Daiana Nascimento Muniz

Estudo Sobre Reconhecimento de Áudio Repetitivo: Desenvolvimento de um Protótipo

São José – SC
fevereiro/2009

Daiana Nascimento Muniz

Estudo Sobre Reconhecimento de Áudio Repetitivo: Desenvolvimento de um Protótipo

Monografia apresentada à Coordenação do Curso Superior de Tecnologia em Sistemas de Telecomunicações do Instituto Federal de Santa Catarina para a obtenção do diploma de Tecnólogo em Sistemas de Telecomunicações.

Orientador:
Prof. Marcos Moecke, M.Eng

CURSO SUPERIOR DE TECNOLOGIA EM SISTEMAS DE TELECOMUNICAÇÕES
INSTITUTO FEDERAL DE SANTA CATARINA

São José – SC
fevereiro/2009

Monografia sob o título "Estudo Sobre Reconhecimento de Áudio Repetitivo: Desenvolvimento de um Protótipo ", defendida por **Daiana Nascimento Muniz** e aprovada em 26 de fevereiro de 2009, em São José, Santa Catarina, pela banca examinadora assim constituída:

_____

Prof. Marcos Moecke, M.Eng
Orientador

_____

Prof. Emerson Ribeiro de Mello, M.Eng
IFSC

_____

Prof. Marcio Henrique Doniak, M.Eng
IFSC

A imaginação é mais importante que o conhecimento.
Albert Eisntein

Agradecimentos

Aos meus pais, por todo apoio e o incentivo oferecido durante a minha vida escolar e acadêmica. Respeitando as minhas decisões e me apoiando nos momentos difíceis.

Aos meus amigos, por todos os momentos de descanso e lazer, pela paciência e principalmente pelas palavras que trouxeram esperança e me impulsionaram a continuar.

Aos diversos educadores do Instituto Federal de Santa Catarina pelo apoio, incentivo e pelas oportunidades.

E a todos aqueles que direta ou indiretamente contribuíram para viabilizar este trabalho.

Resumo

Neste trabalho foi desenvolvido um protótipo que realiza o reconhecimento da repetição de um trecho sonoro armazenado. Para a realização deste, houve o estudo de diversas técnicas de processamento de áudio e a análise de alguns ambientes de desenvolvimento. O protótipo utiliza os atributos do som obtidos pela análise de predição linear (*Linear Predictive Coding*) para realizar o reconhecimento. Os métodos de decisão utilizados para o reconhecimento do padrão acústico geram duas variantes para o protótipo, estas foram: a que utiliza a distância vetorial de Itakura-Saito e a que utiliza a distância vetorial euclidiana. A primeira variante apresentou resultados aquém do esperado e a segunda variante detectou repetição em até 88% dos casos.

Abstract

The present study developed a prototype that performs recognition of sound's repetitions. To achieve this goal, the study of different techniques for audio processing and analysis of some development environments were conducted. After that, the algorithm was implemented using the attributes of the sound obtained by the analysis of Linear Predictive Coding to achieve recognition. The decision-making method used for the recognition of standard acoustic, created two variants to the prototype: (1) to use the distance vector of Itakura-Saito and (2) to use the Euclidean distance vector. The first variant presented results below the expected and the second variant detected recognition in 88% of the cases.

Sumário

Lista de Figuras **10**

Lista de Tabelas **11**

1 Introdução **12**

1.1 Objetivo geral 13

1.2 Motivação 14

1.3 Conteúdo da dissertação 14

2 Fundamentação teórica **15**

2.1 O som e suas características 15

2.2 Conversão analógica/digital do som 16

2.3 Padrões de codificação e compressão de áudio 17

2.4 Reconhecimento de áudio 20

2.5 Extração de parâmetros 22

2.6 Algoritmos de reconhecimento de padrões acústicos 28

3 Áudio Repetitivo **33**

3.1 Estudo dos ambientes de desenvolvimento para processamento de áudio 33

3.2 Sistema de detecção de áudio repetitivo 36

4 Resultados **42**

4.1 Resultados dos testes nos ambientes de desenvolvimento 42

4.2 Metodologia de testes para a aplicação desenvolvida 43

4.3 Testes utilizando a distância de Itakura-Saito 45

4.4 Testes utilizando a distância de euclidiana 48

5 Conclusões **52**
5.1 Trabalhos futuros 53

Lista de Acrônimos **54**

Referências Bibliográficas **55**

Lista de Figuras

Figura 1 – Digitalização do sinal do som. 17
Figura 2 – Sistema genérico de reconhecimento de áudio. 20
Figura 3 – Modelo da produção da fala a partir da predição linear. 24
Figura 4 – Alinhamento não-linear de áudios utilizando DTW 29
Figura 5 – Fluxograma do código *DETECTA_REPETIÇÃO*. 39
Figura 6 – Fluxograma do código *ANALISA_AUDIO*. 40
Figura 7 – Exemplo de áudio de referência 44
Figura 8 – Exemplo de áudio objeto. 44
Figura 9 – Exemplo utilizando distância Itakura-Saito 46
Figura 10 – Relação entre número de coeficientes LPC e média das distâncias mínimas encontradas para distância de Itakura-Saito. 47
Figura 11 – Exemplo utilizando distância euclidiana: 49
Figura 12 – Relação entre número de coeficientes LPC e média das distâncias mínimas encontradas para distância euclidiana. 50

Lista de Tabelas

Tabela 1: Parâmetros utilizados nos testes com distância de Itakura-Saito. 46
Tabela 2: Parâmetros utilizados nos testes com distância euclidiana. 49

1 Introdução

A linguagem oral é a forma mais antiga de comunicação, sendo esta, ainda a maneira mais rápida e fácil de comunicação entre os humanos. Além disso, a linguagem oral natural foi a primeira motivação para a criação de sistemas de telecomunicações, que começaram inicialmente escritos, e no final do século XIX se tornaram orais devido ao feito de Alexander Graham Bell ao transmitir voz através de sinais elétricos, e pela invenção do rádio por Guglielmo Marconi (BERGAMO, 2007).

Os primeiros estudos relacionados com a gravação de áudio datam do início do século XIX, realizado por Thomas Young, seu trabalho conseguia gravar a forma da onda sonora, porém não conseguia reproduzi-la. Somente em 1877, Thomas Alva Edson conseguiu reproduzir um som gravado (SOUND, 2008).

A partir dos estudos de Graham Bell, Marconi e Thomas Edson a evolução de técnicas para transmitir e armazenar áudio foram aprimoradas ao longo dos tempos . Primeiramente, o áudio era armazenado apenas de modo analógico (em discos de vinil ou fita plástica). Porém, o som gravado no modo analógico ocupava muito espaço físico para ser armazenado, além de usar uma ampla largura de banda na radiodifusão e necessitar de um circuito exclusivo na telefonia. Até então, a transmissão do som analógico ocorria através de eletricidade (no sistema telefônico), radiotransmissão (para o sistema de radiodifusão) ou meios físicos gravados (como discos e fitas).

Apenas na década de 1970 começaram os estudos envolvendo a digitalização do áudio (DIGITAL, 2008). A digitalização do áudio também proporcionou novos meios para armazenar e reproduzir tal informação. O que possibilitou a ocupação um menor espaço físico para armazenamento, opções de multiplexação e menor utilização de largura de banda.

Com a digitalização do sinal de áudio, a evolução dos métodos de armazenamento e transmissão ocorreu de maneira acelerada, gerando novas técnicas para compressão de áudio, melhor qualidade sonora sistemas de transmissão, e também a possibilidade de reconhecimento de sons e palavras. Entre as principais técnicas de processamento dos sinais

de áudio estão as seguintes etapas: representação digital do sinal; análise e extração de características do sinal; desenvolvimento de modelos de síntese; e, reconhecimento.

Reconhecer sons consiste em analisar pequenos trechos sonoros e encontrar padrões de repetição, sem analisar seu conteúdo, apenas sua sonoridade. O reconhecimento de palavras trabalha para reconhecer fonemas, sílabas ou palavras de modo a recuperar a mensagem falada. Já, compreender a fala consiste em entender consensualmente o significado da mensagem, visando fazer com que o sistema reconhecedor execute algo. Assim, esta distinção é importante, pois notar a habilidade de um sistema responder inteligentemente à fala é um critério muito significativo para avaliação dos sistemas de processamento de áudio.

Neste trabalho é abordado o reconhecimento do som que também conhecido como reconhecimento de padrões acústicos. A abordagem limita o reconhecimento de padrões acústicos ao reconhecimento de repetições acústicas. Sendo estudadas as técnicas utilizadas para tal, sem levar em consideração a síntese, reconhecimento de palavras ou compreensão de fala.

Portanto, neste trabalho será demonstrado como funciona um protótipo que realiza o reconhecimento de um determinado áudio que se repete, sendo este áudio um som qualquer. Este reconhecimento é feito com base em um algoritmo que realiza a análise do som, compara com um padrão pré-definido e decide se houve ou não o reconhecimento.

1.1 Objetivo geral

A proposta deste trabalho foi desenvolver um protótipo que realizasse o reconhecimento de áudio repetitivo. A construção deste protótipo foi fundamentada no conhecimento adquirido através do estudo de diversas técnicas de processamento de áudio e na análise de diversos ambientes de desenvolvimento.

Tendo também como o objetivo a aprendizagem das diversas técnicas relacionadas ao processamento de áudio, assim como a familiarização com diversos ambientes de desenvolvimento específicos para manipulação de áudio.

1.2 Motivação

A motivação para este estudo é a possibilidade de agregar conhecimento com relação a todas as técnicas estudadas para o processamento de áudio, em especial as técnicas utilizadas para o reconhecimento sonoro.

Atualmente reconhecimento de áudio repetitivo é utilizado como melhoria em diversas soluções, entre estas estão: aplicação nos serviços de publicidade (conferindo a periodicidade e duração de apresentações de propagandas), tradução voz-a-voz (*speech-to-speech*), unidades de resposta audíveis (URA), serviços de identificação pessoal, e possível treinamento fonético para deficientes auditivos, entre outros.

1.3 Conteúdo da dissertação

Este capítulo apresentou a introdução do trabalho proposto, o objetivo e motivação para o mesmo. O capítulo 2 descreve a fundamentação teórica envolvida no reconhecimento do som, analisando o áudio e suas principais características, a digitalização, a codificação, a extração de parâmetros e o reconhecimento de padrões.

O capítulo 3 apresenta o protótipo de reconhecimento de repetição sonora. Iniciando com os testes realizados nos diversos ambientes de desenvolvimento até a construção do protótipo de reconhecimento de repetição sonora. Detalhando seu principio de funcionamento e organização do algoritmo.

O capítulo 4 apresenta os testes e resultados obtidos com o algoritmo implementado. Finalmente no capítulo 5 são sumarizadas as conclusões e sugestões para trabalhos futuros. No final do trabalho é apresentada uma listagem dos acrônimos e abreviaturas utilizados ao longo do texto com o objetivo de facilitar a leitura.

2 *Fundamentação teórica*

2.1 O som e suas características

O som é uma das formas de energia que o universo apresenta, e como tal obedece às leis fundamentais da física. Bonjorno(1992) e Soares(1984) descrevem o som como aumentos e reduções periódicas da densidade do ar, ou seja, a compressão e rarefação do mesmo. A definição da ABNT(1959) para o som é: "toda e qualquer vibração ou onda mecânica em um meio elástico dentro da faixa de audiofrequência". Resumindo o som apresenta as seguintes características básicas:

- Amplitude: valor instantâneo do nível de sinal.
- Frequência: quantidade de vezes na unidade de tempo em que a forma de onda do sinal se repete. Para os sinais compostos por mais de uma frequência, normalmente é levado em consideração dois valores distintos de frequência, um deles é o menor valor de frequência com alta amplitude (frequência fundamental), e o outro é a maior frequência do sinal (utilizado para delimitar a filtragem).
- Fase: ângulo em que o sinal se apresenta.

A voz é um tipo de som, logo, possui todas as características do mesmo. Do mesmo, pela definição de Ferreira(1986, p.1473), "consiste em um som ou um conjunto de sons emitidos pelo aparelho fonador". A voz tem como características principais à frequência fundamental, a intensidade e a qualidade vocal. A frequência fundamental é o menor valor de frequência que apresenta alta amplitude, formando o primeiro "pico" no espectro de frequências. A intensidade vocal é resultado da amplitude das vibrações das cordas vocais (BLOCH, 1958). E, a qualidade vocal, também chamada de timbre, é o termo atualmente empregado para designar o conjunto de características que identificam a voz, estando relacionada à composição de todas as frequências da onda sonora.

Além de tais características o sinal sonoro pode ser classificado de acordo com sua excitação vocal. A excitação vocal representa a maneira com que o áudio é formatado pelo trato vocal. Portanto pode apresentar:

- Excitação vozeada: para esta, a fonte glotal deve estar fechada, o ar força a fonte a abrir e fechar periodicamente. Gerando uma sequência de impulsos periódica, que define a frequência fundamental;
- Excitação não-vozeada: a fonte glotal está aberta e o ar passa livremente até a boca. Isto resulta em uma turbulência que gera um sinal ruidoso. A forma espectral deste sinal define o filtro resultante do trato vocal;
- Excitação transitória: o fechamento da garganta ou da boca aumenta a pressão do ar, ao repentinamente abrir o caminho obstruído, a pressão do ar cai, gerando uma rajada plosiva.

2.2 Conversão analógica/digital do som

Para a transformação do som em sinal elétrico é necessário que este seja capturado por um transdutor. Este consiste em um dispositivo que é excitado pela onda mecânica sonora, gerando na saída de seus terminais uma corrente elétrica proporcional ao volume do som. O sinal elétrico obtido nesse processo é considerado analógico por apresentar variações contínuas no tempo e em amplitude, correspondentes as pressões sonoras.

A digitalização consiste na restrição do sinal analógico, este é um sinal contínuo com infinitas possibilidades de amplitude. Esta restrição gera um sinal discreto com finitas possibilidades de amplitude. Este procedimento é realizado ao utilizar um amostrador e com um quantizador. O primeiro é responsável pela discretização no tempo, e o segundo realiza a discretização das amplitudes de saída. A codificação serve para transformar a onda digital em uma sequência binária, e é definida pelo padrão de armazenamento utilizado. Com estas etapas se transforma certa tensão elétrica em uma informação correspondente digital. A Figura 1 resume o processo de conversão do sinal analógico em digital, e o processo reverso.

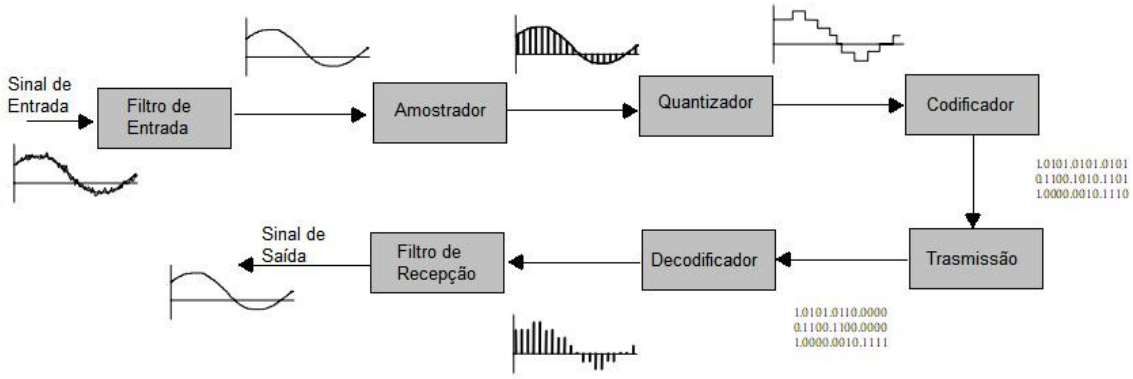

Figura 1 – Digitalização do sinal do som.

Entre as principais características a serem consideradas no processo de digitalização estão: a limitação da frequência máxima do sinal a ser amostrado, para a utilização de uma frequência de amostragem que seja superior ao dobro da frequência máxima do sinal. Este problema é solucionado com o filtro de entrada, este tem o objetivo de evitar a sobreposição do espectro de frequências do sinal após a amostragem.

Além disso, também devem ser considerados no processo de conversão A/D o número de bits usados na codificação, tipo de quantização, os quais juntamente com a taxa de amostragem refletem diretamente na taxa de transmissão de bits resultante.

Na conversão A/D o processamento do áudio é dado primeiramente pelo filtro de entrada, em seguida é realizada a amostragem do sinal filtrado. O próximo passo é quantizar as amostras, e após a codificação da onda digital uma sequência binária será transmitida.

Para a recepção da informação a conversão D/A é constituída primeiramente pela decodificação dos símbolos binários em níveis de amplitude para o sinal. E após este sinal termina de ser convertido para a forma analógica ao utilizar o filtro de recepção, este filtro tem como objetivo eliminar as altas frequências constituintes do sinal discreto. Tal procedimento faz com que a onda digitalizada volte a sua forma original.

2.3 Padrões de codificação e compressão de áudio

Por codificação entende-se a representação do sinal digital em símbolos binários, tendo como objetivo a sua transmissão ou armazenamento, com a qualidade exigida pela aplicação. A compressão é definida pela redução no número de bits utilizados para representar cada amostra.

A codificação em união com a compressão caracteriza os fatores que definem os formatos de áudio (para o armazenamento e para a transmissão). O codificador influencia no tamanho do arquivo e no nível de processamento necessário para decodificar o som. Isto pois o codificador além de transformar as amostras em palavras binárias, faz também todo o procedimento necessário para a compressão.

Assim há diversas características a serem analisadas nos codificadores de áudio (RIBEIRO, 2008), e tais são:

- Qualidades do áudio medida em PSNR (*peak signal noise ratio,* do inglês relação sinal ruído de pico) ou MOS (*mean opinion score,* do inglês pontuação média de opinião);
- Inteligibilidade;
- Possibilidade de reconhecimento do orador;
- Quantização;
- Atraso total do sistema;
- Complexidade do processo de codificação e decodificação;
- Memória necessária para codificação e decodificação;
- Sensibilidade a erros de transmissão;
- Diminuir a quantidade de informação que caracteriza o áudio (compressão binária).

Além de tais características, é necessário que seja feita uma análise do conteúdo sonoro de forma a extrair de seu sinal apenas a informação necessária para a aplicação, assim escolhendo a codificação mais apropriada. Por exemplo, o conteúdo a ser transmitido em uma estação de rádio pode ser apenas falado por um locutor ou pode conter parte musical como em uma propaganda publicitária. Para manter o mesmo nível de qualidade sonora em ambos os casos é necessário codificações distintas para trechos que contém ou não música.

A seguir é apresentada uma descrição sucinta e alguns formatos de arquivos que contém informação sonora:

Modulação por códigos de pulso (*Pulse Code Modulation* - PCM):

É o formato utilizado no sistema telefônico fixo e em CDs. É um método de codificação que demonstra o formato da onda sonora, podendo utilizar quantização linear ou não linear (lei μ ou A). Na telefonia, este formato produz um símbolos de 8 bits, amostrados a uma frequência de 8000 Hz, causando uma taxa de transmissão de 64 Kbps (MOECKE, 2006).

***Waveform audio format* (WAV):**

Utiliza o mesmo princípio de codificação do PCM, porém sua quantização pode utilizar de 8 a 16 bits para cada palavra binária, sendo esta quantizada de maneira linear ou não-linear. Sua frequência de amostragem pode variar entre 8000 e 44.100 Hz. Este padrão foi criado em conjunto pela Microsoft e pela IBM, sendo o pioneiro no armazenamento de som em computadores (WAV, 2008).

MPEG-1 *Audio Layer* 3 (MP3):

É um formato de áudio que utiliza os princípios da psico-acústica para comprimir o sinal. Grande parcelas das frequências que o ouvido humano não percebe são retiradas do sinal sonoro. Consequentemente este padrão resulta em uma compressão com perda de informação, estas perdas são praticamente imperceptíveis ao ouvido humano. Este formato permite até 90% na redução do tamanho do arquivo. Normalmente é utilizado para sons musicais com taxas de amostragem de 32.000 Hz, 44.100 Hz ou 48.000 Hz, resultando em taxas de bits que variam entre 32 Kbps e 320 Kbps (MP3, 2008).

Vorbis (OGG):

Assim como o MP3 é um padrão com alto nível de compressão utilizado principalmente em aplicações musicais. Este padrão possui código aberto e é de uso gratuito. Tendo taxas de bits que variam de 45 a 500 Kbps, e amostragem de 44.100 Hz (VORBIS, 2008).

***Raw* GSM 6.10 *Audio Stream file* (GSM):**

Criado para codificar sons falados, é o padrão utilizado no sistema de telefonia celular GSM (*Global System for Mobile communications*). Foi desenvolvido com o objetivo de atingir uma baixa taxa de bits e a baixa sensibilidade a erros de transmissão, podendo atingir taxas de bits de 5,6 e 13 Kbps com frequência de amostragem de 8000 Hz (FILE, 2008).

Além de todos os outros formatos existentes há também a possibilidade de concatenar alguns tipos de codificação de modo a aumentar o nível de compressão da informação e manter a qualidade do áudio, podendo atingir taxas de transmissão de 2 Kbps para sinais de fala (ETSI, 2008).

Como existe atualmente uma grande variedade de formatos de áudio, a relação de vantagens é dada de acordo com a necessidade da aplicação desejada, seja de qualidade

sonora ou de taxa de transmissão total do sistema, seja esta para conteúdo falado ou musical.

No caso de aplicações de sistemas de reconhecimento de áudio, há preferência por formatos que sejam codificados sem perdas. Neste estudo usaremos para comparação arquivos do tipo WAV com taxa de amostragem de 16000 Hz e símbolos de 16 bits.

2.4 Reconhecimento de áudio

Sistemas de reconhecimento de áudio consistem em algoritmos que identifiquem automaticamente como iguais, sons semelhantes para o ouvido humano. Isto de maneira independente do formato do sinal sonoro. A análise para o reconhecimento de áudio pode ser realizada de duas maneiras distintas, sendo que a primeira considera o áudio apenas como fala, de maneira a tentar identificar as palavras exteriorizadas pelo locutor. E a segunda maneira é a identificação do áudio como um padrão acústico, por exemplo, ao identificar dois acordes musicais como semelhantes. Neste trabalho será considerada apenas a identificação do áudio como um padrão acústico. Contudo alguns exemplos nesta seção utilizam o reconhecimento de palavras para facilitar a compreensão geral da técnica. Um exemplo genérico de reconhecimento de áudio é demonstrado na Figura 2.

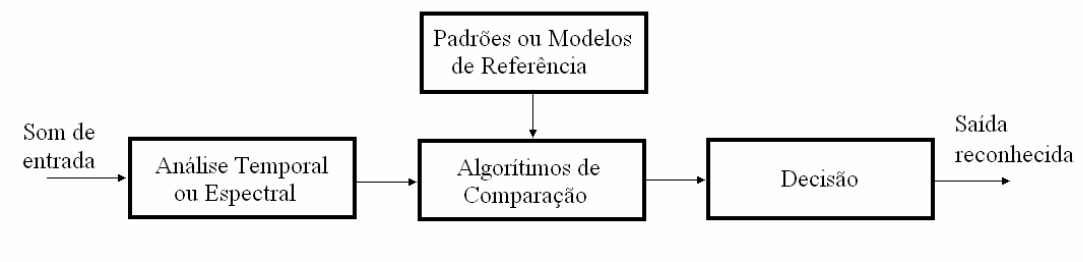

Figura 2 – Sistema genérico de reconhecimento de áudio.

Como pode ser observado na Figura 2, um sistema que realiza o reconhecimento de áudio tem como sinal de entrada o som que será analisado pelo sistema. O som foi previamente digitalizado e dividido em trechos de menor duração, estes trechos são conhecidos como quadros. Na análise temporal ou espectral é feita a extração de parâmetros do áudio. Esta etapa resulta na transferência dos parâmetros encontrados aos algoritmos de comparação. Estes algoritmos desenvolvem toda a comparação do áudio analisado com os padrões de referência. A partir desta comparação é tomada a decisão se houve ou não reconhecimento do áudio analisado, e assim é exteriorizada a resposta do sistema.

Por exemplo, o reconhecimento de palavras segue o seguinte procedimento: primeiramente são extraídos trechos do sinal de voz (quadros tipicamente com duração de 10 a 20 ms). Em seguida, esses quadros são analisados através da extração de parâmetros que podem ser adiquiridos através de critérios temporais ou frequenciais. Esses parâmetros são organizados de maneira vetorial e então comparados, por meio de algoritmos, com parâmetros de referência presentes numa lista de vocábulos (*codebook*). Os vocábulos possíveis resultantes da aplicação dos algoritmos de comparação são processados utilizando de modelos linguísticos e métricas de distância entre vetores, e ao final tem-se a palavra reconhecida.

O reconhecimento de padrões acústicos é realizado de maneira análoga ao reconhecimento de palavras. Assim se divide o sinal sonoro em pequenos quadros, e destes é extraido seus parâmetros. A comparação é realizada utilizando uma lista de sons, independente de fonemas, tornado o reconhecimento de sons mais simples, porém pode apresentar listas de comparação mais extensas que no reconhecimento de fala.

Um sistema que realize reconhecimento de áudio repetitivo consiste na idéia descrita pela Figura 2. Mas neste os dados exteriorizados pelo sistema se referem apenas a reincidência do reconhecimento de determinado trecho no áudio analisado.

Muitas são as propostas de implementação de reconhecimento de áudio feitas nos últimos anos, contudo, todas as propostas feitas até então revelam algum tipo de premissa para seu funcionamento. Estas premissas são impostas para contornar problemas encontrados no reconhecimento de sons e palavras. A seguir são listados alguns dos problemas encontrados nas implementações existentes.

- Interferência devido a ruídos externos;
- Limitações nos equipamentos de manipulação do áudio, como microfones, filtros e amplificadores. Gerando distorção no sinal analisado;
- Segmentação das unidades acústicas devido a sons não vozeados de baixa amplitude.
- Variabilidade da voz;
- Diferenças na velocidade da reprodução do som, como subamostragem ou superamostragem.

Na comparação de palavras, deve se levar em consideração aspectos como sintaxe, semântica, idioma analisado, extensão do vocabulário, coarticulação, entre outros. Uma questão que deve ser analisada é se a pronúncia das palavras é realizada de maneira isolada ou contínua. Outra análise deve ser efetuada com base no sistema utilizado, se este será

dependente ou não de locutor.

Em sistemas onde a análise é feita apenas pela sonoridade, a procura por repetições é feita levando em consideração toda a sequência sonora a ser analisada.

Outro aspecto a ser levado em consideração é a utilização de áudios armazenados previamente ou gerados em tempo real para a análise. Ou seja, a comparação é dada entre dois áudios, sendo um deles sempre gravado. Este som gravado é utilizado como parâmetro de comparação. O outro som pode estar armazenado ou sendo analisado de maneira simultânea ao processamento.

Como exemplos de possíveis aplicações para o reconhecimento do áudio têm-se:

- Serviços de tradução livre de escrita e em tempo real;
- Aprimoramento de unidades de respostas audíveis (URA), possibilitando a interação oral do usuário, assim aumentando a possibilidade de serviços e facilidade de uso;
- Serviços de identificação pessoal, onde a máquina consiga reconhecer o timbre da pessoa de modo que sua fala se torne uma identificação intransferível;
- Treinamento fonético para deficientes auditivos, ao utilizar uma máquina para testar a inteligibilidade de sua fala;
- Tradução voz-a-voz (speech-to-speech);
- Reconhecimento sonoro de senhas, levando em consideração o conteúdo falado e o locutor;
- Serviços de publicidade, conferindo a frequência e os horários da transmissão de propagandas em rádio ou televisão.

2.5 Extração de parâmetros

Uma etapa essencial ao reconhecimento de padrões acústicos é a extração de parâmetros, estes servem para que o som possa ser analisado. Ou seja, é necessário retirar as características do som para realizar a comparação com os modelos de referência. Estes parâmetros devem distinguir os diferentes sons a partir das características físicas do sinal acústico. Existem diversos métodos para parametrizar o sinal em questão, tendo como classificação geral os parâmetros temporais e frequênciais .

Entre os parâmetros temporais têm a análise da energia dos quadros e o número de cruzamentos por zero, estes lidam diretamente com a forma de onda do sinal analisado.

Normalmente parâmetros temporais são fáceis de aplicar, porém não apresentam resultados tão satisfatórios quanto os parâmetros frequênciais, devido a pequenas diferenças na forma da onda de alguns sons, como vogais, que podem aumentar o número de erros.

Os parâmetros frequênciais são compostos pela resposta das análises de LPC ou de Fourier. A análise frequencial é mais complexa por evidenciar características difíceis de serem observadas no domínio do tempo. A abordagem frequencial tem como principais exemplos de análise a LPC (*Linear Predictive Coding*) e a de Fourier, tais ainda são complementadas pela análise Cepstral gerando LPC-Cepstral e Mel-Cepstral.

Além destas, outras análises podem ser realizadas para que o reconhecimento ocorra de maneira eficiente. Alguns exemplos destas são: a estimação da frequência fundamental, a identificação do gênero do orador e a detecção da atividade de voz (RIBEIRO, 2008).

Neste estudo será focado apenas na análise frequencial levando em consideração as técnicas utilizadas para a retirada de parâmetros necessários para o reconhecimento, descritas a seguir.

2.5.1 Análise LPC (*Linear Predictive Coding*)

A predição linear LPC é um dos métodos dominantes para estimar os parâmetros do sinal de voz. A idéia básica da predição linear é a de que o valor de uma amostra do sinal sonoro pode ser aproximado por combinação linear dos valores das amostras anteriores, com base na correlação entre as amostras anteriores.

Nesta análise se pressupõe que a fala é produzida através de um ruído branco no final de um tubo, onde neste ocasionalmente são adicionados sons de assovio ou estouro. Embora aparentemente grosseiro este modelo é uma boa aproximação da produção da voz humana. A glote (o espaço entre as pregas vocais) produz o zumbido inicial, que é caracterizado pela sua intensidade e frequência. O trato vocal (boca e garganta) compõe o tubo, que é caracterizada pela sua ressonância, dando origem a frequência fundamental do som produzido. Silvos e estouros são gerados pela ação da língua, lábios e garganta durante a produção da fala.

Os coeficientes LPC retirados dessa análise estimam as características do sinal falado a partir de um ruído inicial ou de uma sequência periódica de impulsos. Tais coeficientes servem para modelar um filtro que se aproxima do trato vocal. Tais coeficientes são estimados ao minimizar o erro quadrático entre a amostra analisada atualmente e sua predição.

A extração de parâmetros é feita a partir de áudios codificados em PCM, com 16 bits por

amostra. Sendo transmitidos para o algoritmo de reconhecimento de padrões acústicos, após quantificação, o erro ou resíduo de predição e[n] (veja equação 2.3), dado pela diferença entre a amostra presente e a sua predição. O sinal de saída é sintetizado somando a predição com o resíduo após a quantificação inversa. Caso o resíduo não sofra nenhuma degradação durante a quantização, o sinal de saída será igual ao de entrada. A reconstrução do sinal de entrada é demonstrado pelo esquemático apresentado na Figura 3.

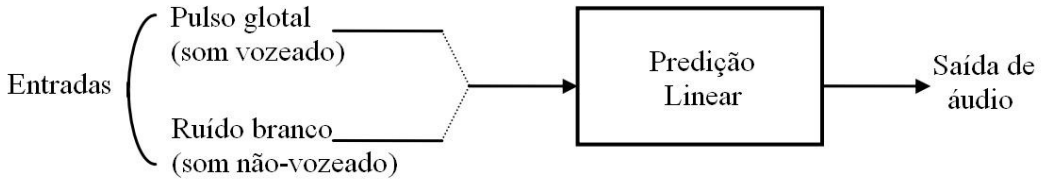

Figura 3 – Modelo da produção da fala a partir da predição linear.

Como os coeficientes representam o envoltório do espectro da informação, são tolerados poucos erros na transmissão destes. Pois um pequeno erro pode provocar toda a distorção do espectro desejado. Assim um erro de quantificação dos parâmetros LPC pode tornar o filtro que modela o trato vocal instável. Para que este problema não ocorra é necessário que os coeficientes LPC sejam transformados para garantir a estabilidade do sistema. Esta transformação pode ser realizada através de coeficientes de reflexão, coeficientes LAR (*Logarithmic Area Ratio*) ou coeficientes LSF (*Line Spectrum Frequencies*). Sendo os coeficientes LSF mais utilizados atualmente e descritos em Ribeiro (2008).

A análise LPC é feita através de um preditor definido pela combinação linear das últimas p amostras analisadas. Sendo estas p amostras correspondentes ao comprimento do vetor de amostras do trecho sonoro analisado, e o preditor é um conjunto (vetor) de coeficientes que definem o formato do sinal a ser analisado. E tem como resultando os parâmetros referentes a um filtro excitado por um sinal denominado erro de predição. Este filtro é obtido através de aplicação de teoremas (Yule-Walker e Wold) e modelos para processos estocásticos (auto-regressivo) e é obtido da expressão:

$$X(z) = H_I(z) \cdot E(z) \tag{2.1}$$

Sendo $H_I(z)$ é o filtro inverso dado por:

$$H_I(z) = \frac{1}{\sum_{k=0}^{P} \alpha_k z^{-k}} \tag{2.2}$$

e α_k é o coeficiente de ganho do filtro, E(z) é a transformada z do erro de predição (ruído brando) dado por:

$$e[n] = x(n) - \sum_{i=1}^{p} \alpha_i x(n-i) \tag{2.3}$$

Assim, a análise LPC tem como parâmetros o vetor de coeficientes do filtro H_I $[1, \alpha_1, \alpha_2, ..., \alpha_p]$ sendo p a ordem da análise. Esses coeficientes podem ser calculados resolvendo o sistema de p equações e p incógnitas. Para isto pode ser utilizado o método de lattice, o método da covariância ou o método da autocorrelação (BECERRA, 1993). Neste trabalho iremos considerar a utilização do último método (autocorrelação), descrito pela solução do sistema de equações de Yule-Walker:

$$\begin{vmatrix} R(0) & R(1) & R(2) & ... & R(p-1) \\ R(1) & R(0) & R(1) & ... & R(p-2) \\ ... & ... & ... & ... & ... \\ R(p-1) & R(p-2) & R(p-3) & ... & R(0) \end{vmatrix} . \begin{vmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_p \end{vmatrix} = \begin{vmatrix} R(1) \\ R(2) \\ ... \\ R(p) \end{vmatrix} \tag{2.4}$$

Sendo:

$$R(j) = \sum_{n=0}^{N-1-j} x(n)x(n+j) \tag{2.5}$$

é a autocorrelação de x(n). Os coeficientes obtidos dessa maneira minimizam a energia do sinal residual.

A abrangência da análise LPC é que $H_I(z)$ incorpora efeitos da resposta em frequência, da irradiação e do filtro conformador do pulso glotal (para fonemas sonoros), enquanto que E(z) permite identificar se o sinal analisado é um som vozeado ou não vozeado.

Neste estudo a predição linear é utilizada para a retirada das características fundamentais do som. Desconsiderando a síntese dos coeficientes para a construção do áudio.

2.5.2 Análise de Fourier

Pode-se afirmar que o sinal de voz é uma sequência de intervalos estacionários dentro dos quais a distribuição espectral de potência é mais ou menos constante, adotando um comportamento não-estacionário. Por outro lado se tem o conhecimento que há uma

correlação entre a densidade espectral de potência e o caráter fonético de muitos fonemas, o que faz a análise espectral de Fourier uma possibilidade para a parametrização do sinal de áudio. Assim os parâmetros da análise de Fourier são dados pelos coeficientes da Transformada Discreta de Fourier (DFT - *Discret Fourier Transform*):

$$X[k] = \sum_{n=1}^{N} x[n] e^{\frac{-j2\pi(k-1)(n-1)}{N}} \quad \text{onde } 1 \le k \le N \tag{2.6}$$

Para esse tipo de parametrização, vale salientar que a resolução em frequência é dada por $\Delta f = 1/NT$ sendo T o período de amostragem. Assim a resolução em frequência aumenta de forma diretamente proporcional ao número de amostras por quadro. Na parametrização temporal, contudo, a resolução aumenta à medida que o tamanho dos quadros diminuem, e o número de amostras por quadro se tornam mais concentradas.

O entendimento da resolução é essencial quando se analisa fonemas não-oclusivos ou oclusivos. Para fonemas não-oclusivos a identidade espectral está fortemente relacionada à densidade espectral sendo mais importante, neste caso, a resolução em frequência. Por outro lado, fonemas oclusivos são melhor identificados pela mudança de estacionariedade do sinal sendo, portanto, mais vantajoso possuir uma resolução maior no domínio do tempo (BECERRA, 1993).

É importante salientar que tal análise deve ser realizada em intervalos de tempo que sejam menores que os intervalos de estacionariedade, isto para garantir uma resolução mínima na representação dinâmica do sinal. Para diminuir as consequências da divisão do áudio em pequenos quadros é feita a filtragem das extremidades destes de modo a evitar descontinuidades entre os quadros, tal técnica é conhecida como janelamento e tem demonstrado resultados eficientes.

Tal análise também pode ser realizada através do algoritmo de Transformada Rápida de Fourier (FFT - *Fast Fourier Transform*), tendo o mesmo tipo de resultado e ganho na diminuição da complexidade computacional.

2.5.3 Análise Cepstral

O sinal de áudio é formado por uma excitação convoluída com a resposta ao impulso do modelo do trato vocal. No entanto, em aplicações de reconhecimento de voz, é desejável separá-las. A idéia da análise Cepstral é levar o sinal para um domínio onde isso seja possível.

O modelo básico para a produção do sinal de voz consiste em um filtro H(z), correspondendo ao trato vocal, multiplicado pela transformada z da excitação glotal E(z). Se, no domínio da frequência a operação de produto for substituída pela operação soma, através da aplicação da função logaritmo em $H(z)E(z)$, os dois sinais podem ser separados pelo cálculo da transformada inversa de Fourier de $\log[H(z)E(z)] = \log[H(z)] + \log[E(z)]$. Esta transformada inversa é denominada cepstrum, sendo um parâmetro pseudo-temporal.

Há dois tipos de análise Cepstral: a baseada em FFT e a aplicada à análise LPC. A análise Mel-cepstral é a forma de análise Cepstral baseada em FFT, já a análise Cepstral LPC à transformada z é aplicada no sinal de voz modelado através da análise LPC.

2.5.4 LPC Cepstral

Consiste basicamente em aplicar logaritmo na análise LPC, ou seja, aplicar logaritmo nos dois lados da equação 2.1:

$$\log[X(z)] = \log[H_I(z)] + \log[E(z)] \tag{2.7}$$

Logo os coeficientes LPC-cepstrais, c_n, correspondem a transformada z inversa de $\log[X(z)]$, sendo que os coeficientes de menor ordem correspondem ao filtro do trato vocal. Os parâmetros dessa análise são apresentados através da equação:

$$c_n = \alpha_n + \sum_{i=1}^{n-1} \frac{n-i}{n} \alpha_i c_{n-1} \quad \text{para } n \geq 1 \tag{2.8}$$

Sendo $\alpha_i = 0$ para $i > p$, sendo p a ordem da análise LPC, α o coefciente LPC, e i a posição do coeficiente (AKISHINO, 2007).

Tais coeficientes mostram a facilidade de separar a convolução da exitação com a resposta ao impulso através de uma soma, facilitando todo o processamento do sinal sonoro.

2.5.5 Análise Mel-Cepstral

A análise cepstral pode ser realizada aplicando o logaritmo diretamente na FFT do sinal de voz ou empregando a análise LPC. A análise Mel-Cepstral aplica a análise Cepstral sobre a FFT do sinao de áudio. Se tratando da FFT o cepstrum corresponde à transformada inversa de Fourier de $\log[X(z)]$, onde X(z) é obtido pela DFT do sinal sonoro.

A idéia consiste em dividir a faixa de frequência útil em filtros passa-faixa cuja largura é proporcional à escala de mel (BECERRA, 1993) tentando simular a resposta em frequência

da membrana basilar. Assim, determina-se a energia de cada filtro e, após é calculado o logaritmo de cada energia, aplicada à transformada inversa de Fourier.

2.6 Algoritmos de reconhecimento de padrões acústicos

Até agora foram apresentadas as técnicas básicas mais comumente utilizadas no processamento do sinal sonoro prévio ao reconhecimento. Como resultado deste processamento, o áudio é dividido em segmentos de igual duração onde se extraem uma série de parâmetros, que devem estar intimamente correlacionados com a identidade do segmento em questão. Sendo assim, cada trecho após a parametrização, fica representada por uma sequência temporal de vetores de parâmetros. Como estes coeficientes são obtidos por meio de medidas físicas sobre o sinal sonoro, podemos dizer também que cada trecho fica representado por uma sequência temporal de vetores de observação.

O reconhecimento de padrões acústicos consiste em associar uma sequência de vetores de parâmetros de observação, relativos a um trecho sonoro, a uma das sequências ou modelos de referência. A elocução a ser reconhecida não pode fazer parte, é claro, da base de dados utilizada para obter as sequências ou modelos de referências (treinamento).

O avanço do reconhecimento sonoro nos últimos anos se deve principalmente a três técnicas básicas: o DTW (*Dynamic Time Warping*), os HMM (*Hidden Markov Models*) e, mais recentemente, as redes neurais. Todos os sistemas de reconhecimento automático trabalham com uma ou mais destas três técnicas. A seguir temos uma explanação de tais técnicas.

Neste trabalho foi optado pela utilização de medida de distância entre quadros. Isto pois o objetivo específico do trabalho é reconhecer sons repetitivos. As técnicas supracitadas apesar de sua grande aplicação, são voltadas para o reconhecimento de palavras. Não apresentando o caráter apropriado para este estudo.

As medidas de distância entre quadros, discutidas na seção 2.6.4 apresentam uma comparação mais genérica entre os parâmetros retirados do áudio analisado.

2.6.1 DTW (*Dynamic Time Warping*)

O algoritmo DTW foi o primeiro algoritmo a realizar reconhecimento de padrões acústicos de maneira satisfatória, ele se destina a alinhar quadros de forma a compensar as

diferentes velocidades de pronúncia das palavras. Este procura eliminar as diferenças de duração dos períodos estacionários dos sons analisados. O algoritmo tem como função encontrar e minimizar a função de distância entre os dois vetores de parâmetros que caracterizam o áudio, alinhando estes de modo não linear. Considerando os vetores como sendo compostos por parâmetros extraídos de cada sinal sonoro. Com vetores X e Y é traçado uma curva de alinhamento não-linear que força os pontos dos vetores a coincidirem, conforme mostra a Figura 4.

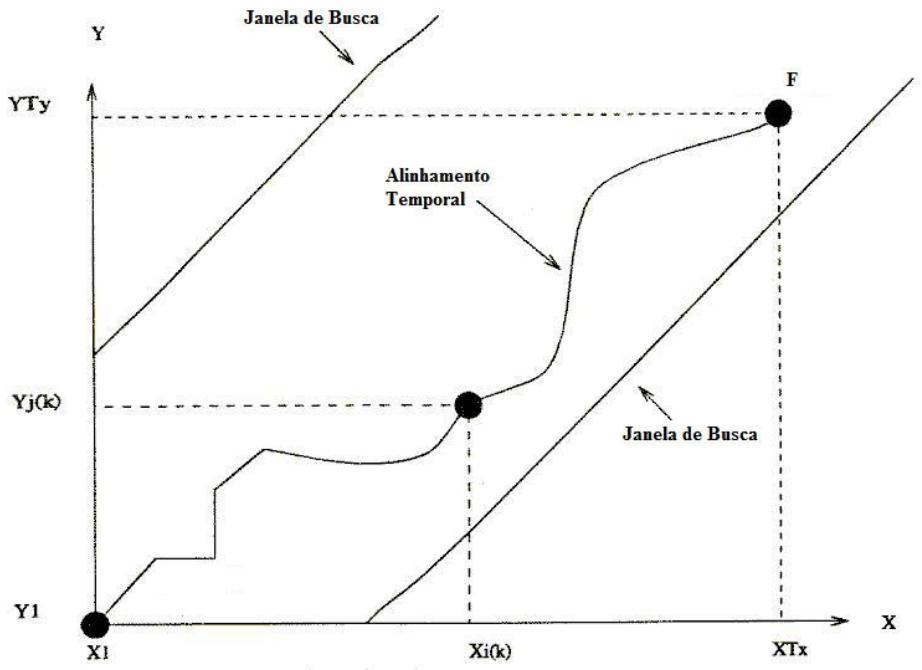

Figura 4 – Alinhamento não-linear de áudios utilizando DTW
(imagem baseada em Becerra, 1993).

Assim o objetivo do método DTW se dá ao encontrar a função F que minimize a distância entre os vetores X e Y. Isto com as restrições de continuidade, monotonicidade e coincidência dos pontos extremos dos vetores. Notando que as sequências de X e Y são posicionadas de modo que o primeiro termo de cada vetor esteja na origem do gráfico.

O motivo pelo qual o método DTW teve pouco uso, comparando a outros métodos, é devido ao seu caráter determinístico. Este caráter implica na necessidade de existir um padrão de referência correspondente a cada palavra pronunciada, o que leva a se ter vários padrões de uma mesma palavra, na intenção de cobrir todas as variações possíveis na sua elocução, por um ou mais locutores (BECERRA, 1993).

2.6.2 HMM (*Hiddem Markov Models*)

O modelo oculto de Markov consiste em um modelamento estocástico no qual algum parâmetro desconhecido é obtido através de parâmetros observáveis. Ou seja, este modelamento pode ser utilizado quando um evento atual é determinado ou diretamente influenciado pelos últimos eventos anteriores. No caso do reconhecimento de padrões acústicos o parâmetro desconhecido é a sequência de estados dos quadros de áudio que por sua vez determinam o som.

No modelo associa-se a cada estado uma função de distribuição de probabilidade de observações. Assim, cada estado pode gerar uma observação, entre um conjunto, de acordo com a distribuição. A mesma sequência de observações pode assim ser gerada, com probabilidades diferentes através de sequências diferentes de estados.

Os maiores problemas para a aplicação deste modelo são: (a) a determinação da probabilidade de uma sequência observável, (b) a determinação da sequência de estados e (c) a estimação dos parâmetros para reconhecimento de um dado modelo. Tais problemas foram contornados respectivamente pelos algoritmos progressivo-regressivo (*Forward-Backward*), Viterbi e Baum-Welch descritos em Becerra (1993).

Com HMM o reconhecimento é efetuado determinando a probabilidade da entidade a ser reconhecida ter sido gerada por cada um dos modelos. Assim, para que haja o reconhecimento, primeiramente é necessário definir o conjunto de classes de sons a serem reconhecidos, depois escolher uma topologia para o tipo de modelo. Posteriormente obter, para cada classe um conjunto com dimensão razoável de dados de treino. E por último, treinar os modelos utilizando o algoritmo de Baum-Welch.

Para o reconhecimento de um som, a princípio é extraída a sequência de observações correspondentes ao sinal de fala. Após é calculada a probabilidade da sequência de observações, dado cada um dos modelos. E assim, atribui-se a sequência de observações ao som associado ao modelo que obteve a máxima probabilidade. (RIBEIRO, 2008).

2.6.3 Redes Neurais

Redes neurais artificiais são técnicas computacionais amplamente utilizadas para o reconhecimento de qualquer tipo de padrão. Esta rede é construída a partir de um modelo matemático inspirado na estrutura neurológica humana. O grande diferencial destas técnicas computacionais e a capacidade de adquirir conhecimento através da experiência. Portanto a

utilização de redes neurais no reconhecimento de áudio é feita com base no reconhecimento de padrões inato a esta técnica.

Para que seja possível realizar o reconhecimento sonoro são seguidos quatro passos básicos (HOSOM, 1999). Primeiramente ocorre a digitalização da fala que se quer reconhecer. Um segundo passo realiza uma computação das características que representam o domínio espectral contido na fala (regiões de alta energia em frequências particulares). Esse passo é computado a cada 10 ms, sendo que cada seção de 10 ms é chamada de quadro. O terceiro passo consiste no uso de uma rede neural para realizar a classificação de um conjunto de características dentro das características fonéticas básicas para cada quadro. Por fim, o quarto passo aplica um método de busca para associar as saídas da rede neural com padrões de palavras, a fim de encontrar a palavra com a qual as saídas da rede mais se assemelham. Um exemplo de método de busca é o Viterbi (HOSOM, 1999).

Utilizando este o melhor caminho encontrado corresponderá a palavra reconhecida. É importante lembrar que a função da rede neural não é determinar se o quadro analisado é ou não pertencente a uma determinada categoria, mas sim informar qual a probabilidade daquele quadro pertencer a uma determinada categoria. Para que haja tal distinção a rede neural possui uma estrutura de forma a reconhecer fonemas, chamando-se, assim, de mapa fonético.

Devido a sua alta complexidade os sistemas baseados em redes neurais estão voltados principalmente a compreensão da fala, não sendo muito utilizada para o reconhecimento.

2.6.4 Medidas de distância entre quadros

Também conhecidas como medidas de distorção ou de semelhança, estas têm como função medir a diferença da identidade fonética entre duas partes do som analisado. Isto comparando trechos de igual duração, por meio de distância numérica, métrica ou não, entre os coeficientes ou parâmetros extraídos do sinal sonoro do intervalo de tempo em questão (BECERRA, 1993). Esta distância é calculada com base em trechos sonoros de duração entre 10 e 30ms, pois neste o sinal sonoro pode ser considerado como estacionário ou quase-estacionário. Entre as medidas de distância de quadros mais utilizadas estão a euclidiana, a de Mahalanobis e a de Itakura-Saito.

O objetivo destas medidas de distância é encontrar o menor valor o possível ao se comparar dois quadros de parâmetros. No caso de sistemas de reconhecimento de áudio repetitivo este irá comparar os quadros que contem os parâmetros do sinal sonoro.

A distância mais conhecida é a euclidiana, esta realiza a diferença ponto a ponto

diretamente aos vetores em questão. Sendo bastante utilizada em sistemas de reconhecimento de fala. Esta medida é feita pela equação a seguir:

$$d^2 = \sum_{i=1}^{N} (X - Y) \tag{2.9}$$

Sendo X e Y os vetores de parâmetros a serem analisados, d a distância encontrada entre X e Y, e N o número de parâmetros a serem comparados de uma única vez.

Outro método, muito utilizado em análise de parâmetros LPC é a de Itakura-Saito, demonstrada a seguir:

$$d = \frac{\sigma_X^2 \cdot \overline{X}^T \cdot \overline{R}_X \cdot \overline{X}}{\sigma_Y^2 \cdot \overline{Y}^T \cdot \overline{R}_Y \cdot \overline{Y}} + \log\left(\frac{\sigma_X^2}{\sigma_Y^2}\right) - 1 \tag{2.10}$$

sendo que R é a matriz de autocorrelação do sinal X e Y , e σ representam os parâmetros de ganho LPC dos mesmos.

O método de Mahalanobis é considerado como uma generalização da distância euclidiana de maneira ponderada. Este método em vez de calcular a diferença entre os vetores, analisa o desvio padrão entre eles.

$$d = \sqrt{(X - Y)^T \cdot W^{-1} \cdot (X - Y)} \tag{2.11}$$

Sendo W^{-1} a matriz de covariância inversa que pondera cada parâmetro de maneira individual. E $(X - Y)^T$ representa a versão transposta de diferença entre X e Y.

Outros métodos de medidas de distância entre quadros não serão abordados neste trabalho, devido a sua pequena aplicação na área de reconhecimento de repetições de áudio.

No próximo capítulo esta a descrição do trabalho realizado nos ambientes de desenvolvimento contendo toda a descrição do protótipo proposto neste trabalho. Para a realização deste foi utilizada dentre as técnicas supracitadas, a digitalização do áudio para o formato WAV. O reconhecimento do áudio foi feito com base na extração de parâmetros via análise LPC, e o método de reconhecimento de padrões acústicos utilizou as medidas de distância de quadros de Itakura-Saito e euclidiana.

3 *Áudio Repetitivo*

Este trabalho foi iniciado com o objetivo de ampliar o conhecimento em processamento de sinais sonoros, fazendo testes em sistemas que realizam reconhecimento de áudio para detectar repetições em sequências de som. Foram estudadas as principais técnicas para parametrização das características do áudio e para a comparação destes parâmetros, de modo a detectar um padrão acústico no arquivo de som analisado. Uma das etapas realizadas neste trabalho foi o teste em diversos ambientes de desenvolvimento utilizados para o processamento do áudio, de maneira que foram estudados e avaliados diversos métodos relacionados com o reconhecimento dos padrões acústicos. A partir dos testes, a continuação dada ao trabalho foi a realização de reconhecimento das repetições através da implementação de um protótipo. Este protótipo detecta a repetição de trechos de um áudio de referência em um arquivo de áudio com maior duração de tempo. Para tal, o sistema realiza uma varredura, buscando em uma determinada sequência de áudio por possíveis repetições da amostra de áudio de comparação.

3.1 Estudo dos ambientes de desenvolvimento para processamento de áudio

Nesta etapa dos estudos dos ambientes de desenvolvimento, foram analisados de modo a melhorar a compreensão do reconhecimento do áudio. Entre as possibilidades foram analisadas os seguintes ambientes: MARF (*Modular audio recognition framework*), HTK (*Hidden Markov Model Toolkit*), e a biblioteca Voicebox do Matlab (*Matrix Laboratory*). Estes ambientes apresentam ferramentas para o processamento de áudio, possibilitando a construção de aplicações específicas.

3.1.1 Modular Audio Recognition Framework (MARF)

O MARF é um ambiente de desenvolvimento que possui um banco de dados próprio, de uso geral, para o reconhecimento de padrões acústicos (MOKHOV, 2008). Ele é organizado na forma de uma biblioteca que dispõe de diversas interfaces de programação de aplicativos (API). Esta plataforma é desenvolvida em JAVA, sendo gratuita e de código aberto (MODULAR, 2007), servindo de suporte para o desenvolvimento de diversas aplicações de processamento de áudio.

O funcionamento básico do MARF pode ser dividido nas etapas:

- **Aquisição do sinal de áudio**: obtido diretamente pela interface de áudio do computador (microfone), ou através de um arquivo em formato WAV, MP3 ou arquivos de texto (com extensão .txt);
- **Pré-processamento**: realiza a normalização do sinal de entrada, filtragem e limitação da amplitude do sinal;
- **Extração de parâmetros**: realizada através de predição linear (LPC) ou análise no domínio da frequência (FFT);
- **Treinamento**: compreende o processo para formar o banco de dados que será utilizado na comparação dos trechos de áudio;
- **Classificação**: utiliza as informações obtidas no treinamento, visando reconhecer palavras isoladas.

A fase de treinamento e a classificação podem ser realizadas por diversos métodos, variando a forma como é realizada a comparação dos padrões acústicos. Esse processo possibilita a utilização de algoritmos de redes neurais e diversos classificadores de distância vetorial, como Chebyshev, euclidiana, Mahalanobis, Diff, Hamming, e Minkowski. (MOKHOV, 2008).

Este ambiente apresenta diversas aplicações prontas para realizar testes de desempenho da plataforma, como exemplo os testes para desempenho da FFT ou testes de funcionamento dos dispositivos de entrada e saída de som. O MARF também possui disponível algumas aplicações que realizassem diretamente processamento de som.

3.1.2 Hidden Markov Model Toolkit (HTK)

O HTK é um ambiente de desenvolvimento que dispõe de ferramentas para a construção e manipulação de Modelos Ocultos de Markov (HMM). O HTK teve como função inicial à

pesquisa para o reconhecimento de fala, apesar de poder ser utilizado em diversas aplicações, como síntese de fala, reconhecimento de locutor, dentre várias funções relacionadas (YOUNG, 2006).

O aplicativo é formado por um conjunto de bibliotecas e ferramentas construídas em linguagem de programação C. Estas ferramentas são manipuladas através de linha de comando, criando a possibilidade de programação através de rotinas em *shell-script*. O HTK possibilita facilitar a análise da fala, realizando o treinamento para o HMM com o intuito de reconhecer sons, além de conter interface para testes e análises de resultados.

Seu funcionamento pode separado em quatro etapas:

- **Preparação dos dados:** é composto pela ferramenta de entrada do som, seguida pela etapa que retira os parâmetros utilizados nas ferramentas de treinamento e para testes;
- **Treinamento**: consiste na utilização de áudios de referência para a construção dos modelos ocultos de Markov;
- **Reconhecimento**: pode ser realizado com base no algoritmo de Viterbi, encontrando o som de maior probabilidade em uma treliça. Ou pode realizar a comparação direta com uma lista (vocábulo) existente;
- **Análise**: é utilizada para avaliar a performance do reconhecimento. Isto é feito de maneira que o áudio de entrada do sistema é comparado diretamente com a saída do reconhecedor, assim alcançando bons resultados de eficiência da aplicação.

Esta aplicação tem o treinamento, o reconhecimento e a análise baseados apenas em HMM e algoritmos auxiliares a este. Porém, a preparação dos dados pode conter diversos tipos de parâmetros extraídos do sinal sonoro, como LPC, Cepstral, Mel-cepstral e outros. Apesar desta comparação ser realizada apenas entre parâmetros do mesmo gênero.

3.1.3 Matlab

O Matlab é um ambiente de desenvolvimento utilizado para programar em linguagem proprietária .m, sendo extremamente dinâmico e atendendo a diversas aplicações. Para processamento de áudio, o *software* possui algumas ferramentas prontas, dando a possibilidade da construção do algoritmo desejado.

A biblioteca Voicebox (2008) foi analisada para o processamento de sons, sendo construída com o apoio a biblioteca básica do Matlab e possuindo licença de utilização aberta. Seu objetivo é facilitar o processamento de áudio ao organizar diversos algoritmos

em funções, utilizados para esta necessidade. Por exemplo, existem funções para calcular a conversão da escala Mel-cepstral em frequência (Hz).

Esta biblioteca e as funções básicas do Matlab/Simulink foram utilizadas na construção do sistema desenvolvido neste trabalho. A contribuição desta biblioteca ocorreu através de uma função que calcula a distância de Itakura-Saito.

3.2 Sistema de detecção de áudio repetitivo

Este protótipo foi desenvolvido para verificar se um trecho de som de referência (A) se repete em um outro arquivo de áudio (B), podendo ou não A estar contido em B. Neste sistema é contabilizado o total de repetições e indicado em quais momentos estas repetições ocorreram. Um exemplo de aplicação deste sistema seria a determinação da quantidade de vezes que uma determinada propaganda é repetida na programação de uma estação de rádio ou canal de televisão.

Este sistema foi desenvolvido usando o ambiente Matlab/Simulink. Esta escolha foi feita para facilitar o desenvolvimento, devido à familiaridade com o ambiente. O sistema realiza todo o processamento de entrada do áudio, realiza a retirada dos parâmetros dos áudios e a análise da repetição. Este protótipo foi dividido entre dois códigos, o *DETECTA_REPETIÇÃO* e o *ANALISA_AUDIO*. O primeiro é o código principal, este é responsável pela entrada do áudio, pela comparação e decisão realizada com relação ao reconhecimento da repetição. O segundo, o código secundário, realiza a parametrização dos áudios envolvidos no processamento.

A primeira etapa do aplicativo ocorre utilizando o código *DETECTA_REPETIÇÃO* e a obtenção dos parâmetros LPC é feita pelo subsistema *ANALISA_AUDIO*. Este subsistema retorna dos parâmetros LPC, e a partir destes é realizada a decisão da ocorrência de repetições com o código *DETECTA_REPETIÇÃO*. O código documentado do sistema está disponível em: *http://www.sj.cefetsc.edu.br/~daiana/Tcc*.

3.2.1 Funcionamento

Este protótipo tem como função o reconhecimento da repetição de um determinado áudio, e este segue o modelo descrito na Figura 2. Sendo que este utiliza a análise LPC para a parametrização e as medidas de distância vetorial para a comparação. O padrão de referência

é construído com base na análise LPC do áudio (A). A decisão contabiliza o número de repetições detectadas e seus respectivos momentos. Assim o som de entrada é o áudio objeto (B) e a saída é a apresentação dos momentos em que houve repetições.

A analise teórica levou a escolha da distância euclidiana devido a sua simplicidade de cálculo, enquanto que a distância de Itakura-Saito foi escolhida teoricamente apresenta um melhor desempenho em relação a outras medidas (distâncias euclidiana e de Malahanobis), em relação a comparação de palavras isoladas em conjunto aos parâmetros LPC (BECERRA, 1993).

O código principal *DETECTA_REPETIÇÃO* realiza uma varredura no arquivo do áudio, objeto (B), para procurar a repetição do áudio de referência (A). Ambos os áudios são processados para obter os parâmetros LPC, usando o código secundário *ANALISA_AUDIO*.

Inicialmente cada arquivo de áudio é subdividido em segmentos de 10ms, onde cada segmento é chamado de quadro, esta divisão é realizada no código principal.

Cada quadro é processado pelo código secundário para efetuar a parametrização do trecho sonoro contido no quadro em questão. Assim, cada quadro gera como retorno ao código principal um vetor de parâmetros LPC. Estes vetores serão utilizados para a realização das comparações que buscam repetições.

O cálculo de comparação é realizado quadro-a-quadro, onde é realizada a análise do primeiro quadro (entre os selecionados) do trecho de áudio (A), com o primeiro quadro do áudio (B). Caso esta comparação indique uma distância maior que o limiar máximo, é analisado o próximo quadro do áudio (B). Se o limiar for respeitado, é efetuada a segunda comparação. Nesta é comparado todos os quadros do áudio (A) com o trecho do áudio (B) consecutivo ao quadro utilizado pela primeira análise. Se a distância entre os trechos respeitar o limiar indica que houve uma repetição. Caso esta distância não respeitar o segundo limiar, é analisado o próximo quadro do áudio (B) com o primeiro quadro do áudio

Para a comparação das medidas de distância entre os quadros dos áudios A e B, foram implementadas duas variantes, uma que utiliza a medida da distância de Itakura-Saito e a outra que utiliza a distância euclidiana. Ambas são usadas para avaliar se há repetição do áudio. Estas medidas de distância servem para indicar que ambos os trechos são similares ou não, concluindo-se a ocorrência da repetição.

No código esta parte é realizada em duas etapas, sendo comparada a dois limiares. Isto foi feito para aumentar o desempenho do protótipo com relação a velocidade de execução e eficiência na detecção.

Na primeira etapa é feita uma análise de um único quadro de A com cada um dos quadros de B, assim, cada distância calculada que obtiver um valor menor que o primeiro limiar, será realizada uma segunda avaliação. Na segunda etapa, os trechos que passaram para a segunda avaliação serão comparados um a um com todo o trecho do áudio de referência com um trecho do áudio objeto. Quando a distância calculada for menor que o segundo limiar de comparação, é considerado que houve repetição do áudio de referência no ponto do áudio sob avaliação.

As etapas de funcionamento dos códigos *DETECTA_REPETIÇÃO* e *ANALISA_AUDIO* estão resumidas na Figura 5 e Figura 6. O código descrito na Figura 6 está contido na atividade "Obter os coeficientes LPC do som" da Figura 5.

O código *DETECTA_REPETIÇÃO* pode ser resumido nas seguintes etapas:

- **Inicialização do programa:** serve para definir os parâmetros do sistema como extensão do quadro analisado, números de coeficientes LPC retirados de cada quadro, quadros a serem comparados, limiares de comparação, tempo de simulação, entre outros;
- **Leitura dos arquivos de áudio:** que consiste em ler os arquivos do tipo WAV, e definir suas características, como tipo de quantização e taxa de amostragem, entre outros;
- **Obter os coeficientes LPC:** este serve para fazer a interface com o código secundário, enviando os quadros de áudio e recebendo os vetores de parâmetros.
- **Detecção das repetições de áudio:** a partir do momento que o primeiro quadro avaliado teve distância inferior ao primeiro limiar, será calculada a distância do trecho subsequente com todo o trecho retirado do áudio de referência. Esta segunda comparação terá como limitação um limiar mais restritivo que o primeiro, de modo a aumentar a precisão da comparação. Caso esta nova distância seja menor que o novo limiar, haverá incremento ao número de repetições e armazenado à sequência de quadros;
- **Mostrar o número de repetições:** consiste em expor os resultados alcançados com o algoritmo, expondo além do número total de repetições indicando o tempo no qual ocorreu a repetição

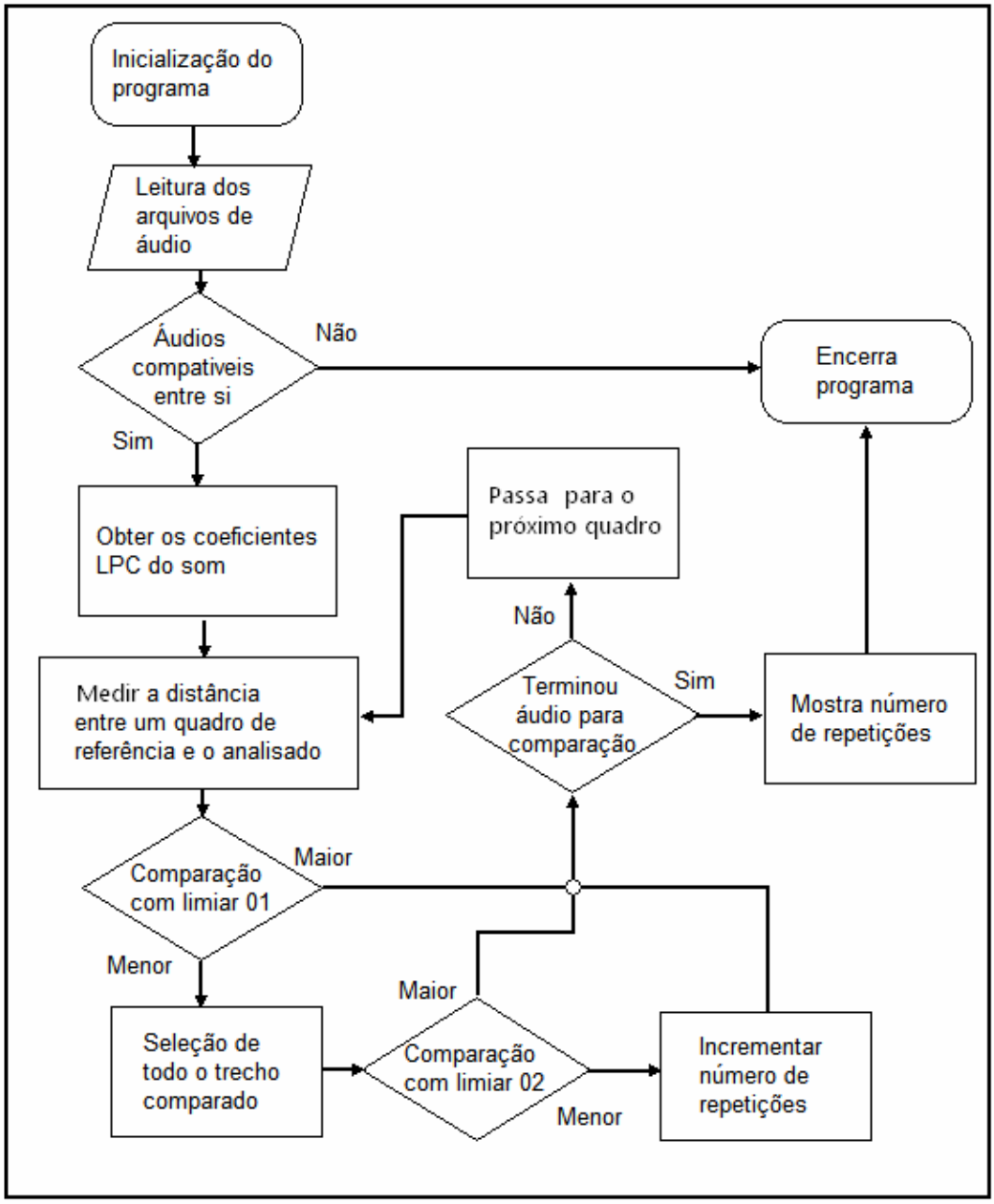

Figura 5 – Fluxograma do código *DETECTA_REPETIÇÃO*.

.

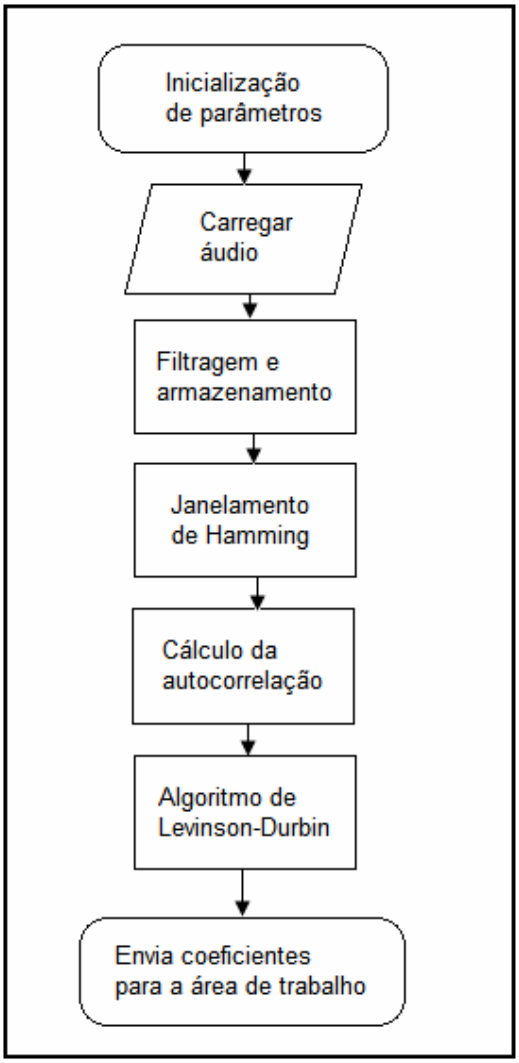

Figura 6 – Fluxograma do código *ANALISA_AUDIO*.

O código *ANALISA_AUDIO* calcula os coeficientes LPC que serão utilizados na comparação das medidas de distâncias. Este está contido no fluxograma mostrado na Figura 6 como sendo parte da atividade "Obter os coeficientes LPC do som". O esquemático detalhado do código é apresentado na Figura 6, e adota a ordem a seguir:

- **Inicialização de parâmetros:** procede a inicialização dos parâmetros e realiza uma cópia da informação contida no arquivo de áudio;
- **Filtragem e armazenamento:** limita o sinal sonoro e o separa em diversos quadros de menor duração;

- **Janelamento de Hamming**: realiza a sobreposição dos quadros, utilizando a janela de Hamming, servindo para diminuir o efeito causado por mudanças bruscas no sinal sonoro na transição de quadros;
- **Cálculo da autocorrelação:** tem como intuito definir o sistema de equações polinomiais que descrevem o comportamento sonoro, este equacionamento é necessário para a retirada dos coeficientes LPC;
- **Algoritmo de Levinson-Durbim:** a partir da sequência de polinômios retirados da etapa anterior, ocorre o cálculo do algoritmo de Levinson-Durbim para encontrar os coeficientes LPC;
- Envia os coeficientes para o termino do processamento do aplicativo *DETECTA_REPETIÇÃO*.

Neste capítulo foram descritas as etapas do trabalho realizado, dos testes nos ambiente de desenvolvimento até a construção do protótipo desenvolvido. No próximo capítulo serão demonstrados os resultados alcançados em cada uma dessas etapas.

4 *Resultados*

4.1 Resultados dos testes nos ambientes de desenvolvimento

Os testes foram realizados com o objetivo de aprofundar os estudos relacionados ao reconhecimento de áudios e possibilitar a utilização de alguma ferramenta para comparação com o protótipo desenvolvido.

No caso do MARF não foram encontradas aplicações que possibilitassem o reconhecimento de sinais sonoros puros (sem a necessidade de fala), ou mesmo que realizassem reconhecimento de repetições de fala. Além de possuir incompatibilidades com versões mais atuais do JAVA ou do sistema operacional testado (Ubuntu 8.04 e Windos XP). Devido a tais problemáticas o MARF foi abandonado.

Devido a abordagem diferenciada o HTK não apresenta aplicativos prontos desde o princípio, devido a sua utilização feita através de linha de comando. Porém, os testes e estudos foram realizados devido a esta ferramenta possuir tutorias que auxiliariam na construção de sistemas para reconhecer a repetição sonora. A tentativa da construção deste sistema explicitado no tutorial (YOUNG, 2008) foi falha. Isto, pois houve problemas com incompatibilidade de formatos de áudios e inclusive antagonismo com formatos internos de informação que não conseguiam ser lidos por nenhum programa interno ou externo. Sua documentação apresentava informações ambíguas, criando uma falta de credibilidade com relação a este pacote. Por estes motivos os com o HTK testes foram abandonados.

A biblioteca Voicebox, do Matlab, não foi completamente testada devido a sua ampla extensão. Porém, as funções que foram testadas apresentaram resultado satisfatório. Uma delas, a '*distisar.m*' foi utilizada em uma das variantes do algoritmo desenvolvido. Esta função calcula a distância de Itakura-Saito a partir de dois vetores. A partir desta foi

facilitado o desenvolvimento da variante do protótipo que utiliza a distância de Itakura-Saito, portanto foi de grande utilidade para o trabalho desenvolvido.

4.2 Metodologia de testes para a aplicação desenvolvida

Neste capítulo é apresentado os testes realizados com protótipo descrito na seção 3.2. Os resultados alcançados com esses testes serão demonstrados nas próximas seções. A partir deste protótipo foram realizados testes que comparam o desempenho no reconhecimento de repetições de áudio de referência (A) em um áudio objeto (B).

Todos os testes foram realizados com arquivos que continham informação musical ou fala natural (contínua). Os arquivos usados estão no formato WAV, mono-canal, com quantização linear (do tipo PCM) de 16 bits e taxa de amostragem de 16000 Hz. Os trechos de áudio utilizados como referência de comparação têm duração de 100ms e os arquivos de áudio utilizados como objeto têm duração que varia entre 1,5 a 5 segundos.

No exemplo apresentado aqui, o áudio de referência (A) tinha duração de 100ms e o áudio objeto (B) tinha duração de 1 segundo, estes tinham respectivamente 20 e 200 quadros. Todos os quadros trabalhados tinham duração de 10ms e se sobrepunham ao quadro anterior em 5ms. Uma das características do protótipo é que o áudio (A) pode ser utilizado para comparação de todos os quadros ou pode utilizar a comparação de alguns quadros escolhidos de maneira manual (indicando-os como variáveis do sistema). A sequência de quadros analisados no áudio B será compatível com a sequência selecionada do áudio A, de modo a manter a comparação equitativa.

Para os testes básicos, de cada quadro foram calculados 12 parâmetros LPC e foi prevista a utilização de 10 quadros para comparação geral.

Os áudios encontrados juntamente com os códigos do protótipo formam o exemplo apresentado para a comparação geral do desempenho das variantes do protótipo. No pacote de arquivos está contido o arquivo *'Mus01.wav'* e *'Mus02.wav'*, respectivamente áudio referência e objeto. Estes áudios apresentam conteúdo musical e são utilizados como exemplo de demonstração neste capitulo. A forma de onda do áudio de referência esta apresentada na Figura 7, e a forma de onda do áudio objeto está demonstrada na Figura 8. A Figura 7 (a) e (b) consistem em um mesmo áudio representado em escalas de tempo diferenciadas, no qual a parte (a) a escala apresentada tem a mesma proporção da escala da Figura 8. A

representação da forma de onda do áudio objeto está contida na Figura 8.

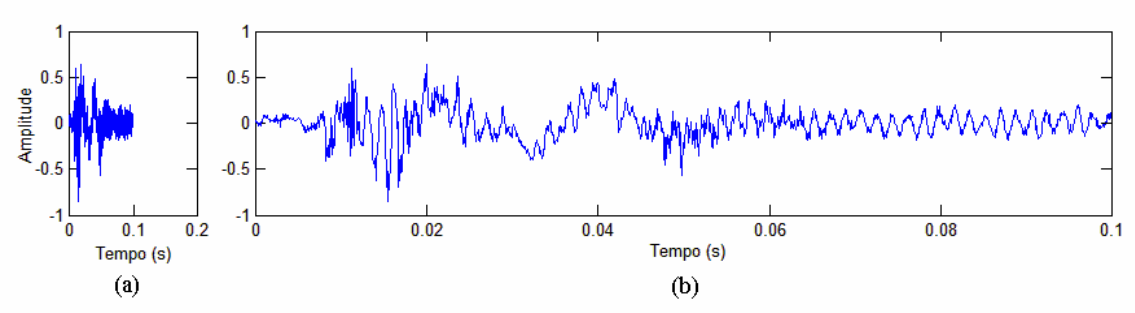

Figura 7 – Exemplo de áudio de referência: (a) e (b) apresentam o mesmo áudio com escalas de tempo diferenciadas.

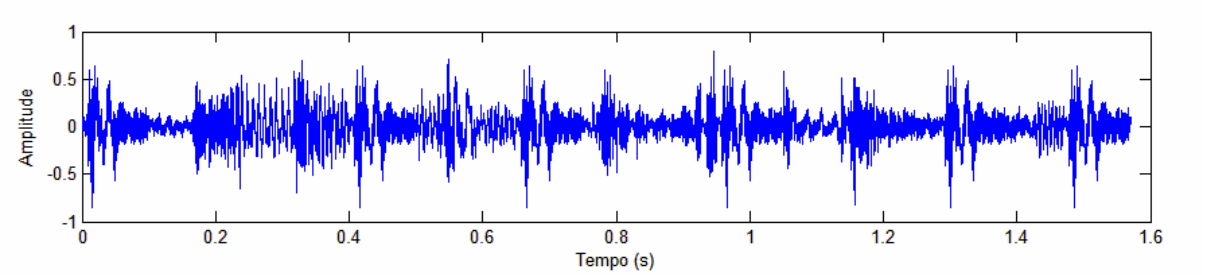

Figura 8 – Exemplo de áudio objeto.

Os testes básicos foram realizados considerando os seguintes aspectos:

- A variação causada pela alteração do método para cálculo da distância;
- A alteração no número de coeficientes LPC;
- A inserção de ruídos ao sinal sonoro;
- O não alinhamento do áudio repetitivo com relação aos quadros de análise.

A variação do método de cálculo de distância vetorial divide o protótipo em duas variantes, uma que utiliza a distância de Itakura-Saito como métrica e a outra que utiliza a distância euclidiana. Esta subdivisão e seus resultados serão melhor explanados nas seções 4.3 e 4.4.

O número de coeficientes LPC implica diretamente no desempenho do sistema, pois quanto maior o seu número, maior a quantia de características do áudio armazenada. A variação destes foi realizada testando o mesmo áudio diversas vezes, onde apenas o número de coeficientes LPC variava. Esta variação levou em conta de 4 à 24 coeficientes.

Para os testes que envolviam a inserção de ruído ao áudio objeto, o ruído foi inserido ao

ser adicionado ao áudio. Este ruído é do tipo gaussiano com amplitude unitária e largura de banda tendendo ao infinito, e foi gerado a partir da função *'awgn'* do Matlab. Este primeiro ruído gerado foi filtrado de modo a atenuar todas as suas componentes espectrais. E a variação dos testes envolvendo a adição de ruído ao sinal sonoro é dada pela atenuação aplicada ao ruído inicial. Foram utilizados ruídos que foram atenuados de 0 à 120dB. A adição do ruído ao áudio objeto tem como objetivo avaliar o desempenho do sistema com relação a ruídos adicionados no percurso entre a geração e recepção do sinal de áudio.

Para o teste que verifica o desempenho com relação a diferença de alinhamento do sinal sonoro com relação ao quadro analisado. Esta diferença de alinhamento consiste na diferença temporal entre um trecho sonoro contido em um quadro e sua cópia deslocada no tempo em um segundo quadro. Os arquivos que continham desalinhamento foram feitos ao incluir um sinal aleatório com duração de 1, 2, 2.5, 3, 4 e 5ms. A inserção destes trechos foi feita entre uma repetição e outra do áudio objeto analisado.

4.3 Testes utilizando a distância de Itakura-Saito

Esta análise alcançou resultados satisfatórios analisando 10 quadros (escolhidos do total de 20) do áudio de referência, sendo estes a metade do áudio com maior importância, ou seja, foram descartados os trechos com silêncio ou de períodos onde o áudio é estacionário. A seleção de tais quadros foi feita de maneira manual, ou seja, foi escolhido a posição inicial e a posição final da sequência de quadros utilizados.

O áudio objeto contido no arquivo *Mus02.wav,* apresenta seis repetições do áudio de referência, encontradas a partir da distância de Itakura-Saito, que tem seus valores mínimos demonstrados na Figura 9 (a), e a parte Figura 9 (b) permite a visualização do resultado da aplicação sobreposto a forma de onda do áudio objeto.

Com base na Figura 9 é possível ver os momentos em que há repetições, isto é percebido visualmente ao notar que em todos os trechos destacados o formato da onda sonora é idêntico.

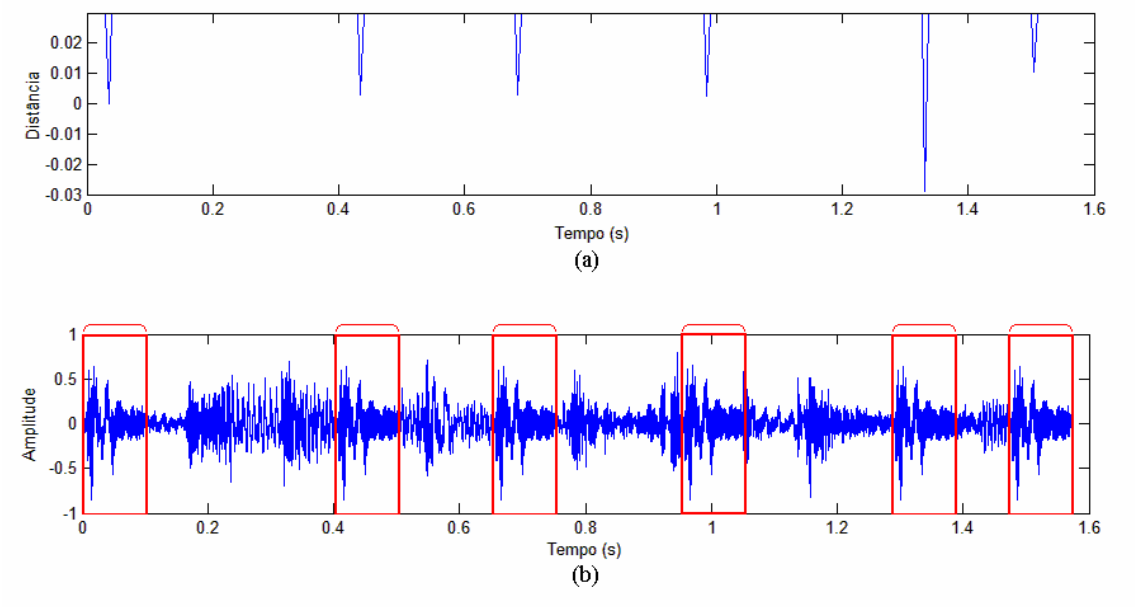

Figura 9 – Exemplo utilizando distância Itakura-Saito: (a) instantes em que houve detecção de repetição; (b) demonstração das repetições detectadas no áudio objeto.

Com relação aos limiares de comparação, resultados satisfatórios foram alcançados com os valores 2 e 0,03, respectivamente para primeira e segunda comparação. Estes valores indicam a distância máxima tolerada na comparação de um quadro, para o primeiro limiar, e de um trecho para o segundo limiar. Estes valores foram encontrados ao realizar diversos testes com áudios distintos. Em três testes realizados com o mesmo áudio, foram analisados a possibilidade de variação dos coeficientes LPC. As características desses testes são demonstradas na Tabela 1.

Tabela 1: Parâmetros utilizados nos testes com distância de Itakura-Saito.

Parâmetros (unidade)	**Teste 01**	**Teste 02**	**Teste 03**
Quadro inicial do áudio de referência (índice)	7	7	7
Número de quadros utilizados (unidade)	10	10	1
Primeiro limiar (adimensional)	2	10	2
Segundo limiar (adimensional)	0,03	0.15	0,03

Entre os parâmetros de entrada principais estão o quadro inicial e o final, e ambos os limiares de comparação. Assim o teste 01 demonstra os parâmetros onde houve

funcionamento ótimo para 12 coeficientes LPC. Para o teste 02, foi alterado o limiar em comparação com o teste 01, e o teste 03 teve apenas um quadro analisado na segunda comparação, enquanto o teste 01 analisou 10 quadros. A partir desses testes foram feitas as variações do número de coeficientes LPC, estas variações estão demonstradas na Figura 10.

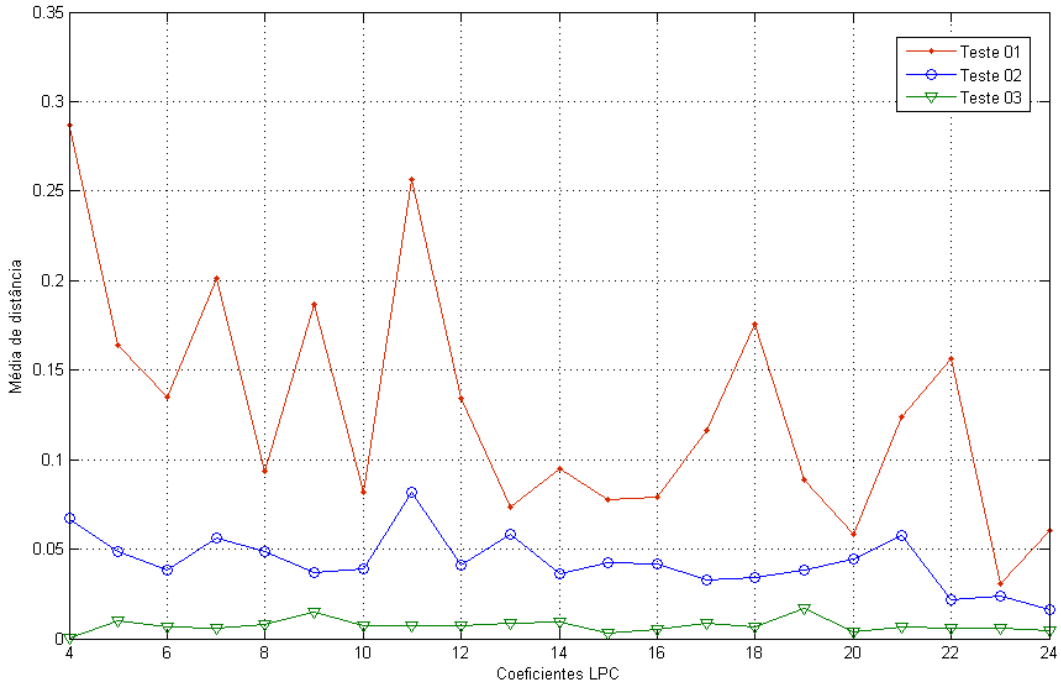

Figura 10 – Relação entre número de coeficientes LPC e média das distâncias mínimas encontradas para distância de Itakura-Saito.

Com esses testes percebeu-se que o limiar de decisão adotado depende diretamente do número de coeficientes LPC. No entanto, não foi possível encontrar um padrão que descreva esta variação. Tal evento é mostrado na Figura 10, nesta cada linha representa uma sequência de testes diferenciados para os mesmos arquivos de áudio. O valor representado pelo eixo da abscissa indica a média da distância encontrada para o teste realizado com determinado número de coeficientes LPC. E a variação no número de coeficientes está representado pelo eixo da coordenada do gráfico. As curvas representadas no gráfico demonstram a media do valor absoluto das distâncias encontradas como repetições. Destes testes todos obtiveram sucesso utilizando 12 coeficientes LPC, para o áudio *Mus*.

Com os valores médios de distância entre os pontos onde houve reconhecimento é possível estabelecer um critério para adotar os limiares de decisão. E a partir dos valores

encontrados, demonstrados na Figura 10, não foi possível encontrar algum tipo de padronização para os limiares adotados em todos os testes. Os valores utilizados como base dos testes (2 e 0,03) foram encontrados de maneira empírica.

A inserção de ruído no áudio objeto torna o sistema muito susceptível a erros. Por exemplo: nos testes, todos os ruídos inseridos que tinham como nível de atenuação de 80dB impossibilitaram o reconhecimento de todas as repetições. Mas a inserção de um ruído com 100dB de atenuação apresentou como resultado todas as repetições existentes e mais uma falsa. Todos os testes com ruídos entre 80 e 100 dB tiveram o resultado correto.

Outro fator analisado é a diferença de desempenho que o desalinhamento do quadro do áudio de referência com relação aos quadros analisados no áudio objeto. Verificou-se que uma diferença de alguns milisegundos no início do quadro temporal do áudio de referência em relação a posição em que este se repete no áudio objeto aumenta a probabilidade de detecção errada ou perda de repetições. Nos testes houve a constatação de erros quando havia deslocamento no tempo de 1 até 3ms.

Portanto, após todos esses testes constatou-se uma grande sensibilidade e instabilidade deste sistema com relação às variações de parâmetros gerais do sistema. Por apresentar falha ao deixar de reconhecer repetições em algumas situações, ou por encontrar repetições inexistentes em condições de qualidade ideais.

4.4 Testes utilizando a distância de euclidiana

Os limiares de decisão utilizados na avaliação foram fixados em 0,9 e 0,5 respectivamente para o primeiro e segundo limiar de comparação. Isto, pois estes valores apresentaram bons resultados em todos os testes. A distância euclidiana encontrada para o exemplo (áudio *Mus*) está demonstrada na Figura 11 (a). A Figura 11 (b) destaca tais distâncias encontradas na forma de onda do som objeto, isto de maneira a comprovar a eficiência do aplicativo.

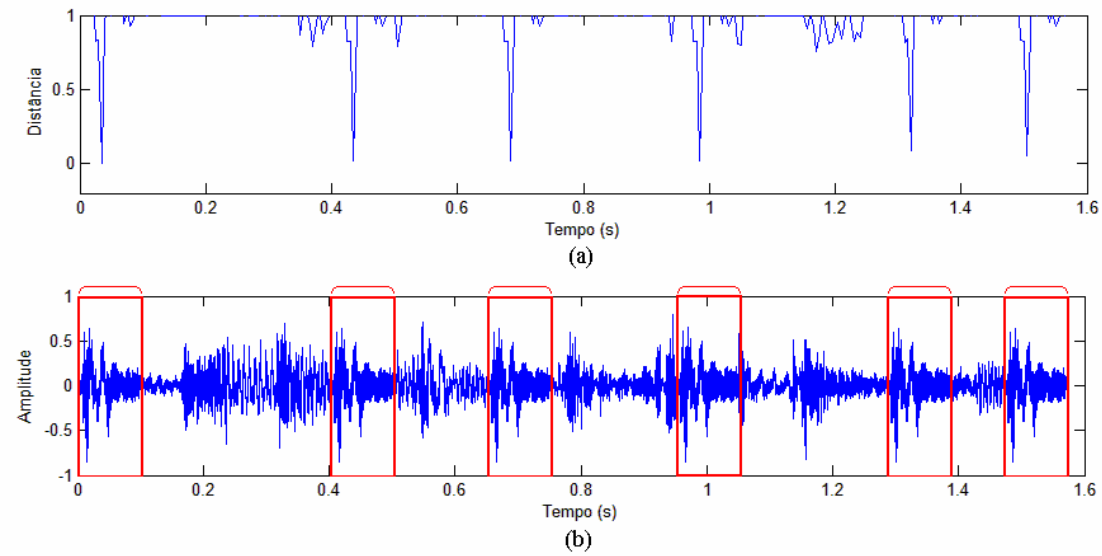

Figura 11 – Exemplo utilizando distância euclidiana: (a) instantes em que houve detecção de repetição; (b) demonstração das repetições detectadas no áudio objeto.

O resultados da distância euclidiana, como mostrados na Figura 11 (a) indica uma pequena variação destes valores. Isto significa que o limiar de decisão não vai influenciar de maneira brusca o resultado do sistema. Para comprovação foi realizado testes que avaliam a média das distâncias encontradas com o número de coeficientes LPC utilizados na análise, e os parâmetros iniciais desses testes estão descritos na e seus resultados estão demonstrados na Tabela 2.

Tabela 2: Parâmetros utilizados nos testes com distância euclidiana.

Parâmetros (unidade)	**Teste 01**	**Teste 02**	**Teste 03**
Quadro inicial do áudio de referência (índice)	7	7	7
Número de quadros utilizados (unidade)	10	10	1
Primeiro limiar (adimensional)	0,9	5	0,9
Segundo limiar (adimensional)	0,5	2	0,5

Os parâmetros de entrada são os mesmos utilizados para a variante que realiza o calculo da distância de Itakura-Saito. Com isso o primeiro teste demonstra os parâmetros onde houve funcionamento ótimo para 12 coeficientes LPC. Para o segundo teste, foi alterado o limiar em

comparação com o primeiro teste, e o terceiro teste teve apenas um quadro analisado na segunda comparação, enquanto o primeiro teste analisou 10 quadros. A partir desses testes foram feitas as variações do número de coeficientes LPC, estas variações estão demonstradas na Figura 12.

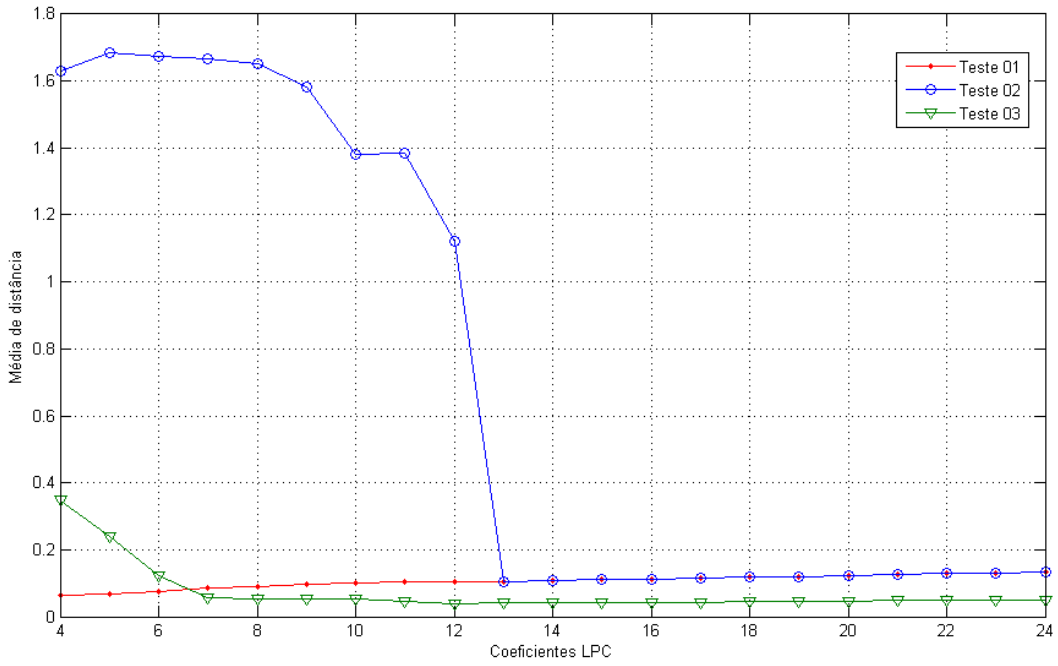

Figura 12 – Relação entre número de coeficientes LPC e média das distâncias mínimas encontradas para distância euclidiana.

As curvas representadas no gráfico demonstram a media do valor das distâncias encontradas como repetições. O valor representado pelo eixo da abscissa indica a média da distância encontrada para o teste realizado com determinado número de coeficientes LPC. E a variação no número de coeficientes está representada pelo eixo da coordenada do gráfico. Destes testes todos obtiveram sucesso utilizando 13 ou mais coeficientes LPC, para o áudio *Mus*. Houve falhas no reconhecimento apenas para o segundo teste com 12 ou menos coeficientes LPC. Estas curvas estão relacionadas a três testes delimitados pelos parâmetros descritos na Tabela 2.

Neste gráfico, todas as situações em que a média de distância foi maior que 0,6 houve erro na detecção de repetição. Portanto o segundo limiar foi definido como 0,5 para garantir o funcionamento do sistema, e o primeiro limiar foi limitado em 0,9 para diminuir

consideravelmente a carga computacional do sistema e mesmo assim continuar com resultados satisfatórios.

A análise visual do gráfico de distância mínima para todos os testes considerados na Figura 12, levou a conclusão que quanto maior o número de parâmetros LPC melhor o desempenho do sistema como um todo, devido a maior precisão alcançada pelo cálculo de distância.

Nos testes de tolerância a adição de ruído gaussiano ao áudio objeto, teve resultados aceitáveis de reconhecimento onde os áudios continham ruídos que com atenuações acima de 45 dB. Ou seja, todo áudio com ruído que tinha atenuação menor que 45 dB teve sérios problemas de detecção. Nestes casos normalmente o sistema não encontrou nenhuma repetição. Porém acima deste valor todas as repetições foram encontradas, e não houve nenhum caso de encontrar repetições inexistentes ou falhas.

A diferença de alinhamento dos quadros com relação ao áudio de referência não causou grande influência no desempenho, havendo sempre o reconhecimento das repetições nos momentos correto. Apesar de graficamente piorar um pouco o resultado, para atrasos de 2 a 3ms, o limiar de decisão auxiliou no reconhecimento correto das repetições.

Estes resultados mostraram que as simulações utilizando a distância de Itakura-Saito não apresentaram resultados coerentes para praticamente todos os testes. Enquanto as simulações que utilizaram a distância euclidiana tiveram resultados coerentes e satisfatórios. Tal fato contrariou a teoria, descrita por Becerra (1993), que indicava que o melhor desempenho da medida de distância vetorial era dada pela métrica de Itakura-Saito.

5 *Conclusões*

Este trabalho teve como objetivo desenvolver um protótipo que realizasse o reconhecimento da repetição de um áudio de referência em um áudio objeto. Para alcançar este objetivo foi realizado o estudo das diversas técnicas utilizadas para a parametrização do áudio. Com base nesses estudos foi decidida a utilização da análise LPC para a retirada dos parâmetros do áudio, isto por esta técnica ser a mais indicada para este tipo de aplicação. Dentre os algoritmos de comparação foi optado pela utilização de duas medidas de distância vetorial, estas medidas são a de Itakura-Saito e a Euclidiana. Ambas foram escolhidas devido a maior facilidade de aplicação e devido ao caráter genérico apresentado por estas, ou seja, estas técnicas eram voltadas para comparação de todo tipo de áudio, e não apenas voz.

Após todos os estudos, foi realizada uma gama de testes envolvendo três ambientes de desenvolvimento, estes foram o MARF, o HTK e o Matlab. Os testes foram feitos com o intuito de desenvolver meios para comparar o desempenho do protótipo a ser desenvolvido. Contudo os testes efetuados com o MARF e o HTK tiveram resultados aquém do esperado. Já os testes com o Matlab não proporcionaram os meios de comparação de desempenho, porém mostraram potencial deste ambiente de desenvolvimento. Assim o Matlab foi escolhido para o desenvolvimento do protótipo, isto com o auxilio da biblioteca Voicebox.

O protótipo desenvolvido, no geral teve um desempenho bom. Foi alcançado o objetivo de reconhecer a repetição de um trecho sonoro determinado.

A utilização da distância de Itakura-Saito trouxe resultados bem positivos para alguns tipos de configuração. Assim podendo ser bem útil para ambientes em que não haja variação da configuração do sistema. Porém, ao variar sua configuração demonstrou grande instabilidade, podendo não reconhecer repetições perfeitas ou reconhecendo reconhecimento de repetições inexistentes. Mostrando assim resultados incoerentes.

O protótipo que utiliza a distância euclidiana também trouxe resultados positivos, além de apresentar um comportamento linear (houve melhora com o aumento do numero de quadros analisados) com relação a variação de suas configurações. Em todos os testes

realizados com ambas as variantes do sistema com os parâmetros básicos estipulados, em nenhum caso o código que utiliza distância euclidiana encontrou repetições inexistentes, ou seja, apenas falhou ao detectar a repetição. Com este protótipo foi possível alcançar resultados completamente corretos em até 88% dos casos (53 testes de 60), para os parâmetros básicos. Nos testes realizados com a distância de Itakura-Saito, este desempenho foi menor, alcançando em torno dos 72% (43 testes de 60).

Com a aplicação desenvolvida, de modo geral, os testes realizados obtiveram um melhor desempenho ao utilizar a métrica de distância euclidiana como meio de decisão da ocorrência das repetições no som, se comparado a distância de Itakura-Saito. Isto, pois houve menor variabilidade dos resultados, comprovando menor sensibilidade do sistema que utiliza distância euclidiana, contradizendo a teoria descrita por Becerra (1993).

5.1 Trabalhos futuros

Diante do trabalho desenvolvido, e dos resultados obtidos, consideramos que em continuidade ao trabalho podem ser estudadas as seguintes sugestões:

- Continuar o desenvolvimento deste protótipo de modo a torná-lo um sistema estável para o reconhecimento de áudio repetitivo.
- Implementar um sistema que utilize outros métodos de parametrização do áudio, como Mel-cepstral , para comparar seu desempenho com a predição linear (LPC);
- Realizar a análise de áudio repetitivo utilizando formatos de áudio que tenham perdas causadas por compressão, como MP3 ou Vorbis;
- Implementar o sistema em *hardware* e aplicá-lo a detecção e medição de áudio de propagandas;
- Utilizar o sistema para reconhecimento de repetições em músicas para demonstração jurídica em casos de plágio;
- Usar correlação cruzada no tempo para verificar se há ou não repetições.
- Desenvolver um sistema com outras medidas de distância;
- Encontrar uma maneira de determinar de modo automático os limiares de comparação.
- Desenvolver alguma aplicação relacionada a reconhecimento de áudio em ambientes de desenvolvimento como o MARF ou o HTK.

Lista de Acrônimos

API – Interface de Programação de Aplicativos (*Application Programming Interface);*

A/D – Conversão do sinal analógico para o sinal digital;

C – Formato de linguagem de programação C;

CD – Disco compacto *(Compact Disc);*

D/A – Conversão do sinal digital para o sinal analógico;

DFT – Transformada discreta de Fourier (*discrete Fourier transform*);

DTW – *Dynamic time warping*;

FFT – Transformada rápida de Fourier (*fast Fourier transform*);

GSM – *Global System for Mobile Communications;*

HMM – Modelo oculto de Markov (*Hidden Markov Model);*

HTK – *Hidden Markov Model Toolkit;*

JAVA – Formato de linguagem de programação JAVA;

LAR – *Logarithmic Area Ratio*;

LPC – Codificação de predição linear (*Linear Predictive Coding);*

LSF – *Line Spectrum Frequencies;*

MARF – *Modular audio recognition framework*;

MP3 – MPEG-1 *Audio Layer* 3;

MPEG – *Moving Picture Experts Group;*

OGG – Container para arquivos do tipo Vorbis;

PCM – Modulação por código de pulso (*Pulse code modulation*);

URA – Unidade de Resposta Audível;

WAV – *Waveform audio format;*

Referências Bibliográficas

ABNT. **Quantidades, unidades e símbolos das grandezas acústicas fundamentais**. Rio de Janeiro, v.7, n.45, p. 50-72, jul./ago. 1959.

AKISHINO, S. A. **Reconhecimento de palavras faladas.** Trabalho da disciplina de Processamento Digital de Sinais do curso de Engenharia Elétrica, Setor de Tecnologia da Universidade Federal do Paraná. Curitiba, 2007.

BECERRA, J. **Reconhecimento Automático de Palavras Isoladas:** Estudo e Aplicação de Métodos Determinísticos. Dissertação apresentada a faculdade de engenharia elétrica da Unicamp como requisito parcial para obtenção do título de mestre. Nov, 1993.

BERGAMO, R. T. . **Introdução as comunicações móveis**. CEFET/SC. São José, 2007.

BLOCH, P. .**Estudos da Voz Humana.** Rio de Janeiro: BRASILUSO, 1958.

BONJORNO, J. R., RAMOS, C. M. .**Física 2** : termologia, óptica geométrica, ondulatória. São Paulo: FTC, 1992.

DIGITAL audio. **Wikipedia the free encyclopedia**. Disponível em: <http://en.wikipedia.org/wiki/Digital_audio>. Acessado em: 11 de nov. de 2008.

ETSI ES 201 980. **Digital Radio Mondiale (DRM).** Disponível em: < http://www.drm.org/uploads/media/ETSI_ES_201_980_v2.3.1.pdf >. Acessado em: 2 de dez. de 2008.

FERREIRA, Aurélio, B. H. . **Novo dicionário da língua portuguesa**. Rio de Janeiro: Nova Fronteira, 1975.

FILE extension GSM. **File extencions.** Disponível em: <http://www.file-extensions.org/gsm-file-extension-raw-gsm-6-10-audio-stream-file>.Acessado em: 2 de dez. de 2008.

HOSOM, J. P.; COLE, R.; FANTY, M. . **Speech recognition using neural networks at the center for spoken language understanding**. 1999. Disponível em: <http://cslu.bme.ogi.edu/tutordemos/nnet_recog/recog.html>. Acessado em: 16 de jan. 2009.

MODULAR audio recognition framework v.0.3.0.6 (0.3.0 final) and its applications. **MARF**, The Modular Audio Recognition Framework, and its Applications, 2007.Disponível em: < http://marf.sourceforge.net/docs/marf/0.3.0.6/report.pdf>. Acessado em: 3 de out. de 2008.

MOECKE, M. . **PCM**: modulação por código de pulso. Apostila para o curso superior de tecnologia em Sistemas de Telecomunicações. São José, 2006.

MOKHOV, S. A. . **Introducing MARF**: a modular audio recognition framework and its applications for scientific and software engineering research. Montreal, 2008.

MP3. **Wikipedia the free encyclopedia**. Disponível em: <http://en.wikipedia.org/wiki /MP3>. Acessado em: 11 de nov. de 2008.

RIBEIRO, C. E. M. .**Processamento digital de fala**. Instituto superior de engenharia de Lisboa. Disponível em: <http://www.deetc.isel.ipl.pt/comunicacoesep/disciplinas/pdf/>. Acessado em: 18 de nov. de 2008.

SOARES, P. A. T., FERRARO, N.G. , SANTOS, J. I. C. . **Aulas de física**. São Paulo: Atual, 1984.

SOUND recording and reprodution. **Wikipedia the free encyclopedia**. Disponível em: <http://en.wikipedia.org/wiki/Sound_recording_and_reproduction> . Acessado em: 16 de dez. de 2008.

VOICEBOX: Speech Processing Toolbox for MATLAB. **Imperial College.** Disponível em: < http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>. Acessado em: 3 de out. de 2008.

VORBIS. **Wikipedia the free encyclopedia**. Disponível em: <http://en.wikipedia.org/wiki/ Vorbis>. Acessado em: 11 de nov. de 2008.

WAV. **Wikipedia the free encyclopedia**. . Disponível em: <http://en.wikipedia.org/wiki/ Wav >. Acessado em: 11 de nov. de 2008.

YOUNG, S. et al. **The HTK book**. 2006 Disponível em: < http://htk.eng.cam.ac.uk/download.shtml>. Acessado em: 3 de out. de 2008.