cada janela de espectro de potencias por um banco de filtros triangulares na escala Mel. Normalmente são usados 20 filtros passa-banda igualmente espaçados. Sendo 10 filtros uniformemente espaçados no eixo da frequência até 1000 Hz e acima de 1000 Hz as faixas são distribuídas segundo uma escala logarítmica. A Figura 25 apresenta um sinal com filtros triangulares.

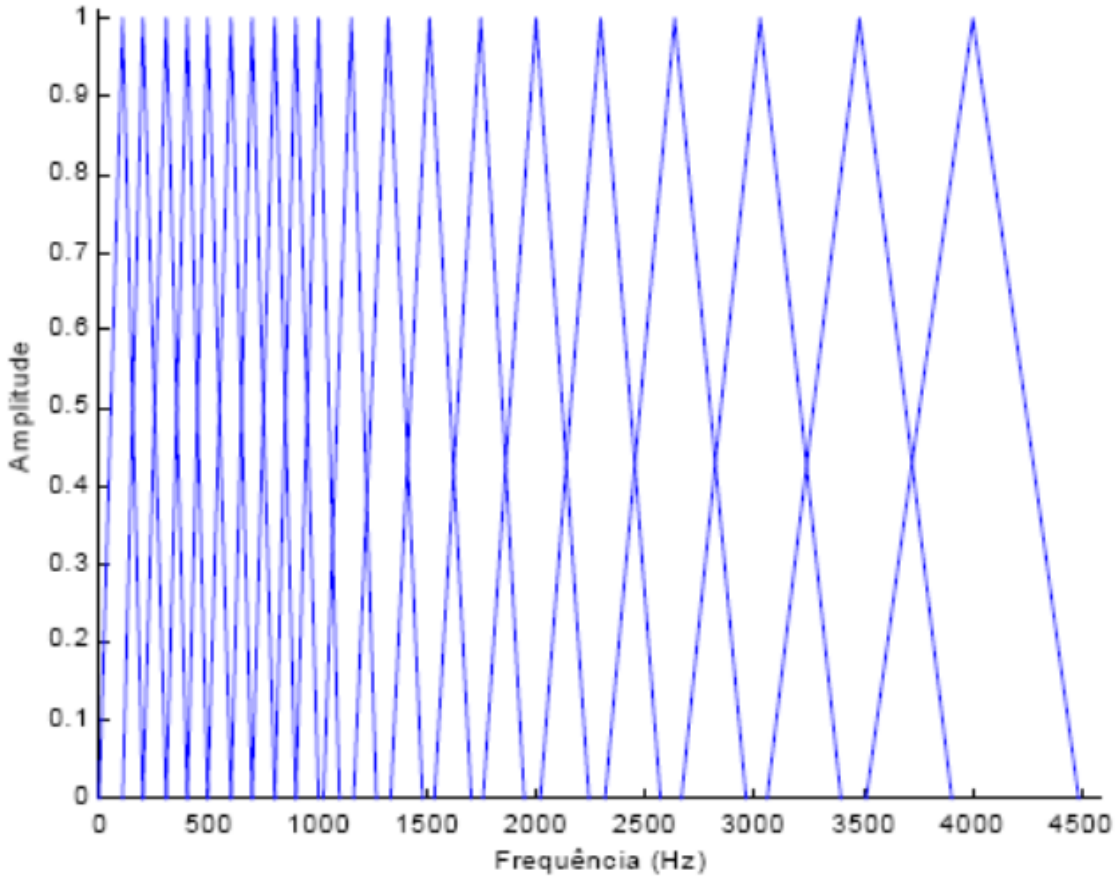

Figura 25: Banco de 20 filtros na escala Mel (SILVA, 2009).

Definidos os filtros, os atributos *MFCC* são obtidos para cada quadro do sinal com as seguintes etapas:

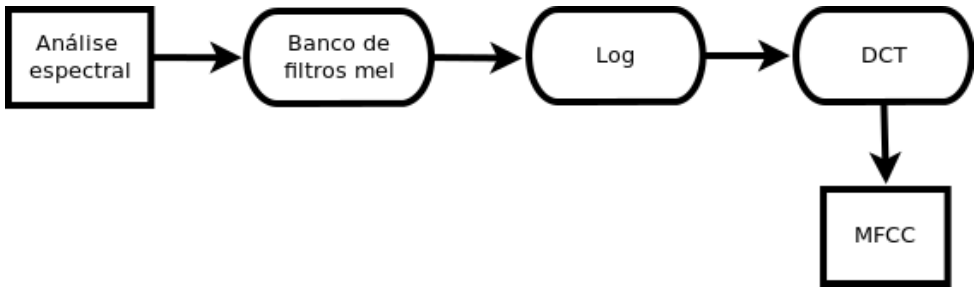

Figura 26: Etapas da extração dos parâmetros

1. Na análise espectral a transformada discreta de Fourier é aplicada a cada quadro resultante do janelamento de hamming, obtendo-se o espectro de potências, que são os parâmetros mais úteis do sinal no domínio de frequências ao invés do domínio do tempo permitindo uma distinção mais detalhada da composição fonética do sinal de som (SILVA, 2009).

2. O espectro é dividido em bandas através dos filtros triangulares na escala Mel. Utilizando 20 filtros no formato triangular passa-faixa, sendo 10 filtros uniformemente espaçados no eixo da frequência até 1 kHz e acima de 1 kHz as faixas são distribuídas segundo uma escala logarítmica, como mostrado na Figura 25.

3. Calcula-se o logaritmo da energia de cada banda, já que esse tipo de não-linearidade é observado no sistema auditivo humano.

4. Por último, a transformada discreta do cosseno (DCT) é aplicada à sequência de logaritmos do item anterior, a fim de descorrelatá-los, gerando os 12 coeficientes *MFCC*. Em sistemas de *RAF* normalmente são descartados alguns dos últimos coeficientes *MFCC*, pois isto provoca uma suavização do sinal. Segundo Silva (2009) em geral são mantidos menos de 15 coeficientes. O primeiro coeficiente é função da soma das energias de todos os filtros e também não costuma ser utilizado, resultando em 11 coeficientes.

4.4 DTW

Como os coeficientes *MFCC* dos padrões de referência estão representados em um arquivo no cartão de memória do dispositivo móvel eles precisam ser recuperados, para facilitar o processo de recuperação foi utilizado uma biblioteca chamada de *GSON* que manipula a estrutura de dados JavaScript Object Notation *(JSON)*. JSON é em formato texto e completamente independente de linguagem, pois usa convenções que são familiares às linguagens C e familiares, incluindo C++, C#, Java, JavaScript, Perl, Python e muitas outras. Estas propriedades fazem com que JSON seja um formato ideal de troca de dados (JSON, 2013). Como resultado do JSON é retornado uma lista de objetos contendo os coeficientes *MFCC* e o nome da cor no atributo palavra de todas as amostras criadas como referência, para utilização da orientação a objetos do Java os coeficientes *MFCC* estando em forma de objetos facilita a manipulação dos dados.

```
1 public class Coeficiente {
2     private float[] mFCC;
3     private String palavra;
4 }
```

Com essa lista de coeficientes é chamado o método que calcula a distância *Euclidiana* para cada objeto e a elocução teste, somando o valor de cada distância em relação a cada padrão. Essas distâncias são organizadas em outra lista em ordem crescente, com

as distâncias e a palavra que referencia a cor do padrão. O primeiro elemento da lista é o padrão com a menor distância, e esse padrão representa a cor reconhecida. Como a técnica *DTW* apresenta a limitação de sempre oferecer uma menor distância mesmo se a elocução pronunciada não estiver presente no dicionário do sistema é necessário aplicar uma outra técnica na lista de distâncias, gerando coeficientes de seletividade o reconhecimento é validado ou descartado. Caso o reconhecimento não atingir um nível mínimo aceitável é retornado o valor "-1" indicando que não houve reconhecimento, caso contrário, é retornado palavra com a cor reconhecida. O método que valida o reconhecimento pode ser visto abaixo, onde x é o vetor com as distâncias encontradas.

```
1  private boolean isValidRecognize(float[] x) {
2
3          if (x.length == 1) {
4              if (x[0] > 10) {
5                  return false;
6              }
7          } else {
8              float s1 = (x[1] - x[0]) / x[0];
9              float s2 = (RecognizerUtil.getMedia(x) - x[1]) / x
                   [0];
10             if (s1 <= 0.1) {
11                 return false;
12             }
13             if (s1 >= 0.25 || s2 >= 0.8) {
14                 return true;
15             } else {
16                 return false;
17             }
18         }
19         return true;
20     }
```

4.5 Aplicação desenvolvida

A aplicação para testes de reconhecimento foi desenvolvida para o sistema operacional *Android* superiores a versão 3.0 *(Honeycomb)*, e foi testada nos seguintes dispositivos:

- Celulares: Samsung Galaxy X, Samsung Galaxy Nexus 4
- Tablet: Motorola Xoom 2 ME

A primeira tela disponível para o usuário pode ser vista na Figura 27 onde possui as funções básicas do sistema.

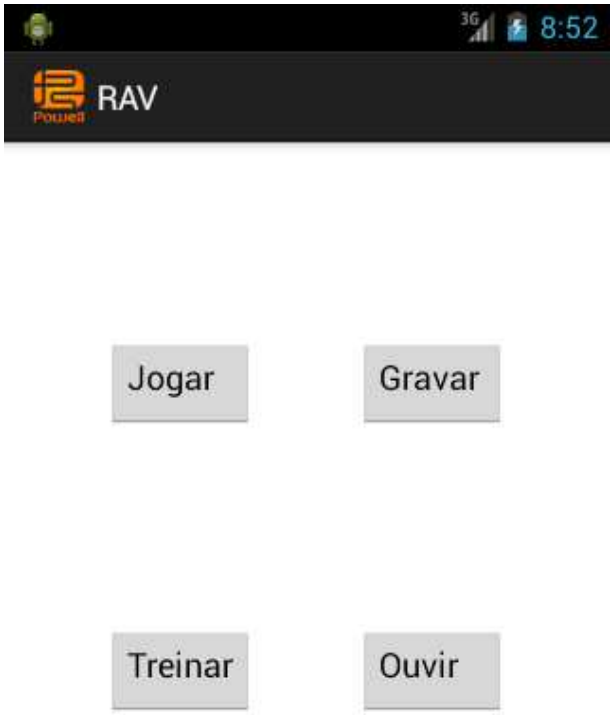

Figura 27: Menu principal da aplicação.

4.5.1 Gravação

A primeira etapa do sistema é a aquisição do sinal de voz, essa função é encontrada no sistema pela função *gravar* no menu, onde são capturadas as amostras que serão usadas como padrões. São capturados a sequência de cores que serão usadas pela aplicação, gerando no cartão de memória um arquivo no formato *WAV*. Esse arquivo tem o tempo máximo de 2 segundos, então no intervalo de 0s a 2s é necessário pronunciar o nome da cor como mostrado na Figura 28.

Figura 28: Gravando padrões para o treinamento.

4.5.2 Ouvir

A função **ouvir** no menu é responsável por listar os padrões salvos no cartão de memória para treinamento, sendo possível removê-los como mostrado na Figura 29.

Figura 29: Ouvindo elocuções de treino.

4.5.3 Treino

Com os arquivos criados na fase de gravação é possível gerar os padrões a serem comparados com elocução de teste na função *treinar* do menu. Nesta etapa, o arquivo de fala gravado no cartão de memória passa pelas fases de pré-processamento e extração de características que já foram apresentadas nos capítulos anteriores, sendo salvo os coeficientes *MFCC* do padrão treinado em um arquivo de texto com o nome da cor. Dependendo da quantidade de elocuções criadas para treino, o treinamento pode demorar um pouco, nesse tempo é mostrada uma mensagem para o usuário que pode ser vista na Figura 30.

Figura 30: Treinando amostras.

Quanto mais padrões para cada cor maiores esses arquivos de texto ficarão, podendo gerar problemas de desempenho. Qualquer variação no ambiente pode alterar o resultado dos padrões então é necessário treinar novos padrões para variados ambientes. No final do treinamento é mostrado a seguinte mensagem que pode ser vista na Figura 31.

Figura 31: Treino concluído.

4.5.4 Jogo

Na função *jogar* são listadas todas as cores disponíveis no dicionário da aplicação como pode ser visto na Figura 32 e o usuário deve escolher uma cor para ser mostrada na tela.

Figura 32: Lista de cores definidas no dicionário.

O objetivo da aplicação é apenas reconhecer a palavra pronunciada, por isso a simplicidade do jogo. Quando o usuário clicar na opção **falar**, um tempo de 2 segundos é fornecido para pronuncia do nome da cor escolhida, o arquivo resultante dessa gravação é chamado de elocução de teste, passando pelo mesmo processo do treinamento com o pré-processamento e a extração de características, o resultado desses processos é comparado com os padrões criados no treinamento usando a distância euclidiana entre os vetores, o nome desse método é DTW. Geralmente a menor distância entre a elocução de teste e os padrões é a resposta correta para cor, mas quando é dita outra palavra fora do dicionário de padrões também se encontra uma distância mínima, para contornar esse problema é feito um cálculo de seletividade, validando os falsos resultados. Caso o sistema reconheça a elocução de teste como um padrão do sistema, é mostrado na tela a cor pronunciada, como apresentado na Figura 33.

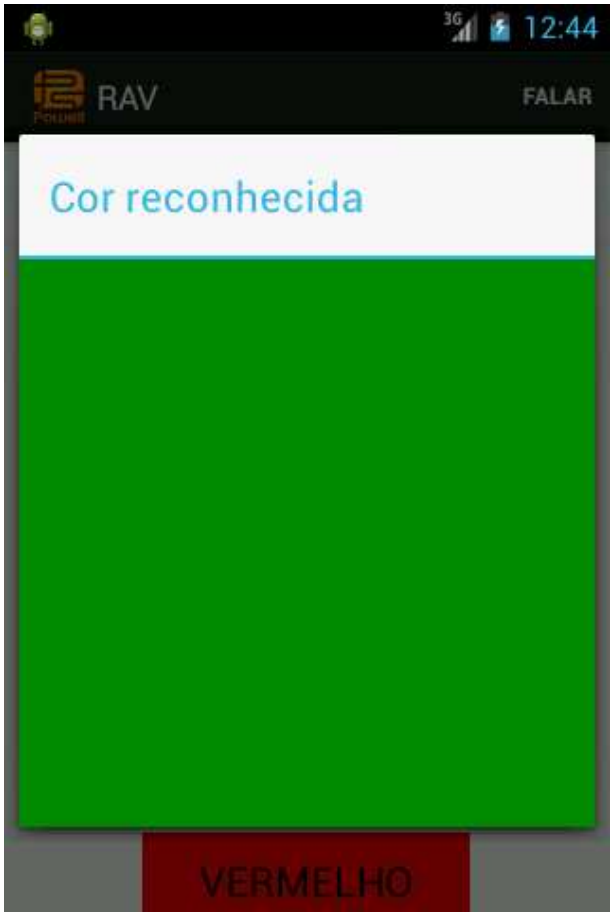

Figura 33: Cor reconhecida.

Clicando no botão de finalizar é mostrado um *placar* com as cores que foram reconhecidas pelo sistema, como na Figura 34.

Figura 34: Resultado do reconhecimento.

4.6 Resultados

Foram feitos testes para o reconhecimento de seis palavras: **amarelo, azul, branco, preto, verde, vermelho**. Como o sistema funciona em dispositivos móveis a variação do ambiente é um fator que dificulta o reconhecimento já que se um padrão for gravado em um ambiente silencioso e a elocução teste for dita em um ambiente ruidoso o reconhecimento não será eficiente. Outra questão importante no reconhecimento é a questão do tempo de pronunciamento, o sistema oferece 2 segundos para elocução das palavras, tempo mais que suficiente para pronunciar cada palavra isoladamente, o problema ocorre quando existe um retardamento na pronuncia da palavra, na Figura 35 a palavra **Amarelo** é dita na forma correta.

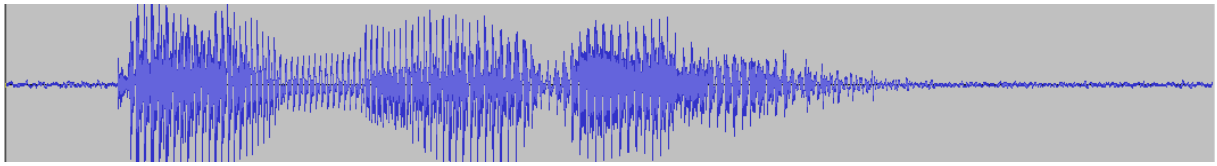

Figura 35: Espectro da palavra amarelo.

Já na Figura 36 a palavra foi pronunciada com atraso, gerando a falsa impressão para o usuário que ela foi capturada de forma correta, mas como o tempo de 2 segundos já havia se esgotado o sinal sonoro fica deformado em relação a amostra correta pois uma pequena parte do sinal foi capturada.

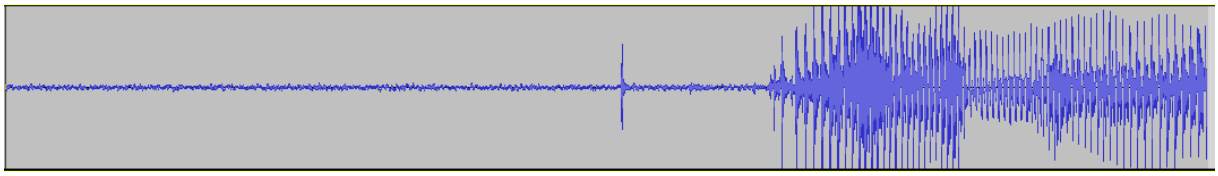

Figura 36: Espectro da palavra amarelo com a fala retardada.

O cálculo de distância entre essas 2 amostras dificilmente apontará **Amarelo** como a resposta certa, já que uma pequena parte da amostra pronunciada com retardo será analisada. Nos testes realizados as palavras pronunciadas com retardo geralmente são associados as menores amostras como *preto* ou *azul*.

O sistema proposto tinha como característica ser independente do locutor, porém essa característica não foi alcançada, devido a necessidade de inúmeros treinamentos com vários locutores e o risco da aplicação ter seu desempenho comprometido, já que as limitações de hardware são evidentes em dispositivos móveis.

As elocuções para treinamento e para teste foram feitas em ambiente com ruídos, fator que compromete o desempenho da aplicação, abaixo serão mostradas os sinais de alguns padrões usados no treinamento, a amplitude do sinal varia de -1 a 1 e o intervalo de tempo real das palavras é de no máximo 1 segundo:

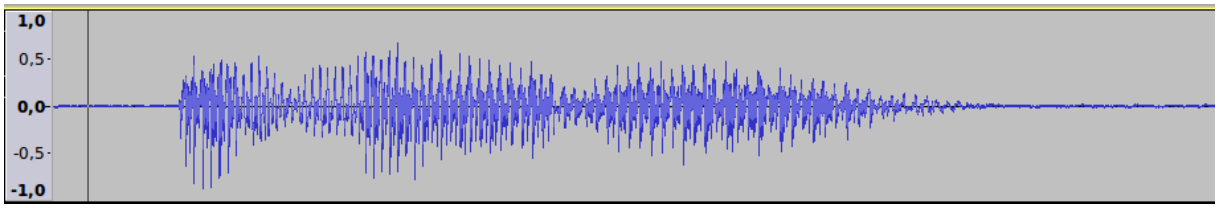

Figura 37: Amarelo.

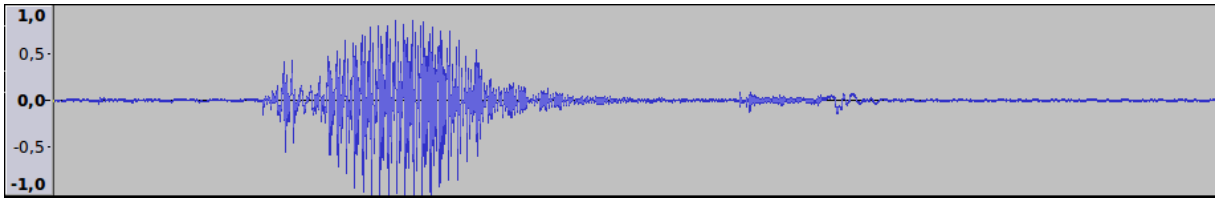

Figura 38: Preto.

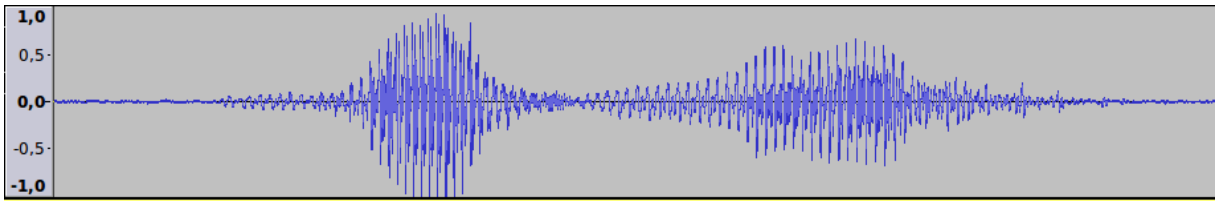

Figura 39: Vermelho.

Essas amostras foram geradas no mesmo ambiente então sofreram dos mesmo ruídos e interferências. O sinal de voz varia a cada gravação, devido a velocidade da pronúncia, acentuação etc. Na Figura 40 é apresentada uma elocução de teste com a pronúncia da palavra vermelho, comparando visualmente com os padrões de referência, a amostra que mais se aproxima é a Figura 39, quando comparada pelo método DTW o resultado também é Figura 39, resultando em um reconhecimento correto.

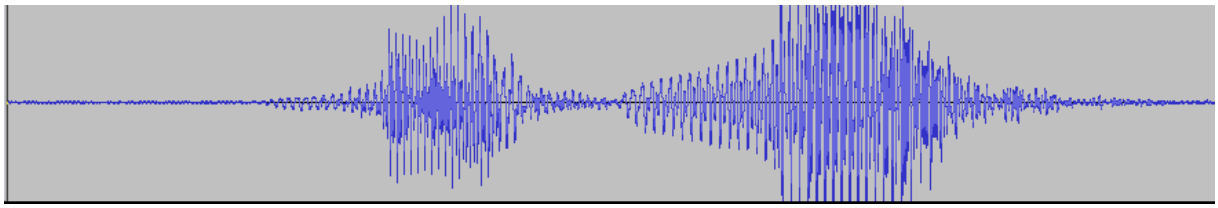

Figura 40: Elocução vermelho para teste.

Como exemplo de um teste mal realizado a Figura 41 representa a cor **Azul**, mas omitimos do dicionário do sistema a cor azul, o reconhecimento tenderia para o padrão da Figura 38, resultando em um reconhecimento errado se os coeficientes de seletividade não existissem.

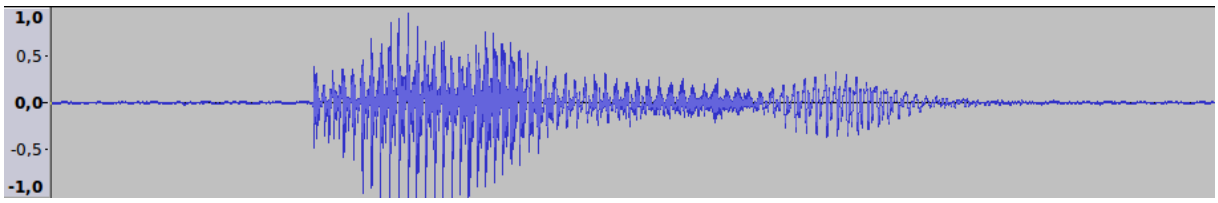

Figura 41: Elocução azul para teste.

Os resultados do reconhecimento foram abaixo do esperado, já que todas as técnicas propostas foram as mais utilizadas nos trabalhos referentes. Na Tabela 3 é apresentado resultados para execução do sistema treinado por 4 pessoas, 2 do sexo feminino representadas por **M1** e **M2** e 2 do sexo masculino representados por **H1** e **H2**, cada locutor repetiu a elocução teste para cada palavra 5 vezes, obtendo:

	Amarelo	Azul	Branco	Preto	Verde	Vermelho
H1	80%	60%	60%	60%	60%	80%
M1	60%	40%	40%	40%	40%	60%
H2	40%	60%	40%	60%	40%	40%
M2	40%	40%	40%	60%	40%	60%

Tabela 3: Tabela com resultados do reconhecimento.

4.7 Dificuldades

Cada etapa do sistema de reconhecimento automático de voz abrange grandes áreas de estudos criando uma variedade muito grande de combinações das técnicas de implementação, dificultando a escolha desses métodos. A maioria dos estudos na área abordam e usam como base a teoria ou sistemas já prontos, dificultando a implementação dos métodos na criação de novos sistemas. Existem muitos estudos na área de reconhecimento de voz, mas poucos mostram detalhadamente as técnicas usadas. Como sistemas móveis podem estar presentes nos mais variados ambientes cria-se um grande problema na precisão do reconhecimento, então é preciso treinar o sistema para os diferentes ambientes melhorando assim a precisão do reconhecimento. Os algoritmos utilizados para detecção de extremos não funcionaram perfeitamente e foi preciso fazer alterações para melhorar o seu funcionamento. Existem uma grande variação de hardware em dispositivos *Android* e o fato da aplicação ser rodada em qualquer ambiente dificulta a precisão de reconhecimento.

5 Conclusão

O trabalho realizado tinha como objetivo a pesquisa e desenvolvimento de um sistema de reconhecimento de voz, tendo como enfoque um sistema de reconhecimento de palavras isoladas. Além de resultar em uma plataforma inicial facilitando trabalhos futuros também foi obtido conhecimento teórico e prático na área de reconhecimento da fala.

Nos capítulos iniciais são mostrados as aplicações de sistemas de *RAF*, histórico e características dos sistemas de reconhecimento. Nos capítulos seguintes são mostrados os métodos de implementação do sistema.

Na aquisição do sinal de voz fora utilizados métodos específicos para captura do som em dispositivos móveis com sistema operacional Android criando um arquivo *.WAV* na memória do dispositivo, foram abordadas técnicas de remoção de ruído, detecção de extremos da fala, pré-enfase, janelamento do sinal em quadros, extração de características utilizando *MFCC* e reconhecimento dos padrões. As técnicas utilizadas na aquisição do sinal de voz, pré-processamento e extração de características deste projeto são as técnicas mais utilizadas na construção de sistemas de reconhecimento de voz.

A técnica utilizada para comparação de padrões foi a *DTW*, que se mostra eficiente apenas para um vocabulário pequeno, já que é feita uma comparação de distâncias entre cada padrão criado anteriormente e a elocução de teste, ou seja, um vocabulário grande geraria muitos padrões para serem comparados diminuindo a eficiência e desempenho do sistema. Como esse trabalho tem como proposta uma introdução as técnicas de reconhecimento de *RAF* o resultado da implementação foi satisfatório.

Foi construído um sistema de *RAF* em java, linguagem necessária para criação de aplicativos no sistema operacional *Android*, o sistema pode ser considerado dependente de locutor, já que um número considerável de amostras teria de ser criado no treinamento para possibilitar a utilização de qualquer locutor, diminuindo a eficiência do sistema e com reconhecimento para palavras isoladas pois o objetivo da aplicação é reconhecer comandos.

5.1 Trabalhos futuros

Devido a complexidade do trabalho não foi possível desenvolver todas as funcionalidades desejas como: independência do locutor, classificação utilizando *HMM*, taxa de acerto maior que 90%, mas é possível uma implementação futura, além de:

- Testar combinações de valores na fase de aquisição de voz, como taxas de amostragem diferentes, número de canais, número e taxa de sobreposição das janelas;
- Os arquivos de voz gerados para treinamento e execução do jogo são salvos no cartão de memória no formato *PCM* que possui como subclasse o formato de arquivo de formato **wave** *(WAV)*, ocupando muita memória e diminuindo o desempenho;
- Modificar o algoritmo de detecção de extremos para melhor obtenção do sinal útil para o reconhecimento;
- Ajustar novos parâmetros para extração de características;
- Testar outros métodos de extração de características como: *SSCH* e *PNCC*;
- Mudar o método de reconhecimento de padrões para *HMM*, *RNA* ou um modelo híbrido para aumentar o tamanho do dicionário de palavras.

Referências Bibliográficas

ALVARENGA, R. J. *Reconhecimento de comandos de voz por redes neurais*. Monografia (Graduação), Taubaté-SP, 2012.

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.

BRAGA, P. de L. *Reconhecimento de voz dependente de locutor utilizando Redes Neurais Artificiais*. Monografia (Graduação) — UFPE, Recife, 2006.

BRESOLIN, A. de A. *Estudo do Reconhecimento de Voz para o Acionamento de Equipamentos Elétricos via Comandos em Português*. Monografia (Mestrado), Joinville, 2003.

BRESOLIN, A. de A. *Reconhecimento de voz através de unidades menores do que a palavra, utilizando Wavelet Packet e SVM em uma nova Estrutura Hierárquica de Decisão*. Monografia (Doutorado), Natal, 2008.

CARVALHO, P. R. N. *Sistema de Reconhecimento de Apito em Ambientes Ruidosos*. Monografia (Mestrado), Braga, 2008.

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

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

COLE, R. *Survey of the state of the art in human language technology*. [S.l.]: Cambridge University Press, 1997.

CUADROS, C. D. R. *Reconhecimento de voz e de locutor em ambientes ruidosos, comparação das técnicas MFCC e ZCPA*. Monografia (Graduação), Niterói, 2007.

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.

FARIAS, M. C. Métodos de codificação de voz, uma introdução. 2011.

FILHO, M. da C. S. *Classificação Automática de Gêneros de Áudio Digital*. Monografia (Graduação), Recife-Pe, 2006.

FILHO, R. B. Disciplina: Princípios de Comunicações I, *Princípios de Comunicações I*. 2013. Disponível em: <http://www.decom.fee.unicamp.br/˜baldini/EE881.htm>. Acesso em: 19/01/2014.

FURTUNĂ, T. F. Dynamic programming algorithms in speech recognition. 2008.

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

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

GRACA, E. *A divertida e melancólica atração virtual de Ela, filme que estreia em fevereiro*. O globo, 2014. Disponível em: <http://oglobo.globo.com/cultura/a-divertida-melancolica-atracao-virtual-de-ela-filme-que-estreia-e Acesso em: 13/01/2014.

INCE, A. N. *Digital speech processing: speech coding, synthesis, and recognition*. [S.l.]: Kluwer Academic Publishers, 1992.

JSON. *Introdução ao JSON*. 2013. Disponível em: <http://www.json.org/json-pt.html>. Acesso em: 20/01/2014.

LEMOS, D. R.; RODRIGUES, G. J.; HERNANDEZ, E. D. M. Reconhecedor de voz via redes neurais. 2004.

LOUZADA, J. A. *Reconhecimento automático de fala por computador*. Monografia (Graduação), 2010.

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

MULATINHO, G. M.; MESQUITA, L. http://www.acervodigital.unesp.br/handle/123456789/63524, *Reconhecimento automático de voz para aplicações em automação implementado em FGPA*. 2011. Disponível em: <http://www.acervodigital.unesp.br/handle/123456789/64571>. Acesso em: 20/01/2014.

OBLIVIONMOVIE2013. *Oblivion*. oblivionmovie2013, 2013. Disponível em: <http://www.oblivionmovie2013.com/>. Acesso em: 13/01/2014.

OLIVEIRA, K. M. de. *Reconhecimento de voz através de reconhecimento de padrões*. Monografia (Graduação), Salvador-BA, 2002.

PAULA, M. B. de. *Reconhecimento de palavras faladas utilizando Redes Neurais Artificiais*. Monografia (Graduação), Pelotas, 2000.

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

RABINER, L. R. *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.

RAMIRO, P. H. de O. *Sistema de acionamento de dispostivos comandado por voz*. Monografia (Graduação), Brasília, 2010.

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.

RUARO, M. Obtenção da frequência de um sinal de som por meio da fft em java me. Santo Angelo - RS, 2010.

RUARO, M. *SRM: Framework para Reconhecimento de Som em Dispositivos Móveis*. Monografia (Graduação), Santo Ângelo, 2010.

SAHA, G.; CHAKROBORTY, S.; SENAPATI, S. A new silence removal and endpoint detection algorithm for speech and speaker recognition applications. India, 2005.

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) — UFPA, Belém, 2010.

SIQUEIRA, J. K.; ALCAIM, A. Comparação dos atributos mfcc, ssch e pncc para reconhecimento robusto de voz contínua. 2011.

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