

Márcio Geovani Jasinski

*Comparação entre metodologias de análise  
de sinal aplicadas ao reconhecimento de  
voz utilizando um vocabulário restrito*

Projeto de Pesquisa para elaboração do Trabalho de Conclusão de Curso apresentado como exigência para a obtenção do título de Bacharel em Ciências da Computação à Universidade Federal de Santa Catarina - UFSC, no curso de Ciências da Computação.

Orientador:

Prof. Dr. rer.nat. Aldo von Wangenheim

BACHARELADO EM CIÊNCIAS DA COMPUTAÇÃO  
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA - INE  
CENTRO DE TECNOLÓGICO - CTC  
UNIVERSIDADE FEDERAL DE SANTA CATARINA - UFSC

Florianópolis - SC

Julho / 2006

Trabalho de conclusão de curso sob o título “*Comparação entre metodologias de análise de sinal aplicadas ao reconhecimento de voz utilizando um vocabulário restrito*”, defendido por Márcio Geovani Jasinski sob aprovação, em Florianópolis, Santa Catarina, pela banca examinadora constituída:

---

Prof. Dr. rer.nat. Aldo von Wangenheim  
Departamento de Informática e Estatística - INE  
Orientador

---

M.Sc. Rafael Simon Maia  
Departamento de Informática e Estatística - INE  
Coorientador

---

Prof. Dr. Mauro Roisenberg  
Departamento de Informática e Estatística - INE

---

M.Sc. Rafael Andrade  
Departamento de Informática e Estatística - INE

---

Prof. Dr. João Candido Dovicchi  
Departamento de Informática e Estatística - INE

*Para pessoas especiais que me ensinaram a crescer, viver e estudar*  
*Minha família*

# *Agradecimentos*

Aos membros do Projeto Cyclops que sempre estiveram dispostos em discutir idéias e resolver problemas. Ao meu orientador, Prof. Aldo von Wangenheim e membros da banca, pelas críticas e sugestões que melhoraram em muito o resultado final deste trabalho. Um agradecimento especial aos revisores do texto que contribuíram para correções e mudanças estruturais no texto: Rafael Andrade, Rafael Simon Maia, Antônio da Luz Junior, Daniel Duarte Abdala, Éverton Fabian Jasinski e Larissa da Veiga.

Ao professor Antonio Carlos Mariani, com quem trabalhei durante dois anos e meio e tive enormes aprendizados nas ciências da computação. Aos membros do Laboratório de Sistemas de Conhecimento, grandes amigos sempre dispostos a trocar informação e ensinamentos.

Agradeço aos meus amigos de curso, companheiros dos trabalhos de aula, e todos os membros da CCO022, em especial a todos que compartilharam noites programando para cumprir os prazos de entrega dos trabalhos.

Aos meus tios, Lindomar Varella e Ilse Varella, pela confiança depositada em mim e pelo apoio recebido durante toda a faculdade.

A minha namorada, Larissa da Veiga, pela paciência, pelo incentivo, por acreditar em mim e por ser tão maravilhosa. Minha fonte de inspiração e perspectiva de um ótimo futuro.

Ao meu irmão Éverton Fabian, pela valorização e proteção desde que éramos crianças. Obrigado pelos exemplos de caráter, personalidade e inteligência. Um irmão incrível e o melhor amigo que se pode ter. Ao meu irmãozinho Vinícius César Jasinski, pelo carinho e por me mostrar a alegria que existe em simplesmente brincar.

Aos meus pais pela educação, e pelo apoio, ensinamentos e correções que sempre recebi nas inúmeras decisões da vida. Pais carinhosos e exemplares, meus eternos tutores na vida pessoal e profissional.

# *Resumo*

O processo de geração de laudos pode ser dinamizado com a aplicação de novas tecnologias em informática médica. Inovações na telemedicina tornaram acessíveis via *web*, recursos como imagens e laudos estruturados DICOM, ou, ainda, laudo em áudio digital. Este trabalho aborda tecnologias que possibilitam melhorias no processo de laudo médico.

As pesquisas são realizadas no Projeto Cyclops, que atua desde 1998 no desenvolvimento de *software* para medicina. Entretanto, o projeto enfrenta resistência por parte dos médicos, devido a usabilidade limitada inerente às interfaces de *mouse* e teclado.

O foco do trabalho está em formas de reduzir esta resistência, neste caso, através do laudo ditado em formato digital e processamento de sinal. Assim, aborda técnicas em reconhecimento de voz sobre o áudio com vocabulário restrito e apresenta o estado da arte, com as principais aplicações existentes no mercado para reconhecimento de voz. Também são descritas as teorias existentes como Redes Neurais e Modelos Ocultos de Markov que permitem a transcrição do áudio para texto.

A metodologia proposta para *softwares* de laudo ditado considera os processos de captura de áudio, armazenamento, recuperação, aplicação de *tags* e reconhecimento de voz. Assim, o trabalho avalia o a utilização de arquivos em diferentes formatos de áudio para armazenamento em banco de dados com acesso via *web*.

Os protótipos desenvolvidos para dinamizar o processo de laudo médico são descritos ao final deste trabalho. A proposta aborda uma metodologia modular com soluções livres e portáteis onde a independência dos módulos desenvolvidos permite o reuso em diferentes aplicações.

**Palavras Chaves:** Reconhecimento de voz, Áudio digital, laudo ditado, telemedicina, Modelos Ocultos de Markov.

# *Abstract*

The report generating process time can be optimized by using new computer technologies on medical area. Breakthroughs in telemedicine have made access possible from web , resources, such as, images and DICOM structured reports, or even, reports in digital audio format. This monograph approaches technologies to optimize medical report process.

All research has support of Cyclops Project that have experience on medical area since 1998, however the project faces resistance of the some medical teams because of restricted usability inherent of keyborad and mouse interfaces.

This project is focused in ways to reduce medical teams resistance using speech report stored on digital format. Moreover, its approaches techniques on speech recognition with a restricted vocabulary and this research show the state of art with some applications on market to do speech recognition. Theories about how can be done speech recognition to text are described such as Hidden Markov Models and Neural Networks.

The metodology propused for speech report software consider the processes of capture of audio, storage, recovery, TAGS application and voice recognition. Thus, it's evaluates some kinds of archives in different formats to audio storage that have access from web.

The report process can be improved using speech recognition that offers a natural succession from dictation-transcription process. In this project some applications that offer speech recognition are commented and some theory involved on Hidden Markov Models Toolkit is presented.

The prototypes developed to improve medical report process will be described on last part of this monograph. It propose modular methodology do implement portable systems using free solutions. Independent modules allow integration with differents kinds of applications.

**Keywords:** Speech Recognition, Digital Audio, Speech Report, telemedicine, Hidden Markov Models

# *Lista de Figuras*

1	Laudo manuscrito (esq.), fitas usadas na em laudo ditado (cen.) e laudo impresso para confirmação (dir.) . . . . .	p. 17
2	Laudo ditado durante o exame (dir.) e revisão do laudo após reconhecimento de voz automático (esq.) . . . . .	p. 18
3	Laudo estruturado no padrão DICOM . . . . .	p. 26
4	Sistema PACS adaptado ao HU-UFSC . . . . .	p. 27
5	Abstração PortAudio para aplicações . . . . .	p. 33
6	Três arquivos de 20 segundos concatenados no mesmo arquivo. Da esquerda para a direita o mesmo áudio em WAV, MP3 e OGG respectivamente(AMMOURA; CARLACCI, 2002) . . . . .	p. 34
7	Modelo Oculto de Markov(YOUNG et al., 2005) . . . . .	p. 49
8	Rede Neural Aprendendo a função seno(YOUNG et al., 2005) . . . . .	p. 51
9	Abstração para implementar diferentes bibliotecas de áudio . . . . .	p. 57
10	Abstração de reconhecimento de voz HTK . . . . .	p. 58
11	Portal de telemedicina sem suporte a laudo ditado . . . . .	p. 59
12	Portal de telemedicina com suporte a laudo ditado . . . . .	p. 60
13	Fluxo de informação no sistema SR . . . . .	p. 60
14	Comunicação entre os módulos para reconhecimento de voz com o HTK. . . . .	p. 62
15	Comparação no armazenamento dos formatos WAV, MP3, OGG, Speex . . . . .	p. 64
16	Redução do tamanho do arquivo entre MP3, OGG, Speex . . . . .	p. 65
17	Proposta de implementação para laudo ditado e reconhecimento de voz. . . . .	p. 67
18	Programa Cycreport reproduzindo arquivo MP3 . . . . .	p. 68
19	Digitação de um exames obtido do servidor PACS . . . . .	p. 69

20	Aviso de forma visual para o digitador . . . . .	p. 70
21	Autômato gerado a partir da gramática de 5 palavras . . . . .	p. 72
22	Marcando partes do áudio com o <i>HSLab</i> . . . . .	p. 73
23	Diagrama UML simplificado . . . . .	p. 87



## *Lista de Tabelas*

1	Comparação entre formatos MPEG-1 . . . . .	p. 39
2	Parâmetros que caracterizam um sistema ASR(MIT, 2003) . . . . .	p. 44
3	Perplexidades em diferentes domínios(YNOGUTI, 1999) . . . . .	p. 44
4	Vantagens do formato OGG sobre o MP3 . . . . .	p. 63
5	Tamanho em Kbytes dos laudos gravados em diferentes formatos de áudio	p. 65
6	Formatos com suporte do programa Cyclops Áudio Report . . . . .	p. 70
7	Gramática com 5 palavras . . . . .	p. 71

# *Lista de acrônimos e abreviações*

AAC - Codificação avançada de áudio

AIFF - Audio Interchange File Format

ALSA - Advanced Linux Sound Architecture ou Arquitetura de Som Avançada Linux

API - Application Programming Interface ou Interface de Programação de Aplicativos

AU - Audio for Unix

BSD - Berkeley Software Distribution

CELP - Code-Excited Linear Prediction

CODEC - Codificador, Decodificador

DICOM: Digital Imaging and Communications in Medicine

GNU - Gnu is not unix

GPL - Gnu public license

HMM - Modelo Markoviano Oculto

HU - Hospital Universitário da UFSC

ID3 - abreviação de Identify an MP3

LGPL - GNU Lesser General Public License

LIRC - Linux Infrared Remote Control

MFCC - Mel Frequency Cepstral Coefficient

MIT - Massachusetts Institute of Technology

MLF - Master Label File

MP3 - MPEG Layer-3

OGG- Nome do formato de áudio, vídeo e metadados da Xiph.org

OSALP -Open Source Áudio Library

OSS - Open Sound System

PACS - Picture Archiving and Communication Systems

PCM - Pulse-code modulation

PortAudio - Portable cross-platform Audio API

SGI - Silicon Graphics, Inc

SR - Speech Recognition ou Structured Report (de acordo com o contexto)

TAG- Do inglês, etiqueta

UFSC - Universidade de Santa Catarina

VQF - Transform-domain Weighted Interleave Vector Quantization (Vector Quantization Format ).

WAV ou WAVE - Waveform audio format

WMA - Windows Media Audio

# *Sumário*

<b>1</b>	<b>Introdução</b>	p. 16
1.1	Projeto Cyclops . . . . .	p. 18
1.2	Motivação . . . . .	p. 19
1.3	Objetivos . . . . .	p. 20
1.3.1	Objetivo Geral . . . . .	p. 20
1.3.2	Objetivos Específicos . . . . .	p. 20
1.4	Justificativa . . . . .	p. 21
1.5	Estrutura do trabalho . . . . .	p. 22
<b>2</b>	<b>Fundamentação Teórica</b>	p. 24
2.1	Tecnologias de Laudo Médico . . . . .	p. 24
2.1.1	Padrão DICOM . . . . .	p. 25
2.1.1.1	DICOM Structured Report . . . . .	p. 26
2.1.2	PACS . . . . .	p. 26
2.1.3	Telemedicina . . . . .	p. 27
2.1.4	Áudio Digital . . . . .	p. 28
2.1.5	Reconhecimento de voz . . . . .	p. 29
2.2	Manipulação de áudio . . . . .	p. 30
2.2.1	Open Source Audio Library . . . . .	p. 31
2.2.2	PortAudio . . . . .	p. 32
2.3	Tecnologias de compactação de áudio . . . . .	p. 32
2.3.1	Taxa de bits (Bitrate) . . . . .	p. 35

2.3.2	Frequência . . . . .	p. 35
2.3.3	Amostragem (Sampling) . . . . .	p. 35
2.3.4	Qualidade do áudio . . . . .	p. 36
2.3.5	Classificação dos CODECs . . . . .	p. 36
2.3.6	Tags ID3v2 . . . . .	p. 37
2.3.7	Formato de áudio MP3 . . . . .	p. 37
2.3.8	Formato de áudio OGG . . . . .	p. 39
2.3.9	Speex . . . . .	p. 40
2.4	Sistema de reconhecimento automático de voz . . . . .	p. 40
2.4.1	Flexibilidade do sistema . . . . .	p. 41
2.4.2	Classificação do sistema de reconhecimento de voz . . . . .	p. 41
2.4.3	Obstáculos no reconhecimento de voz . . . . .	p. 42
2.5	Modelos para reconhecimento de voz . . . . .	p. 44
2.5.1	Análise da Fala . . . . .	p. 45
2.5.1.1	Formantes . . . . .	p. 45
2.5.2	Modelos Markovianos . . . . .	p. 46
2.5.3	Modelos Ocultos de Markov . . . . .	p. 46
2.5.4	Reconhecimento de palavras isoladas . . . . .	p. 47
2.5.5	Reconhecimento de voz contínuo . . . . .	p. 49
2.5.6	Redes Neurais . . . . .	p. 49
2.5.7	Estado da arte . . . . .	p. 51
<b>3</b>	<b>Metodologia</b>	p. 56
3.1	Ferramentas de desenvolvimento . . . . .	p. 56
3.1.1	Biblioteca áudio . . . . .	p. 57
3.1.2	Biblioteca de reconhecimento de voz . . . . .	p. 58
3.1.3	PHP e PostgreSQL . . . . .	p. 58

3.2	Metodologia para reconhecimento de voz . . . . .	p. 59
3.3	Análise dos formatos de áudio . . . . .	p. 62
3.4	Implementação . . . . .	p. 65
3.4.1	Cyclops Report Recorder . . . . .	p. 66
3.4.1.1	Funções do programa . . . . .	p. 68
3.4.2	Reconhecimento de voz com HTK . . . . .	p. 71
<b>4</b>	<b>Conclusão</b>	p. 74
	<b>Referências</b>	p. 76
	<b>Anexo A – Configuração acústica</b>	p. 81
	<b>Anexo B – Arquivo HMM base</b>	p. 82
	<b>Apêndice A – Documentação do Protótipo Cycreport</b>	p. 84
A.1	Levantamento de Requisitos . . . . .	p. 85
A.1.1	Objetivo . . . . .	p. 85
A.1.2	Domínio . . . . .	p. 85
A.1.3	Requisitos Funcionais . . . . .	p. 85
A.1.3.1	O que o sistema deve permitir ao usuário . . . . .	p. 85
A.1.3.2	Diagrama de classes e Casos de Uso . . . . .	p. 85
A.1.3.3	Documentação . . . . .	p. 86
A.1.4	Requisitos não funcionais . . . . .	p. 86
A.1.4.1	Confiabilidade . . . . .	p. 86
A.1.4.2	Desempenho . . . . .	p. 86
A.1.4.3	Portabilidade . . . . .	p. 86
A.1.4.4	Usabilidade . . . . .	p. 86
A.1.5	Digrama de classes . . . . .	p. 86

Apêndice B - Redução de informação em arquivos de áudio	p. 88
Apêndice C - Artigo	p. 89
Apêndice D - Cabeçalho formato WAV	p. 96

# 1 *Introdução*

O sucesso e a disseminação dos programas para diagnóstico médico por imagem incentivou a pesquisa no reconhecimento de voz contínuo na medicina. Porém, no Brasil, o atual método de geração de laudo radiológico pouco explora as vantagens dos avanços em processamento de sinais. O sistema de laudo manuscrito com imagens em filmes precisa ser adaptado aos programas médicos atuais e futuros. Existem esforços nessa área onde é feito o armazenamento das informações pertinentes ao exame com laudo digitado, por exemplo PACS<sup>1</sup>(CYCLOPS, 2006). Entretanto, a digitação do texto pode ser substituída pelo reconhecimento de voz que permite reduzir custos operacionais e o tempo do processo(WHITE, 2005).

A geração de laudos médicos em clínicas e hospitais difere entre instituições desde a metodologia até a tecnologia disponível na instituição. Por exemplo, procedimentos do HU/UFSC com laudo manuscrito e posterior digitação, diferem da clínica DMI<sup>2</sup>, com laudo gravado em fitas micro-cassetes transcritas por uma equipe especializada. Além disso, pode variar entre as instituições, a tecnologia dos aparelhos para exames e o volume de atendimentos diários.

O primeiro cenário, é o de laudo em formulário manuscrito, utilizado em setores do HU/UFSC fica sujeito aos seguintes erros: rasura, letra ilegível, troca de paciente entre exames e documento perdido. O processo de geração do laudo é lento e o armazenamento destes documentos torna inviável a recuperação para consulta, dado o volume de exames realizados a cada dia.

Uma alternativa ao processo manuscrito, utilizada na Clínica DMI, é ditar o laudo e armazená-lo em fita para posterior transcrição do mesmo. Embora a estratégia seja mais dinâmica (Figura 1), alguns entraves ainda existem. As fitas são regravadas o que

---

<sup>1</sup>Picture Archiving and Communications System, do inglês, Sistemas de arquivamento e comunicação de imagens

<sup>2</sup>Diferenças entre o Hospital Polydoro Ernani de São Thiago e Clínica DMI - Diagnóstico Médico por Imagem, ambos de Florianópolis-SC



elimina o registro para posterior consulta e deteriora a qualidade do áudio. Os nomes dos pacientes são marcados nas fitas por etiquetas, ocasionando problemas de letra ilegível e troca de paciente entre exames. A utilização desse procedimento demanda uma equipe de digitadores a ser mantida para a transcrição do laudo, que precisa ser revisado pelo médico para corrigir erros de interpretação.

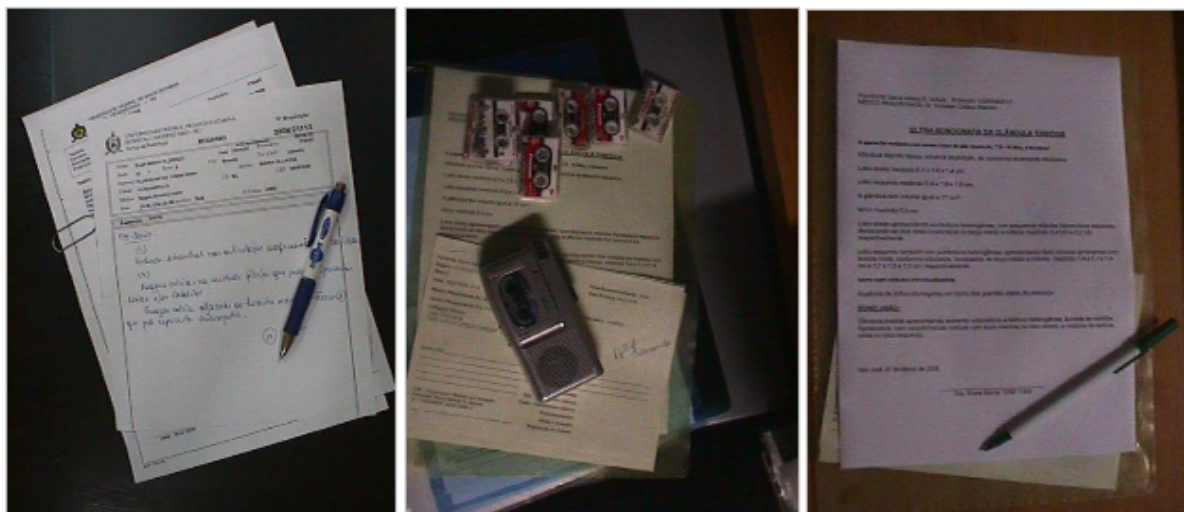


Figura 1: Laudo manuscrito (esq.), fitas usadas na em laudo ditado (cen.) e laudo impresso para confirmação (dir.)

O ambiente ideal deve utilizar os benefícios da tecnologia da informação para geração de laudo. O uso de PACS e gravação do laudo em áudio digital permite corrigir deficiências nos métodos tradicionais citados. Dessa maneira, um áudio de alta qualidade com fácil acesso, possibilita consultas do laudo via *software/web site* de forma mais rápida e segura. Além disso, elimina a possibilidade de trocas de pacientes entre diferentes exames, pois o laudo ditado está sempre associado a exame e paciente corretos. Nesse ambiente (Figura 2) os problemas de documentos perdidos, rasurados, com letra ilegível e transcrição de ditados são solucionados. O laudo é armazenado em áudio digital e a transcrição é feita por um sistema de reconhecimento de voz. Assim, o médico faz a revisão do laudo via *web*, onde estão as informações do exame: imagens, dados do paciente, laudo ditado e em formato de texto.

O uso de sistemas de informação para melhorar a qualidade e produtividade nos serviços médicos é um fato amplamente aceito. Segundo a PHILIPS ELECTRONICS, nomeado Líder Europeu de Tecnologia pela Frost Sullivan, 2006:

*“Um dos fatores mais cruciais na área da saúde a nível europeu é a necessi-*

*dade do utilizador final obter maior eficiência e produtividade. As vantagens combinadas dos sistemas de reconhecimento de voz e dos sistemas de informação para os serviços de saúde, tal como o registo médico eletrónico (EMR - Electronic Medical Record) ou os sistemas de informação hospitalares (HIS - Hospital Information Systems), são importantes para alargar a capacidade de produção das instituições de saúde, bem como dos indivíduos.”*

Este trabalho aborda o uso das principais tecnologias médicas computacionais afim de dinamizar o processo de laudo. De acordo com DELLANI, 2001, o sistema PACS em conjunto com os sistemas de informação radiológica (RIS) e de informação hospitalar (HIS) formam a base informatização de hospitais e clínicas. A implantação de um desses sistemas traz melhorias em acesso, consulta e integração das informações, pela vinculação de imagens ao registo médico eletrónico do paciente.



Figura 2: Laudo ditado durante o exame (dir.) e revisão do laudo após reconhecimento de voz automático (esq.)

## 1.1 Projeto Cyclops

Criado como um projeto de longo prazo pelos professores Dr. Aldo von Wangenheim e Dr. Michael M. Richter na universidade de Kaiserslautern em 1992, o Projeto Cyclops é uma cooperação bilateral entre o Brasil e a Alemanha com o intuito em desenvolver métodos, técnicas e ferramentas na área médica. Os principais campos de pesquisa na

área computacional são inteligência artificial, reconhecimento de padrões, computação gráfica e redes wireless. A organização atual do projeto é composta de 4 grupos:

- PACS compatíveis com DICOM;
- Processamento de imagens médicas e análise de sinal;
- Gerenciamento de workflow médico;
- Telemedicina e Tecnologia Wireless.

O sistema proposto neste trabalho pertence a análise de sinal associado a telemedicina. Esta área procura adaptar trabalho e recursos médicos em sistemas computacionais e de telecomunicações no diagnóstico e tratamento médico. Telemedicina a área interdisciplinar que utiliza tecnologias para aproximar profissionais da saúde e pacientes, visando realizar ações médicas à distância (BASHSHUR; SANDERS; SHANNON, 1999).

## 1.2 Motivação

A manipulação de áudio digital para reconhecimento de voz<sup>3</sup> permite a implantação de um sistema de laudo médico que fornece segurança, agilidade e conforto aos médicos na concepção do laudo. Isto, aliado a necessidade em reduzir a resistência dos usuários nos programas para laudo médico, motiva a pesquisa e o desenvolvimento da tecnologia de reconhecimento de voz. A relevância da área pode ser observada pelo investimento de soluções comerciais como a Philips (PHILIPS ELECTRONICS, 2006), Nuance (NUANCE COMMUNICATIONS, INC., 2006), StructRad (STRUCTRAD LLC, 2006) e a MacSym (MACSYM, 1985). Isto é confirmado por YNOGUTI, 1999:

*“Algumas das principais áreas de aplicação comercial para os sistemas de reconhecimento automático de fala são: ditado, interfaces para computadores pessoais, serviços de telefonia automáticos e aplicações industriais especiais. A principal razão para o sucesso comercial tem sido o aumento na produtividade proporcionado por estes sistemas que auxiliam ou substituem operadores humanos.”*

---

<sup>3</sup>Durante o trabalho, reconhecimento de voz será abreviado por SR - Speech Recognition. É a mesma abreviação para o termo que será utilizado adiante Dicom SR - Structured Report. As abreviações serão usadas em diferentes contextos.

A pesquisa em armazenamento de laudos digitais também é motivada pelo desafio em desenvolver uma aplicação inédita no país, laudo ditado com transcrição automática por meio de reconhecimento de voz. Os resultados do uso dessa aplicação em hospitais pode reduzir custos e agilizar o processo de laudo médico. Segundo LEPANDTO et al., 2005:

*“Estudos realizados sobre o impacto da implantação de PACS em hospitais, mostram que o tempo necessário para emissão de laudos pode ser reduzido. Com essa tecnologia, o tempo necessário para manipulação e produção de filmes é eliminado. Embora em alguns setores dos hospitais o sistema não aumente a produtividade, ele permite a implantação de um programa para reconhecimento de voz.”*

Um PACS associado a SR deve auxiliar hospitais com alta demanda de atendimento a aumentar a eficiência no atendimento aos pacientes. O Projeto Cyclops, responsável pela Rede Catarinense de Telemedicina, vê no armazenamento de laudos digitais e sistemas de SR, a possibilidade de ampliar os hospitais atendidos e aumentar a aceitação dos médicos nos programas desenvolvidos.

## 1.3 **Objetivos**

### 1.3.1 **Objetivo Geral**

Implementar um sistema de gravação, reprodução, codificação, armazenamento e recuperação de áudio, visando aplicar sistemas de reconhecimento de voz em laudo ditado.

### 1.3.2 **Objetivos Específicos**

- Desenvolver uma metodologia de protótipo para gravação e reprodução de áudio para laudos médicos em sistemas distintos como PC e Palmtops;
- Utilizar TAGs<sup>4</sup> para identificação do áudio independente do exame;
- Avaliar formatos de áudio quanto a usabilidade e tamanho resultante de compactação e codificação;

---

<sup>4</sup>TAG do inglês é etiqueta. No texto o uso de TAGs está associado a campos pré definidos que são criados em arquivos afim de identificar atributos do mesmo.

- Implementar uma base de laudos em áudio para armazenar de forma segura e confiável os laudos. Adaptar o portal de exames(WALLAUER, 2005) para consulta e edição dos laudos por digitadores.
- Analisar métodos e ferramentas para o futuro desenvolvimento de um protótipo de reconhecimento de voz que faça transcrição automática áudio - texto;
- Desenvolver procedimentos para avaliação, testes de usabilidade e validação dos protótipos.

## 1.4 Justificativa

Os diferentes cenários existentes no processo de confecção de laudo citados na introdução, descrevem formas distintas para confecção do laudo. O tempo e a confiabilidade do processo estão diretamente relacionados com as tecnologias em informática médica adotadas. Os formulários manuscritos estão sujeitos a falhas como:

- Letra ilegível - Preenchimento do formulário que impossibilita entendimento do digitador ou paciente;
- Trocas - Podem ocorrer trocas onde um paciente recebe o laudo que pertence a outro paciente, iniciando assim, o tratamento indevido;
- Perdas - O volume de documentos de um dia pode ocasionar perdas imperceptíveis pelos técnicos do setor. Somente na hora que o paciente precisa do laudo médico, o problema é detectado;
- Histórico - Mesmo que sejam guardados, é inviável o uso de histórico para acompanhamento dos exames de todos os pacientes atendidos na rede pública;
- Lentidão - O processo manual torna a confecção do laudo final demorada sobretudo se erros como os citados acima ocorrem.

O processo de laudo ditado em fitas é uma evolução do processo manual, entretanto, está sujeito a problemas que podem ser evitados:

- Trocas - Ainda existe o risco de trocas entre laudo de diferentes pacientes pois os exames e pacientes são identificados de forma manuscrita nas fitas;
- Perdas - O transporte das fitas pode danificá-las e as mesmas podem ser perdidas;

- Histórico - A regravação das fitas elimina o histórico do áudio ditado e deteriora a qualidade do ditados subsequentes;
- Lentidão - Embora o ditado em fitas forneça boa usabilidade, o processo ainda é lento pois depende de transcrição manual e correção do médico na confirmação do laudo.

Um sistema de laudo ditado em formato digital com reconhecimento de voz apresenta soluções para os problemas acima citados. Além disso, o Projeto Cyclops(CYCLOPS, 2006) encontra resistência dos médicos na adaptação dos *softwares*, uma vez que estes não desejam digitar os laudos no computador. O presente trabalho pesquisa formas de sanar estes problemas, ou seja, permitir que os sistemas de laudo médico elaborados pelo Cyclops Project tenham suporte ao diagnóstico ditado. Para isso deve-se elaborar um sistema de captura de áudio que armazene o laudo em um banco de dados seguindo padrões específicos(DICOM, 2006). Este laudo pode ser repassado ao digitadores ou gerar o diagnóstico automaticamente via reconhecimento de voz.

## 1.5 Estrutura do trabalho

O trabalho está estruturado em quatro capítulos introdução, fundamentação teórica, metodologia e conclusão. Documentos relacionados ao trabalho de outros autores estão em anexo após as referências e arquivos extras elaborados pelo autor encontram-se em apêndice. A introdução objetiva contextualizar o leitor sobre o objetivo do trabalho e onde este será aplicado. Estão presentes nesta seção os motivos que levaram a escolha do tema, as justificativas do desenvolvimento deste trabalho e quais problemas serão atacados.

O segundo capítulo, fundamentação teórica, trata dos conceitos pesquisados para o desenvolvimento de ferramentas e análise das tecnologias existentes. A primeira parte desta seção apresenta as técnicas para auxílio a exames médicos existentes, e que podem ser exploradas na área de reconhecimento de voz. Em seguida, são apresentadas soluções para captura de áudio nos sistemas operacionais Windows e Linux. Essas pesquisas tem como objetivo encontrar soluções confiáveis com código aberto. Pois assim, o programa pode ser alterado e analisado de acordo com as necessidades do Projeto Cyclops. O terceiro tópico levanta os formatos de arquivo mais relevantes em aplicações de captura e reprodução de áudio. A teoria de reconhecimento de voz é analisada na última parte da fundamentação teórica, onde conceitos de reconhecimento de voz são abordados de

---

forma sucinta. Finalmente são apresentadas algumas ferramentas científicas e comerciais da área, o que representa o estado atual da arte.

No terceiro capítulo, é apresentada a metodologia de trabalho utilizada. Nesta seção os resultados obtidos pela análise das opções levantadas na fundamentação teórica são descritos. Além disso, neste capítulo as ferramentas implementadas durante o trabalho são apresentadas.

A última parte deste projeto, discute trabalhos futuros e os resultados obtidos. É nesse capítulo que a conclusão se encontra e o trabalho é finalizado.

## 2 *Fundamentação Teórica*

### 2.1 **Tecnologias de Laudo Médico**

Na introdução, foram apresentados os atuais métodos de geração de laudo. Como foi observado, esses métodos podem ser substituídos por formas mais dinâmicas e confiáveis. Isto é confirmado por ALEXANDRINI; BORTOLUZZI; WANGENHEIM, 2005, onde fala dos profissionais de saúde que fazem registros seguindo determinações do Conselho Federal de Medicina (CFM) e das instituições que visam maior agilidade no processo clínico e hospitalar e por isso, procuram adotar sistemas de registro clínico digital. Além disso, DAVID; ZUCHERMAN; JR., 2005, comenta que laudos produzidos de forma manuscrita ou pelo ditado em fitas dificultam análises subseqüentes e exigem mais cuidados na sua confirmação.

As inovações digitais disponíveis permitem automatização para a geração de laudos, maior segurança no armazenamento e a possibilidade de armazenar o histórico do paciente. As vantagens desses ganhos refletem na quantidade de pacientes atendidos como também na rapidez em que são detectados casos graves. Dentre as tecnologias mais relevantes para a edição e geração de laudos destacam-se:

- PACS - Sistema para arquivamento e comunicação em diagnóstico por imagem. Permite acesso, em qualquer setor, de imagens médicas em formato digital(CENTRO DE CIÊNCIAS DAS IMAGENS E FÍSICA MÉDICA, 2006);
- DICOM Structured Report - Padrão DICOM para laudos estruturados;
- Telemedicina - Aplicações desenvolvidas e disseminadas com tecnologias que possibilitam a medicina acessível de qualquer lugar(CYCLOPS, 2006);
- Áudio Digital - Permite gravar laudo com qualidade superior as fitas. Além disso, a informação armazenada não é perdida como ocorre com fitas regravadas. O uso dessa



tecnologia é a primeira etapa para a utilização de um sistema de reconhecimento de voz.

- Reconhecimento de voz - É o processo de obter palavras faladas como entrada em um programa de computador e transformá-las em texto(JIM BAUMANN, 1993). O reconhecimento automático de voz permite a transcrição do laudo ditado para texto livre e/ou DICOM Structured Report.

Os dois últimos itens citados são o principal foco de pesquisa deste trabalho. As demais tecnologias podem ser integradas a recursos de armazenamento de áudio e reconhecimento de voz. A descrição de cada tecnologia é apresentada a seguir, os itens Áudio Digital e Reconhecimento de voz recebem capítulos mais detalhados no decorrer do trabalho.

### 2.1.1 Padrão DICOM

Inicialmente criado para especificar o padrão de comunicação e armazenamento de imagens digitais na área médica, o DICOM acabou tornando-se padrão para armazenamento de informações como áudio e vídeo em exames e estudos. Além disso, o DICOM(NEMA, 2004):

- É aplicável a ambiente de rede;
- É aplicável a mídias off-line;
- Especifica como os dispositivos reagem aos comandos recebidos e dados que estão sendo trocados;
- Especifica o nível de conformidade;
- É um documento estruturado *multi-part*;
- Introduz informações explícitas além de imagens como *waveforms*, relatório e impressão;
- Especifica a forma de identificar unicamente qualquer objeto;

### 2.1.1.1 DICOM Structured Report

Introduzido em 2000 pelo Comitê de Padrões de Imagens Digitais e Comunicação na Medicina, o laudo estruturado (CLUNIE, 2000) é um padrão digital e formal para laudos e dados textuais com capacidade de referenciar outros objetos DICOM. É um formato de laudo que evita ambigüidades aumentando a consistência de acesso à um mesmo laudo criado em diferentes sistemas

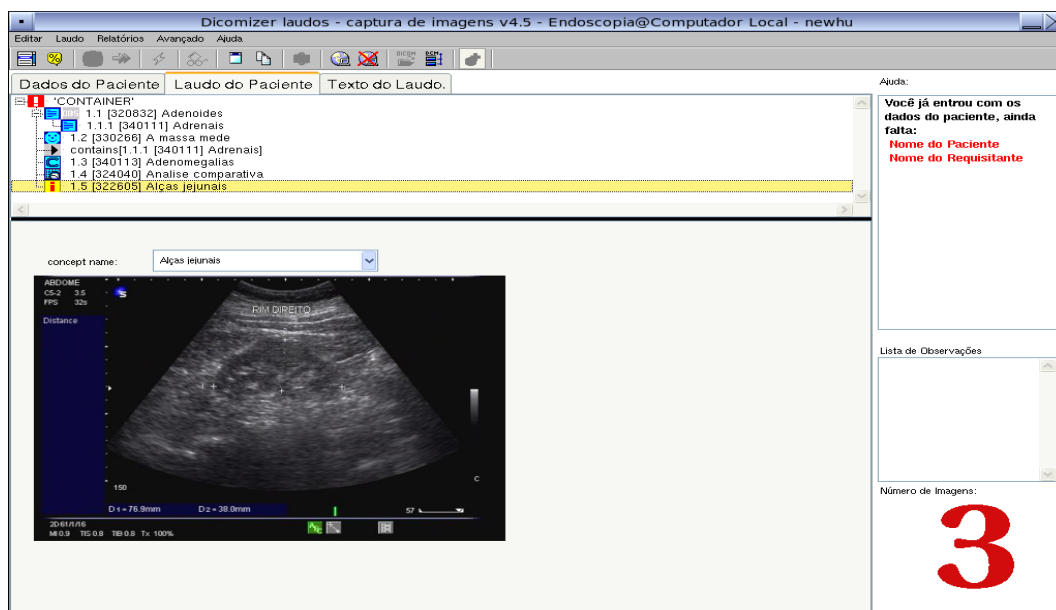


Figura 3: Laudo estruturado no padrão DICOM

Laudo estruturado é definido como a criação da informação padronizada em modelos navegados por menus dentro de uma linguagem natural de laudo (Figura 3). O propósito do laudo estruturado é prover um método de laudo formal que possa substituir o tradicional método de ditado e transcrição que exige interpretação de exames médicos (DAVID; ZUCHERMAN; JR., 2005).

### 2.1.2 PACS

PACS (Picture Archiving and Communication System) é uma abordagem que surgiu para aperfeiçoar o diagnóstico médico por imagem. O Sistema para arquivamento e comunicação em diagnóstico por imagem permite acesso, em qualquer setor do hospital, das imagens médicas em formato digital. Um PACS pode ser dividido em quatro processos: aquisição, exibição, disponibilização e armazenamento de imagens (AZEVEDO-MARQUES; TRADD; JUNIOR, 2001).

O Projeto Cyclops junto ao Hospital Polydoro Ernani de São Thiago, utilizam um sistema de aquisição de dados PACS, denominado Dicomizer, em aparelhos não compatíveis com DICOM. Embora o projeto já tenha um PACS DICOM compatível, a falta do padrão DICOM nos aparelhos do HU/UFSC fez do Dicomizer a solução para imagens fora do padrão. O software foi desenvolvido em Smalltalk(CINCOM, 2006) e obtém a imagem dos aparelhos através de placas para captura de vídeo. Através do sistema, o exame pode ser laudado ou postergado com acesso aos dados através do portal de telemedicina do HU(WALLAUER, 2005).

O fluxo de trabalho PACS dentro do HU é apresentado na Figura 4, onde o PACS é composto do *software* de captura, base de dados para armazenamento das informações e portal *web* para acesso aos exames. O médico pode capturar as imagens durante o exame e enviá-las para a base de dados. O laudo do exame pode ser feito no momento do exame ou através do portal de telemedicina disponível na internet com acesso restrito.

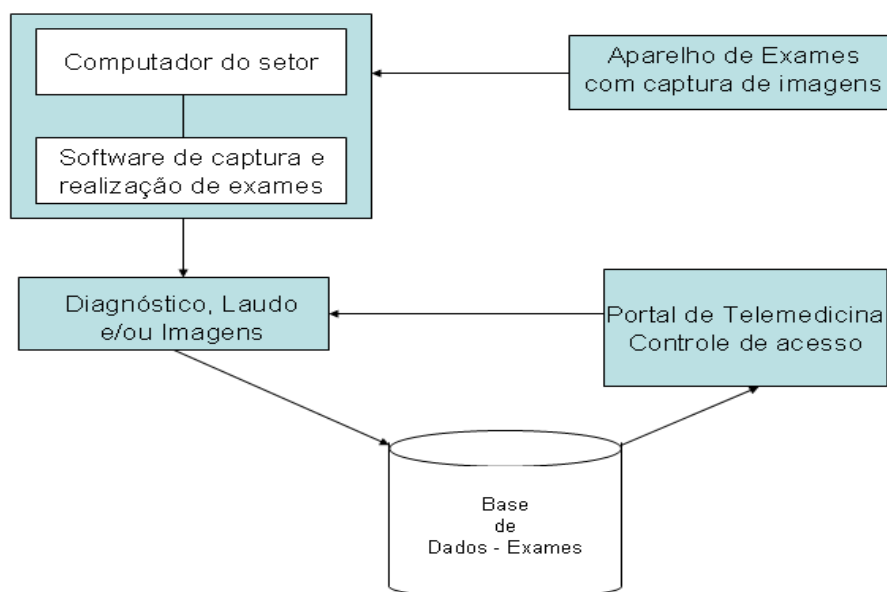


Figura 4: Sistema PACS adaptado ao HU-UFSC

### 2.1.3 Telemedicina

Telemedicina é um termo genérico que combina áreas distintas como medicina, engenharia e computação. O termo é usado para descrever o uso de tecnologias de telecomunicação e de informação para suportar serviços, métodos, treinamento e informação em saúde para provedores de assistência médica e pacientes. A essência dessas áreas é a oferta de serviços e informação para médicos e pacientes à distância eliminando a necessidade

de locomoção para os centros de referência.(LOPES et al., 2005).

O principal objetivo da telemedicina é desenvolver e disseminar tecnologias para tornar a medicina de ponta acessível a todos e em todos os lugares. Oferecer aos pacientes e médicos rapidez para realização de exames e diagnóstico. Assim, os benefícios da telemedicina trazem melhorias na assistência primária, aumento da disponibilidade de recursos para educação médica e informação em saúde em comunidades desprovidas de pessoal qualificado(LOPES et al., 2005). Entretanto, para que o serviço de telemedicina e tele diagnóstico funcione em qualquer ambiente de produção, fazem-se necessárias:(DELLANI, 2001)

1. Existência de infraestrutura de conectividade entre os centros de saúde;
2. Padrão para comunicação e arquivamento da informação médica: imagens e laudos;

O Projeto Cyclops, em conjunto com a Secretaria de Saúde do Estado de Santa Catarina já conta com um sistema de telemedicina e o telediagnóstico no Estado(WALLAUER, 2005). A existência de uma boa infraestrutura no sistema de saúde de Santa Catarina permitiu a instalação da tecnologia de telemedicina desenvolvida no Cyclops. Mesmo enfrentando resistência de parte dos médicos em adotar os sistemas de informação, os resultados são positivos. No HU/UFSC, foram contabilizados os exames de dois setores, um de baixa e outro de alta demanda no período de março a maio de 2006. No setor de baixa movimentação, o portal do HU recebeu em média dois exames por dia. No mesmo período, os exames realizados no setor de maior quantidade de atendimentos, foram atingidos a média de 21 exames por dia.

Os programas para captura de imagens do exame são compatíveis com o portal de telemedicina. Cada exame realizado por um cliente é enviado ao servidor, assim um exame realizado no oeste do estado pode receber laudo por médicos do litoral. Este processo permite maior agilidade e evita o deslocamento do paciente devido a falta de recursos humanos capacitados em regiões afastadas dos grandes centros.

#### **2.1.4 Áudio Digital**

O áudio digital consolidou-se em diversos formatos como tecnologia computacional(FRAUNHOFER GESSELLSCHAFT INSTITUTEN, 2006). Entretanto essa tecnologia bastante difundida não vem sendo utilizada em benefício de hospitais públicos. A codificação dos laudos em áudio digital oferece diversas vantagens(MACSYM, 1985):

- Evita o uso de gravadores micro-cassetes e as fitas destes aparelhos;
- Alta qualidade de áudio e recursos avançados de manipulação de áudio;
- Acesso ao laudo gravado de qualquer computador conectado a *Internet*;
- Redução nas perdas de informações por fitas reutilizadas ou em más condições;
- Primeira etapa necessária ao reconhecimento de voz automatizado;
- Aumento de produtividade de aproximadamente 30%;
- Possibilita a gravação do laudo de qualquer computador e até mesmo Palmtops;

### 2.1.5 Reconhecimento de voz

Sistemas de reconhecimento automático de voz (SR<sup>1</sup>) estão gradualmente substituindo os serviços tradicionais de transcrição nos departamentos de radiologia na Europa e na América do Norte. Esta revolução na radiologia ocorre principalmente por dois fatores (WHITE, 2005):

- Reduzir os custos operacional;
- Reduzir o tempo necessário para confecção dos laudos;

Utilizando áudio digital, laudos serão gravados e digitados posteriormente pela secretária do médico ou equipe de digitadores. Uma vez o laudo digitado, este pode ser enviado ao portal de telemedicina ou para o e-mail do médico. O acesso ao portal permite confirmar o laudo ou modificá-lo se necessário, além de oferecer a visualização das imagens e o laudo no formato de áudio.

Um sistema de reconhecimento de voz é um programa de computador que recebe a fala como entrada através de um microfone e gera como saída o texto relativo a entrada. Desde 1981 existem sistemas de SR simples porém funcionais. No entanto, essas tecnologias utilizavam algoritmo de reconhecimento discreto, que obriga o orador a falar palavras individuais com pausas. Além disso esses sistemas eram lentos, dependiam do orador e do ruído externo (LANGER, 2002). A evolução dos computadores e da pesquisa em SR permite hoje o desenvolvimento de aplicações comerciais viáveis para reconhecimento da fala contínua. De acordo com YNOGUTI, 1999:

---

<sup>1</sup>SR - Speech Recognition, do inglês reconhecimento de voz

*“Depois de vários anos de pesquisa, a tecnologia de reconhecimento de fala está passando o limiar da praticabilidade. A última década testemunhou um progresso assombroso na tecnologia de reconhecimento de fala, no sentido de que estão se tornando disponíveis algoritmos e sistemas de alto desempenho. Em muitos casos, a transição de protótipos de laboratório para sistemas comerciais já se iniciou.”*

É importante ressaltar que o uso de ferramentas com reconhecimento de voz devem ser usadas com critério de usabilidade. As interfaces atuais, teclado, mouse e pedal são interfaces extremamente eficientes(YNOGUTI, 1999). O uso de SR deve combinar as melhores características dos dispositivos afim de obter procedimentos de trabalho mais dinâmicos.

## 2.2 Manipulação de áudio

A comunicação da placa de som com o sistema operacional é feita através de *drivers* disponibilizados pelo fabricante do hardware de áudio. Devido a complexidade dessas interfaces, os sistemas operacionais utilizam bibliotecas que fazem uma camada de mais alto nível com as aplicações. Assim, essas bibliotecas fornecem soluções independentes de hardware.

As próximas seções mencionam diversos formatos de áudio relevantes ao trabalho. Para melhor compreensão, as extensões de cada formato são explicada a seguir:

1. **WAV** - Formato de áudio criado pela Microsoft e IBM lançado em 1995 para Windows. O formato não é codificado e a extensão WAV vem do nome *Waveform*, do inglês 'forma de onda';
2. **AU** - Formato de áudio criado pela Sun Microsystems e NeXT. A extensão AU vem de áudio e é o padrão acústico para a linguagem Java;
3. **MPEG** - Abreviação de *Moving Picture Experts Group*. É o nome de uma família de padrões definidos para o uso de codificação das informações de áudio e vídeo digital compactados;
4. **MP3** - Um dos primeiros formatos de arquivos a comprimir áudio com perda de dados e com eficiência. A abreviação MP3 vem de *MPEG Audio Layer-3*;

5. **OGG** - Formato de compressão livre, código aberto e livre de patentes desenvolvido pela Xiph.org Foundation;
6. **WMA** - Formato desenvolvido pela Microsoft, WMA é o formato para *Windows Media Audio*, padrão de codificação em Windows;
7. **VQF** - Extensão da tecnologia de áudio compactado criado pela Yamaha chamado de *TwinVQ*.

### 2.2.1 Open Source Audio Library

O acrônimo OSALP abrevia *Open Source Audio Library*, uma biblioteca para manipular funções de áudio no paradigma orientado a objetos em C++. Implementado para ser compatível com plataformas UNIX, a versão 0.7.3 funciona em Linux, FreeBSD e Solaris. O projeto surgiu da necessidade de Bruce Forsberg em manipular e editar arquivos de áudio de forma simples e atualmente é mantido por Darrick Servis(OSALP, 2002).

A biblioteca funciona através do OSS<sup>2</sup> no BSD e Linux. Isto é uma desvantagem no projeto pois esse sistema de áudio foi oficialmente substituído pelo ALSA<sup>3</sup> no kernel do Linux. Embora o OSALP funcione sobre o OSS, a biblioteca possibilita reproduzir, editar e gravar áudio nos formatos:

- WAV - Áudio linear manipulado pelo OSALP;
- AU - Manipulado de forma linear pelo OSALP;
- MP3 - Utiliza bibliotecas externas: LAME, blade, splay e mpg123;
- OGG - Através do codificador e decodificador OGG-Vorbis;

Embora a biblioteca utilize uma arquitetura de áudio descontinuada no Linux, este conjunto de classes implementa conversão da taxa de amostragem (*sample rate*), modificação do passo (*pitch*), reprodução e gravação de áudio, edição e mistura de arquivos(*mixing*), filtros e conversões de formatos. Por estar nos termos da LGPL<sup>4</sup> possui código fonte disponível para ser usado e modificado.

---

<sup>2</sup>OSS é abreviação de *Open Sound System*

<sup>3</sup>ALSA é abreviação de *Advanced Linux Sound Architecture*

<sup>4</sup>LGPL - *Lesser General Public License*, significa 'Licença Geral Pública Leve'

## 2.2.2 PortAudio

PortAudio é uma biblioteca para manipulação de entrada e saída de áudio. O projeto pretende facilitar o desenvolvimento de softwares para síntese e reprodução de áudio em diferentes plataformas, a versão 1.8 compila e executa nos sistemas operacionais (PORTAUDIO, 2001):

1. Windows - Sistema da Microsoft;
2. Macintosh - Desenvolvido pela Apple;
3. Linux - Distribuições baseadas em UNIX projetado inicialmente por Linus Trovalds;
4. BeOS - Desenvolvido pela Be Incorporated;
5. Irix - Distribuição da Silicon Graphics;

O PortAudio possui uma API<sup>5</sup> mais simples que bibliotecas nativas de diferentes plataformas. Assim, para fins pedagógicos, o conjunto de funções escritas em C do PortAudio, possibilita aos estudantes criar programas áudio de forma simples, onde o mesmo código funciona em diferentes sistemas (Figura 5). Por exemplo, no Linux usa-se o sistema de som OSS se este existe, caso contrário, é usado o subconjunto OSS da especificação ALSA (BENCINA; BURK, 2001).

O projeto tem suporte a sincronização de áudio em tempo real e suporte a restrições temporais para aplicações com computação gráfica (BENCINA, 2003). Dentre as aplicações mais conhecidas para manipulação de áudio que utilizam a biblioteca, destacam-se:

- Audacity - Editor de áudio (AUDACITY, 2005);
- Winamp (plugin) - Reprodução de áudio (WINAMP PORTAUDIO, 2006);
- JSyn - API Java para síntese de música modular em tempo real (JSYN, 2005);
- LIRC - Captura de sinais infra-vermelhos pelo canal de som (LIRC, 2006);

## 2.3 Tecnologias de compactação de áudio

Em um sistema de armazenamento de áudio ditado o tamanho do arquivo precisa ser reduzido, se possível, sem alterar a essência do áudio. Isto é necessário pois, o custo do

---

<sup>5</sup>API - *Application Programming Interface* ou Interface de Programação de Aplicativos



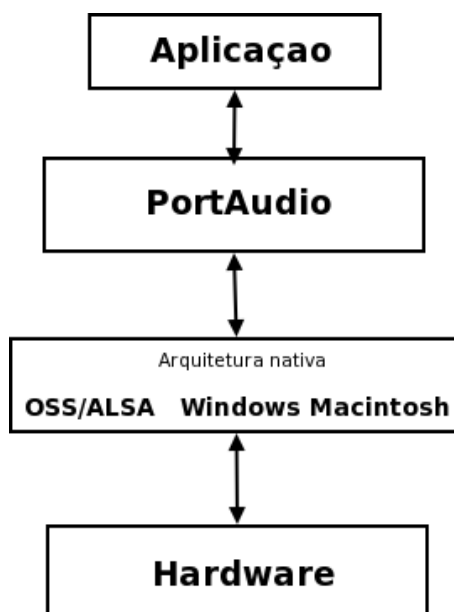


Figura 5: Abstração PortAudio para aplicações

espaço em disco para cada arquivo de áudio ditado é alto em um sistema do cotidiano. Além disso, devido as características da voz humana (Ítem 2.3.2) é possível reduzir a representação digital da fala mantendo a consistência das informações.

Os diferentes algoritmos de codificação buscam eliminar as partes que não podem ser detectadas por humanos reduzindo o tamanho do arquivo em até 26 vezes (Resultados apresentados no ítem 3.3). A forma mais simples de verificar a redução da informação é ouvir o áudio original e o codificado. Comparando ambos é possível detectar perda inerente ao processo. Existem muitos CODECs<sup>6</sup> para diversos formatos já desenvolvidos, destacam-se:

- LAME - Ferramenta usada para o ensino de codificação MP3, usado na maioria do software de código aberto(LAME, 2005);
- Xing - CODEC para formato MP3 em Windows, de excelente velocidade porém o código é proprietário(XING, 2005);
- Bladeenc - CODEC para formato MP3 gratuito dentro da Licença GPL<sup>7</sup>(BLADEENC, 2005);

<sup>6</sup>CODEC é a abreviação de codificador - decodificador

<sup>7</sup>GPL - *General Public License*, significa 'Licença Geral Pública' - Termo padrão para distribuições de código aberto e livre.

- Vorbis - CODEC para formato OGG codificador de código aberto desenvolvido pelo grupo Xiphophorous(AMMOURA; CARLACCI, 2002);
- TwinVQ - CODEC para o formato proprietário da Yamaha, VQF;
- Windows Media Audio - CODEC criado pela Microsoft para áudio com redução de informação para o formato WMA;
- RA, RAM, RM - Arquivos do Real Audio, tecnologia usada para *streaming* (rádio e/ou vídeo on-line) na Internet. Conseguem uma boa taxa de compactação, mas a qualidade é depreciada.

O formato mais famoso de compactação de áudio é o *MPEG Layer 3* conhecido como MP3, porém este formato é patenteado por *Thomson Consumer Electronics*(THOMSON CONSUMER ELECTRONICS, 2005) e motiva o desenvolvimento de CODECS como o OGG-Vorbis livre de patentes.

Arquivos codificados fornecem uma representação de dados de áudio através de conversões de pulsos em códigos por modulação-codificação (PCM). O som capturado através de um microfone é gravado em formato WAV que preserva todas as características do som no momento da gravação. Embora os programas utilizados possam codificar e decodificar arquivos de som, uma vez codificado o arquivo nunca mais será o mesmo<sup>8</sup> (Procedimento no Apêndice B). Para verificar a verdadeira diferença entre os arquivos devem ser usadas ferramentas especializadas como o espectrograma da figura 6:

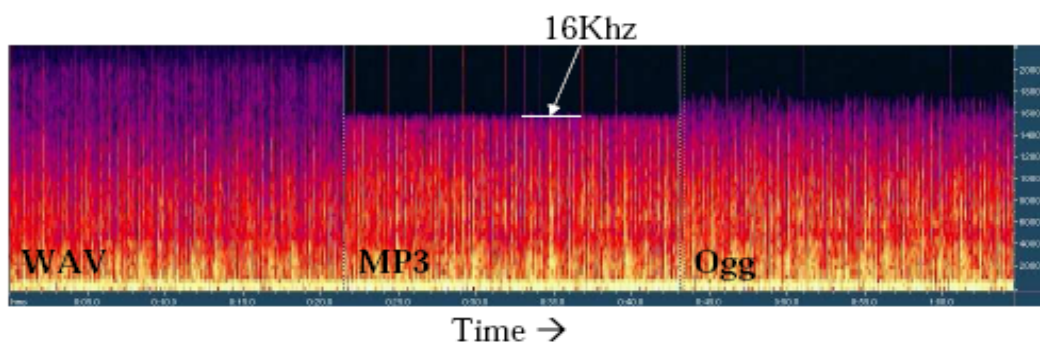


Figura 6: Três arquivos de 20 segundos concatenados no mesmo arquivo. Da esquerda para a direita o mesmo áudio em WAV, MP3 e OGG respectivamente(AMMOURA; CARLACCI, 2002)

<sup>8</sup>Considera-se os CODECS com perda somente neste caso. Existem CODECS sem perda que evitam este problema que serão abordados na seção 2.3.5

Os formatos codificados apresentam vantagens em relação ao tamanho do arquivo resultante, porém, estes arquivos tem menor qualidade de som e é necessário decodificar os dados antes de tocá-lo. Por isso somente computadores e aparelhos com suporte a mp3 podem processá-lo.

### 2.3.1 Taxa de bits (Bitrate)

O bitrate é a quantidade de bits usados para armazenar cada segundo de áudio em arquivo. A taxa é medida em kb/s, ou seja 1024 bits por segundo. Dessa forma um arquivo armazenado a 128 kb/s, utiliza 128\*1024 bits por segundo de informação, total 131072 b/s. Em geral, quanto mais informação necessita ser mantida, maior deve ser o bitrate. É possível também variar o bitrate durante a codificação de um áudio, dividindo-o em partes de diferentes taxas para codificação(FRAUNHOFER GESSELLSCHAFT INSTITUTEN, 2006).

Com a possibilidade de variar o bitrate (VBR) os melhores codificadores baixam a taxa de amostragem nos trechos complexidade de sinal e aumentam o bitrate a medida que a complexidade aumenta.

### 2.3.2 Frequência

Em Física, frequência é definida como número completo de ciclos ou oscilações por segundo de um sistema. Em Áudio, refere-se à capacidade do áudio em possuir mais/menos representação para sons graves ou agudos, de acordo com quantidade de ondas (ciclos) completas por segundo(ALBUQUERQUE, 2006). Na fala, o alcance de frequências varia de acordo com a idade, sexo e condições físicas.

### 2.3.3 Amostragem (Sampling)

Para armazenar o sinal analógico, é necessário discretizar a informação contínua em um conjunto de dados digital através da medição do sinal a intervalos regulares. A cada medida, a amplitude da onda sonora é determinada pela soma das amplitudes de todas as frequências que a compõe naquele instante. Isto resulta em uma seqüência de medições que compõe o sinal digitalizado. Neste processo, é inerente a perda de informação, já que o fenômeno real é contínuo e a representação discreta.

A qualidade da representação digitalizada do sinal, é definida pela taxa de medição do

sinal, denominada taxa de amostragem, e seu valor mínimo depende de que aspectos do sinal se pretende capturar. Assim, a taxa de amostragem está bastante ligada a frequência, onde a amostragem corresponde à medida da amplitude do sinal tomada em intervalos fixos de tempo. A relação com a frequência está na quantidade que o sinal é medido por cada segundo. Áudio codificado a uma frequência de amostragem de 44.1 kHz, 16 bits por amostra, com dois canais (*stereo*) possui taxa de amostragem de:  $44.100 \times 16 \times 2 = 1.411.200$  bits por segundo.

### 2.3.4 Qualidade do áudio

A redução em informação característica do formato codificados fornece opções diferentes na qualidade e tamanho do arquivo. Isto é definido pela taxa de bits (bitrate), normalmente usa-se taxas entre 128 a 256 kbit/s (kilobits por segundo). A qualidade equivalente a um CD de áudio sem compressão requer taxas entre 1378 kbit/s a 1411 kbit/s com frequência de 44.1 kHz.

Os algoritmos de codificação, reduzem a informação explorando o fato do ouvido humano ser capaz de captar sons entre 20 Hz a 20 KHz. Assim as frequências fora desse intervalo são ignoradas pelo CODEC reduzindo a qualidade e o tamanho do áudio. Arquivos de música normalmente são codificados em 128 kbits stereo 44khz, nessa configuração o resultado é até doze vezes menor que o original(FRAUNHOFER GESSELLSCHAFT INSTITUTEN, 2006).

A taxa de bits por segundo influi diretamente na qualidade do áudio comprimido. Taxas baixas geram arquivos de má qualidade, e podem gerar sons não existentes originalmente, denominados artefatos de compressão. Outro fator que influencia a qualidade do som compactado é a dificuldade do sinal ser codificado(FRAUNHOFER GESSELLSCHAFT INSTITUTEN, 2006).

### 2.3.5 Classificação dos CODECs

Para reduzir o tamanho dos arquivos originais um codificador pode reduzir algumas informações e compactar o arquivo com os dados relevantes. Os codificadores mais comuns para os formatos MP3, OGG e o WMA possuem perda<sup>9</sup> pois eliminam partes do arquivo de acordo com a qualidade desejada.

---

<sup>9</sup>Do inglês Lossy Format

Existe outra alternativa, codificar com baixa perda<sup>10</sup>, para reduzir o tamanho do arquivo sem remover informações do arquivo original. Nesse caso o arquivo original pode ser apenas compactado e o áudio codificado tem a mesma qualidade do original.

Embora a qualidade seja boa, os tamanhos dos arquivos são grandes e por isso os codificadores de baixa perda são menos difundidos. Para profissionais que necessitam evitar qualquer perda no som original, existem boas alternativas:

- Lossless-Audio - CODEC com interface amigável para Windows e Linux e distribuição gratuita(BEVIN, 2004);
- FLAC - CODEC de código aberto desenvolvido como alternativa aos formatos com perda(FLAC, 2000);
- Monkey's Audio - Programa para Windows código aberto com o objetivo de ser utilizado em outros programas(MONKEY, 2004);

### 2.3.6 Tags ID3v2

Os arquivos codificados em formato digital podem conter informações extras que identificam o áudio. Essas informações são chamadas de *TAGs* e podem incluir dados do artista e da música até imagens do álbum. Os aplicativos e dispositivos que reproduzem o áudio podem usar essas *TAGs* para mostrar dados adicionais do arquivo ao usuário.

Para adicionar as informações foi desenvolvido em 1996 por Eric Kemp, um padrão que determina como as *TAGs* devem descrever um arquivo de áudio. Denominado ID3 é a abreviação de “Identify an MP3”. Após a primeira versão do ID3, duas revisões foram feitas, sendo a última denominada de ID3v2(ID3, 1998).

### 2.3.7 Formato de áudio MP3

Formato de áudio digital de alta compressão desenvolvido e padronizado em 1991 por uma equipe de pesquisadores na Universidade de Hannover, membros do Comitê de Audio ISO/IEC MPEG coordenado pelo Professor Hans Musmann(INSTITUT TNT, 2005).

O CODEC definido no padrão do MP3 permite representar áudio em arquivos de tamanho reduzido sem perder de forma significativa a qualidade original. Dessa forma,

---

<sup>10</sup>Do inglês Lossless.

é possível obter redução de até doze vezes em relação ao tamanho do arquivo original (INSTITUT TNT, 2005).

O formato é definido pelo padrão “*ISO/IEC 11172-3 Layer 3*”, também conhecido como “*MPEG-1 Audio Layer 3*”, que define como arquivos armazenados MP3 devem ser salvos. O mesmo vale para o o padrão “*MPEG-2 Audio Layer 3*”, evolução do formato proposto por FRAUNHOFER GESSELLSCHAFT INSTITUTEN, 2006.

Para transformar sinais de tempo em frequência, é utilizada uma transformação híbrida do sinal no domínio do tempo dentro da frequência no domínio do tempo:

- Filtro de quadratura multifase com 32 faixas;
- Fluxo de 32 ou 12 MDCT (Modified Discrete Cosine Transform);
- Redução de alising<sup>11</sup>.

Em dezembro de 2004 foi lançado a versão *sorround* do formato mp3 com suporte aos 5 canais típicos do áudio sorround. Os arquivos são similares e o novo formato é compatível com o padrão *stereo*.

Na especificação MPEG, o sucessor do formato mp3 é o AAC (Codificação avançada de áudio) do padrão “MPEG-4”. No entanto, nem o sucessor do mp3 e as alternativas acima citadas devem modificar o quadro atual, devido a popularidade que o formato mp3 conquistou. Aparelhos de som de para casa e carro, DVD players e celulares já são produzidos com suporte a mp3.

Embora a grande vantagem do MP3 seja a popularidade e difusão ampla em software e hardware, este formato está sob uma patente, obrigando a qualquer programa que use codificação em mp3 pagar royalties à *Thomson Consumer Electronics* (THOMSON CONSUMER ELECTRONICS, 2005). Foi contra essa patente que surgiu o formato OGG Vorbis, que pode ser utilizado sem pagamento de *royalties*.

Na tabela 1, é apresentada uma comparação entre os formatos MPEG-1 Layer 1, 2 e 3 feita pelo Instituto Fraunhofer Gesellschaft (FRAUNHOFER GESSELLSCHAFT INSTITUTEN, 2006).

Outra vantagem do formato reside na facilidade de edição. Não é necessário recodificar um arquivo em MP3 após editá-lo, cortando seções e adicionando outros trechos de áudio

---

<sup>11</sup> “Distorção estática em áudio causado por baixa taxa de amostragem”

Camada	Bit Rate	Compressão
Layer 1	384 kb/s	Até 4 vezes
Layer 2	192 a 256 kb/s	De 6 a 8 vezes
Layer 3	112 a 128 kb/s	De 10 a 12 vezes

Tabela 1: Comparação entre formatos MPEG-1

no arquivo. O software Audacity, é uma excelente opção para editar áudio, é gratuito, *opensource* e multi-plataforma(AUDACITY, 2005).

### 2.3.8 Formato de áudio OGG

O nome OGG surgiu de um jargão criado no jogo para computador Netrek. O termo significava ataque de *kamikaze*, e mais tarde, passou a ser usado com sentido de fazer algo forçado sem preocupação com as conseqüências. O projeto do OGG foi planejado de forma ambiciosa, substituir todos os formatos de áudio e vídeo proprietários. OGG Vorbis é código aberto, livre de patentes, sem proprietários e livre de royalties.

O formato define como será o encapsulamento do conteúdo em um arquivo. Este conteúdo pode ser música, vídeo ou ambos codificado por um ou mais CODECs. O projeto mantido pela Xiph.org, desenvolveu e contribuiu com vários CODECs para diferentes funções. Compressão com perdas:

- Speex - Voz (8-32kbps);
- Vorbis - Música (16-256kbps).

Compressão sem perdas:

- FLAC - Áudio de alta fidelidade;
- Squish - Áudio de alta fidelidade.

Ogg Vorbis possui flexibilidade para armazenar som com alta qualidade com performance superior ao MP3 e WMA. O formato é definido pela RFC 3533 e o tipo de dados MIME pela RFC 3534. O nome Ogg é utilizado para denotar a extensão dos arquivos enquanto que o nome Vorbis é o codificador para músicas. Outros CODECs importantes são o Theora (vídeo) e o Speex para captura de voz(XIPH, 1994).

### 2.3.9 Speex

Speex é um CODEC projetado para compressão de áudio falado. O projeto tem como objetivo reduzir as barreiras de entrada para aplicações de voz provendo uma alternativa livre que evita os gastos com CODECs proprietários. O Speex é parte do GNU Project, e o CODEC está disponível sob uma variante da licença BSD para a Xiph.org que permite alterações no código se necessário(XIPH, 2006).

Este CODEC é baseado no método CELP<sup>12</sup> e é projetado para comprimir voz em taxas que variam de 2 a 44 kbps, o que permite a redução de ruídos por filtros de forma eficaz(VALIN; MONTGOMERY, 2006) e também:

- Faixas de 8Khz a 32Khz de compressão no mesmo bitstream;
- Operação com bitrate variável (VBR);
- Detecção de atividade de voz (VAD);
- Transmissão descontínua (DTX);

## 2.4 Sistema de reconhecimento automático de voz

O reconhecimento automático de voz, é o processo pelo qual ocorre a transformação ou o mapeamento de um fluxo acústico da fala em um texto escrito para posterior análise de um sistema de processamento de Linguagem Natural. Mais especificamente, é um processo de vários níveis em identificação de padrões, onde os sinais acústicos são examinados e estruturados dentro de uma hierarquia de unidades de fonemas, palavras, frases e sentenças. Cada nível adiciona restrições temporais como o conhecimento das palavras pronunciadas e a ordem correta das mesmas, compensando erros e incertezas dos níveis mais baixos. Essa hierarquia pode ser melhor explorada combinando decisões probabilísticas nos níveis inferiores e que decide de forma discreta somente nos níveis mais altos(TEBELSKIS, 1995).

A linguagem natural provê facilidade no uso por ser eficiente e flexível, um médico pode ditar o laudo ao mesmo tempo em que manipula um equipamento de imagens e analisa as informações na tela do computador. A tecnologia de reconhecimento de voz em sistemas de informação em saúde começou com a utilização de sistemas voltados para

---

<sup>12</sup>CELP - Code-Excited Linear Prediction



automação de laudos. Um estudo de caso mostra que gerar laudos e corrigi-los com ASR<sup>13</sup> é 30% mais rápido que o método de digitação baseado em fitas(MACSYM, 1985).

### 2.4.1 Flexibilidade do sistema

A implementação de um sistema de SR deve levar em consideração características operacionais do ambiente onde será implantado o *software*. A flexibilidade de um sistema de reconhecimento de voz pode ser dividida em três grupos:

1. **Dependente de Locutor** - É um sistema configurado com as características de pronúncia e timbre de voz de um único locutor. Nesse caso não existe flexibilidade, e por isso são os sistemas mais fáceis de se desenvolver;
2. **Independente de Locutor** - O sistema independente de locutor permite operar com qualquer locutor de um determinado grupo lingüístico, por exemplo “Português do Brasil”. Para operar nesse modo, o sistema necessita de uma vasta base de dados com pronúncias e timbres de voz de pessoas diferentes. Embora seja a solução mais flexível, é a mais difícil de desenvolver e também a menos confiável;
3. **Adaptativo** - Utiliza mecanismos de inteligência artificial para “ensinar” o *software* a responder corretamente e adaptar às características da fala de cada usuário. É a abordagem mais utilizada no desenvolvimento de aplicações comerciais.

### 2.4.2 Classificação do sistema de reconhecimento de voz

A forma como o locutor interage com o sistema SR está relacionada na usabilidade que o programa fornece. Assim, um sistema SR pode ser classificado em:

1. **Palavra Isolada** - Realiza o reconhecimento de uma palavra de cada vez e exige do locutor intervalos longos entre as palavras. Esta técnica é a forma mais simples de implementação pois o início e final de cada palavra são fáceis de determinar, a pronúncia não afeta palavras vizinhas e são reduzidas as comparações de verossimilhança da palavra-candidata com a base de conhecimento;
2. **Palavras Conectadas** - Similar ao sistema de palavras isoladas, porém, aceita seqüências curtas de palavras, como comandos ao sistema operacional seguidos de parâmetros;

---

<sup>13</sup>ASR - Automatic Speech Recognition, Reconhecimento automático de voz.

3. **Fala Contínua** - Tem flexibilidade para reconhecer um fluxo de áudio com longas seqüências de palavras, estas podem ser conectadas sem pausa em função das características de pronúncia. Esta abordagem possui alta complexidade na análise pois:

- Há dificuldade em detectar o início e o final das palavras;
- Fonemas de início e fim das palavras são influenciados por palavras vizinhas;
- A velocidade do ditado é variável.

4. **Fala Espontânea** - Neste sistema, é possível filtrar “ruídos de preenchimento” que ocorrem durante a expressão verbal natural de uma pessoa quando pensa o que falar. Um sistema de fala espontânea deve ser capaz de processar a fala naturalmente e não como um ditado;

**Identificação de Voz** - Tem como objetivo identificar o usuário através das características da voz de cada locutor. Pode ser usada em segurança ou para carregar o perfil do usuário em um sistema adaptativo.

### 2.4.3 Obstáculos no reconhecimento de voz

Para reconhecer um ditado, é necessário identificar fonemas, sílabas e palavras afim de formar a mensagem e em seguida analisar se o que foi produzido faz sentido no contexto do locutor. Existem limitações na comunicação falada que dificultam o reconhecimento em um sistema computacional. Em MIT, 2003, algumas dessas dificuldades são discutidas:

- **Fonética** - Homônimos, por exemplo:

1. “conserto” - “concerto”;
2. “sessão” - “seção” - “cessão”;

Embora os problemas dos homônimos sejam um problema para o sistema de SR, estes podem ser contornados através da linguagem e dicionários definidos. Além disso é possível contextualizar as palavras com recursos de inteligência artificial.

- **Fonológica** - Palavras com início e final de fonemas semelhantes:

1. “exames experimentais”;

Este problema é pertinente no SR independente de locutor com fala natural/-contínua, pois diferentes pessoas tem formas bem distintas de falar, onde a velocidade é um fator de destaque. A solução nesse caso é utilizar o maior número de pessoas possível no treinamento do sistema. Assim o SR tem menor probabilidade de unir os fonemas entre as duas palavras.

- **Fonotáctica** - Palavras de difícil pronúncia:

1. “stvamtkish”;

A definição da gramática e do dicionários podem solucionar esse problema restringindo em poucas palavras com pronúncia difícil.

- **Sintática** - Frases montadas de forma incorreta:

1. “Irei acompanhá-la até sua casa.”;
2. “Casa até acompanhá-la sua irei.”;

A eficiência de um sistema SR é altamente dependente do dicionário e da gramática. Existe um limite na expansão do dicionário, onde a adição de novas palavras pode resultar em uma grande expansão no conteúdo permitido do ditado(WHITE, 2005).

- **Semântica** - Palavras sem sentido na mesma frase:

1. “Nenhuma anormalidade encontrada.”;
2. “Nenhum anormal idade encontrada.”;

- **Contextual** A mesma sugestão mencionada no problema de sintática pode ser aplicado no caso da semântica e do contexto.

Para o desenvolvimento de um sistema ASR existem diversas dificuldades tais como sotaque do locutor, a velocidade da fala, ruídos no ambiente, diferença de voz entre pessoas e palavras novas para o sistema. De acordo com MIT (2003), os parâmetros que caracterizam as capacidades de um sistema ASR são apresentados na tabela 2.

Projetar um ASR significa decidir a forma de representar o sinal, de modelar as restrições e definir como será feita a pesquisa pela melhor resposta. Um sistema de SR deve possuir capacidade para realizar no mínimo as seguintes tarefas:

1. Entrada do discurso - Capturar o ditado via microfone e armazená-lo em formato digital;

Parâmetros	Abrangência
Modo de falar	De palavra isolada a conversação contínua
Estilo de falar	Voz de leitura a voz espontânea
Dependência	Depende locutor a Não depende do locutor
Vocabulário	Pequeno (<20 palavras) a grande (>50.000 palavras)
Modelo de Linguagem	Estado finito a sensível ao contexto
Perplexidade	Pequena (<10) a grande (>200)
SNR	Alta (>30dB) a baixa (<10dB)
Transdutor	Cancelamento de ruído no microfone ao telefone celular

Tabela 2: Parâmetros que caracterizam um sistema ASR(MIT, 2003)

2. Preparação de dados - Converter os dados para o formato de processamento em reconhecimento de voz e eliminar os ruídos;
3. Reconhecimento - Para cada entrada, o sistema deve selecionar a palavra que melhor representa a informação de acordo com o dicionário definido e dados obtidas por treinamento;
4. Decodificador - Conversão dos resultados obtidos para o formato texto do usuário;

Dadas as dificuldade de implementação de um sistema SR, é importante verificar a viabilidade do uso dessa tecnologia no domínio desejado (Tabela 3). Uma forma de medir a dificuldade do reconhecimento de voz, é a perplexidade. Esta combina o tamanho do vocabulário e o modelo de linguagem, e pode ser definida como a média do número de palavras que pode seguir uma palavra depois que o modelo de linguagem foi aplicado(YNOGUTI, 1999).

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

Tabela 3: Perplexidades em diferentes domínios(YNOGUTI, 1999)

## 2.5 Modelos para reconhecimento de voz

Existem duas abordagens mais difundidas para SR. A primeira utiliza conceitos de inteligência artificial com redes neurais. A segunda utiliza conceitos de estatística com Modelos Markovianos. Em YNOGUTI afirma-se que:

*“Nos sistemas que constituem o estado da arte na área de reconhecimento de fala predominam os modelos estatísticos, notadamente aqueles baseados em Modelos Ocultos de Markov (Hidden Markov Models, HMM). Os HMM’s são estruturas poderosas pois são capazes de modelar ao mesmo tempo as variabilidades acústicas e temporais do sinal de voz.”*

As seções seguintes apresentam alguns conceitos importantes para a teoria de reconhecimento de voz. Em seguida, são abordadas técnicas de redes neurais e modelos ocultos de Markov para processamento em SR.

### 2.5.1 Análise da Fala

As características acústicas da voz podem ser modeladas como uma seqüência de fonte, filtro do trato vocal, e características de radiação. Assim, o sistema de reconhecimento de voz pode representar o sinal falado como um conjunto de coeficientes *cepstral*, calculados a uma taxa fixa de quadro. Além disso, é importante observar que a construção de fonemas depende muitas vezes das sílabas e que vogais e consoantes diferem no grau de contração(MIT, 2003).

Para fazer a análise da fala, é necessário estimar os formantes gerados no trato vocal do locutor. Os formantes são fundamentais para inteligibilidade da fala e por isso são feitas pesquisas nas formas de sintetizá-los(MENESES; MOREIRA; SANTOS, 2001).

#### 2.5.1.1 Formantes

No momento da fala, um sinal é gerado pelas cordas vocais e é distorcido pelas cavidades laríngea, bucal e nasal que, acusticamente, variam sua freqüência de ressonância de acordo com a forma adquirida durante a articulação. Os formantes ocorrem em intervalos de freqüência de 1kHz onde cada um corresponde a uma freqüência de ressonância do trato vocal. que cria os formantes F1, F2 e F3, sendo que F0 é definido pela freqüência fundamental de vibração das cordas vocais(NOHAMA, 2001), (MENESES; MOREIRA; SANTOS, 2001).

A fala é produzida em faixa de freqüências variadas, com diferentes harmônicos e a conformação do trato altera suas características, reforçando algumas faixas e dissipando outras. Assim, cada fonema se enquadra em formantes específicos, por exemplo, o reconhecimento de voz é obtidos pelas seguintes análises em: (SIMAS, 2002), (MIT, 2003):

1. Primeiro Formante (F1) - Alto/Baixo;
2. Segundo Formante (F2) - Anterior/Posterior;
3. Terceiro Formante (F3) - Retroflexão.

## 2.5.2 Modelos Markovianos

O Modelo de Markov, é um modelo ou simulação baseado em cadeias Markovianas. Estas são, em tempo discreto, processos de Markov onde o espaço de estados é finito e o conjunto de índices varia de 0 a  $n$  ( $n|n \in N, 0 \leq n < \infty$ ). Assim, um Modelo Markoviano é uma representação para modelar sinal através de uma seqüência de observações. Para cada Cadeia de Markov, existe uma fonte que gera saídas observáveis, denominada de Fonte de Markov. Os símbolos que a fonte gera depende unicamente da observação anterior, assim, a Ordem da Cadeia de Markov é o número de seqüências anteriores necessárias para uma determinada saída. Cada estado de uma Cadeia de Markov representa um símbolo (observação) de um evento correspondente. É possível computar a partir de uma dada seqüência de símbolos quais foram os estados que geraram tal seqüência (BRITTO et al., 2001).

## 2.5.3 Modelos Ocultos de Markov

Se em um modelo Markoviano, cada estado representar a probabilidade sobre o conjunto total de símbolos, então ele é denominado de Modelo Oculto de Markov - HMM<sup>14</sup>. A estrutura do HMM é a mesma da Cadeia de Markov e permite computar a seqüência de estados com maior probabilidade de gerar a seqüência observada (BRITTO et al., 2001).

O sistema de reconhecimento da fala assume que o sinal *waveform* gerado a partir da recepção do áudio, é uma mensagem codificada em uma seqüência de um ou mais símbolos. Para efetuar a operação de reconhecimento, a seqüência de símbolos subjacentes, obtidos de uma interação falada é, inicialmente, convertida em uma seqüência igualmente espaçada de vetores com parâmetros discretos. As representações paramétricas mais utilizadas são *smoothed spectra* ou coeficiente de predição linear (YOUNG et al., 2005). A seqüência gerada é assumida como a representação exata do *waveform* usando como base a duração de um vetor. Assim, o waveform do discurso ditado pode ser considerado estacionário, ou seja, sem mudanças no tempo.

---

<sup>14</sup>Do inglês, Hidden Markov Model - Modelo Oculto de Markov

O objetivo do reconhecedor é mapear as seqüências dos vetores do ditado e as seqüências subjacentes de símbolos possíveis. Dois problemas tornam este processo complexo, conforme apresentado em Young et al. (2005, p. 3):

1. O mapeamento dos símbolos para a fala não é um para um. Diferentes símbolos subjacentes podem causar sons semelhantes no discurso. Além disso, há variações no *waveform* provenientes de sotaque e ruído do ambiente;
2. Os limites entre os símbolos não podem ser identificados explicitamente a partir do *waveform*. Assim, não é possível tratar o *waveform* do discurso como uma seqüência de padrões estáticos concatenados.

#### 2.5.4 Reconhecimento de palavras isoladas

Segundo Young et al. (2005, p. 4), no reconhecimento isolado, cada palavra ditada pode ser representada por uma seqüência de vetores ou observações  $O$ :

$$O = o_1, o_2, \dots, o_t$$

onde  $O_t$  é o vetor falado no tempo  $t$ . O reconhecimento de uma palavra isolada pode ser assumido como o cálculo de:

$$\arg \max_i P(w_i|O)$$

assim,  $w_i$  é a  $i$ -ésima palavra do vocabulário. Esta probabilidade é calculada usando a regra de Bayes:

$$P(w_i|O) = \frac{P(O|w_i)P(w_i)}{P(O)}$$

para um conjunto de probabilidades iniciais  $P(w_i)$ , a palavra falada mais provável depende somente do cálculo de  $P(O|w_i)$ . Dada a dimensão da seqüência de observação  $O$ , a estimativa direta da junção condicional de probabilidade  $P(o_1, o_2, \dots | w_i)$  de todas as palavras faladas não pode ser usada para reconhecimento de palavras. Entretanto, se um modelo paramétrico das produções das palavras como o Modelo de Markov é usado, a estimativa dos dados é possível. Desde que a estimativa da densidade de classe condicional  $P(o_1, o_2, \dots, o_t | w_i)$  for replicada pelo problema de estimativa dos parâmetros do modelo Markoviano.

No reconhecimento de voz com HMM, um modelo de Markov é gerado para cada seqüência de vetores de observação da fala correspondente a cada palavra (Figura 7). O Modelo Markoviano é uma máquina de estados finitos que muda de estado:

- A cada unidade de tempo;
- Sempre que um estado  $j$  é inserido em um tempo  $t$ .

Como apresentado na Figura 7, a cada mudança de estado, um vetor de fala  $o_t$  é gerado a partir da função densidade de probabilidade  $b_j(o_t)$ , e a transição do estado  $i$  para  $j$  depende da probabilidade discreta  $a_{ij}$ . A figura 7 mostra uma máquina de estados finitos onde seis modelos de estado movem-se através da seqüência de estados  $X = 1, 2, 2, 3, 4, 4, 5, 6$  em ordem para gerar  $o_1$  a  $o_6$ . O probabilidade de junção de 0, é calculada pelo modelo Markoviano movendo a seqüência de estados  $X$ . O cálculo é o produto das probabilidade de transição e probabilidade de saída. Para a seqüência de estados  $X$ , na figura 7:

$$P(O, X|M) = a_{12}b_2(o_1) * a_{22} * b_2(o_2) * a_{23} * b_3(o_3) \dots a_{45} * b_5(o_6)$$

No reconhecimento prático, as informações conhecidas são: a seqüência de observação  $O$  e os parâmetros  $a_{ij}$ ,  $b_j(o_t)$  para os modelo  $M_i$ . A seqüência de estados subjacente  $X$  é oculta e precisa ser calculada. Assim, a probabilidade dos estados seguintes pode ser obtida por operações sobre as seqüências de estados possíveis de  $X = x(1), x(2), x(3), \dots, x(T)$ :

- Pelo somatório:

$$P(O|M) = \sum a_{x(0)x(1)} \prod_{t=1}^T b_{x(t)}(o_t) a_{x(t)x(t+1)}$$

- Através do máximo valor de  $x$ , para obter a seqüência mais parecida:

$$P'(O|M) = \max_{x(0)x(1)} \prod_{t=1}^T b_{x(t)}(o_t) a_{x(t)x(t+1)}$$

Note que  $x(0)$  é o estado de entrada e  $x(T + 1)$  é o estado de saída (YOUNG et al., 2005).

As Cadeias de Markov devem ser treinadas para realizar o reconhecimento. A estimação dos parâmetros dos HMMs, como em todos os sistemas estocásticos, é baseada em exemplos de treinamento e é geralmente feita utilizando o algoritmo *forward-backward*, também conhecido como algoritmo Baum-Welch. O critério utilizado para a reestimação dos parâmetros é o de máxima verossimilhança ML (*Maximum Likelihood*), que consiste em aumentar, a cada época de treinamento, a probabilidade *a posteriori*, ou seja, a probabilidade do modelo gerar a seqüência de observações.



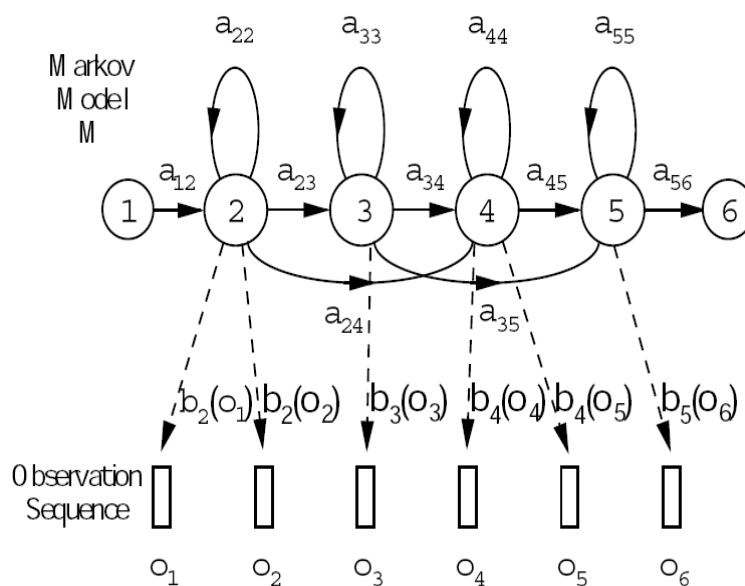


Figura 7: Modelo Oculto de Markov (YOUNG et al., 2005)

### 2.5.5 Reconhecimento de voz contínuo

O processo de reconhecimento envolve a conexão de HMMs em seqüência. Cada modelo da seqüência corresponde diretamente ao suposto símbolo subjacente. Estes podem ser palavras inteiras ou apenas fonemas para o reconhecimento contínuo. A inexistência de saídas dos estados de entrada e saída é feita com o objetivo de unir vários modelos em seqüência (YOUNG et al., 2005).

O treinamento para ditado contínuo necessita de muitas iterações e os limites que dividem os segmentos da fala correspondente a cada fonema não serão conhecidos. Em alguns casos é interessante marcar manualmente os limites dos dados. Entretanto este processo limita o modelo e prejudica boas estimativas. A solução para este problema é utilizar ferramentas auxiliares no treinamento do modelo.

### 2.5.6 Redes Neurais

Existe outra abordagem para obtenção de reconhecimento de voz que utiliza uma técnica da inteligência artificial chamada Redes Neurais. Esta seção vai explicar brevemente algumas das características desse campo de pesquisa, porém sem avançar nos aspectos internos. Os trabalhos de TEBELSKIS (1995) e ANDRADE; JR. (1998) são ricos em detalhes e experimentos e devem ser consultados para maiores detalhes.

Apesar do bom desempenho obtido no atual estado da arte dos HMMs, este modelo faz um número muito grande de suposições que limitam o seu potencial de eficácia. O uso de redes neurais em SR evita muita destas suposições através do aprendizado, permitindo até mesmo ambientes com mais ruído. Embora as redes neurais possam ser usadas nos modelos acústicos, ainda não está bem claro como podem ser exploradas as restrições temporais (TEBELSKIS, 1995).

As redes neurais procuram aproximar a arquitetura do cérebro humano para se beneficiar dos aspectos naturais das interconexões dos neurônios e por isso são caracterizadas por três componentes:

1. Padrões de conexão entre os neurônios;
2. Algoritmo de aprendizagem;
3. Função de ativação.

Esta proximidade de um cérebro humano permite às redes neurais algumas vantagens como ferramenta no reconhecimento de voz (ANDRADE; JR., 1998):

- Quantidade reduzida de informações armazenadas;
- Não há necessidade de manutenção de novas palavras;
- Permitem maior flexibilidade pelo aprendizado.

A figura 8, apresenta um exemplo do resultado obtido com aprendizado em redes neurais. Neste caso, a rede foi treinada para aprender a função seno e cosseno, onde resultado em azul foi obtido pela rede e a linha vermelha é a função correta. Para isso os parâmetros usados foram:

- Algoritmo: *back-propagation*
- Função: Bipolar  $\text{tg}(x)$ ;
- Épocas: 15000;
- Precisão: 0.0001;
- Neurônios na camada escondida: 10;
- Taxa de aprendizado: 10.

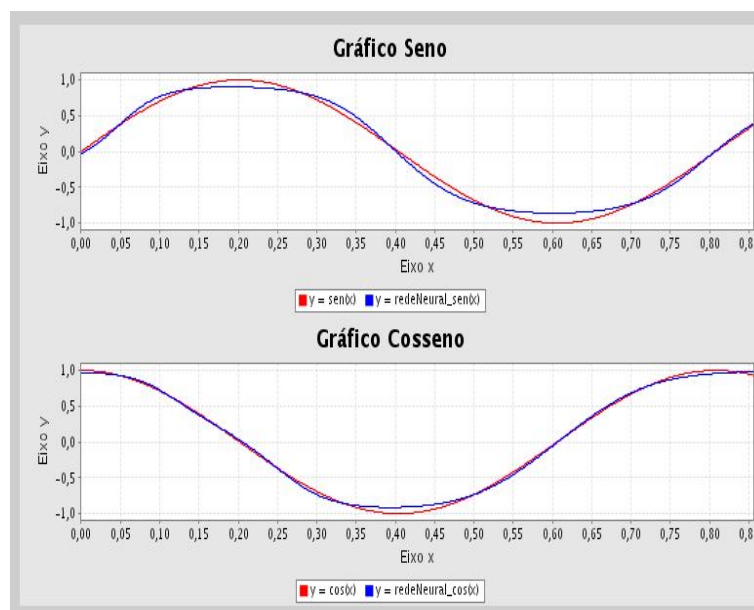


Figura 8: Rede Neural Aprendendo a função seno(YOUNG et al., 2005)

No reconhecimento de voz, para uma boa generalização, o conjunto de treinamento deve se aproximar do universo do discurso. Além disso, para melhores resultados pode ser realizado um treinamento supervisionado onde a rede neural aprende de acordo com respostas do usuário(ANDRADE; JR., 1998).

### 2.5.7 Estado da arte

A tecnologia de reconhecimento de voz é tema de pesquisas em diversos domínios. Algumas das aplicações e projetos existentes na área devem ser analisados para definir o atual estado da arte. A seguir são descritas as ferramentas encontradas com utilidade para a manipulação de áudio e reconhecimento de voz.

#### 1. Praat: doing phonetics by computer

Software para análise de fonemas com recursos de análise de áudio. Embora o programa não tenha suporte a reconhecimento de voz, a ferramenta é útil na análise de fonemas e padrões. O programa possui versões para Windows, Linux, Macintosh e demais sistemas UNIX. Além disso, a licença é gratuita e o código fonte é disponível para modificações. O programa possui documentação consistente onde as principais análises sobre áudio são explicadas: espectral (espectrogramas), formal, por intensidade e análise por pitch. Para o estudo de padrões fonéticos o sistema oferece técnicas de inteligência artificial como redes neurais(BOERSMA, 2001).

## 2. Spoken Language Systems Group

O SLS têm como objetivo tornar possível a interação entre homem e máquina através da fala. O laboratório possui o foco em inteligência artificial, e desenvolve aplicações de reconhecimento de voz pela fala natural. Desde 1989, o grupo desenvolveu aplicativos para diferentes contextos na área:(SLS GROUP, 2006)

- MERCURY - Interface de conversação que permite navegar pelas informações de vôo via telefone;
- PEGASUS - Interface de conversação para prover o estado do vôo e portão de embarque, também via telefone;
- PENATES - Permite interação pela fala para consulta de informações sobre restaurantes em Boston;
- VOYAGER - Sistema que fornece informações sobre o tráfego usando reconhecimento de voz pelo telefone;
- ORION - Agente de agendamento que executa tarefas e entrega avisos por interface de conversação;
- JUPITER - Programa de consulta de tempo (temperatura, umidade, ventos) em diferentes cidades através da fala.

## 3. XVoice

XVoice é uma aplicação que faz o reconhecimento de voz através do ViaVoice SDK (Software Development Kit) da IBM. O programa envia a entrada para o sistema da IBM e redireciona a saída para outras aplicações(XVOICE, 1999). A maior desvantagem desse programa é que a IBM retirou os pacotes disponíveis para linux de seu site. Por isso, sem o pacote ViaVoice SDK que era distribuído separadamente, o XVoice não tem utilidade pois fica sem o módulo para reconhecimento de voz.

## 4. Open Mind Speech

O projeto surgiu com o objetivo de desenvolver aplicações de reconhecimento de voz livre, na licença GPL, e coletar dados de voz de pessoas pela internet. O programa foi projetado para ser integrável com diversas aplicações para Linux. O projeto está sem novidades desde maio de 2000, e por isso o uso dessa ferramenta é desencorajado. Apesar disso a versão beta possui implementação para processamento de sinal com reconhecimento de voz e identificação do locutor por sexo e idade(OPEN MIND SPEECH, 2000).

### 5. StructRad

Ferramenta comercial desenvolvida para dinamizar o processo de laudo radiológico. O StructRad permite o reconhecimento de voz e a integração com outras ferramentas da empresa StructRad LLC. Além disso, possibilita otimizações pelo usuário em macros, templates, ditado, edição por palavras, e laudo estruturado e a integração com RIS/PACS(STRUCTRAD LLC, 2006).

### 6. The Festival Speech Synthesis System

O *framework* Festival oferece uma estrutura em sistemas de síntese de ditado. Através de diferentes especificações, é possível construir discursos de texto em voz sintetizada. Existe suporte oficial para inglês americano, britânico e para espanhol. O sistema é todo desenvolvido em C++ sob licença do X11 que permite uso comercial e não comercial de forma irrestrita. O projeto Festival não possui sistema de reconhecimento de voz incluso e fica limitado a síntese de voz(THE CENTRE FOR SPEECH TECHNOLOGY RESEARCH, 1999).

### 7. Edictation

Software comercial que possui sistema de reconhecimento de voz em inglês associado a laudo estruturado e assinatura digital. A aplicação possui sistema que recupera automaticamente termos e frases únicos para criar e distribuir laudos radiológicos(EDICTATION INC., 2004).

### 8. Philips SpeechMagic

O SpeechMagic é um software consolidado no mercado clínico que permite ditado com reconhecimento de voz. A grande desvantagem deste sistema é que ele proprietário. O programa da Philips, SpeechMagic, recebeu o prêmio europeu de liderança em tecnologia *European Technology Leadership Award*, atribuído pela Frost Sullivan, no mercado de reconhecimento de voz para serviços de saúde na Europa(PHILIPS ELECTRONICS, 2006).

### 9. CSphinx-4

Desenvolvido em conjunto pela Universidade Carnegie Mellon, Sun Microsystems Laboratories, Mitsubishi Electric Research Labs e Hewlett Packard (HP), com contribuições da Universidade da Califórnia (UCSC) e o Massachusetts Institute of Technology (MIT). O CSphinx-4 é um sistema de reconhecimento de voz escrito em Java que permite reconhecimento de voz em tempo real e em *batch*(OPEN SOURCE TECHNOLOGY GROUP, 2004).

## 10. **Hidden Markov Models Toolkit**

A biblioteca HTK é um conjunto de ferramentas desenvolvidas para manipulação de Modelos Ocultos de Markov (HMM). O projeto foi inicialmente usado em pesquisas de reconhecimento de voz, porém já foi utilizado em síntese de voz e reconhecimento de caracteres. O HTK possui um conjunto de módulos que facilitam a análise da fala, treinamento dos HMMs, testes e análise de resultados(YOUNG et al., 2005).

## 11. **Dragon Systems**

Conjunto de ferramentas para reconhecimento de voz desenvolvido pela Nuance Communications, Inc. A empresa comercializa aplicações na área com destaque para:

- Dragon Naturally Speaking Mobile - Aparelho de reconhecimento de voz móvel que faz a transcrição ao ligar o sistema em um computador;
- Dragon Naturally Speaking - Um dos primeiros sistemas de reconhecimento de voz natural. O programa tem suporte para o inglês falado de forma natural, onde o processo de treinamento demora em média 18 minutos. Assim, o usuário está apto a ditar o texto para máquina de uma maneira natural e as palavras ditadas tem 95% de acerto(SILVA; ROSA; VIANA, 1999).

A empresa possui um software para área médica em desenvolvimento, a empresa já oferece o programa em pré venda pela página web da Nuance(NUANCE COMMUNICATIONS, INC., 2006).

## 12. **IBM ViaVoice**

Famoso software de reconhecimento de voz da IBM com suporte a diversas línguas entre elas o português do Brasil. O tempo de treinamento leva em média 15 minutos, com a leitura de um texto padrão. Em seguida, pode ser feito o ditado citando as palavras e a pontuação. Atualmente a IBM está reestruturando a comercialização do Via Voice mundialmente, a venda deste produto está sendo realizada pela Nuance.

## 13. **Projeto REVOX**

Pesquisa aplicações em reconhecimento automático de voz com assinatura vocal para aplicações industriais, com desenvolvimento de software e hardware específicos controladores mecânicos como elevadores(ADAMI; DORNELES, 1997).

## 14. **CVoiceControl**

Sistema de reconhecimento de voz que habilita o usuário a mapear comando de voz

para comandos UNIX. O *software* automaticamente detecta a entrada no microfone e processa este sinal e se o reconhecimento tiver sucesso é executado o comando correspondente(KIECZA, 2002).

#### 15. ISIP Speech Recognition Toolkit

Ambiente de pesquisa para aplicações em análise da fala, reconhecimento e verificação de voz. O objetivo primário do projeto, denominado “*Internet-Accessible Speech Recognition Technology*”, é desenvolver módulos independentes com o estado da arte em reconhecimento de voz afim de simplificar o uso e avanços nas pesquisas da área(TEAM, 2006).

#### 16. Transcriber

Ferramenta para assistência na segmentação, marcação e transcrição de sinais de voz. O *software* é portátil, e utiliza dados em formato estruturado no padrão XML com suporte a transcrição em múltiplas línguas. Através do *framework* implementado, a adaptação da ferramenta em processar novas tarefas e suportar diferentes formantos ficam mais flexíveis(BARRAS et al., 2001).

## 3 Metodologia

### 3.1 Ferramentas de desenvolvimento

O desenvolvimento do trabalho foi feito com tecnologias livres e de código aberto, com sistema operacional Linux. Porém, os módulos usados e implementados são portáveis para Microsoft Windows. Todos os protótipos de manipulação de áudio foram implementados em C++, linguagem padrão do Projeto Cyclops e que permite facilidade na integração com as bibliotecas de captura, reprodução e conversão do áudio:

- Bibliotecas de captura e reprodução de áudio (OSALP e PortAudio);
- Bibliotecas para conversão de áudio Vorbis e LAME;
- Biblioteca de reconhecimento de voz HTK;

Para o desenvolvimento de GUI<sup>1</sup> foi utilizada a biblioteca wxWidgets, que permite desenvolver aplicações para Win32, Mac OS X, GTK+, X11, Motif, WinCE, recompilando o código para cada ambiente gráfico. O uso do wxWidgets pode ser feito através das linguagens C++, Python, Perl e C#/.NET e independente da linguagem escolhida o programa mantém a aparência da plataforma onde está executando(WXWIDGETS, 2006). A escolha do wxWidgets é fundamentada por ser *cross-plataform*, livre e *open-source*. A biblioteca teve início em 1992 na Universidade de Edinburgh e já possui amadurecimento para aplicações confiáveis:

- KICAD - Aplicação para projetos eletrônicos para Windows e Linux;
- Kirix Strata - Aplicação para banco de dados para Windows e Linux;
- Transcribe - Aplicação gravação de músicas para Windows, Mac OS X e Linux;

---

<sup>1</sup>GUI - Graphical User Interface, Interface Gráfica do Usuário



- Cn3D - Aplicação para visualização de estruturas 3D para Windows, Mac OS X e Linux;

### 3.1.1 Biblioteca áudio

Para a gravação e reprodução dos laudos a biblioteca PortAudio oferece a solução mais flexível para diferentes sistemas operacionais em manipulação de entrada e saída de áudio. No entanto a manipulação de formatos definidos de áudio não é suportado pela biblioteca PortAudio. A solução para sistemas linux nesse caso foi o uso da OSALP que permite a utilização de formatos comprimidos ou compactados para reprodução como MP3, OGG e WAV. Com o objetivo em tornar a aplicação mais flexível, foi definida uma interface que estende as classes de manipulação de áudio para cada biblioteca(Figura 9).

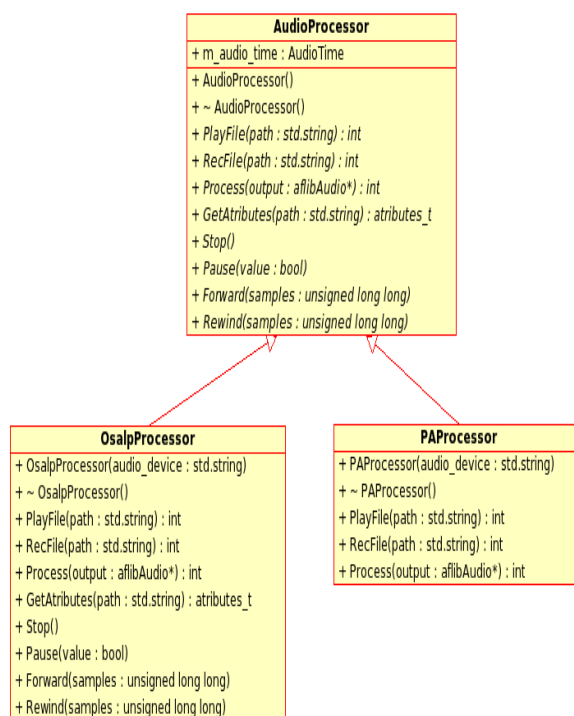


Figura 9: Abstração para implementar diferentes bibliotecas de áudio

O desenvolvimento de uma versão portátil entre Linux e Windows requer o uso do PortAudio como biblioteca oficial do programa para manipulação de áudio. Para isso, após a gravação com essa biblioteca, foi implementado o cabeçalho de arquivos WAV (Apêndice D) e a partir deste fica mais viável converter o arquivo nos formatos OGG e MP3.

### 3.1.2 Biblioteca de reconhecimento de voz

Dentre as ferramentas pesquisadas, o HTK (C++) e o CSphinx-4 (Java) destacaram-se pela maturidade do projeto e boa documentação. A escolha pelo HTK foi feita com base na linguagem C++ usada no desenvolvimento da biblioteca que mantém o padrão usado no Cyclops.

A figura 10 apresenta os dois principais estágios de processamento envolvidos no reconhecimento de voz pelo HTK. Primeiro é realizado o treinamento com um ditado conhecido e transcrição definida. Isto permite estimar os parâmetros do conjunto de HMMs nas iterações de treinamento e associá-los a transcrições. No segundo estágio, ditados desconhecidos formam a entrada do sistema que irá executar o reconhecimento e transcrição com base nos parâmetros definidos pelo treinamento.

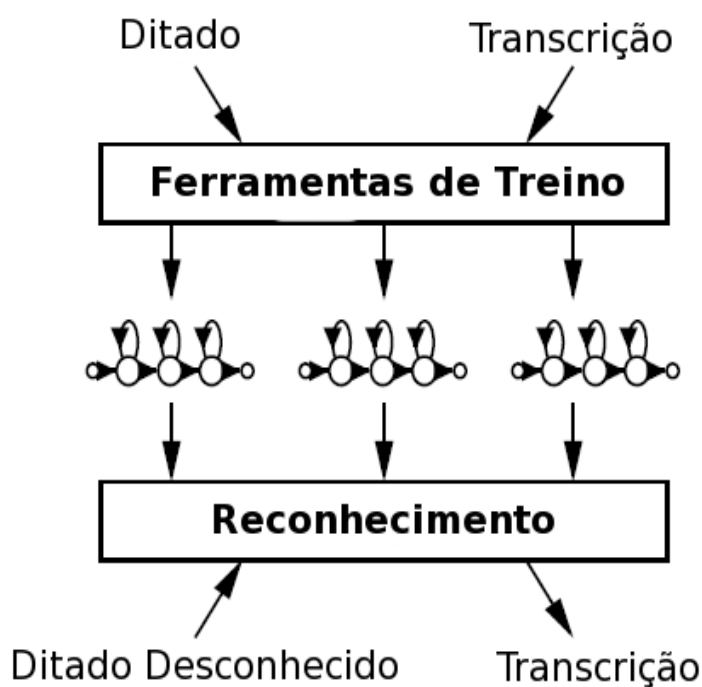


Figura 10: Abstração de reconhecimento de voz HTK

### 3.1.3 PHP e PostgreSQL

O sistema de gravação de laudo tem como requisito o acesso de laudo digital pelo portal de telemedicina desenvolvido pelo Cyclops. Cada ditado gravado deve ser associado a um exame, seja por método de reconhecimento de voz ou pela ação de digitadores o

laudo pode ser transcrito em texto. É importante nesse caso, manter as ferramentas utilizadas pelo projeto para o desenvolvimento web. As escolhas pelo PHP e PostgreSQL são fundamentadas no trabalho de WALLAUER, 2005.

Com o objetivo de incrementar a usabilidade do portal com as propostas deste trabalho, foram propostas novas funcionalidades no portal de telemedicina. A figura 11 apresenta as informações dos exames no portal sem suporte a laudo ditado. As alterações sugeridas podem ser vistas na figura 12. A listagem de exames recebeu dois ícones novos:

1. Microfone e texto - Sinaliza que o laudo já passou pelo processo de transcrição e aguarda as correções finais do médico;
2. Caixa de som - Indica que esse exame possui um laudo ditado. O áudio armazenado pode ser reproduzido para facilitar a correção do texto;

**HOSPITAL UNIVERSITÁRIO**  
Prof. Polydoro Ernani de São Thiago

Portal de Telemedicina - Executor >> Lista de exames para laudo

Bem vindo Cyclops - Hoje é 19 de Julho de 2006. [Ajuda ?](#) [Alterar minha senha](#) [Sair](#)

Busca:  [ 1 ]  
Requisição  Setor: Broncoscopia  
Mostrar: Todos

Requisição	Origem	Data	Paciente	Requisitante	Executor	Exame
✓ HU18848542	Florianópolis	07/07/2006 00:00:00	BRONCO-REMOTO	Dr. Cyclops	Dr. Cyclops	Broncofibroscopia
✓ HU6521509	Florianópolis	06/03/2006 00:00:00	Cyclops_Bronco	Dr. Cyclops	Dr. Cyclops	Broncofibroscopia

[Ajuda ?](#) [Alterar minha senha](#) [Sair](#)

Desenvolvimento - Equipe de Telemedicina © 2005 - 2006 Cyclops Team

Figura 11: Portal de telemedicina sem suporte a laudo ditado

## 3.2 Metodologia para reconhecimento de voz

Empresas de tecnologia na área médica investem no reconhecimento de voz com sistemas de informação na saúde. A Philips, Dragon e a MacSym possuem programas de SR



Figura 12: Portal de telemedicina com suporte a laudo ditado

na área médica, no entanto, as soluções existentes são comerciais e nem sempre permitem integração com a telemedicina ou PACS. Esta proposta considera a integração com os sistemas de telemedicina, PACS, gravação de laudo digital e reconhecimento de voz.

A solução descrita a seguir é o uso de SR para processamento posterior, ou seja, não serão consideradas as restrições de tempo real. Essa solução é conhecida como processamento em *batch*.

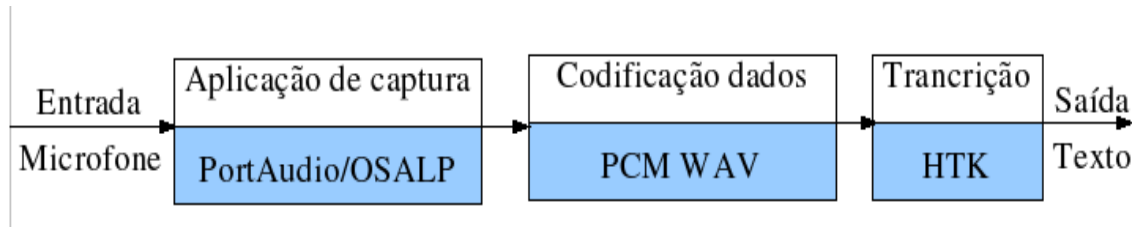


Figura 13: Fluxo de informação no sistema SR

A figura 13 apresenta o fluxo da informação, desde o ditado pelo microfone até a geração de texto. Cada módulo por onde a informação é transferida, utiliza programas e bibliotecas open-source. A entrada de áudio é realizada por um microfone, controlada por uma interface gráfica com o usuário. O fluxo de áudio é manipulado com o PortAudio e o resultado é convertido em um arquivo no formato WAV. Este arquivo é enviado para o sistema de reconhecimento de voz HTK para processamento acústico e transcrição do áudio em formato texto.

A forma de reconhecimento em *batch* oferece independência à aplicação onde o laudo é gravado. Como o processo de reconhecimento e transcrição é executado em um servidor, é possível enviar o laudo ditado gravados em dispositivos móveis, tais como computadores de mão.

O HTK oferece um conjunto de ferramentas que auxiliam no processo de reconhecimento de voz. Com base nos trabalhos de MOREAU e YOUNG et al., foram executados os seguintes procedimentos para implementar um sistema mínimo de reconhecimento de voz:

- **Definição da Gramática e do Dicionário** - Regras usadas no reconhecimento. No dicionário são definidas as palavras aceitas pelo sistema no contexto que a aplicação será utilizada, neste caso, da Radiologia;
- **Definição do Modelo Acústico** - Caracterização da forma como os sons das palavras devem ser representados;
- **Definição do corpo de treinamento** - Frases que utilizam palavras do dicionário para realização do treinamento do sistema de reconhecimento;
- **Definição dos modelos HMMS** - Arquivos que definem o modelo e a forma de transcrição para cada palavra do dicionário;
- **Configurações de codificação** - Etapa onde os algoritmos de codificação e parâmetros de reconhecimento e treinamento são definidos.
- **Treinamento** - Processo usado para balancear os valores dos HMMs a fim de identificar a entrada do usuário.
- **Reconhecimento** - Etapa onde o sistema é utilizado para gravação de ditado em áudio digital. O resultado da gravação é processado pelo módulo SR que retorna a transcrição em texto livre da entrada em áudio;
- **Avaliações** - Análise dos resultados de acerto do texto gerado pelo SR em relação ao texto ideal.

O funcionamento e a comunicação das informações mencionadas é apresentado na figura 14. A definição da gramática permite gerar a rede de navegação das frases permitidas pela liguagem. A construção do modelo de reconhecimento é obtida pelo dicionário, pelo autômato referente a gramática e pelas informações de treinamento. Com essas informações, os HMMs são montados e fazem a transcrição do áudio de entrada para texto na saída.

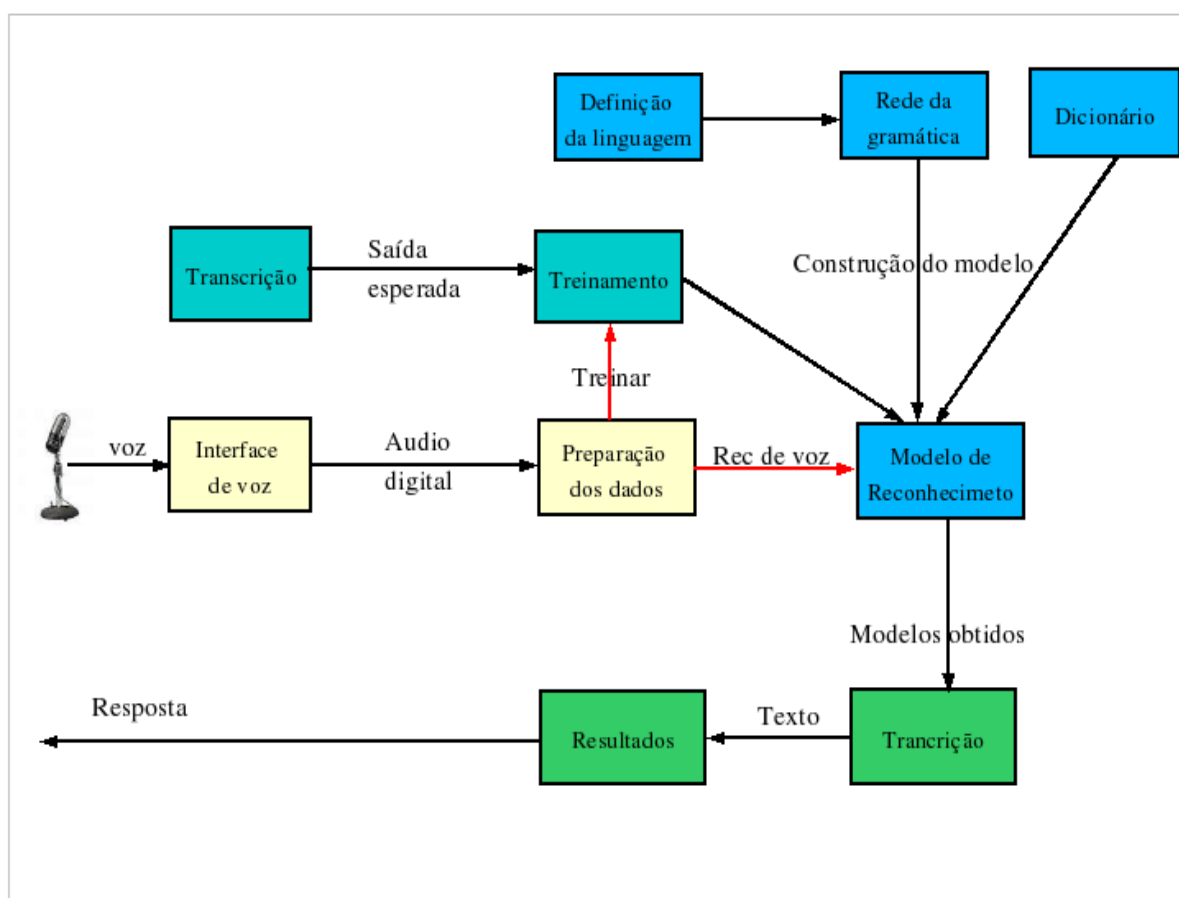


Figura 14: Comunicação entre os módulos para reconhecimento de voz com o HTK.

### 3.3 Análise dos formatos de áudio

Três formatos de áudio foram analisados para o uso no sistema de telemedicina. A escolha se torna difícil pois os formatos possuem qualidades em direções opostas. O MP3, é mais difundido, possui programas para reprodução e gravação em computador, celular e palmtops. No entanto, está vinculado a royalties e perde qualidade para uso em reconhecimento de voz. O OGG é um formato livre com boa difusão em ambientes linux, porém o uso desse formato é inviável para aparelhos móveis dada a pouca usabilidade desse tipo de arquivo. A qualidade do formato OGG é maior, como afirma AMMOURA; CARLACCI:

*“Ogg Vorbis provê bitrate variável o que preserva mais detalhes e definição quando necessário. Uma das principais razões que motiva o desenvolvimento do Ogg Vorbis está relacionado as patentes. Diferente dos CODECs mais populares, o Ogg Vorbis é código aberto, não proprietário e é livre de*

*patentes e de royalties. É um formato de áudio de propósito geral para uso de áudio e música com alta qualidade. Uma das principais características que faz o OGG bem aceito e a possibilidade de ajustar o bitrate de acordo com parâmetros de qualidade ou média. Isto coloca o Vorbis no mesmo patamar de representação de áudio que o MPEG-4 (AAC) e semelhante a, mas com melhor performance que o MP3, WMA e PAC.”*

A tabela 4 apresenta as vantagens do formato livre em relação ao MP3. No entanto, a reprodução dos laudos em cada cliente fica restrita se o formato definido for OGG/Speex.

Atributo	MP3	OGG
Livre de proprietários	-	X
Livre de patente	-	X
Código Aberto	-	X
CODEC para voz	-	Speex
Suporte em palmtops	X	-

Tabela 4: Vantagens do formato OGG sobre o MP3

O áudio sem codificação torna-se proibitivo para armazenamento e reforça a necessidade de um formato de áudio que compacte a informação. A imagem 15, mostra o gráfico dos tamanhos do laudo nos formatos de áudio estudados (Tabela 5) onde cinco laudos médicos foram gravados e em seguida codificados em OGG, MP3 e Speex. É possível observar que a melhor taxa de compactação é obtida pelo CODEC específico para voz: Speex.

A comparação entre o OGG, MP3 e Speex para obtenção da tabela 5 e gráfico foi realizada com os seguintes parâmetros:

1. **MP3** - CODEC: LAME

**Taxa Amostragem:** 44100 Hz

**Taxa de bits:** 128 Kbps

**Canais:** 2

**Bits por amostra:** 16

2. **OGG** - CODEC: Vorbis

**Taxa Amostragem:** 44100 Hz

**Taxa de bits:** 128 Kbps

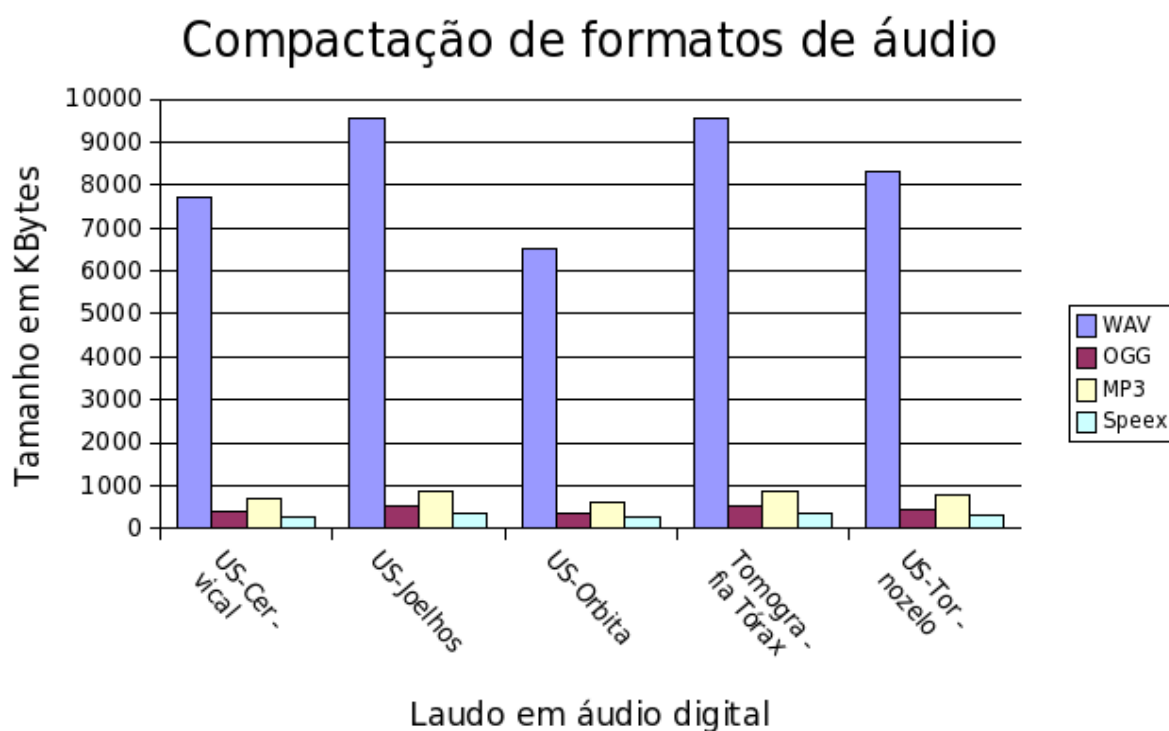


Figura 15: Comparação no armazenamento dos formatos WAV, MP3, OGG, Speex

**Canais: 2**

**Bits por amostra: 16**

3. **Speex - CODEC: Speex**

**Qualidade: 8/10**

**Taxa de bits fixa: 48 Kbps**

**Complexidade de codificação: 3**

**Frames por pacote OGG: 1**

A figura 16 mostra a redução de cada formato em relação ao arquivo original. O formato de áudio específico para voz, Speex, obtém em média redução de 26,05 vezes em relação ao áudio original. Em segundo, o formato OGG teve redução média de 14,08 vezes enquanto o MP3 com média 10,91 fica em terceiro. Esses valores podem mudar de acordo com os parâmetros utilizados.

É importante esclarecer que o uso das TAGs ID3v2 fica restrito aos formatos OGG e MP3. As tags permitem a associação do exame com laudo ditado somente pelo arquivo do ditado; pois este possui na tag a informação necessária para referenciar ao exame que



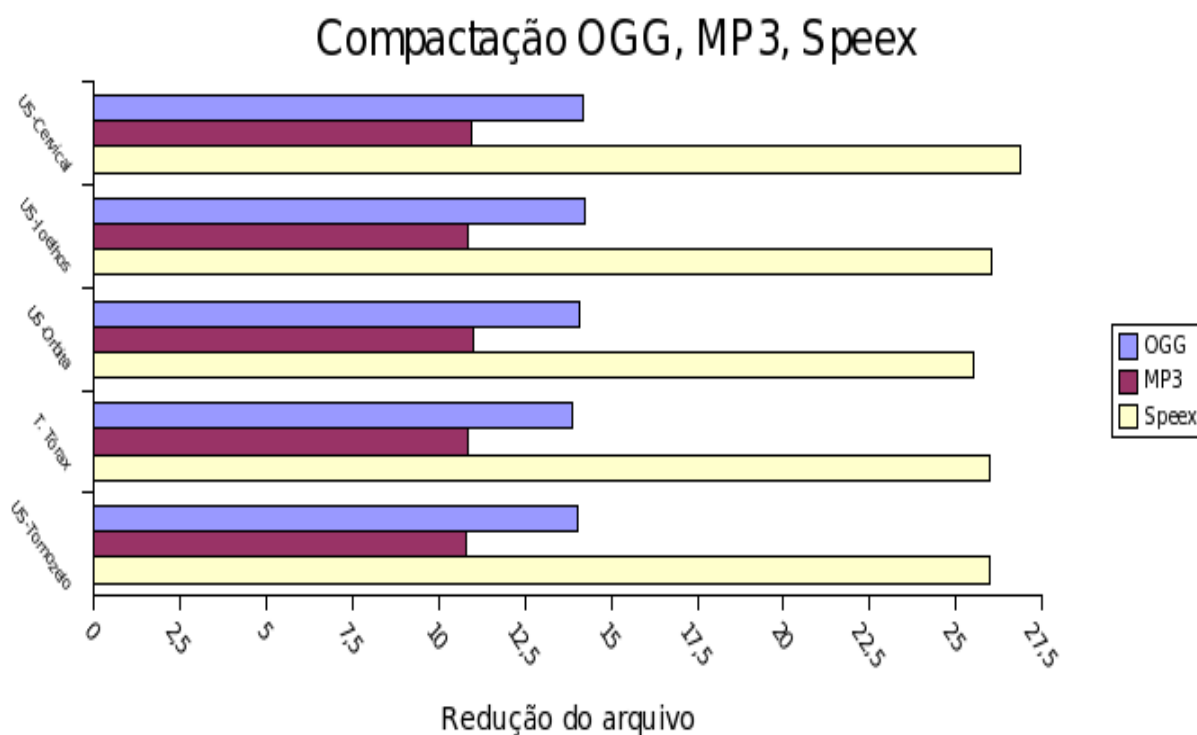


Figura 16: Redução do tamanho do arquivo entre MP3, OGG, Speex

Laudo	WAV	OGG	MP3	Speex
US-Cervical	77728	544	704	288
US-Joelhos	9568	672	880	368
US-Órbita	6528	464	592	256
Tomografia Tórax	9552	688	880	368
US-Tornozelo	8304	592	768	320

Tabela 5: Tamanho em Kbytes dos laudos gravados em diferentes formatos de áudio

pertence. A solução ideal é um modelo híbrido, que compacte as informações no banco de dados, minimizando o custo de armazenamento, e forneça usabilidade para os usuários. Para o cliente que irá reproduzir os laudos, o MP3 é o formato que fornece a melhor usabilidade pois todos os programas de áudio tem suporte ao formato.

### 3.4 Implementação

A implementação do protótipo de gravação e reprodução de áudio foi feita com análise de requisitos e documentação em UML (Apêndice A). Por tratar-se de um programa para validação deste trabalho, somente os diagramas necessários para um protótipo inicial

foram implementados. O processo de desenvolvimento adotado foi prototipagem modular em ciclos. O objetivo desse modelo híbrido é contornar as limitações da paralisação dos requisitos antes do projeto do sistema ou da codificação pela prototipagem. Segundo, Pressman (1997), a modularização permite soluções genéricas e reutilizáveis, e o processo de ciclos incentiva a evolução dos módulos e protótipo.

Foram consideradas duas abordagens para o desenvolvimento do protótipo. A primeira, considera o processo de reconhecimento de voz embutido no programa de gravação. Nesse caso o reconhecimento é possível planejar um sistema em tempo real. A segunda, utiliza o servidor PACS para realizar o reconhecimento dos laudos ditados por um processo em batch executado em um servidor. O sistema adotado utiliza a última abordagem, como é mostrado na figura 17, pois esta oferece maior independência do programa e é mais simples para implementação.

As informações do exame são enviadas juntamente com o áudio do laudo ditado. O servidor inicia o processo de reconhecimento e publica o exame com laudo em formato de áudio digital. Quando o módulo HTK encerra a transcrição, o texto é adicionado ao exame e aguarda a correção do médico. O fluxograma da figura 17 permite o uso de aparelhos móveis para envio de laudo ditado, desde que, este esteja associado a um exame.

### 3.4.1 Cyclops Report Recorder

Um protótipo de manipulação de áudio foi desenvolvido para análises e futura integração com PACS existentes no projeto Cyclops. O programa, chamado de Cyclops Audio Recorder, pode ser usado independente da maturidade de um sistema reconhecimento de voz. Assim, o programa pode ser implantado em clínicas que utilizam laudo ditado com gravação em fitas e transcrição por equipe de digitação.

A entrada de áudio realizada através de um microfone é controlada por uma interface gráfica com o usuário (Figura 18). O fluxo de áudio é manipulado com o PortAudio e o resultado é convertido no formato WAV. Este arquivo é enviado ao servidor que armazena o ditado junto ao exame. A transcrição do áudio em texto é realizada pelo sistema de SR ou por um digitador. O médico realiza as correções escutando o laudo gravado por ele e finalmente confirma o exame. O procedimento descrito pode dinamizar o fluxo de trabalho sobretudo se tiver boa usabilidade. Por exemplo, o sistema usado pelo digitador pode sinalizar quando um novo exame está pendente, e o médico recebe por email o aviso de que o laudo aguarda correção.

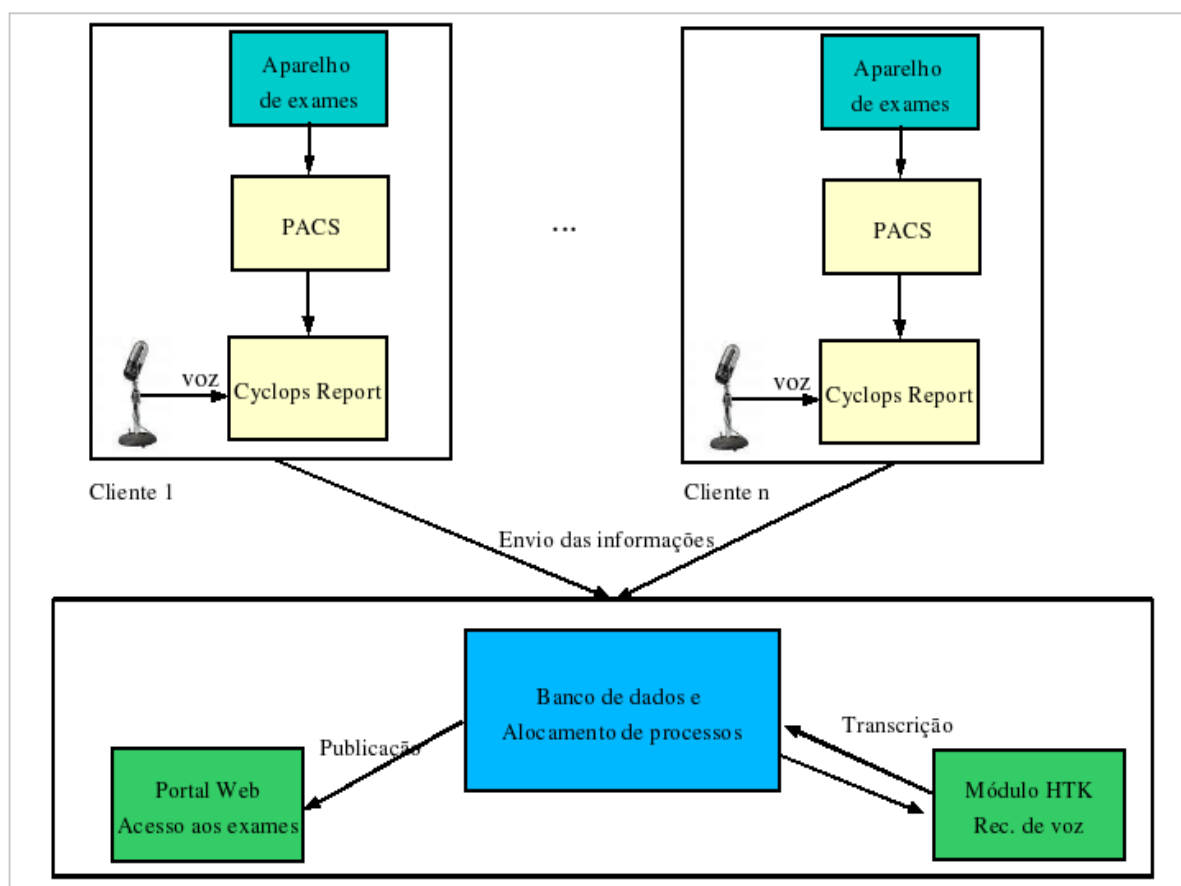


Figura 17: Proposta de implementação para laudo ditado e reconhecimento de voz.

Embora o *software* esteja em fase de desenvolvimento e adaptações, a metodologia de projeto adotada permite apresentar alguns resultados. Com a biblioteca OSALP o programa reproduz áudio codificados em WAV, OGG e MP3 em ambientes linux. A reprodução de uma música em MP3 pode ser visualizado na figura 18.

Ainda na figura 18, podem ser observados os campos adicionados para suporte às tags no padrão ID3v2 com o objetivo em tornar mais independentes os arquivos de laudos ditados. As tags ID3v2 permitem adicionar informações extras a arquivos de áudio formatados (MP3, OGG). A vantagem do uso desse mecanismo é que é possível consultar e gravar informações extras no próprio arquivo de áudio. As informações suportadas pelo programa são:

- Título - Nome do exame que foi realizado;
- Hospital - Armazena o nome do hospital onde o laudo foi gravado;
- Médico - Armazena o nome do médico que gravou o laudo ditado;

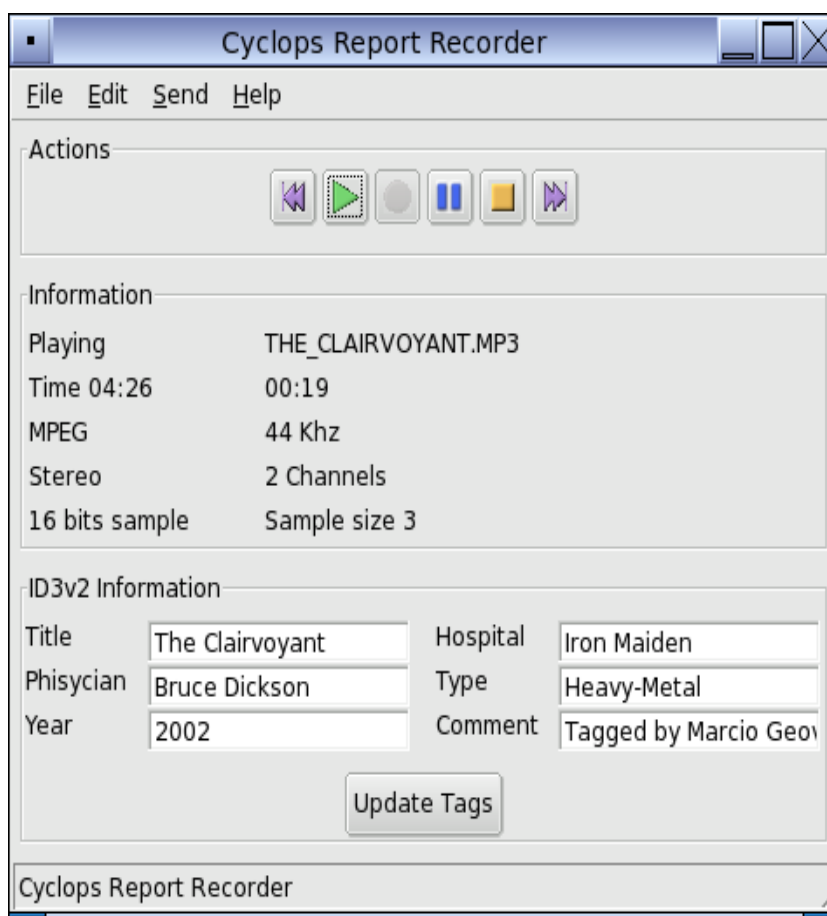


Figura 18: Programa Cycreport reproduzindo arquivo MP3

- Tipo - Tipo do exame que o laudo pertence;
- Ano - Ano em que o exame foi feito e o laudo armazenado;
- Comentário - Informações sobre setor do hospital nome do paciente entre outras;

Outras informações podem ser adicionadas seguindo o padrão ID3v2. Neste trabalho o objetivo é apresentar a viabilidade do uso destas tags para os laudos médicos. O uso deste mecanismo pode ser mais explorado com informações lidas de arquivos XML configurados para os diferentes Hospitais e Setores.

#### 3.4.1.1 Funções do programa

Como mencionado 3.4.1, o protótipo foi planejado para ser usado com reconhecimento de voz ou por digitadores. Por isso, conforme a figura 19, campos para edição do laudo foram incluídos em conjunto com funções triviais para uso de laudo em áudio digital:

1. Gravação - Armazenar um laudo ditado em arquivo para posterior manipulação;
2. Reprodução - Permitir que um arquivo gravado pelo programa seja reproduzido;
3. Interromper - Parar a gravação ou a reprodução em um momento e permitir que a ação continue do instante desejado;
4. Avançar e Retroceder - Essas funções são úteis para digitadores que fazem o processo de transcrição dos exames em áudio digital para texto puro. As duas ações podem estar associadas a pedais que quando pressionados avançam ou retrocedem o tempo do áudio. No protótipo deste trabalho essas opções funcionam nos botões avançando 10 segundos quando pressionados.
5. Parar - Cancelar uma reprodução ou gravação;
6. Editar texto - Campo para adição de texto referente a um laudo sem formatação.

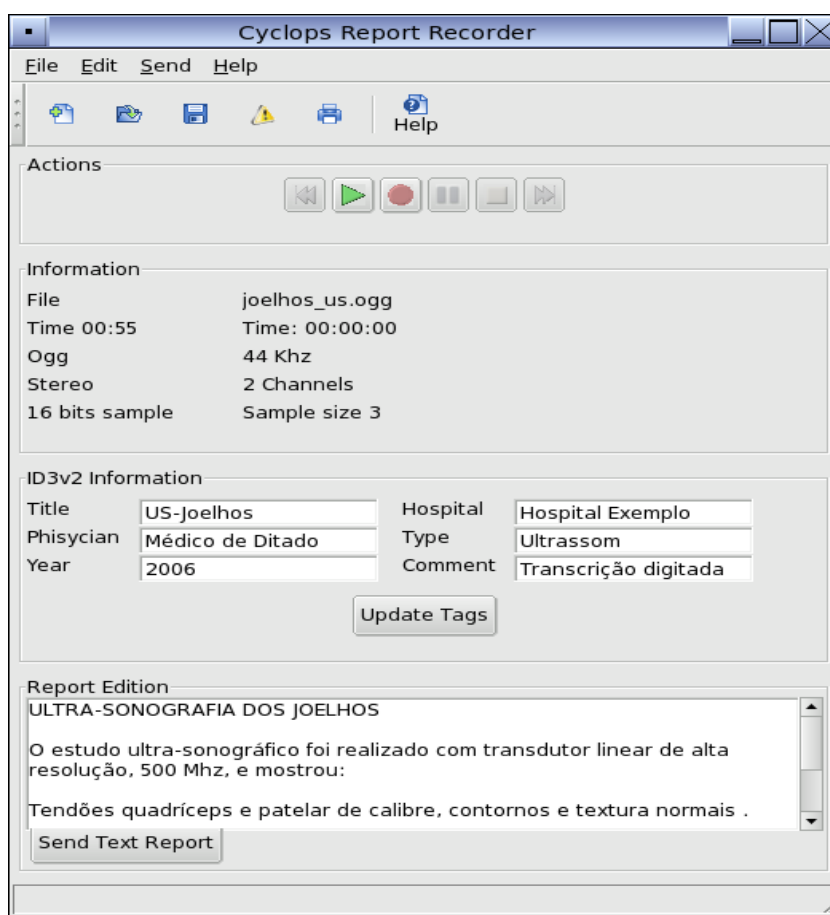


Figura 19: Digitação de um exames obtido do servidor PACS

Através do cycreport é possível fazer conversões de áudio, desde que as bibliotecas de cada CODEC estejam instaladas no computador em que o programa está rodando. A tabela 6 mostra os formatos suportados para gravação e reprodução.

Ação	MP3	OGG	WAV	RAW
Gravar	-	-	X	X
Reproduzir	X	X	X	X
Converter	X	X	X	X

Tabela 6: Formatos com suporte do programa Cyclops Áudio Report

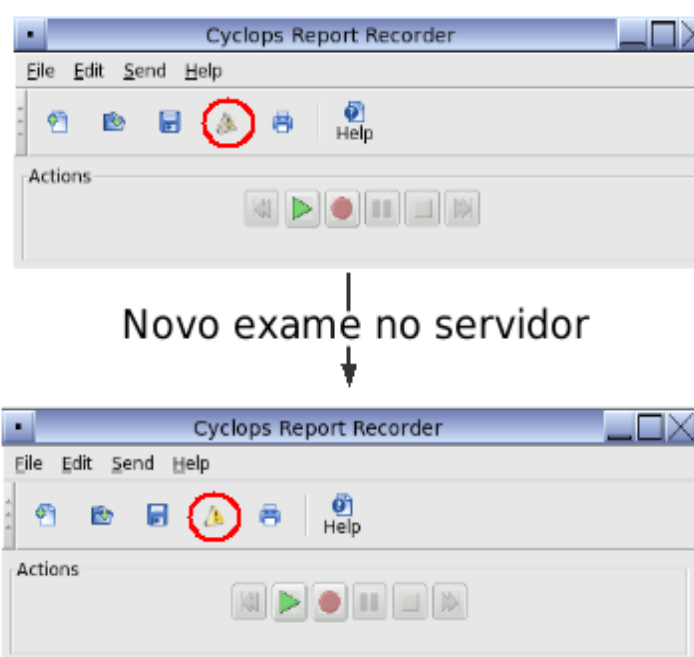


Figura 20: Aviso de forma visual para o digitador

O protótipo permite o envio dos laudos digitais para o portal de telemedicina, um banco de dados PostgreSQL. O envio destes laudos em formato de áudio digital, permite que o médico ouça o laudo de qualquer computador conectado a internet. Além disso, o programa verifica a base de dados por novos laudos gravados sem transcrição e avisa ao digitador, conforme a figura 20.

Os parâmetros de gravação usados no programa são definidos estaticamente e futuramente podem ser definidos em tempo de execução do programa. Um áudio em formato WAV gravado pelo cycreport possui as características suficientes para garantir boa qualidade de gravação:

- Frequência de 44100 Hz (amostras por segundo);

- 2 Canais;
- 16 bits com sinal de tamanho de cada amostra;

### 3.4.2 Reconhecimento de voz com HTK

De acordo com a metodologia adotada, foram realizados testes com a ferramenta HTK no reconhecimento de palavras contínuo em português. O fato de nenhuma base de dados na língua portuguesa ter sido encontrada dificultou o trabalho pois todas as palavras definidas na gramática tiveram que ser gravadas várias vezes.

```
$tipo= TORAX | JOELHO;  
$pnom= DO $tipo;  
$exam= ESTUDO $pnom | EXAME $pnom;  
( START_SIL [ $exam ] END_SIL )
```

Tabela 7: Gramática com 5 palavras

Para a execução dos procedimentos definidos em 3.2, foram utilizadas ferramentas disponíveis no HTK e scripts Unix. As primeiras análises foram feitas com uma gramática simples, de cinco palavras, como mostra a figura 7. A partir dessa definição, é possível gerar o autômato que representa a linguagem a ser reconhecida. O programa que realiza este procedimento é o *HDparse*, que gera em um arquivo o equivalente a figura 21. Note que o início e término de cada iteração completa no autômato é feito pela detecção do silêncio: *START-SIL*, *END-SIL*.

O dicionário pode ser construído manualmente em casos simples, mas torna-se inviável a medida que o número de palavras cresce. A definição do dicionário pode ser feita de duas formas diferentes. Na primeira, as palavras são listadas ao lado da subdivisão de cada fonema que a constrói. Já na segunda, usa-se uma marcação (*label*) para cada palavra que está associada a uma lista de arquivos gravados. Estes *labels* são definidos na gravação do áudio pela ferramenta *HSLab*, como mostra a figura 22

Cada palavra da gramática foi gravada dez vezes pelo processo manual, ou seja, cinquenta gravações seguidas de marcações em um sistema dependente de locutor. Uma vez gravadas todas as palavras, é necessário realizar a preparação acústica do áudio para o processamento do HTK. O programa *HCopy* realiza esta tarefa recebendo como entrada

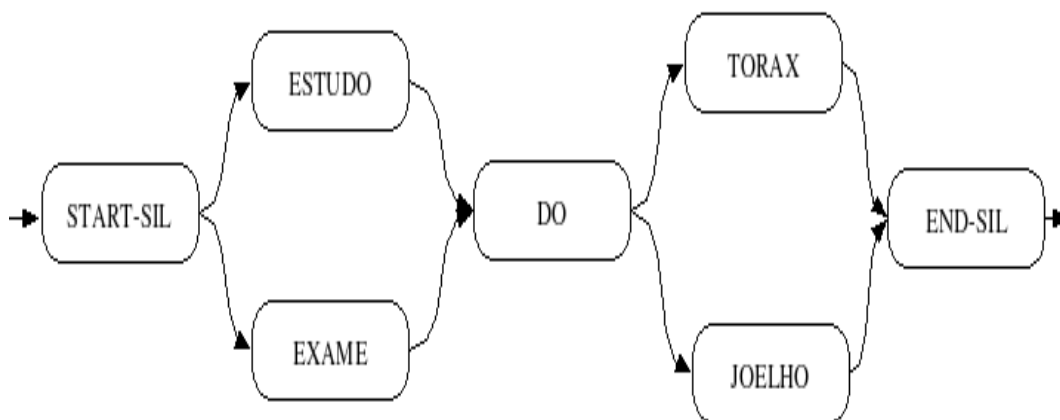


Figura 21: Autômato gerado a partir da gramática de 5 palavras

a configuração acústica (Anexo A) e gera arquivos no formato MLF (*Master Label File*), que contém a correta transcrição de todo corpo de testes

Com o programa *HsGen*, é possível verificar a validade da gramática de acordo com o dicionário. Passando como entrada o autômato da linguagem e o dicionário utilizado, a saída será uma lista das construções possíveis com a gramática definida. Estas frases são úteis na realização de treinamentos em sistemas independentes de locutores.

Para cada palavra, HMMs sem valores definidos (Anexo B) foram criados com posterior preparação dos dados sobre os modelos. Os modelos usados tem 6 estados, entrada, saída e quatro estados ativos. A definição dos valores de cada HMMs é obtida através de três programas:

1. **Hinit** - Usado para inicializar os valores de cada modelo de Markov com alinhamento dos dados no tempo através de algoritmos específicos. Para o procedimento descrito a inicialização foi feita com o algoritmo de Viterbi(YOUNG et al., 2005);
2. **HCompV** - Tem a mesma função do HInit, é chamada de inicialização plana. Mesmo com o uso do *HInit*, o *HCompV* é útil pois gera um arquivo chamado vFloors que evita erros originados por pequenas variações na estimativa de treinamento;
3. **HRest** - Realiza a reestimação dos parâmetros existentes nos modelos e cria uma evolução do HMM otimizando as probabilidade de transição, variações nos vetores de cada função de observação.



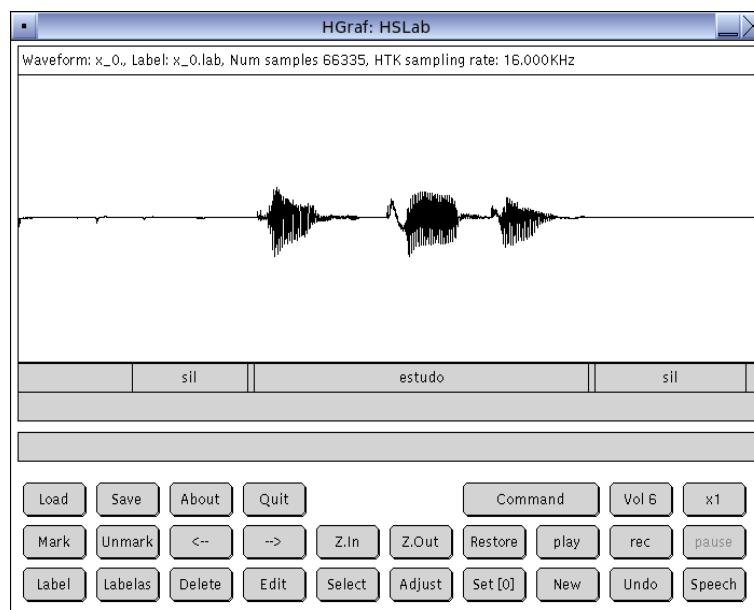


Figura 22: Marcando partes do áudio com o *HSLab*

O reconhecimento de voz é realizado pelo *HVite* onde pode ser verificado o funcionamento do modelo construído. O HTK oferece ainda, uma ferramenta de análise de resultados chamada *HResult*. No modelo descrito nesta seção, a taxa de acerto média obtida entre os dois programas foi de 80%. Porém, é necessário observar que a gramática simples facilita o acerto de 95% das frases obtido pelo *HVite*. A comparação realizada pelo *HResults*, entre sentenças completas e palavras isoladas, obteve somente 65% de acerto.

## 4 *Conclusão*

Este trabalho propõe uma metodologia para dinamizar e aumentar a confiabilidade do processo de geração de laudo médico. Uma proposta inédita na rede pública de saúde brasileira, que permite reduzir resistência dos recursos humanos na adaptação do sistemas informatizados para hospitais e clínicas.

A introdução do ditado digital fornece qualidade de áudio e mantém histórico da gravação junto às imagens e demais informações do exame. Além disso, adiciona uma interface mais amigável em relação ao tradicional sistema de digitação. No entanto, é importante estar ciente que as interfaces de mouse, pedal e teclado são eficientes e o laudo ditado deve complementar o uso desses equipamentos e não substituí-los.

As análises realizadas com os formatos codificados de áudio mostra que o armazenamento de áudio sem compactação é inviável por dois motivos. Primeiro pela usabilidade, arquivos grandes demoram para ser carregados, principalmente se a conexão utilizada não for banda larga. E segundo pelo custo, os arquivos codificados ocupam entre 10 a 26 vezes menos espaço em disco. A escolha do formato de áudio deve levar em consideração a usabilidade para os usuários e a viabilidade do áudio em reconhecimento de voz.

As pesquisas em reconhecimento indicam um importante passo na otimização do processo de confecção do laudo, pois elimina a transcrição manual. Para a aplicação deste sistema, é necessário ter definida uma gramática e modelos acústicos na língua portuguesa para cada domínio. A baixa perplexidade existente na radiologia indica que este é o domínio mais indicado para uma primeira implantação de um sistema de reconhecimento de voz.

Os resultados obtidos com o HTK indicam que o uso de uma gramática contribui para a confiabilidade de acertos do sistema de reconhecimento de voz. Nesse ponto, o uso de DICOM Structured Report pode ser uma excelente alternativa, já que a forma não ambígua do laudo estruturado pode ser representado nas gramáticas.

A telemedicina permite que laudos médicos possam ser dados à distância. Através

---

dos conteúdo explorado nesse trabalho, é possível expandir o uso de gravação de laudo para dispositivos móveis como notebooks ou palmtops. Assim, o médico grava o laudo no palmtop e envia pela internet ao servidor para posterior reconhecimento de voz.

Embora o trabalho não tenha gerado ferramentas para uso imediato em hospitais e clínicas, as ferramentas e metodologias aqui propostas são o primeiro passo para uma implantação definitiva. Além disso, o módulo de manipulação de áudio independente do reconhecimento de voz permite a uma adaptação gradual do sistema de telemedicina com laudo ditado em formato digital até que o sistema SR esteja com maturidade suficiente para ser implantado.

## *Referências*

- ADAMI, A. G.; DORNELES, R. V. Revox - sistema de reconhecimento de voz aplicado ao controle industrial. *Simpósio de Ciência e Tecnologia*, v. 2, 1997.
- ALBUQUERQUE, P. [S.l.], 2006. Disponível em: <<http://www.mpc.com.br/audiespresso>>. Acesso em: 21 julho 2006.
- ALEXANDRINI, F.; BORTOLUZZI, M. K.; WANGENHEIM, A. von. Improving content based recovery on a radiological reports database. German Research Center for Artificial Intelligence DFKI GmbH, Kaiserslautern, Germany, p. 309–315, April 2005.
- ALSA. Disponível em: <<http://www.portaudio.com>>. Acesso em: 01 maio 2006.
- AMARAL, M. A.; BARRIVIERA, R.; TEIXEIRA, E. C. Reconhecimento de voz para automação residencial baseado em agentes inteligentes. *Revista Eletrônica de Sistemas de Informação*, n. 04, Novembro 2004.
- AMMOURA, A.; CARLACCI, F. Ogg vorbis and mp3 audio stream characterization. University of Alberta, Setembro 2002.
- ANDRADE, A. de O.; JR., C. G. P. *Reconhecimento de voz utilizando redes neurais*. Trabalho de conclusão de curso — Universidade Federal de Goiás, 1998.
- AUDACITY. *The Free, Cross-Platform Sound Editor*. Dominic Mazzoni, 2005. Disponível em: <<http://audacity.sourceforge.net>>. Acesso em: 10 novembro 2005.
- AZEVEDO-MARQUES, P. M. de; TRADD, C. S.; JUNIOR, J. E. Implantação de um mini-pacs (sistema de arquivamento e distribuição de imagens) em hospital universitário. *Radiol Bras*, v. 34, n. 4, p. 221–224, Julho-Agosto 2001. ISSN 0100-3984.
- BARRAS, C. et al. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, v. 33, n. 1-2, 2001.
- BASHSHUR, R. L.; SANDERS, J. H.; SHANNON, G. W. *Telemedicine Theory and Practice*. Illinois: Springer, 1999. 435pag p.
- BENCINA, R. Portaudio and media synchronisation. Australasian Computer Music Conference - ACMC2003, Perth - Australia, p. 13–20, Julho 2003.
- BENCINA, R.; BURK, P. Portaudio - an open source cross platform audio api. International Computer Music Conference - ICMC2001, Havan - Cuba, p. 263–266, Julho 2001.
- Michael Bevin*. Lossless Audio Home, 2004. Disponível em: <<http://www.lossless-audio.com>>. Acesso em: 3 abril 2006.

- BLADEENC. *Tord's Home - Bladeenc*. Tord Jansson, 2005. Disponível em: <<http://bladeenc.mp3.no>>. Acesso em: 10 novembro 2005.
- BOERSMA, P. Praat, a system for doing phonetics by computer. *Glott International*, v. 5:9/10, p. 341–345, 2001.
- BRITTO, A. de S. et al. Técnicas em processamento e análise de documentos manuscritos. *Edição Especial Computação Gráfica e Processamento de Imagens*, n. 2, p. 158, Outubro 2001.
- CENTRO DE CIÊNCIAS DAS IMAGENS E FÍSICA MÉDICA. *CCIFM - Projetos*. Disponível em: <<http://www.cci.fmrp.usp.br/projetos/radiologiadigital.html>>. Acesso em: 29 março 2006.
- CHARNIAK, E. *Statistical Language Learning*. Cambridge, Massachusetts: MIT Press, 1993.
- CINCOM. *Cincom Smalltalk*. Disponível em: <<http://smalltalk.cincom.com>>. Acesso em: 15 agosto 2006.
- CLUNIE, D. A. *DICOM Structured Reporting*. [S.l.]: PixelMed, 2000.
- CYCLOPS. *The Cyclops Project*. Florianópolis: Laboratório de Telemedicina - UFSC, 2006. Disponível em: <<http://cyclops.telemedicina.ufsc.br>>. Acesso em: 06 junho 2006.
- DAVID, L.; ZUCHERMAN, M.; JR., W. B. T. Six characteristics of effective structured report and inevitable integration with speech recognition. *Journal of Digital Imaging*, v. 0, n. 0, p. 1–7, 2005.
- DELLANI, P. *Desenvolvimento de um Servidor de Imagens Médicas Digitais no Padrão Dicom*. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, 2001.
- DICOM. *Digital Imaging and Communications in Medicine*. NEMA, Suite 1847 - 1300 North 17th Street - Rosslyn, VA - USA: American College of Radiology and National Electrical Manufacturers Association, 2006. Disponível em: <<http://medical.nema.org>>. Acesso em: 06 junho 2006.
- DOXYGEN. Disponível em: <<http://www.stack.nl/~dimitri/doxygen/>>. Acesso em: 3 abril 2006.
- EDICTATION INC. *eDictation*. Disponível em: <<http://www.edictation.com>>. Acesso em: 21 julho 2006.
- FLAC. Free Audio Codec, 2000. Disponível em: <<http://flac.sourceforge.net>>. Acesso em: 3 abril 2006.
- FRAUNHOFER GESSELLSCHAFT INSTITUTEN. *MP3: MPEG Audio Layer-3*. Disponível em: <<http://www.iis.fraunhofer.de/amm/techinf/layer3>>. Acesso em: 10 julho 2006.
- Formatos de áudio*. [S.l.]: Valle de la Pascua, 2004.
- HERRE, J. et al. An introduction to mp3 surround. *Fraunhofer Institute for Integrated Circuits IIS*, p. 9, 2005.

- ID3. *ID3v2*. Disponível em: <<http://www.id3.org/>>. Acesso em: 16 novembro 2005.
- INSTITUT TNT. *Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung*. Hanover: UNI Hannover, 2005. Disponível em: <<http://www.tnt.uni-hannover.de>>. Acesso em: 10 outubro 2005.
- JIM BAUMANN. *Voice Recognition*. The Encyclopedia of Virtual Environments, 1993. Disponível em: <<http://www.hitl.washington.edu/scivw/EVE>>. Acesso em: 29 março 2006.
- JSYN. *JSyn - Java Audio Synthesis*. Softsynth, 2005. Disponível em: <<http://www.softsynth.com/jsyn>>. Acesso em: 10 novembro 2005.
- CVoiceControl - command and control for Linux!* Kieczka CVoiceControl, 2002. Disponível em: <<http://www.kieczka.net/daniel/linux>>. Acesso em: 17 agosto 2006.
- LAME. *LAME Ain't an Mp3 Encoder*. OSTG Open Source Technology Group, 2005. Disponível em: <<http://lame.sourceforge.net>>. Acesso em: 10 novembro 2005.
- LANGER, S. Radiology speech recognition: workflow, integration and productivity issues. *Curr Probl Diagn Radiol*, v. 31, p. 95–104, Maio 2002.
- LEPANDTO, L. et al. Ogg vorbis and mp3 audio stream characterization. *Journal of Digital Imaging*, v. 0, n. 0, p. 1–6, 2005.
- LIRC. *Linux Infrared Remote Control*. Christoph Bartelmus, 2006. Disponível em: <<http://www.lirc.org>>. Acesso em: 10 julho 2006.
- O que é Telemedicina?* Universidade Federal de São Paulo, Escola Paulisa de Telemedicina, 2005. Disponível em: <<http://www.unifesp.br/dis/set/telemedicina.php>>. Acesso em: 15 agosto 2006.
- MACSYM. *Medical Systems*. MACSYS Tecnologia Médica, 1985. Disponível em: <<http://www.macsym.com.br>>. Acesso em: 18 novembro 2005.
- Processamento Digital da Fala*. Instituto Superior de Engenharia de Lisboa, Departamento de Engenharia de Electrónica e Telecomunicações e de Computadores, 2001. Disponível em: <<http://www.deetc.isel.ipl.pt/comunicacoesep/disciplinas/pdf>>. Acesso em: 17 agosto 2006.
- Automatic Speech Recognition*. MIT Open Course, 2003. Disponível em: <<http://ocw.mit.edu>>. Acesso em: 1 dezembro 2005.
- MONKEY. *Monkey's Audio*, 2004. Disponível em: <<http://www.lmonkeysaudio.com>>. Acesso em: 3 abril 2006.
- MOREAU, N. *HTK (v.3.1): Basic Tutorial*. [S.l.], Fevereiro 2002.
- NEMA. *Digital Imaging and Communications in Medicine*. Ps 3.1-2004. 1300 N. 17th Street Rosslyn, Virginia 22209 USA, 2004. Part 1: Introduction and Overview.
- Novas Tecnologias a Serviço dos Portadores de Necessidades Especiais*. Pontifícia Universidade Católica do Paraná, 2001. Disponível em: <<http://www.ppgia.pucpr.br/percy>>. Acesso em: 17 agosto 2006.

- NUANCE COMMUNICATIONS, INC. *Nuance*. Disponível em: <<http://www.nuance.com>>. Acesso em: 17 julho 2006.
- OPEN MIND SPEECH. *Open Mind Speech - Free Speech Recognition for Linux*. OSTG Open Source Technology Group, 2000. Disponível em: <<http://freespeech.sourceforge.net>>. Acesso em: 17 julho 2006.
- OPEN SOURCE TECHNOLOGY GROUP. *Sphinx-4*. Carnegie Mellon University, Sun Microsystems Inc., Mitsubishi Electric Research Laboratories, 2004. Disponível em: <<http://cmusphinx.sourceforge.net/sphinx4>>. Acesso em: 17 julho 2006.
- OSALP. OSTG Open Source Technology Group, 2002. Disponível em: <<http://osalp.sourceforge.net>>. Acesso em: 30 maio 2006.
- PHILIPS ELECTRONICS. *Philips SpeechMagic*. Ulrike Oswald, 2006. Disponível em: <<http://www.speechrecognition.philips.com>>. Acesso em: 09 junho 2005.
- PORTAUDIO. *PortAudio - portable cross-platform Audio API*. Disponível em: <<http://www.id3.org/>>. Acesso em: 16 fevereiro 2005.
- PRESSMAN, R. S. *Software Engineering: A Practitioner's Approach*. 4. ed. [S.l.]: McGraw-Hill, 1997.
- Reconhecimento de voz*. Universidade Federal de Santa Maria - Centro de Tecnologia, 1999. Disponível em: <<http://www-usr.inf.ufsm.br/~santos/diversos/voz>>. Acesso em: 06 julho 2006.
- SIMAS, T. *Voice recognition using a fuzzy multiple attribute model*. Dissertação (Mestrado) — The Soft Computing and Autonomous Agents, CRI/UNINOVA, Campus da FCT/UNL - PORTUGAL, 2002.
- SLS GROUP. *Spoken Language Systems*. Disponível em: <<http://groups.csail.mit.edu/sls>>. Acesso em: 17 julho 2006.
- STRUCTRAD LLC. *StructRad*. Disponível em: <<http://www.structrad.com>>. Acesso em: 21 julho 2006.
- ASR - Automatic Speech Recognition*. Mississippi State University, 2006. Disponível em: <<http://www.cavs.msstate.edu/hse/ies/projects/speech/software>>. Acesso em: 22 agosto 2006.
- TEBELSKIS, J. *Speech Recognition using Neural Networks*. Tese (Doutorado) — Carnegie Mellon University, 1995.
- THE CENTRE FOR SPEECH TECHNOLOGY RESEARCH. *The Festival Speech Synthesis System*. Disponível em: <<http://www.cstr.ed.ac.uk/projects/festival>>. Acesso em: 17 julho 2006.
- THOMSON CONSUMER ELECTRONICS. *Thomson Worldwide portal*. Disponível em: <<http://www.thomson.net>>. Acesso em: 11 outubro 2005.
- UMBRELLO. OSTG Open Source Technology Group, 2001. Disponível em: <<http://uml.sourceforge.net>>. Acesso em: 3 abril 2006.

- VALIN, J.-M.; MONTGOMERY, C. Improved noise weighting in celp coding of speech - applying the vorbis psychoacoustic model to speex. *Audio Engineering Society 120th Convention*, Maio 2006.
- WALLAUER, J. *Sistema de Gerencia de Exames via Web*. Trabalho de conclusão de curso — Universidade Federal de Santa Catarina, 2005.
- WHITE, K. S. Speech recognition implementation in radiology. *Springer-Verlag*, Maio 2005.
- WINAMP PORTAUDIO. *Winamp PortAudio v18.1 Output Plugin*. OSTG Open Source Technology Group, 2006. Disponível em: <<http://out-pa18-1.sourceforge.net>>. Acesso em: 10 fevereiro 2006.
- WXWIDGETS. wxWidgets Cross Platform GUI Library, 2006. Disponível em: <<http://www.wxwidgets.org/>>. Acesso em: 07 junho 2006.
- XING. *Xing Software*. Xingtech, 2005. Disponível em: <<http://www.xingtech.com>>. Acesso em: 10 novembro 2005.
- XIPH. *Xiph.org*. Disponível em: <<http://www.xiph.org>>. Acesso em: 3 dezembro 2005.
- XIPH. *Speex a free codec for free speech*. Disponível em: <<http://www.speex.org>>. Acesso em: 06 junho 2006.
- XVOICE. *XVoice: Linux Text To Speech Recognition and Integration*. Disponível em: <<http://www.zachary.com/s/xvoice>>. Acesso em: 17 julho 2006.
- YNOGUTI, C. A. *Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov*. Tese (Doutorado) — Universidade Estadual de Campinas, 1999.
- YOUNG, S. et al. *The Htk Book*. [S.l.]: Cambridge University Engineering Department, 2005.



## *ANEXO A – Configuração acústica*

Arquivo de configuração acústica para análise acústica com método MFCC(MOREAU, 2002):

```
#  
# Acoustical analysis configuration file  
#  
SOURCEFORMAT = HTK # Gives the format of the speech files  
TARGETKIND = MFCC_0_D_A # Identifier of the coefficients to use  
# Unit = 0.1 micro-second :  
WINDOWSIZE = 250000.0 # = 25 ms = length of a time frame  
TARGETRATE = 100000.0 # = 10 ms = frame periodicity  
NUMCEPS = 12 # Number of MFCC coeffs (here from c1 to c12)  
USEHAMMING = T # Use of Hamming function for windowing frames  
PREEMCOEF = 0.97 # Pre-emphasis coefficient  
NUMCHANS = 26 # Number of filterbank channels  
CEPLIFTER = 22 # Length of cepstral liftering  
# The End
```

## *ANEXO B – Arquivo HMM base*

Modelo Oculto de Markov com 6 estados onde o primeiro é de entrada(MOREAU, 2002):

```

~o <VecSize> 39 <MFCC_0_D_A>
~h "yes"
<BeginHMM>
<NumStates> 6
  <State> 2
    <Mean> 39
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
    <Variance> 39
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  <State> 3
    <Mean> 39
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
    <Variance> 39
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  <State> 4
    <Mean> 39
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

```

```
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 39
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
<State> 5
<Mean> 39
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 39
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
<TransP> 6
0.0 0.5 0.5 0.0 0.0 0.0
0.0 0.4 0.3 0.3 0.0 0.0
0.0 0.0 0.4 0.3 0.3 0.0
0.0 0.0 0.0 0.4 0.3 0.3
0.0 0.0 0.0 0.0 0.5 0.5
0.0 0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

## *APÊNDICE A – Documentação do Protótipo Cycreport*

O projeto pretende implementar um sistema de captura de áudio para auxiliar o processo de geração de laudos viabilizando a análise de sistemas de reconhecimento de voz. Além disso, é importante que o laudo ditado seja disponível dentro do padrão DICOM. O projeto deve prover as seguintes funcionalidades:

- Programa de captura (microfone) e codificação de áudio;
- Sistema web compatível com o Portal de Telemedicina(WALLAUER, 2005) com suporte ao laudo ditado;
- Armazenamento de áudio digital compatível com DICOM utilizando Waveform;
- Programa para digitação utilizando Portal de Telemedicina.
- Reconhecimento de voz para geração automatizada de laudo.

Cada ítem acima será implementado como um módulo com funcionalidades independentes que permitam maior reusabilidade. Ao final do trabalho espera-se o seguinte funcionamento do sistema com os módulos integrados:

- 1.O médico abre o programa e inicia a operação gravar laudo, em seguida dita o diagnóstico;
- 2.Quando o laudo está pronto, o médico envia o laudo para o portal de telemedicina e ao DCMServer com as imagens do exame;
- 3.O Portal de telemedicina notifica os digitadores que há um novo laudo para ser editado, estes iniciam o processo de reconhecimento de voz e em seguida fazem as correções necessárias atualizando o exame no portal;

4.O médico recebe a notificação que o exame está pronto e assina o laudo se o mesmo estiver correto;

## A.1 Levantamento de Requisitos

### A.1.1 Objetivo

Desenvolver um sistema de armazenamento de áudio para laudos médicos ditados. Aplicar sistemas de reconhecimento de voz a fim de obter o laudo em formato de texto.

### A.1.2 Domínio

O programa é focado na área médica, ou seja, tem como principal objetivo atender Hospitais e Clínicas. A meta é produzir um protótipo que permita analisar técnicas de reconhecimento de voz que envie os laudos em áudio e texto para uma base de dados acessível para consulta posterior.

### A.1.3 Requisitos Funcionais

#### A.1.3.1 O que o sistema deve permitir ao usuário

- Gravar através do microfone um laudo médico;
- Ouvir laudos gravados através do ítem acima;
- Iniciar o processo de reconhecimento de voz;
- Codificar o áudio em um formato que reduza o tamanho original do arquivo;
- Enviar laudos em áudio para portal de exames;
- Produzir um Waveform e enviar para o DCMServer;

#### A.1.3.2 Diagrama de classes e Casos de Uso

O desenvolvimento dos casos de uso foi feito utilizando templates HTML com o intuito em facilitar a publicação destes documentos. O projeto de digrama UML foi criado com a ferramenta Umbrello(UMBRELLO, 2001) e será também apresentada com a documentação gerada pelo programa Doxygen(DOXYGEN, 1997) que atualiza de forma mais simples o diagrama. Detalhes específicos da documentação estão em anexo.

### **A.1.3.3 Documentação**

Para documentação do código será usada a ferramenta Doxygen, um sistema de documentação sob licença GNU para C++, C, Java, Objective-C, Python e IDL com extensões para PHP, C# e D. A ferramenta gera toda API em HTML para documentação online e formatos RTF, PDF, Unix Man Pages para documentação off-line. A documentação é inserida no código tornando a API atualizada e consistente. Além disso o programa gera gráficos das relações entre classes, digrama UML e de colaboração.

## **A.1.4 Requisitos não funcionais**

### **A.1.4.1 Confiabilidade**

Laudos médicos devem ser gravados com boa qualidade e necessitam ser armazenados de forma segura para permitir o uso posterior do mesmo.

### **A.1.4.2 Desempenho**

É importante encontrar um algoritmo eficiente na geração dos laudos e rápido o suficiente evitando sobrecarga no servidor onde o sistema irá rodar.

### **A.1.4.3 Portabilidade**

Após a fase de protótipo o programa deverá ser portátil para rodar em sistemas operacionais Linux, Windows e Windows Mobile.

### **A.1.4.4 Usabilidade**

O programa tem como usuário final o médico. Este normalmente não utiliza programas de computador muito complexos, portanto, a usabilidade é um fator de grande importância em sistemas médicos.

## **A.1.5 Digrama de classes**

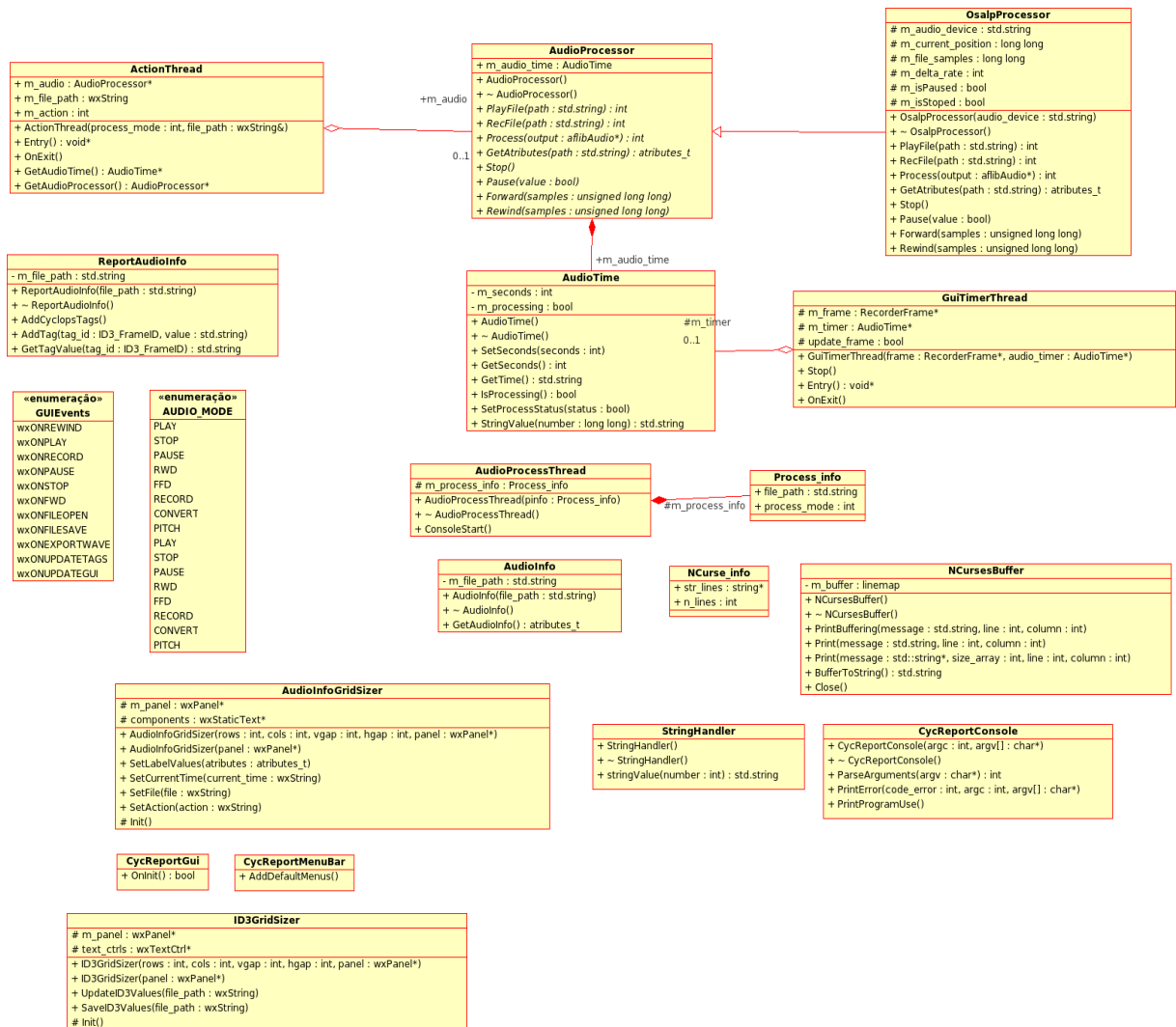


Figura 23: Diagrama UML simplificado

## *APÊNDICE B – Redução de informação em arquivos de áudio*

Embora os programas utilizados possam codificar e decodificar arquivos de som, uma vez codificado o arquivo nunca mais será o mesmo. Pode-se verificar a perda da informação seguindo os passos:

1. Gravar um discurso utilizando o microfone e salvar em arquivo;

```
cycreport -r discurso_orig.wav
```

2. Converter o WAV gravado em OGG ou MP3;

```
oggenc discurso_orig.wav -o discurso.ogg
```

3. Converter os arquivos OGG e MP3 novamente para WAV;

```
oggdec discurso.ogg -o discurso_mod.wav
```

4. Comparar o arquivo WAV original como novo arquivo decodificado;

```
diff discurso_orig.wav discurso_mod.wav
```

Utilizando o programa *diff* no linux entre os dois arquivos WAV, a resposta será que os arquivos são diferentes. Porém eles ao ouvido humano são o mesmo som e podem ter o mesmo tamanho em bytes.



## *APÊNDICE C – Artigo*

O artigo a seguir foi submetido para o CBIS 2006, X Congresso Brasileiro de Informática em Saúde, que recebeu 428 trabalhos para submissão e divulga os artigos aceitos a partir de 7 de agosto de 2006. O evento é organizado pela Sociedade Brasileira de Informática em Saúde.

## Sistema de Reconhecimento de voz na Radiologia com vocabulário restrito

Márcio Geovani Jasinski<sup>1</sup>, Rafael Andrade<sup>2</sup>, Rafael Simon Maia<sup>3</sup>, Aldo von Wangenheim<sup>4</sup>

<sup>1, 2, 3, 4</sup> Projeto Cyclops, Departamento de Informática e Estatística (INE),  
Universidade Federal de Santa Catarina (UFSC), Brasil.

**Resumo** - O processo de geração de laudos pode ser dinamizado com a aplicação de novas tecnologias em informática para a área médica. Inovações como a telemedicina tornaram acessíveis via web, recursos como imagens e laudos estruturados DICOM, ou, ainda, laudo áudio digital. A partir do áudio digital é possível realizar reconhecimento automático de voz, sucessão natural do método de ditado e transcrição. Este trabalho apresenta a proposta de uma metodologia de reconhecimento de voz modular. A independência dos módulos desenvolvidos permite que estes sejam integrados em diferentes tipos de aplicações. Integramos o sistema de gravação de áudio no cliente PACS para o envio de laudo ditado ao servidor PACS. Ao receber os dados, o servidor realiza a transcrição semi-automática do áudio digital para texto livre. Através da comparação do tempo gasto com o processo manual e com o sistema automatizado de laudo falado, conclui-se que um sistema eficiente de reconhecimento de voz reduz o tempo de emissão dos laudos em um sistema de telemedicina.

**Palavras-chave:** PACS, Telemedicina, DICOM, Reconhecimento de voz, Modelos Ocultos de Markov.

**Abstract** - Radiology report turnaround time can be optimized using new computer technologies on medical area. The telemedicine made access possible from web to DICOM images and structured report or on digital audio format. From digital audio report its possible do speech recognition that offers a natural succession from dictation-transcription process. This paper describes a methodology of modular speech recognition system where each module is independent and was developed to integrate many applications. We have integrated recording audio into PACS client to send digital audio report to PACS server. The server will execute semi automatic speech recognition to transcribe digital audio to free text. Comparing turnaround time between manual and automatized process we prove that the voice recognition system reduce report turnaround time under telemedicine system.

**Key-words:** PACS, Telemedicine, DICOM, Speech Recognition, Hidden Markov Models.

### Introdução

O atual método de geração de laudo radiológico, não explora vantagens dos avanços digitais e precisa ser adaptado aos programas médicos atuais e futuros. Dessa forma, além das taxas de erros serem altas, o tempo do processo não é otimizado. Laudos produzidos dessa forma dificultam análises subseqüentes e exigem mais cuidados na sua confirmação [3]

A geração de laudos médicos em clínicas e hospitais difere entre instituições desde a metodologia até a tecnologia disponível na instituição. Laudo em formulário manuscrito é utilizado com frequência e está sujeito a erros: rasura letra ilegível, troca de paciente entre exames e documento perdido. O processo é também lento para a geração do laudo. Além

disso, o armazenamento destes documentos torna inviável a recuperação para consulta; dado o volume de exames realizados a cada dia.

Uma alternativa ao processo manuscrito é laudo ditado em fitas e posterior transcrição. Embora a estratégia seja mais dinâmica (Figura 1), alguns entraves ainda existem, pois as fitas são regravadas o que elimina o registro para posterior consulta e deteriora a qualidade do áudio. Os nomes dos pacientes são marcados nas fitas por etiquetas, ocasionando problemas de letra ilegível e troca de paciente entre exames. A utilização desse procedimento demanda uma equipe de digitadores a ser mantida para a transcrição do laudo e o laudo precisa ser revisado pelo médico para corrigir erros de interpretação.

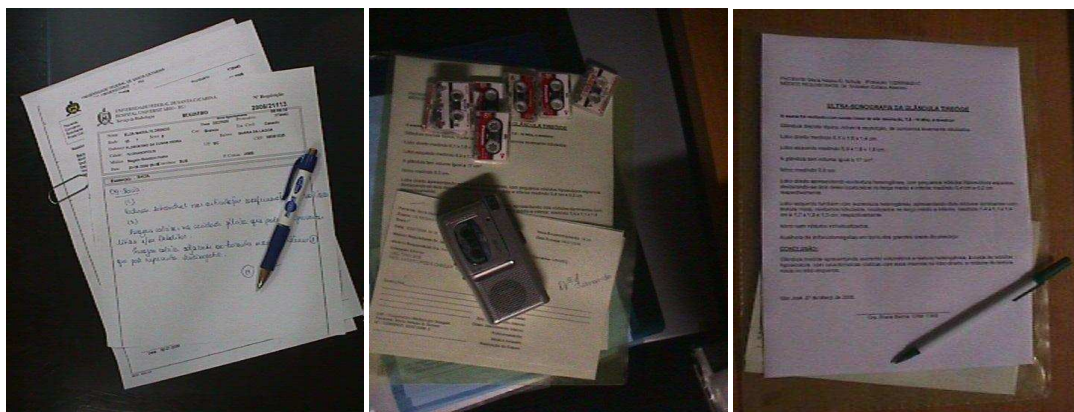


Figura 1: Laudo manuscrito (esq.), fitas usadas na em laudo ditado (cen.) e laudo impresso para confirmação (dir.)

O ambiente ideal deve utilizar os benefícios da tecnologia da informação para geração de laudo. O uso de PACS (*Picture Archiving and Communications System* – do inglês, Sistemas de arquivamento e comunicação de imagens) e gravação do laudo em áudio digital permitem corrigir deficiências nos métodos tradicionais de laudo. Dessa maneira, um áudio de alta qualidade com fácil acesso, possibilita consultas do laudo via *software/web site* de forma mais rápida e segura. Diminuindo assim a possibilidade de trocas de pacientes entre diferentes exames, pois o laudo ditado está sempre associado a exame e paciente corretos. Reduz também os problemas com documentos perdidos, rasurados, com letra ilegível e transcrição de ditados. [5]

No sistema ideal (Figura 2), o laudo é armazenado em áudio digital que é processado por um sistema de reconhecimento de voz para posterior revisão do médico.

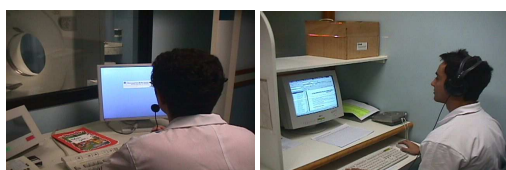


Figura 2: Laudo ditado durante o exame (dir.) e revisão do laudo após reconhecimento de voz automático (esq.)

As inovações digitais disponíveis: DICOM, Telemedicina, PACS/RIS, Áudio Digital, permitem automatização para a geração de laudos e maior segurança no armazenamento dados. Além disso, sistemas informatizados mantêm o histórico de atendimento do paciente. Essas vantagens refletem no tempo necessário edição do laudo e na rapidez em que são detectados reincidências através do histórico. A dinamização da geração

de laudos pode ser obtida com o uso de algumas tecnologias de informática médicas como:

- **DICOM<sup>1</sup> Structured Report (SR)** - Padrão DICOM para laudos estruturados. O DICOM SR define a formação, o armazenamento e a transferência de documentos estruturados que podem representar laudos, ou qualquer tipo de observação clínica. Estes documentos compreendem informações de contexto, tais como procedimentos que devem ser executados para o sucesso de um tratamento, e dados sobre profissionais de saúde envolvidos [1]
- **PACS** - Sistema para arquivamento e comunicação em diagnóstico por imagem. Permite acesso em qualquer setor de uma unidade hospitalar, de imagens médicas em formato digital [4]
- **Telemedicina** - Aplicações desenvolvidas e disseminadas com tecnologias que possibilitam a medicina acessível de qualquer lugar, reduzindo a frequência de deslocamentos de pacientes [2]
- **Áudio Digital** - Permite que os laudos sejam gravados e armazenados em bases de dados para digitação ou da utilização do processo de reconhecimento de voz. Além disso, o áudio do laudo, armazenado digitalmente, evita os problemas com fitas e aparelhos de gravação de laudos [5]
- **Reconhecimento de voz** - Processo de obter palavras faladas como entrada em um programa de computador e transformá-las em texto [6]

Estudos de caso mostram que a geração de laudos com reconhecimento de voz e PACS reduz o tempo médio para geração dos laudos. A utilização de linguagem natural permite que o médico dite o laudo no mesmo instante em que manipula o equipamento de imagens e analisa as

<sup>1</sup> DICOM - Digital Imaging and Communications in Medicine

informações do exame. Além disso, o uso de PACS elimina o tempo para manipulação e preparação de filmes. O resultado do uso dessas técnicas é a dinamização e melhor qualidade no processo de geração de laudos [4]

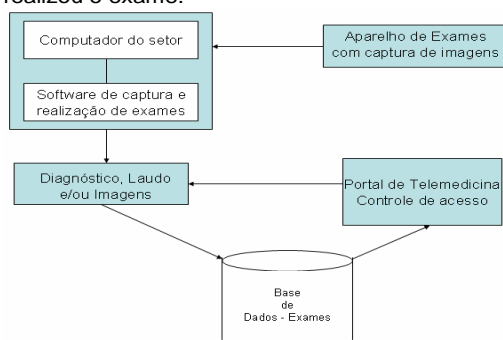
Relatos apresentados em [8] mostram que as novas tecnologias de laudo têm bons resultados quando efetivamente implantadas. O armazenamento de laudos em áudio digital e sistemas de reconhecimento de voz possibilitam melhor aceitação de médicos e redução no tempo de transcrição de laudos.

A metodologia proposta tem como objetivo possibilitar a geração de laudos médicos em áudio e utilizar reconhecimento de voz automático. O texto livre do laudo é gerado pelo servidor e disponível na *web* com a gravação do áudio para as correções necessárias.

### Metodologia

Em um projeto piloto implantado no Hospital Universitário da UFSC, foi desenvolvido um PACS que permite a captura de imagens e possibilita a publicação de exames na rede de telemedicina do hospital. A estrutura de funcionamento deste PACS é apresentada na Figura 3.

O PACS é usado em cinco setores do hospital: e cada setor possui um equipamento, conectado ao aparelho de exames, que é responsável pela captura das imagens. Com as imagens do exame no sistema, o médico envia o exame para a rede de telemedicina. O exame é laudado via *web site*, onde imagens e informações ficam disponíveis para o médico que realizou o exame.



**Figura 3:** PACS implantado no HU-UFSC

Foi observada uma resistência grande dos médicos ao digitarem eles mesmos os laudos no PACS/Web. Alguns setores optaram por continuar a emitir laudos manuscritos. A possibilidade de oferta de um sistema de laudo ditado irá reduzir a resistência à utilização de um sistema totalmente computadorizado

Foram contabilizados os exames de dois setores, um de baixa e outro de alta demanda no período de março a maio de 2006. No setor de baixa movimentação, o PACS é utilizado para exame e o laudo editado via portal de telemedicina, onde são feitas alterações no modelo pré-definido. Nos casos mais simples, o paciente sai da sala de exames com o laudo do exame. No mesmo período, os exames realizados no setor de maior quantidade de atendimentos, foram atingidos a média de 21 exames por dia. Esse volume de exames fez com que os médicos utilizassem o PACS somente para impressão das imagens capturadas. Mesmo com modelos de laudo previamente definidos no sistema, o tempo necessário para a digitação dos laudos foi maior que a emissão de laudos manuais.

Nesse contexto, a tecnologia de reconhecimento de voz pode dinamizar o processo de geração de laudos, reduzir o tempo necessário para geração destes laudos e aumentar a confiabilidade do processo. Um estudo de caso [5] mostra que gerar laudos e corrigi-los utilizando o processo de reconhecimento de voz, é 40% mais rápido que o método de digitação e transcrição com gravadores de fitas. Porém, a implantação de um sistema para emissão de laudo via reconhecimento de voz, deve ser feito respeitando o processo de trabalho de cada setor, caso contrário, produzirá um efeito contrário do esperado, aumentando o tempo para geração de laudos [7].

O desenvolvimento do trabalho foi feito com tecnologias livres e de código aberto. O sistema operacional utilizado foi Linux, porém os módulos usados são portáveis para Microsoft Windows.

Para a gravação e reprodução dos laudos, a solução utilizada foi a biblioteca PortAudio. Escolhida, pois oferece a solução mais flexível em diferentes sistemas operacionais na manipulação de entrada e saída de áudio. A biblioteca foi projetada para facilitar o desenvolvimento de softwares de síntese e reprodução de áudio em diversas plataformas [9]

O HTK foi a solução livre e open-source para definição do vocabulário treinamento e reconhecimento de voz. A biblioteca HTK é um conjunto de ferramentas desenvolvidas para manipulação de Modelos Ocultos de Markov<sup>2</sup> (HMM). O projeto foi inicialmente usado em pesquisas de reconhecimento de voz, porém já utilizado em síntese de voz e reconhecimento de caracteres. O HTK possui um conjunto de módulos que facilitam a análise da fala, treinamento dos HMMs, testes e análise de resultados [10]

A proposta de desenvolvimento de uma metodologia modular para a geração de laudos

<sup>2</sup> HMM – Hidden Markov Model, Modelo Oculto de Markov.

médicos utilizando reconhecimento de voz considera soluções não comerciais. Existem bibliotecas *open-source* que oferecem uma solução para cada parte de um sistema de reconhecimento de voz. A solução descrita a seguir é o uso de reconhecimento de voz para processamento posterior, ou seja, não serão consideradas as restrições de tempo real.

A entrada de áudio é realizada através de um microfone e é controlada por uma interface gráfica com o usuário. O fluxo de áudio é manipulado com o PortAudio e o resultado é convertido no formato WAV. Este arquivo é enviado para o sistema de reconhecimento de voz para processamento do áudio e geração do laudo em formato texto. O arquivo obtido deve ser convertido no formato WAV para ser processado pela biblioteca de reconhecimento de voz

O vocabulário médico é bem definido, formal e com perplexidade<sup>3</sup> baixa [11] como mostra a Tabela 1. Por isso, o uso de vocabulário restrito para a Radiologia é a proposta de uso do sistema de reconhecimento de voz..

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

**Tabela 1:** Perplexidades típicas em diferentes domínios [11].

O sistema proposto para reconhecimento de voz é constituído de [13] , [10]

- **Definição de Gramática e do Dicionário** - Regras usadas no reconhecimento. No dicionário são definidas as palavras aceitas pelo sistema dentro contexto no qual a aplicação será utilizada, neste caso, da Radiologia;
- **Definição do Modelo Acústico** – Caracterização da forma como os sons das palavras devem ser representados;
- **Definição do corpo de treinamento** – Frases que utilizam palavras do dicionário para realização do treinamento do sistema de reconhecimento;
- **Definição dos modelos HMMS** (Hidden Markov Models) – Arquivos que definem o modelo e a forma de transcrição para cada palavra do dicionário;
- **Configurações de codificação** – Etapa onde os algoritmos de codificação e parâmetros de reconhecimento e treinamento são definidos.
- **Treinamento** - Processo usado para balancear os valores dos HMMS a fim de

<sup>3</sup> Em [11] , a definição de perplexidade é dada como o número médio de palavras possíveis depois que o modelo de linguagem foi aplicado.

identificar a entrada do usuário. A grande vantagem em utilizar os HMMS é a presença de algoritmos eficientes para o treinamento e o reconhecimento. HMM possibilita também a integração de vários tipos de conhecimento (sintático, lingüístico) em um modelo único.

- **Reconhecimento** – Etapa onde o sistema é utilizado para gravação de ditado em áudio digital. O resultado da gravação é processado pelo módulo de reconhecimento de voz que retorna a transcrição em texto livre da entrada em áudio;
- **Avaliações** – Análise dos resultados de acerto do texto gerado pelo reconhecimento de voz em relação ao texto ideal.

## Resultados

Empresas de tecnologia na área médica investem no de reconhecimento de voz em sistemas de informação na saúde. A Philips[16] e a MacSym[5] possuem programas de reconhecimento na área médica. No entanto, as soluções existentes são comerciais e não tem integração com a telemedicina ou com um PACS. Esta proposta considera a integração com os sistemas de telemedicina, PACS, gravação de laudo digital e reconhecimento de voz.

A baixa perplexidade observada no vocabulário médico radiológico é um fator relevante para a realização de testes com o protótipo de reconhecimento de voz neste setor.

A Figura 4, apresenta o uso do sistema de gravação e reprodução do laudo médico. O exame realizado no PACS (Figura 3) é enviado ao banco de dados e acessado via *web*. O laudo em áudio digital pode ser gravado durante o exame ou após o envio do mesmo para o banco de dados, com acesso pelo portal de telemedicina. Após o envio do laudo em áudio, o servidor inicia o processo de reconhecimento de voz e conversão para texto. Realizada a transcrição, o laudo temporário fica disponibilizado, com acesso restrito ao locutor do laudo para correções e aprovação final. Somente com a confirmação do médico é que o laudo é publicado no prontuário do paciente.

Utilizando um dicionário de palavras reduzido, foram realizados testes preliminares utilizando a metodologia apresentada neste trabalho. O experimento consistiu na gravação de um conjunto de laudos simplificados e os seus envios a um servidor de reconhecimento de voz, onde foi realizada a conversão do áudio digital para texto. O processo de ditado e transcrição por reconhecimento automático obteve bons acertos na conversão do áudio para texto livre, sendo viável com correção do texto gerado automaticamente.

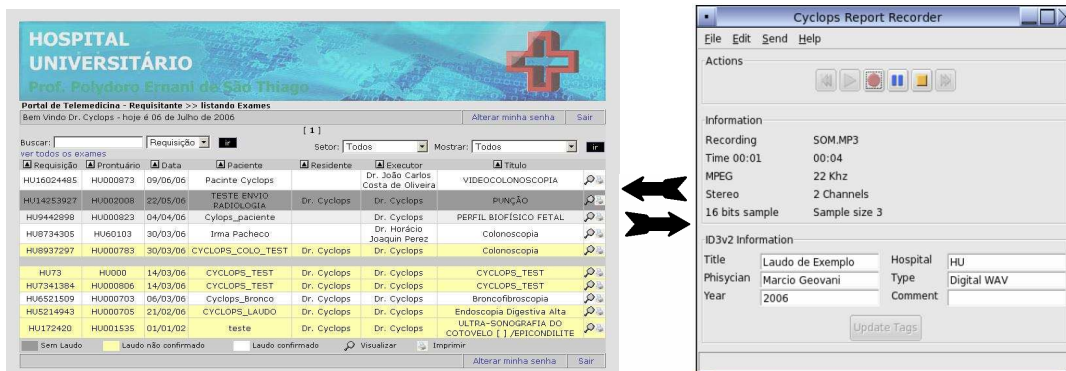


Figura 4: Informações do exame: imagens médicas obtidas com PACS e o laudo ditado em formato digital publicados no sistema de telemedicina para processamento de reconhecimento de voz no servidor.

necessitaria realizar uma verificação e confirmar o laudo.

**Discussão e Conclusões**

A metodologia proposta é uma solução para problemas enfrentados pelo Projeto Cyclops, onde o uso do sistema de reconhecimento de voz reduziu a resistência dos médicos em utilizar sistemas médicos. Problemas encontrados no processo de ditado em fitas e transcrição também são solucionados. Com um sistema de PACS e de reconhecimento de voz, é possível manter todos os laudos ditados em uma base de dados. Isto aumenta a qualidade dos documentos e evita erros com etiquetas em papel.

Os resultados obtidos com a realização dos testes e posterior contato com os especialistas que participaram do processo, indicam que a integração deste sistema terá a aceitação dos médicos para a realização de ditado de laudos médicos.

Uma gramática na língua portuguesa e os modelos acústicos precisam ser definidos para que o uso de um sistema de reconhecimento de voz seja passível de ser utilizado nos Hospitais Brasileiros. O avanço da pesquisa em reconhecimento de voz nos EUA e na Europa tiveram um grande crescimento após a criação das bases centralizadas [11]. Uma forma de se de obter bons resultados é definir uma base de reconhecimento de voz para o português onde a contribuição de diferentes instituições seja convergida.

A telemedicina pode ser expandida de modo que laudos médicos possam ser dados à distância, através de aparelhos móveis como notebooks ou Palmtops (computadores de mão) conectados a internet. Dessa forma o médico grava o laudo e envia para o servidor executar o reconhecimento de voz. Porém, quando o servidor terminar este processamento, o médico

Outra área onde a pesquisa em reconhecimento de voz pode ser expandida é no uso desses sistemas no padrão DICOM de laudos estruturados (DICOM SR). A forma mais organizada do conteúdo dos laudos estruturados facilita a manipulação dos dados e o reconhecimento de voz oferece uma evolução natural dos processos de ditado e transcrição [3].

Formas de armazenamento dos dados precisam ser consideradas. Devido ao formato sem compactação, arquivos em WAV ocupam muito espaço nas bases de dados. Uma solução seria codificar o áudio em formatos diferentes como MP3 ou OGG. Em [15] são feitas análises sobre o uso desses formatos e a perda de informação inerente na codificação. A pesquisa nessa área pode ser feita, também, com formatos específicos para voz, como o Speex [13] que mantém a qualidade da voz após a codificação.

**Referências**

[1] DICOM Official Website. [http://medical.nema.org] 10 Julho 2006.

[2] DELLANI, P., *Desenvolvimento de um Servidor de Imagens Médicas Digitais no Padrão Dicom*, Dissertação de Mestrado - Universidade Federal de Santa Catarina – 2001.

[3] DAVID, L.; ZUCHERMAN, M.; JR., W. B. T., “Six characteristics of effective structured report and inevitable integration with speech recognition”, *Journal of Digital Imaging*, v. 0, n. 0, p. 1–7, 2005.

[4] CCIFM - Centro de Ciências das Imagens e Física Médica - Projetos. [http://www.cci.fmrp.usp.br/projetos/radiologia

- digital.html] 29 Março 2006.
- [5] MACSYM. Medical Systems. MACSYM Tecnologia Médica, 1985. [http://www.macsym.com.br] Novembro 2005.
- [6] BAUMANN, J., “Voice Recognition Human Interface Technology Laboratory”, *The Encyclopedia of Virtual Environments* [http://www.hitl.washington.edu/scivw/EVE] 07 Junho 2006
- [7] LEPANDTO, L., PARÉ G., AUBRY D., ROBILLARD P., “Impact of PACS on Dictation Turnaround Time and Productivity”. *Journal of Digital Imaging*, v. 0, n. 0, p. 1–6, 2005.
- [8] WALLAUER, J. *Sistema de Gerencia de Exames via Web*, Trabalho de conclusão de Curso, Universidade Federal de Santa Catarina, 2005.
- [9] PortAudio, Portable cross-platform Audio API. Disponível em: [http://www.id3.org] 07 Junho 2006
- [10] HTK Hidden Markov Model Toolkit, 2006. Disponível em: [http://htk.eng.cam.ac.uk]. 07 Junho 2006.
- [11] YNOGUTI, C. A., *Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov*, Tese de Doutorado, Universidade Estadual de Campinas, 1999.
- [12] WXWIDGETS Cross Platform GUI Library. wxWidgets, 2006. [http://www.wxwidgets.org/] 07 junho 2006.
- [13] Speex a free codec for free speech, 2006. [http://www.speex.org] 07 Junho 2006
- [14] BREGA, J. R. F. ; SEMENTILLE, A. C. ; RODELLO, I. A. ; MELO, W. C. C. . “Uma Interface de Reconhecimento de Voz para Movimentação de Agentes e Avatares Humanóides em Ambientes Virtuais” . In: SVR 2003 - VI SYMPOSIUM ON VIRTUAL REALITY, 2003, Ribeirão Preto. Proceedings of SVR 2003 - VI Symposium on Virtual Reality, 2003. v. 1. p. 419-419.
- [15] SON, R. J. J. H. von, “Can standard analysis tools be used on decompressed speech?”, Institute of Phonetic Sciences/ACLCL University of Amsterdam, 2002
- [16] Philips Speech Processing. [http://www.speech.philips.com] 06 Julho 2006

#### Contato

##### Márcio Geovani Jasinski

Laboratório de Telemedicina – HU/UFSC  
Email: [marciogj@inf.ufsc.br](mailto:marciogj@inf.ufsc.br)  
Fone: (048) 3331-9166

##### Rafael Andrade

Laboratório de Telemedicina – HU/UFSC  
Email: [andrade@telemedicina.ufsc.br](mailto:andrade@telemedicina.ufsc.br)  
Fone: (048) 3331-9166

##### Rafael Simon Maia

Laboratório de Telemedicina – HU/UFSC  
Email: [simon@inf.ufsc.br](mailto:simon@inf.ufsc.br)  
Fone: (048) 3331-9166

##### Aldo Von Wangenheim

Depto. de Informática e Estatística da UFSC.  
Email: [awangenh@inf.ufsc.br](mailto:awangenh@inf.ufsc.br)  
Fone: (048) 3331-9516

## *APÊNDICE D - Cabeçalho formato WAV*

```
typedef struct {  
    unsigned short format_tag;  
    unsigned short channels;  
    unsigned long samplerate;  
    unsigned long bytes_per_second;  
    unsigned short blockalign;  
    unsigned short bits_per_sample;  
} WaveChunk;
```