UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

INVESTIGAÇÃO DO RECONHECIMENTO DE FALA BASEADO EM UM AMBIENTE TELEFÔNICO

Cleunio Bezerra de França Filho

Recife, Dezembro de 2010

Universidade Federal de Pernambuco

Centro de Informática

Cleunio Bezerra de França de Filho

INVESTIGAÇÃO DO RECONHECIMENTO DE FALA BASEADO EM UM AMBIENTE TELEFÔNICO

Trabalho de Graduação sob o título “Investigação do reconhecimento de fala baseado em um ambiente telefônico”, defendida por Cleunio Bezerra de França Filho e aprovada em 21 de Dezembro de 2010, em Recife, Pernambuco, pela banca examinadora constituída pelos doutores:

Prof. Dr. Edson de Barros Carvalho

Centro de Informática - UFPE

Orientador

Prof. Dr. Germano Crispim Vasconcelos

Centro de Informática – UFPE

AGRADECIMENTOS

Dedico meus sinceros agradecimentos para:

– Deus, por tudo;

– Dona Rouuuse, minha mãe, por ter me educado, por ter me acompanhado durante minha infância nos médicos e passar noites acordadas comigo e por todas as lutas que só a mulher com a sua força consegue superar;

- ao meu pai, por ter ajudado a concluir meus estudos;

- as minhas irmãs, Priscilla e Pollyana, por estarem presente durante toda a minha vida e por me dá forças;

- a Bela, minha querida namorada, por ter acompanhado minha trajetória e ter que aturar minhas viradas de noite no CIN;

– Prof. Dr. Edson de Barros Carvalho, meu orientador nesta monografia. Obrigado por ter me mostrado a área de Inteligência Artificial e acreditar no meu trabalho;

– Profª Drª Marcília Andrade Campos, por ter sido minha “mãe” durante o curso e por me orientar durante o período de monitoria;

- ao Alexandre, pelas pesquisas e o acompanhamento direto nos projetos, nas escritas de artigos e no compartilhamento de idéias de novos projetos

- ao Weber, por ter iniciado a minha co-orientação, ao Carlos (Sambarilove) e ao Daniel (Poste), por ter me ajudado a tirar os bugs do projeto da monografia;

- ao pessoal do meu projeto: Profª Drª Magdala, Claudinalle e Marília, por fazer parte do meu crescimento profissional, assim como, todo o pessoal do Nutes e da Tissue Bond;

- aos amigos do Colégio Damas e do Bosque das Sequóias. Estes por me acompanhar durante toda minha vida e vivenciar momentos de alegrias e tristezas;

- aos amigos do CIN, em especial, ao pessoal da entrada 2006.1 e do grupo eng-da-computacao-20061-ufpe@googlegroups.com, por ter virado noites, compartilhados finais de semana e feriados;

- as pessoas que ajudaram a construir a base de dados: o pessoal do Nectar, Eduardo e Patrícia Franklin, Edgar, Mari, Rodrigo, Lucas, o Preto, Márcio, Bola,

Estevão, Fábio, Morato, Fernando, Gabriel, Helder, Jonathan, Leyla, Petrônio, Renan;

- Enfim, a minha família e a todos que compartilharam comigo mais esta conquista.

"Recife
Não a Veneza americana
Não a Mauritsstad dos armadores das Índias Ocidentais
Não o Recife dos Mascates
Nem mesmo o Recife que aprendi a amar depois
- Recife das revoluções libertárias"

Trecho do poema Evocação do Recife

Manuel Bandeira (Escritor Pernambucano,1886 – 1968)

Resumo

O rápido crescimento das aplicações na área de reconhecimento de voz tem proporcionado a disseminação de uma grande quantidade de produtos que passam a fazer parte do dia a dia da população. Como exemplo, podemos citar os Call Centers. Nos dias atuais estes sistemas são automáticos, diminuindo o tempo de espera e, conseqüentemente, a satisfação do usuário. Este trabalho tem por finalidade promover a integração de um servidor de voz sobre IP para desenvolvimento de uma aplicação de atendimento telefônico utilizando recursos de reconhecimento de fala. Veremos, inicialmente, as técnicas de reconhecimento de voz. Concomitantemente a esta etapa, uma base de dados será criada. Esta base usará o telefone como receptor e gravador da voz. Em seguida, será realizada a implementação do sistema e, por fim, os resultados serão refinados e otimizados e novas sugestões para pesquisa futuras serão apresentadas. Os resultados foram eficazes com taxas de 93%.

Palavras-chave: Asterisk; FIVE; Reconhecimeto de Fala; Reconhecimento de voz

Abstract

The rapid growth of applications in the area of speech recognition has provided to disseminate a large amount of products that become part of everyday life of the population. For example, we can cite the Call Centers. Nowadays these systems are automatic, that aim to decrease the waiting time and consequently the user satisfaction. This work aims to promote the integration of a voice over IP Server for application development the telephone attendance using resources speech recognition. Initially, the techniques of speech recognition will be presented. Concomitantly with this phase, a database will be created. This database will used the phone as a receiver and voice recorder. Then will be realized the implementation of the system and the results will be refined and optimized and new suggestions for future research will be presented. The results were effective with sucess rates of 93%.

Keywords: Asterisk; FIVE; Speech Recognition; Voice Recognition

Sumário

Glossário

ASR - Automatic Speech Recognition

ASV - Automatic Speaker Verification engine

API - Application Programming Interface

APP - Application

DSP - Digital Signal Processing

FXO - Foreign eXchange Office

Foreign eXchange Subscriber (FXS)

FIVE - Framework for an Integrated Voice Environment

FGMM - Fuzzy Gaussian Mixture Models

GMM - Gaussian Mixture Models

HMM – Hidden Markov Models

LPC - Linear Predictive Coding

LPCC - Linear Predictive Coding Coefficient

MFCC – Mel-Frequency Cepstral Coefficients

NLP - Natural Language Processing

PABX – Private Automatic Branch Exchange

SIP – Session Internet Protocol

TTS - Text-To-Speech

URL - Uniform Resource Locator

VAD - Voice Activity Detection

VoIP – Voice over Internet Protocol

Lista de Figuras

Lista de Tabelas

Capítulo 1

INTRODUÇÃO

Na computação as mudanças e evoluções têm que ser rápidas e constantes para poder satisfazer as necessidades do mercado que surgem a cada momento. Isso acontece, também, na área de reconhecimento de padrões. Essa área tem sido cada vez mais explorada e pesquisada. Essas pesquisas originaram diversas áreas como: reconhecimento de voz, reconhecimento de faces, reconhecimento de escrita, entre outros.

1.1 Motivação

O crescimento da quantidade de operações dos sistemas correspondentes ao processamento de voz e de suas inúmeras aplicações atuais deu origem a vários sistemas, como por exemplo, os Call Centers. Antigamente, estes sistemas eram gerenciados manualmente, o que provocava demora nos atendimentos e, conseqüentemente, insatisfação dos clientes.

Gans, Koole e Mandelbaum (2003) definem Call Centers como um conjunto de recursos que permitem a prestação de serviços através do telefone. O ambiente de trabalho de um grande Call Centers pode ser visualizado como um espaço sem fim, no qual as pessoas sentam-se na frente do computador, proporcionando, assim, teleserviços para clientes invisíveis.

Os Call Centers atuais têm várias vantagens no atendimento às necessidades dos clientes. Entre elas cita-se: a utilização de recursos finitos (atendentes), que conseguem maximizar os resultados de performance (diminuição de filas de espera...) e redução de custos; uso inteligente das árvores de atendimento, que visa diminuir o tempo das chamadas aumentando a rapidez do sistema; e a possibilidade de paralelização das chamadas aumentando a eficiência.

Com a evolução da Internet e da tecnologia VoIP (Voice over Internet Protocol), os sistemas, cada vez mais, tem usado a rede que conecta os computadores mundiais, ou seja, a rede que conecta milhões de computadores em todo o mundo (Kurose, Ross, 2003).

Diante disso, os Call Centers tem utilizado esta rede para baratear, ainda mais, seu uso e permitir a pesquisa em novas áreas como, por exemplo, o reconhecimento de voz.

1.2 Objetivos

O objetivo geral deste trabalho é promover a integração de um servidor de voz sobre IP para desenvolvimento de uma aplicação de atendimento telefônico utilizando recursos de reconhecimento de fala.

Isto irá provocar um amadurecimento dos sistemas telefônicos, quanto aos aspectos de diversificação de suas aplicações e integração com um sistema de reconhecimento de voz.

Para isto, os seguintes passos serão necessários:

- Estudo da teoria e das diversas técnicas existentes no processo de reconhecimento de voz;
- Estudo do Asterisk (Meggelen, Madsen, 2007) (Spencer, 2003) (tecnologia VOIP) e do studo do FIVE (Framework for an Integrated Voice Environment) (Maciel, Carvalho, 2010);
- Uma base de dados será criada. Esta base usará o telefone como receptor;
- Uma implementação do sistema será desenvolvida;
- Os resultados serão refinados e otimizados.

1.3 Organização

Este documento está dividido em seis capítulos. No capítulo 2, será apresentada uma introdução sobre a história do reconhecimento de voz; serão apresentadas as definições de Reconhecimento de Fala, Reconhecimento de Locutor e Síntese de Fala. Neste capítulo, ainda, na seção de Reconhecimento de locutor, será apresentada a diferença entre Verificação de Locutor e Identificação de Locutor. No capítulo 3, será apresentado um resumo sobre VoIP, assim como, será dada uma explanação sobre o Asterisk (arquitetura, protocolos suportados, entre outros), que utiliza a tecnologia VoIP. O capítulo 4 é dedicado ao estudo do FIVE. São descritos sua arquitetura, alguns algoritmos utilizados. O capítulo 5 será destinado aos experimentos. Os resultados da implementação serão apresentados neste capítulo, bem como a base de dados, o critério utilizado na extração de características e classificação dessa base. E por último, um capítulo com a conclusão do que foi apresentado em todo o documento e possíveis trabalhos futuros.

Capítulo 2

Tecnologia de Voz

Neste capítulo mostraremos um breve histórico do desenvolvimento da área de Reconhecimento de Voz, e apresentaremos seus conceitos básicos e áreas de pesquisa.

2.1 Histórico

Desde a segunda metade do século XVIII, tentativas de desenvolver uma máquina capaz de se comunicar com o ser humano através da voz têm sido realizadas (Juang, Rabiner, 2004). O primeiro interesse não foi reconhecer ou entender a voz do ser humano, mas sim criar uma máquina capaz de falar, provavelmente pelo conhecimento existente na época sobre a acústica de tubos ressonantes.

Na primeira metade do século XX, cientistas do Laboratório Bell observaram o relacionamento existente entre o espectro de voz, e as características do som que são percebidas pelo ouvinte. Devido a esse trabalho, a maioria dos sistemas modernos de reconhecimento de voz é baseada na medição da potência do espectro de voz (Juang, Rabiner, 2004).

Nos anos 50 surgiu o primeiro algoritmo de classificação. Este algoritmo foi proposto por Fix e Hodges (1951). Este algoritmo é o k-NN ou k Nearest Neighbours ou, em português, k vizinhos mais próximos. Mais somente com os resultados de (Aha, 1992) que esse algoritmo ficou conhecido como método de classificação nas áreas de aprendizado de máquina e mineração de dados (Borsato, Merschmann, Plastino, 2007). A idéia é associar a instância, que deseja classificar, a maior ocorrência dos k mais próximos (vizinhos).

No final dos anos 60, Atal e Itakura formularam independentemente o conceito do “Codificador de Predição Linear” (LPC), sigla em inglês (Juang,

Rabiner, 2004), que simplificou bastante a estimativa dos parâmetros de voz e as técnicas para análise de voz. A partir desta técnica geram os Coeficientes de Predição Linear, ou ainda, Linear Predictive Coding Coefficient (LPCC).

Em meados de 1970, a idéia de usar conceitos de reconhecimento de padrões em reconhecimento de voz, baseados em métodos LPC foram propostos. Aproximadamente, nesse mesmo período, a primeira companhia de reconhecimento de voz foi fundada Threshold Technology (Juang, Rabiner, 2004). O LPC é um extrator de característica da voz. Ele extrai as características que se deseja classificar. Em seguida surgiu o Mel-Frequency Cepstral Coefficient (MFCC) (Maciel, Carvalho, 2010) (Tran, Wagner, 1999). Este é o mais usado hoje em dia. Tais métodos são colocados na escala *mel e geram* os coeficientes *cepstral.* O LPCC (Juang, Rabiner, 2004) (Tran, Wagner, 1999) tem a vantagem de ser mais simples e, conseqüente, precisa de menos processamento. Porém, com a evolução do processamento dos dados o MFCC tornou mais utilizado (Mashao, Skosan, 2006).

Um grande Avanço nas pesquisas ocorreu em meados dos anos 80, quando surgiu a idéia de usar “Modelos de Markov Escondidos” (HMM) para o reconhecimento de voz (Juang, Rabiner, 2004). A popularidade de HMM permaneceu pelas duas últimas décadas. Outra tecnologia introduzida nos anos 80, e que também se tornou muito popular nos sistemas de reconhecimento de voz, foi a utilização de redes neurais artificiais.

Nos anos 90 e na 1º década do século XXI surgiram outras técnicas de reconhecimento de voz como, por exemplo, o GMM (Gaussian Mixture Models) e o FGMM (Fuzzy Gaussian Mixture Models). Estes algoritmos podem ser estudados mais em (Reynolds, Rose, 1995) e (Tran, Wagner, 1999).

2.2 Áreas de Pesquisa

Como foi dito anteriormente, a necessidade do uso de reconhecimento de padrões na identificação do usuário, de uma forma geral, tem aumentado as

áreas de pesquisas. Nesta seção, veremos três ramos de processamento de voz (França, 2007): reconhecimento de locutor, síntese de fala e reconhecimento da fala.

2.2.1 Reconhecimento de Locutor

O Reconhecimento de Locutor é divido em duas áreas (Reynolds, Rose, 1995): verificação de locutor e identificação de locutor. Nesta subseção serão apresentadas estas duas subáreas.

2.2.1.1 Verificação de Locutor

A verificação de locutor é feito a autenticação de uma pessoa, ou seja, faz uma verificação para ter certeza que o locutor é mesmo quem diz ser, apenas com uso da voz. Para isso, a resposta do sistema é binária e o indivíduo tem que fornecer um sinal de voz, para a entrada, e outros sinais de voz, gravados previamente num banco de dados, para servirem de comparação e caracterização da base de dados. O indivíduo, também, terá que fornecer uma forma prévia de identificação, como uma senha, para que o sistema não compare com todas as vozes do banco de dados, mas com apenas a voz necessária. Isso favorece o rápido processamento da comparação e, consequentemente, um rápido tempo de resposta (Campbell, 1997).

A arquitetura da verificação do locutor é feita da seguinte forma: Primeiramente, são gravadas e armazenadas as locuções de cada usuário. Na maioria das vezes essas locuções são feitas offline, ou seja, o sistema não está atuando em tempo real. Em seguida, o sistema de treinamento é desenvolvido a partir de técnicas que serão apresentadas no próximo tópico. E por último, é feito a classificação e a verificação da autenticidade da elocução de entrada

(Carvalho, Santos, 2008). A figura abaixo representa o texto apresentado anteriormente.

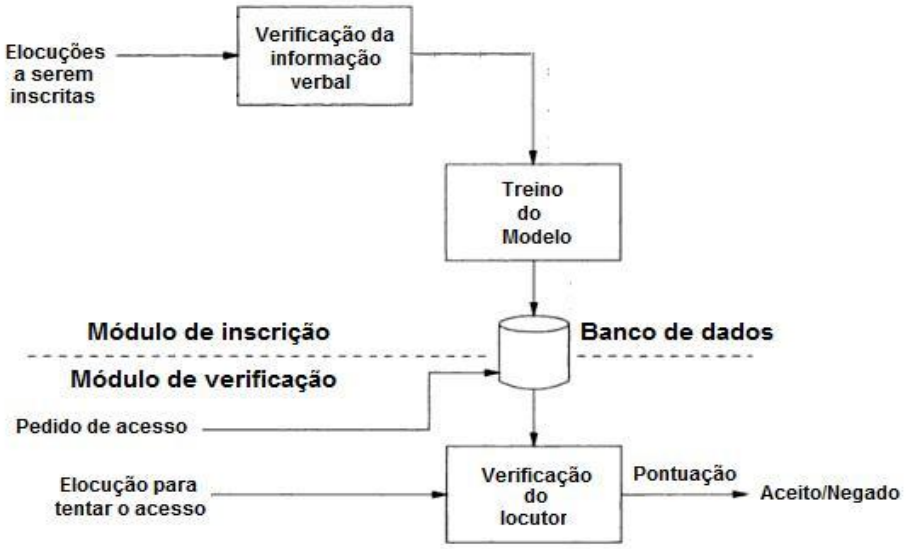

Figura 1 - Esquemático da Verificação do Locutor (Carvalho, Santos, 2008)

2.2.1.2 Identificação de Locutor

A identificação de locutor é um ramo de reconhecimento de voz que usa a voz do indivíduo com o objetivo de reconhecer sua identidade, ou seja, seu objetivo principal é reconhecer a pessoa que fala. Isso é feito com uma comparação do sinal de entrada com uma base de dados presente no sistema, determinando se a pessoa pertence a determinado grupo ou se ela é reconhecida no sistema. Essa comparação, muitas vezes, requer muito tempo de processamento, se a base de dados for muito grande, o que inviabiliza o processo (França, 2007).

O sistema de reconhecimento de locutor é consistido, basicamente, de duas partes como mostrado na figura 2 (Mashao, Skosan, 2006). No front-end está à parte geração de característica e no back-end temos o mecanismo de geração dos modelos e de classificação. Durante a fase de treino a chave é

posta para cima e o sistema gera modelos e, durante a fase de testes ou avaliação, os modelos são usados para checar a identidade do locutor comparando-as com um sinal de fala que será usado como input (Mashao, Skosan, 2006).

.

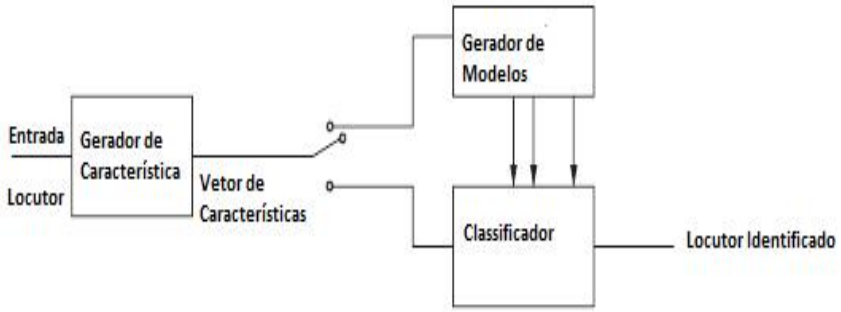

Figura 2 - Sistema de identificação de Locutor (Mashao, Skosan, 2006)

2.2.2 Síntese de Fala

O sintetizador Text-To-Speech (TTS) é um sistema que deve ser capaz de ler qualquer texto em voz alta, ou seja, a partir de uma entrada qualquer de texto, o texto é reproduzido de forma audível.

A figura de Dutoit (1997) representa como atingir a técnica descrita anteriormente. Tal figura pode ser visualizada logo abaixo:

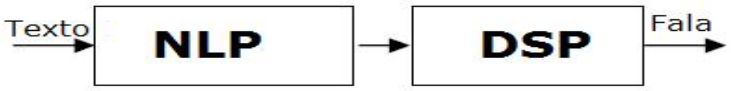

Figura 3 - Sintetizador TTS (Dutoit, 1997)

O primeiro bloco representa o Natural Language Processing (NLP). O NLP é o processo capaz de receber uma entrada textual e transcrever em uma representação lingüística. Tal representação pode, ainda, ser transcrita em uma fonética. No TTS temos ainda o Digital Signal Processing (DSP), no qual a

saída acústica é produzida a partir das informações fonéticas e prosódicas encontradas na transcrição.

Segundo Taylor (2009), o objetivo da Síntese de Fala é alcançar duas características desejas: Inteligibilidade e Naturalidade. A primeira tem o objetivo fazer que um sistema seja capaz de elaborar e emitir uma mensagem com clareza e, a segunda, é transmitir a mensagem gerada com a máxima proximidade da voz humana, ou seja, evitar que a voz se pareça com que esta fala sintética se pareça ao máximo com uma voz robótica.

2.2.3 Reconhecimento da Fala

Reconhecimento de fala é conhecido como reconhecimento automático de fala. É definido como o processo de conversão de um sinal de fala a uma seqüência de palavras, por meio de um algoritmo implementado como um programa de computador (Anusuya, Katti, 2009).

A fala é a forma mais natural da comunicação dos seres humanos. A pesquisa tem sido motivada pelo desejo de se construir modelos mecânicos para emular a comunicação verbal humana. Isto unido a crescente curiosidade tecnológica e o desejo de automatizar tarefas simples têm proporcionado o avanço de diversas áreas, inclusive a área de reconhecimento de fala (Dat, 2000). Com essa motivação, tem se conseguido êxito na evolução de tecnologias juntamente com estudos mais aprofundados na área de modelagens estatísticas. Uma dessas áreas é a área de telefonia e operações de serviços (Anusuya, Katti, 2009).

De modo geral, o reconhecimento de padrão pode ter quatro componentes. A figura 4 (Duda, 2000) mostra estes componentes:

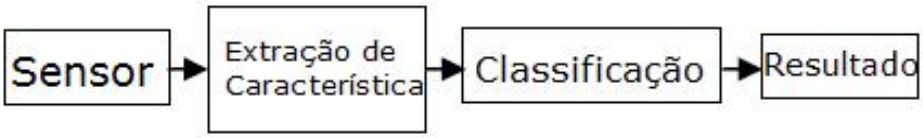

Figura 4 - Esquemático do Reconhecimento de Padrões (Duda, 2000)

Inserindo no contexto do reconhecimento de fala temos: a primeira fase, a fase do sensor, pode ser utilizada um telefone para gravar as locuções, que foi uma das tarefas deste trabalho, ou, até mesmo, um simples microfone. Em seguida, temos fase de extrair as informações e classificar de acordo com nossa base de treino e teste. E por fim, temos o resultado que varia de acordo com nossa aplicação.

A área de reconhecimento automático de fala pode ser dividida em várias classes diferentes dependendo da aplicação e do tipo locuções que tem a capacidade de reconhecer. São eles (Anusuya, Katti, 2009):

- Palavras Isoladas - reconhecedores de palavras isoladas geralmente exigem que entre cada elocução tenha um tempo para o reconhecimento ocorrer. Normalmente, o sistema faz tratamento da elocução gravada entre as pausas;

- Palavras Conectadas – são semelhantes às palavras isoladas, as que permite separar declarações de ser unidas com uma pausa mínima entre elas;

- Palavras Contínuas - reconhecedores de fala contínua permite que os usuários falem quase naturalmente, enquanto o computador determina o conteúdo;

- Palavras Espontâneas - Em um nível básico, pode ser pensado como discurso, que é emitido naturalmente e o sistema não tem um conhecimento prévio deste som.

Além disso, podemos classificar quanto (Reynolds, 2005): ao tipo de locutor:

- Dependente de Texto, Independente de texto ou Adaptativo;

ao tipo do tamanho da base:

- Pequena, Média ou Grande;

e ao estilo de falar:

- Ditado ou espontâneo.

Essas classificações se concentram nas áreas de aplicações do reconhecimento de fala. Dentre essas aplicações podemos citar: domésticas todos os eletrodomésticos controlados por voz; medicina – laudos médicos sendo transcritas enquanto o médico examina o paciente; telecomunicações – sistemas Call Centers automáticos; dentre outros.

Capítulo 3

Tecnologia VoIP

Neste capítulo explanaremos uma breve descrição de VoIP e mostraremos a tecnologia utilizada no projeto. Apresentaremos seus conceitos básicos, arquitetura e ferramentas e aplicações suportadas.

3.1 VoIP

Segundo Tanenbaum (2003), VoIP é, simplesmente, a telefonia da internet. É a inserção dos dados, mais especificamente, a voz sobre o IP (Internet Protocol). O IP pode ser mais bem estudado no RFC791.

Este autor relata que, com o aumento sistemático do trafégo de dados, as empresas perceberam um importante meio de obter recursos usando a rede das redes, a Internet.

Ele especifica ainda, que em 1999, a quantidade do tráfego de dados se igualou a quantidade de dados trafegados na forma de voz. Em 2002, essa diferença foi multiplicada por 10 e, com este aumento, foi perceptível a pequena exigência dos dados neste formato em comparação com aquele. Diante disto, as empresas aproveitaram o crescente aumento da largura de banda para usar a Internet como para a telefonia aumentando, assim, seus lucros.

Com a evolução dos sistemas de compressão dos dados, os dados de voz ficaram cada vez menores, facilitando ainda mais a utilização da Internet pelas empresas de telefonia (Kurose, Ross, 2003). Um exemplo de algoritmo de compressão é o PCM (Pulse Code Modulation).

3.2 Asterisk

Asterisk (Spencer, 2008) é um dos mais conceituados engine e toolkit de telefonia de código aberto. Oferecendo flexibilidade sem precedentes no mundo de propriedades de comunicações, Asterisk permite os desenvolvedores a criar avançadas soluções de comunicação gratuitamente.

Ele foi criado pelo Digium e por uma base de usuários que se encontra em crescimento contínuo. A Digium investe tanto no desenvolvimento de código fonte quanto no hardware de telefonia de baixo custo que funciona com o Asterisk. Um exemplo desse desenvolvimento de hardware é o projeto Zapata. Este projeto foi conduzido por Jim Dixon e um desenho da biblioteca de hardware/driver e interface permitiu que o Asterisk evoluísse para um PABX real (Spencer, 2008) (Dixon, 2002)

Essa tecnologia oferece recursos tais como: correio de voz, conferência, IVR e distribuição de chamada automática. Originalmente desenvolvida para Linux, Asterisk também roda em diferentes sistemas operacionais incluindo NetBSD, OpenBSD, FreeBSD, Mac OS X e Solaris. Uma porta para Microsoft Windows é conhecida como AsteriskWin32. A arquitetura do Asterisk é muito simples. Em essência o Asterisk atua como um middleware, conectando tecnologias telefônicas on the bottom com aplicações telefônicas on the top, criando um consistente ambiente de telefonia ligada. Mais informações sobre a arquitetura podem ser encontradas em (Meggelen, Madsen, 2007).

O núcleo do Asterisk contém alguns mecanismos em que cada um assume um importante papel na operação do software. O Asterisk Gateway Interface, ou AGI, fornece uma interface padrão tal que programas externos podem controlar o Asterisk dialplan. Usualmente, os scripts do AGI são usados para fazer lógica avançada, comunicar com banco de dados relacionais e acessar outros recursos externos. Para aplicações em Java, a maneira mais fácil de interafir com Asterisk é através do protocolo FastAGI. FastAGI

basicamente fornece um recipiente que recebe conexões do servidor do Asterisk, analisa a solicitação e chama por scripts mapeados para chamar a URL (Spencer, 2003).

A estrutura do Asterisk pode ser visualizada na figura 5 (Gonçalves, 2005):

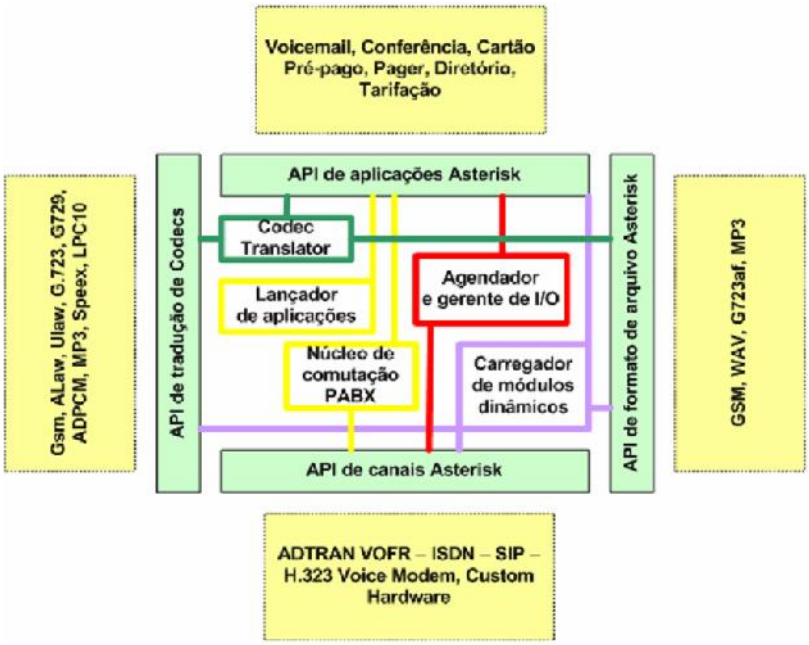

Figura 5 - Arquitetura do Asterisk (Gonçalves, 2005)

A partir desta figura pode-se notar, a gama de tecnologias suportadas pelo Asterisk. Primeiramente, podem-se a grande quantidade de Codecs suportados por tal tecnologia. Segundo Tanenbaum (2003), o codecs é fundamental na construção de uma rede robusta de VoIP e quanto mais codecs a tecnologia suportar mais robusta é a aplicação. Em seguida, nota-se a presença de quatros tipos de formatos de áudio. A variedade desses formatos favorece na escolha da aplicação e do tipo de compactação dos dados necessários. Por fim, temos as API's de aplicações e de Canais do Asterisk. Essas API's permitem vários tipos de aplicações, como foi explicitado anteriormente, e de sinalização de chamada como, por exemplo, SIP e H.323.

As vantagens de usar o Asterisk são mostradas em (Gonçalves, 2005):

- redução de custos extrema;

- ter controle do seu sistema de telefonia;
- ambiente de desenvolvimento fácil e rápido;
- rico e abrangentes em recursos;
- é possível prover conteúdo dinâmico por telefone;
- plano de discagem flexível e poderoso.

Limitações do Asterisk (Gonçalves, 2005):

- Não possuir um driver para acesso R2 Brasil com código aberto;
- Usar o CPU e não o DSP para processar os sinais. Apesar da redução do custo, o sistema ficou muito dependente da qualidade da CPU.

Na figura 6 (Gonçalves, 2005), será apresentado um exemplo de cenário do uso do Asterisk e como ela se encaixa no modelo de telefonia atual.

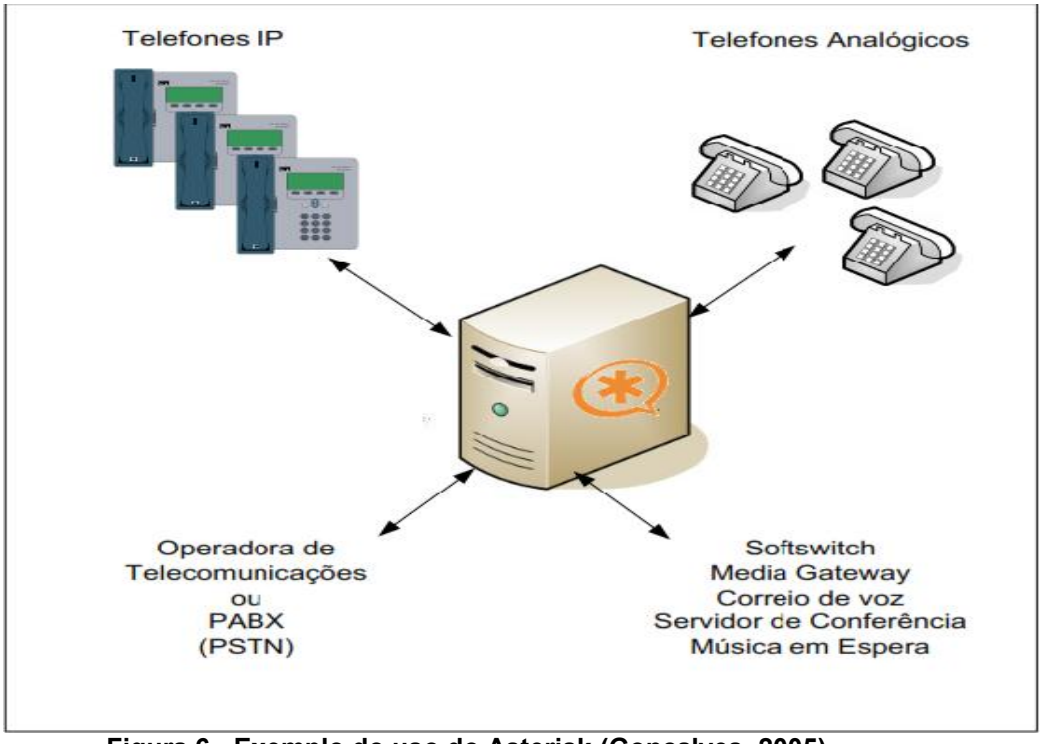

Figura 6 - Exemplo do uso do Asterisk (Gonçalves, 2005)

Nesta figura pode ser verificado um computador-servidor central, com o Asterisk instalado que pode unir todas as conexões existentes em uma empresa, por exemplo. Esta aplicação contém:

- Uma conexão com telefonia IP. Esta conexão é feita pela interface que recebe a linha analógica. Tal interface é chamada de Foreign eXchange Office (FXO) (Gonçalves, 2005);

- Conexões com aparelhos telefônicos analógico. Esta conexão é realizada, também, pela interface FXO;

- Um cabo diretamente da operadora de telefonia ou vindo de uma central de redirecionamento de chamadas. Esta conexão é realizada pela interface Foreign eXchange Subscriber (FXS) (Gonçalves, 2005). É a interface que fornece a linha analógica ao assinante;

- Algumas tecnologias suportadas pelo Asterisk. Por exemplo: voicemail, música em espera, entre outros.

Capítulo 4

Framework for an Integrated Voice Environment

O Framework for an Integrated Voice Environment (FIVE) (Maciel, Carvalho, 2010), como o próprio nome sugere, é um framework integrado de voz baseado em desenvolvimento de interfaces. O FIVE, não de forma coincidente, é composto de cinco módulos. Os módulos de reconhecimento de fala, de síntese de fala e de verificação de locutor representam as áreas de processamento digital de sinais totalizando, assim, três módulos. Os outros dois módulos são o de integração e o de aplicação. A figura 7 (Maciel, Carvalho, 2010) mostra a arquitetura do FIVE:

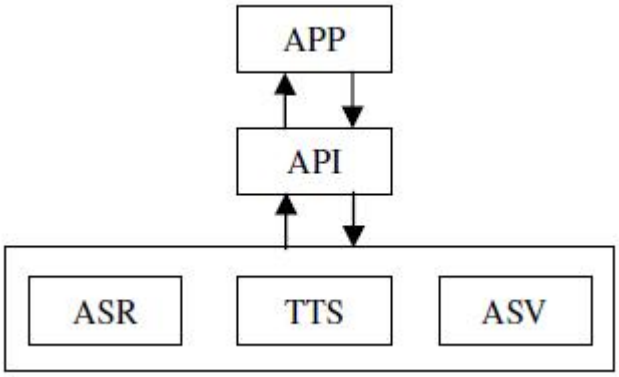

Figura 7 - Arquitetura do FIVE (Maciel, Carvalho, 2010)

As abreviações estão em inglês e representam: ASR (Automatic Speech Recognition), Text-To-Speech engine (TTS), Automatic Speaker Verification engine (ASV), Application Programming Interface (API), Application (APP).

O FIVE (Maciel, Carvalho, 2010) foi escrito em paradigma orientado a objeto, mais especificamente, na linguagem Java. Tal linguagem foi escolhida

porque é extremamente portátil e fácil de integrar com frameworks existentes na literatura.

Podemos citar alguns frameworks utilizados que permitiram a integração proposta pelo FIVE:

- O framework Neuroph, que usa a técnica MLP, e o Weka framework, que usa a técnica SVM. Estes frameworks fazem parte do módulo de reconhecimento de fala.

- O framework FurbSpeech cria o processamento natural da linguagem através do TTS (Test to speech). Estes frameworks fazem parte do módulo de síntese de fala.

- O framework Sautrela, que implementa a quantização vetorial, e PyMix framework, que implementa o GMM. Estes frameworks fazem parte do módulo de reconhecimehhnto de locutor.

O módulo de API tem o objetivo de oferecer aos desenvolvedores um controle em tempo real de todos os módulos. Estes controles se estendem desde as interfaces até a entrada e saída de áudio.

O módulo de aplicação se inseri no contexto da abstração de possíveis aplicações. Tal módulo permite, ainda, a criação de instâncias da API em qualquer tipo de aplicação.

O presente trabalho se insere, basicamente, em três dos cinco módulos. O módulo de reconhecimento de fala, que contém toda a fase de inserir as locuções (apenas inserir, pois as locuções foram feitas com o Asterisk, apesar do FIVE suportar esta opção), inserir os locutores, extrair os parâmetros MFCC e classificar com o HMM.

4.1 HMM

Processos do mundo real produzem resultados que podem ser caracterizados por sinais de natureza discreta ou contínua. Dentre esses processos, temos os sinais do mundo real sendo caracterizado por esses processos, ou seja, caracterizados por modelos de sinais. Esta caracterização pode ser ilustrada por vários fatores como, por exemplo: um modelo de sinal pode prover a base para uma descrição teórica de um sistema de processamento de sinal, o qual pode ser usado para processar o sinal, bem como prover uma saída desejada; os modelos de sinais trabalham extremamente bem na prática, e nos permite realizar importantes sistemas práticos. Ex: sistema de reconhecimento (Rabiner, 1989); entre outros.

Para tal, durante os últimos 15 anos, HMM (Hidden Markov Model) tem sido largamente aplicado em várias áreas, incluindo reconhecimento de voz (Bohlenius, 2005) (Braga, Coelho, Freitas, 2005), modelagem de linguagens (Black, Heiga Zen, 2007), reconhecimento de palavras manuscritas (Black, Lenzo, 2007) (Braga, Freitas, Ferreira, 2003) (Braga, JR, 2007), verificação on-line de assinatura (Braga, Marques, 2007) aprendizado de ações humanas (Braga, Mato, 2006) e detecção de falhas em sistemas dinâmicos (Braga, Freitas, Barros, 2002).

Os modelos estatísticos tentam caracterizar somente propriedades estatísticas dos sinais. É o caso do HMM. Este algoritmo é duplamente estocástico. Um processo estocástico não é visível e consiste de um conjunto de estados conectados por transições de probabilidade representada por autômatos finitos. O outro processo estocástico é observável por um conjunto de saída da função de densidade probabilidade (fdp).

No HMM existem dois tipos de estruturas: os modelos ergóticos e os modelos esquerda-direita (Oliveira, 2010). Na figura 8 (Oliveira, 2010), a seguir, iremos apresentar estes dois tipos de arquitetura do HMM.

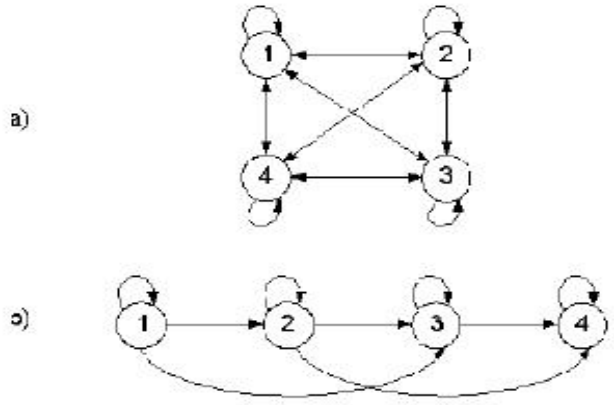

Figura 8 - Tipos de estrutura do HMM. a) Modelo ergóticos b) Modelo esquerda-direita (Oliveira, 2010)

Na figura 8-a) é apresentada um HMM sem restrição ou ergótico. Neste tipo de arquitetura todas as transições são autorizadas, como se pode visualizar. Na figura 8-b) temos o modelo sequencial. Este modelo arquitetural funciona como uma evolução em série através de seus estados. Apesar de ser sequencial Os estados podem ser saltados. O modelo sequencial faz parte modelo esquerda-direita. Como o próprio nome diz, ele começa a esquerda e termina na direita.

4.2 MFCC

A figura 9 (Mashao, Skosan, 2006) mostra todos os passos do MFCC:

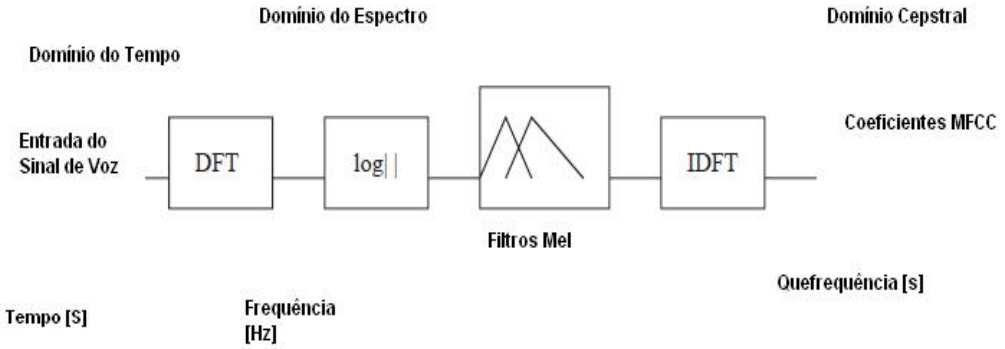

Figura 9. Cepstral Coeficientes de frequência Mel (Mashao, Skosan, 2006)

Primeiro o sinal da fala é adquirido no domínio temporal e depois ocorre a aplicação da transformada de Fourier discreta (Discrete Fourier transform, DFT), transformando, então, para o domínio de freqüência. Em seguida, é aplicada uma magnitude de log (tomando os logaritmos da magnitude de um sinal complexo) do sinal no domínio da freqüência (espectro). Após esse procedimento, o sinal é dito estar no "domínio cepstral".

O sinal no domínio cepstral é "medido em *quefrencies*". *Quefrencies* (Mashao, Skosan, 2006) (Tran, Wagner, 1999) é a transformada inversa de fourrier. As *quefrencies* de baixa ordem contêm informações referentes ao conteúdo da fala e, portanto carrega informações sobre o que está sendo dito. Ao passo que altas quefrencies referem-se ao tom e, portanto caracterizam o locutor (Mashao, Skosan, 2006).

Para obter os coeficientes MFCC o sinal de fala são "janelas" e é convertido para o domínio de freqüência usando-se DFT. No domínio de freqüência o logaritmo da magnitude do sinal complexo é obtido. É executado, então mel-scaling ou mel-warping usando-se filtros. O método comum de implementar esses filtros é usar filtros triangulares que possuem espaçamento linear de 0 até 1kHz e então a não linearidade é imposta de acordo com as aproximações do mel-scaling (Mashao, Skosan, 2006).

Capítulo 5

Experimentos

Neste capítulo é apresentado os passos dos experimentos realizados e os resultados encontrados.

5.1 Base de Dados dos Locutores

Os experimentos descritos nesta seção foram realizados com o objetivo de integrar a plataforma de telefonia Asterisk a um processo de reconhecimento da fala do locutor usando a modelagem HMM. Com esse propósito, foi desenvolvida uma base de dados com uma pequena quantidade de amostras para treinos e testes. Esta foi escolhida para o sistema, pois um banco de dados de tal magnitude torna a aplicação real. A dificuldade em obter quantidade suficiente de usuários dispostos a gravar é grande. Isso se deve ao tempo e paciência que esta tarefa demanda. Nesse sentido, tivemos que solicitar a contribuição de amigos e companheiros de trabalho que entendessem a necessidade da pesquisa para o desenvolvimento da ciência. Em (Maciel, Campos, França, Carvalho, 2009) explicita toda a etapa para gravação dos áudios, utilizada na literatura, e todas as configurações necessárias ao Asterisk para que esta etapa seja realizada com sucesso.

A base de dados foi gravada e processada pela equipe e foi composta, inicialmente, de 22 locutores (14 homens e 8 mulheres) com idades entre 20 e 30 anos. Porém, devido à baixa taxa que iremos apresentar na subseção de resultados, foram gravadas mais locuções de 20 locutores (16 homens e 4 mulheres). Totalizando, assim, 42 (30 homens e 12 mulheres) locutores em nossa base de dados. A tabela a seguir resume essas informações:

Tabela 1 - Tabela de Locuções

1º Base de Dados	Locuções Gravadas	2º Base Dados	Locuções Gravadas
Vozes Masculinas	14	**Vozes Masculinas adicionadas**	16
Vozes Femininas	8	**Vozes Femininas adicionadas**	4
Total	22	**Total**	42

As frases ditas (em português) foram escolhidas de modo que simulasse um ambiente de central telefônica. Esta lista contém palavras de dois tipos: nomes de pessoas, simulando pessoas que trabalha na empresa, e nomes de setores, que simula os ambientes operacionais da empresa. As palavras podem ser visualizas nas tabelas 2 e 3, respectivamente:

Tabela 2 - Locuções Referentes a setores da Empresa

Locuções referentes a setores operacionais da empresa		
RH	**Administração**	**Central telefônica**
Pessoal	**Administrativo**	**Central**
Recursos Humanos	**Financeiro**	**Direção**
Contabilidade	**Secretaria**	**Diretoria**
Contábil	**TI**	**Compras**
Fiscal	**Informática**	**Vendas**
Recepção	**Telefonista**	

Tabela 3 - Locuções Referentes a pessoas da Empresa

Locuções referentes a pessoas da empresa:	
Edson Carvalho	Carmem
Edson	Vanessa
Jordano	VocalLab
Jairson	AgilWare
Rodrigo	Juliana
Alexandre	Polisa

A gravação foi feita em um ambiente sem ruídos, com um telefone. O material gravado foi armazenado em wav mono, 16 bits. A tabela 4 resume as informações da base de dados adquirida.

Tabela 4 - Descrição da base de dados dos locutores

Locutores	42 (30 homens/12 mulheres)
Seção	1
Conteúdo	Palavras das tabelas 2 e 3
Canal	Telefone fixo ou móvel
Ambiente acústico	Vários
Idioma	Português
Tamanho da amostra de áudio	16 bits
Taxa da amostra de áudio	16 KHz
Formado do arquivo	Wave

Depois dessa gravação utilizamos o FIVE para cadastrar os locutores e as locuções e para executar as outras atividades descritas a seguir.

5.2 Extração de Características

O processo de extração de características começou com o algoritmo de detecção da atividade de voz usado para descartar frames de silêncio e/ou ruído. Este tal algoritmo é o Voice Activity Detection (VAD) (Maciel, Carvalho, 2010). O detector de atividade de fala é alto-normalizado, baseado na detecção de energia que rastreia o ruído do sinal e pode adaptar-se a mudanças de ruído.

Na Segunda etapa dos experimentos foi realizado um pré-processamento na base de dados. Locuções foram cortadas e receberam filtros para evitar processamento desnecessário e melhorar o desempenho do sistema. Isso ocorreu devido à diferença de tempo de gravação de cada locução ter sido diferente.

Em seguida, a fala foi fragmentada em frames de 30ms evoluindo numa taxa de frame de 20ms e os vetores característicos mel-scale cepstral foram extraídos dos frames da fala. O cepstrum de mel-scale é a discreta transformada do cosseno da energia logspectral do segmento da fala, como foi descrito na seção do MFCC. Este cálculo é feito para analisar a alteração do espectro de curto prazo e suavizar o espectro de potência modificado.

As características consistem em 39 coeficientes Mel-Frequency cepstral coefficients, amplamente utilizados, usando o espectro de magnitude, com número de filtros igual a 22, com janela de pré ênfase de 0.97 e janela de Hamming para modificar o sinal e transformar para o domínio do MFCC, e computada a cada 10 milissegundos (ou seja, 10 ms é o deslocamento do frame) para um quadro de 20 ms.

Os parâmetros foram escolhidos de forma a maximizar as baixas freqüências, segundo exemplificado na subseção do MFCC, e os outros parâmetros foram escolhidos analisando os artigos na área.

5.3 Classificação

O processo de classificação começou com a divisão do banco de dados em duas bases diferentes para o treinamento e testes. Com a escolha do HMM como algoritmo de treino e classificação, é necessário determinar os parâmetros deste algoritmo. O primeiro parâmetro escolhido foi o número de gaussianas. O número escolhido foi 12. Logo em seguida, foi escolhida a quantidade de iterações. Para tal parâmetro, foi escolhido a quantida de 10. Por último, o número de HERest foram outras variáveis a serem escolhidas. Tal parâmetro foi setado com o valor 3.

Com os parâmetros determinados, foi escolhido a taxa de 75% para teste e 25% para treino. A escolha dos vetores de treinamento foi feita aleatoriamente. Essa aleatoriedade assegura uma boa variabilidade e um consistente do modelo de referência.

Os parâmetros foram escolhidos tanto de forma empírica quanto analisando os artigos na área.

Então, o próximo passo foi calcular os modelos de referência de cada palavra da base de dados. Este cálculo é executado na fase de treinamento. Esta medição é armazenada para servir de cálculo de decisão na fase de testes.

O resultado foi a comparação dos modelos gerados no treinamento com os modelos gerados das locuções pertencentes ao teste. Para apresentar os resultados obtidos foram usados dois tipos de análise dos dados: as tabelas de resultados e a matriz de Confusão. Os dados serão analisados na próxima subseção e as possíveis explicações dos resultados apresentados no trabalho.

5.4 Resultados

As Tabelas a seguir mostram um resumo dos resultados da fase de teste executadas logo depois da fase de treino. Os testes foram executados 5 vezes com cada base de dados.

Na primeira fase do primeiro teste, com a base de dados de 22 locutores, os testes foram executados com todas as locuções. Com isso, a taxa de acerto com maior porcentagem de acerto foi de 83,75% e a pior taxa de acerto foi de 81,88%, como pode ser visualizada nas tabelas a seguir.

Tabela 5 - Resultado com base de 22 locutores

Melhor Resultado	Pior Resultado
83,75%	81,88%

Na segunda fase do primeiro teste, as locuções que tiveram resultados negativos foram analisadas e, alguns, foram retirados. Tais locuções podiam estar viciando os resultados e os seguintes resultados foram atingidos:

Tabela 6 - Resultado com a base de 22 locutores ajustada

Melhor Resultado	Pior Resultado
90,32%	86,25%

Foi percebido, ao longo do estudo, que esta não seria a melhor saída. Então, foi decidido gravar novas locuções para a base de dados. Uma segunda bateria de teste foi realizada e os resultados são apresentados logo abaixo. Uma tabela 7 com o melhor resultado e com o pior resultado é apresentado.

Tabela 7 - Resultado com base de 42 locutores

Melhor Resultado	Pior Resultado
95,00%	91,18%

No gráfico a seguir, os resumos das informações podem ser visualizados.

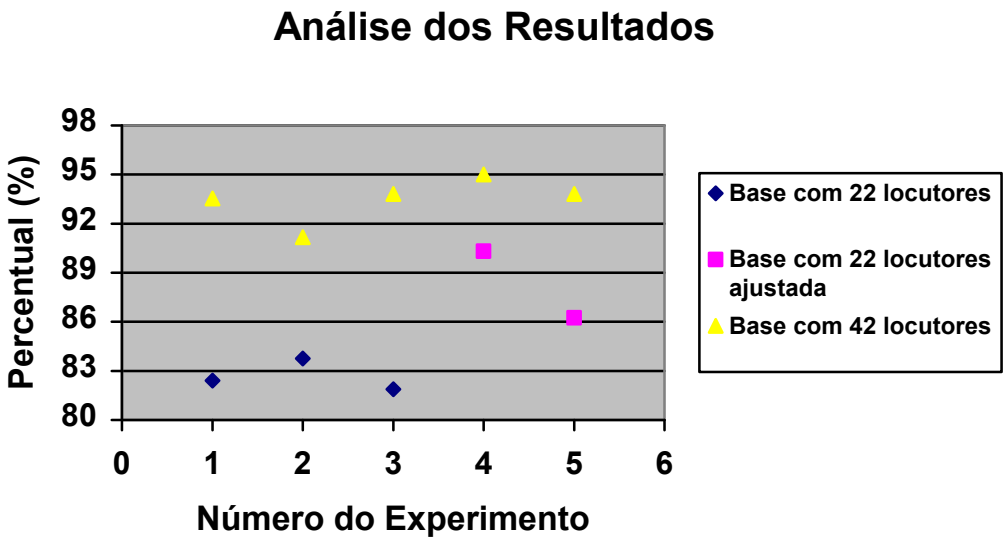

Figura 10 - Resumo dos Resultados

Na próxima página é apresentada a matriz de confusão do melhor resultado encontrado.

Tabela 8 - Matriz de Confusão

	p	r	c	C	F	a	A	f	s	T	i	c	C	t	d	d	c	v	r	e	E	j	j	r	a	c	v	v	a	j	p	d	d	r		
	e	e	o	O	I	d	D	i	e	I	n	e	E	e	i	i	o	e	e	d	D	o	a	o	l	a	a	o	g	u	o	a	a	h		
	s	c	n	N	S	m	M	n	c		f	n	N	l	r	r	m	n	c	s	S	r	i	d	e	r	n	c	i	l	l	n	n			
	s	u	t	T	C	i	l	a	r		o	t	T	e	e	e	p	d	e	o	O	d	r	r	x	m	e	a	l	i	i	i	i			
	o	r	a	A	A	n	N	n	e		r	r	R	f	t	c	r	a	p	n	N	a	s	i	a	e	s	l	w	a	s	e				[%c / %e]
pess	**9**	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[90.0/0.3]
recu	0	**10**	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	
cont	0	0	**10**	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	
cont	0	0	0	**9**	0	0	**1**	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	[90.0/0.3]
fisc	0	0	0	0	**8**	0	0	0	0	0	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[80.0/0.6]
admi	0	0	0	0	0	**7**	**3**	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	[70.0/0.9]
admi	0	0	0	0	0	0	**10**	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	
fina	0	0	0	0	0	0	0	**10**	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	
secr	0	0	0	0	0	0	0	0	**10**	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	
ti	0	0	0	0	0	0	0	0	0	**10**	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	
info	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
cent	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
cent	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
tele	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

dire	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
dire	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[90.0/0.3]
comp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
vend	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[90.0/0.3]
rece	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
edso	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
edso	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
jord	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[90.0/0.3]
jair	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	
rodr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	**9**	0	0	0	0	0	0	0	0	0	0	0	0	[90.0/0.3]
alex	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	
carm	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**	0	0	0	0	0	0	0	0	0	0	[90.0/0.3]
vane	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	**9**	0	0	0	1	0	0	0	0	0	[90.0/0.3]
voca	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	**10**	0	0	0	0	0	0	0	0	
agil	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	**10**	0	0	0	0	0	0	0	
juli	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	**10**	0	0	0	0	0	0	
poli	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	**8**	0	0	0	0	0	[80.0/0.6]
dani	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	**9**	0	**1**	0	0	[90.0/0.3]
dani	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	**9**	**1**	0	0	0	[90.0/0.3]
rh	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	**10**	0			
Ins	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			

Legenda: pesso = pessoa; recurs = recursos humanos; conta = contabilidade; CONTA = Contábil; FISCA = fiscal; admin = administração; ADMIN = administrativo; finan = financeiro; secre = secretaria; TI = ti; infor = informática; centr = central telefônica, CENTR = central; telef = telefonista; diret =diretoria; direc = direção; compr = compras; venda = vendas; recep = recepção; edson = Edson; EDSON = Edson Carvalho; jorda = Jordano; jairs = jairson; rodri = Rodrigo; alexa = Alexandre; carme = Carmem; vanes = Vanessa; vocal = Vocallab; agilw = agilware; julia = Juliana; polis = polisa; danie = Daniele; dani = Dani; rh = rh.

5.5 Análise dos Resultados

Na matriz de confusão apresentada podemos capturar as seguintes informações:

- A palavra “pessoal” é classificada erradamente como a palavra “central”;
- A palavra “contábil” é classificada erradamente como a palavra “administrativo”;
- A palavra “administração” é classificada erradamente como a palavra “administrativo”;
- A palavra “fiscal” é classificada erradamente como a palavra “central”;
- A palavra “direção” é classificada erradamente como a palavra “recepção”;
- A palavra “vendas” se é classificada erradamente como a palavra “recepção”;
- A palavra “jordano” é classificada erradamente como a palavra “contábil”;
- A palavra “Rodrigo” é classificada erradamente como a palavra “Jordano”;
- A palavra “Carmem” é classificada erradamente como a palavra “Contábil”;
- A palavra “Vanessa” é classificada erradamente como a palavra “Polisa”;
- A palavra “Polisa” é classificada erradamente como a palavra “Telefonista” e como a palavra “Vanessa”;
- A palavra “Danielly” é classificada erradamente como a palavra “rh” ;

- A palavra “Dani” é classificada erradamente como a palavra “rh”.

Diante do resultado apresentado e dos resultados encontrados durante a realização do projeto, chegamos a algumas conclusões:

As palavras “pessoal”, “central” e “fiscal”, muitas vezes, são classificadas erroneamente entre si. Isto ocorre devido ao prefixo existente nas palavras. O sufixo “-al”.

As palavras “administrativo” e “administração”, muitas vezes, são classificadas erroneamente entre si. Isto ocorre devido ao prefixo existente nas palavras. O prefixo “-administr”.

As palavras “recepção” e “direção”, muitas vezes, são classificadas erroneamente entre si. Isto ocorre devido ao prefixo existente nas palavras. O sufixo “-ção”.

A palavra “Polisa” é classificada erradamente como “Vanessa” e o inverso também é verdade. Isso se deve a entonação da sílaba tônica. A entonação da vogal “a” é muito forte provocando, assim, o erro na classificação. As locuções classificadas erradas foram observadas e foi verificada que os locutores dão uma maior ênfase para a sílaba tônica. Isto ocorre, também, com a palavra ”telefonista”. Esta palavra se confunde com aquelas.

As palavras curtas como: “rh”, “ti” e “dani” se confundem com as outras palavras. Isto se deve ao tamanho do arquivo de áudio necessário para estas palavras o que, em muitas vezes, se confunde com características de outras locuções.

A palavra “contábil” destoou das outras palavras. Esta foi classificada erradamente com diversas palavras da base de dados.

5.6 Desenvolvimento da Aplicação

A aplicação foi desenvolvida em plataforma Linux. Isto se deve a maturidade do Asterisk no Linux. Existe Asterisk para Windows, MacOS, entre outros (Maciel, Campos, França, Carvalho, 2009); mas, ainda, são plataformas não muito desenvolvidas para tal aplicação.

O Asterisk necessita de uma placa própria com entradas FXO e FXS. No nosso projeto, utilizamos a placa ALO-IP. Esta placa contém uma interface FXS que permite a conexão analógica com o computador. O Asterisk tem um ambiente de configuração própria, mas foi decidido utilizar a linguagem de programação JAVA e todas as suas vantagens.

Para isso, conectamos o Asterisk ao JAVA com Asterisk Gateway Interface utilizando o protocolo FastAGI. Tal protocolo ajuda a distribuir carga para programas de servidores remotos e reduzir a latência de inicialização do programas. Esta conexão permitiu que utilizássemos várias aplicações suportadas ao JAVA, o que não seria possível utilizando o ambiente padrão do Asterisk.

Entre essas aplicações está o FIVE. O FIVE foi desenvolvido inteiramente em JAVA. Sua arquitetura, como foi analisado no estudo e apresentado anteriormente, permite a conexão com vários tipos de plataforma, inclusive com o Asterisk Gateway Interface. O FIVE ainda gera uma API para favorecer tal conexão. Isto permitiu o desenvolvimento do trabalho e seu sucesso.

Capítulo 6

Conclusões e Trabalhos Futuros

O principal objetivo desta monografia é promover a integração de um servidor de voz sobre IP para desenvolvimento de uma aplicação de atendimento telefônico utilizando recursos para reconhecimento de fala. Tal objetivo é prover um avanço na ciência e na pesquisa em questão.

Além disso, foi apresentada uma revisão sobre o histórico de reconhecimento de padrões e, mais especificamente, na área de reconhecimento de voz. Nesta pesquisa foi apresentado as técnicas utilizadas e características inerentes de cada processo.

O FIVE e o Asterisk também foram estudados. Tal estudo permitiu a integração destas tecnologias e utilização no processo de reconhecimento de fala. Esta parte do estudo foi importante, pois uniu tecnologia VoIP com um framework em processo de finalização no meio científico. Esta integração permitiu a criação de uma base de dados com característica própria. Esta se encontra disponível no meio acadêmico para pesquisas futuras.

Por fim, o desenvolvimento da aplicação que produziu resultados satisfatórios para o trabalho proposto.

Os resultados deste trabalho influenciam positivamente a área da pesquisa. As taxas apresentadas na seção experimentos foram superiores a muitos artigos pesquisados na área, apesar do tamanho da base de dados não favorecer o estudo.

A gravação da base de dados foi o principal problema encontrado no estudo realizado. Tivemos que contar com a boa vontade de amigos, companheiros de trabalho e de estudantes do centro de informática. Outra dificuldade foi à instalação e a adaptação ao Asterisk, pois esta tecnologia era desconhecida. Isto teve uma grande influência no tempo de desenvolvimento deste estudo. Por último, a integração do Asterisk com o FIVE era uma

incógnita. A incerteza desta integração e do seu, posterior, sucesso fez com que surgisse a necessidade de efetuar diversos testes para testar tal integração. Por isso, esta foi principal contribuição do nosso trabalho.

O sistema implementado foi finalizado com êxito e se encontra em funcionamento favorecendo, assim, a otimização dos processos corriqueiros do dia a dia no local instalado.

Trabalhos Futuros

O uso de VoIP está em constante evolução. Assim, esta monografia certamente pode servir como base para estudos futuros na área de VoIP, além disso, servir como base inicial para o estudo do histórico dos algoritmos e do reconhecimento de voz. Destacando assim: pode-se propor um estudo do trabalho apresentado utilizando algoritmos mais recentes como, por exemplo, o GMM e o FGMM. Pode-se propor, ainda, um estudo da influência do tempo processamento destes algoritmos no congestionamento da rede. Um passo foi dado, mas, conscientemente, um caminho longo há de ser caminhado.

Referências Bibliográficas

AHA, D. (1992). Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. International Journal of Man-Machine Studies, 36(2):267–287]

Anusuya. M. A., Katti. S.K. Speech Recognition by Machine: A Review. International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.

Black A. W., and Heiga Zen, K. T. Statistical parametric speech synthesis. 32nd IEEE International Conference on Acoustics, Speech, and and gnal Processing (ICASSP) (2007).

Black, A. W., AND Lenzo, K. A. Building Synthetic Voices, 2007.

Bohlenius, J. A speech synthesis for classical latin. Master's thesis, Göteborg University, 2005.

Borsato, B. Merschmann, L. Plastino, A. k-NN: Estimando o Valor do Parâmetro k. III Workshop em Algoritmos e Aplicações de Mineração de Dados, 2007. Rio de Janeiro – Brasil.

Braga, D., Coelho, L., Freitas, D. Transcribing prosody using syntax and pragmatics. III Congreso de Fonética Experimental 2005 (2005).

Braga, D., Freitas, D., Barros, M. J. A drt approach for subjective evaluation of intelligibility in european portuguese synthetic speech. International Conference on SYSTEMS SCIENCE – ICOSYS 2002 (2002).

Braga, D., Freitas, D., Ferreira, H. Processamento linguístico aplicado à síntese da fala. 3º Congresso Luso–Moçambicano de Engenharia (2003).

Braga, D., JR, F. G. V. R. Módulos de processamento de texto baseados em regras para sistemas de conversão texto–fala em português europeu. XXI Encontro da Associação Portuguesa de Linguística (2007).

Braga, D., Marques, M. A. Desambiguação de homógrafos para sistemas de conversão texto–fala em português. Diacrítica (2007), 25–50.

Braga, D., Mato, X. R. F. Algoritmos de conversão grafema–fonema em galego para sistemas de conversão texto–fala. VIII Congreso Internacional de Estudos Galegos – AIEG – Galicia do Outro Lado do Atlántico (2006).

Campbell, J.P. Speaker Recognition: A Tutorial. Proceedings of IEEE (1997).

Carvalho, G., Santos, T. V. Impressão vocal. Online, acesso em 26/09/2009 na URL: http://www.gta.ufrj.br/grad/09_1/versao-final/impvocal/index.html

Dat T. T., Fuzzy Approaches to Speech and Speaker Recognition , A thesis submitted for the degree of Doctor of Philosophy of the university of Canberra, 2000.

Dixon, J. Zapata Telephony, 2002. Última atualização em Fevereiro de 2009. Disponível em: <http://www.zapatatelephony.org>. Acesso em: Maio de 2009.

Dutoit, T., In High-quality text-to-speech synthesis an overview, ICASSP. 1997.

Duda, R., et al, 2000. Pattern Classification, 2nd Edition.

França, V. VoicePrints. Online, acesso em 21/09/2009 na URL: http://www.gta.ufrj.br/grad/07_2/viviane/

Fix e Hodges Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties, 1951.

Gans, N. Koole, G. Mandelbaum, A. Telephone Call Centers: Tutorial, Review, and Research Prospects. 2003

Gonçalves, F.E.A. Como construir e configurar um PABX com software livre versão 1.4, 2005

Juang, B.H., Rabiner, L.R . Automatic Speech Recognition – A Brief History of the Technology Development, 2004. Online, acesso em 30/09/2009 na URL: http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf

Kurose, J. F.; Ross, K. W. Redes de Computadores e a Internet: uma nova abordagem. Tradução de Arlete Simille Marques. São Paulo: Addison Wesley, 2003. cap. 1, 6 e 7..

Maciel, A., Carvalho, E. FIVE - Framework for an Integrated Voice Environment. IWSSIP 2010 - 17th International Conference on Systems, Signals and Image Processing.

Maciel, A., Campos, W., França, C. B., Carvalho, E. Speaker Verification System through telephone channel. An integrated system for telephony plataform Asterisk. SIGMAP, 2009.

Mashao, D., Skosan, M. Combining Classifier Decision for Robust Speaker Identification .In Pattern Recognition(2006).

Meggelen, J. V., Madsen, L., Smith, J., 2007. *Asterisk – The Future of Telephony,* O'REILLY. USA, 2nd edition.

Oliveira, L. E. S. e Morita, M. E. Introdução aos Modelos Escondidos de Markov (HMM), 2010.

Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77, 2 (1989), 257–286.

Reynolds, D., Rose, R. Robust text-independent speaker identification using gaussian mixture speaker models. In IEEE Trans. Speech Audio Process (1995).

Spencer, M. Digium Founder Mark Spencer Recounts the History of Open Source Asterisk PBX. 2008. Matéria publicada em 16 de Março de 2008. Disponível em:<http://socializedsoftware.com/2008/03/16/digium-founder-mark-

spencer-recounts-the-history-of-open-source-asterisk-pbx>. Acesso em: Maio de 2009

Spencer, M., et al., 2003. *The Asterisk Handbook*, Digium, Inc. USA, 2nd version.

Tanenbaum, A. S.: Redes de Computadores, Editora Campus, 4º Edição.2003. Cap 7, 729 – 737.

Taylor, P. Text–to–Speech Synthesis. Cambridge University Press, 2009. overview, ICASSP.

Tran, D., Wagner, M. A robust clustering approach to fuzzy Gaussian mixture models for speaker identification. Third International Conference on Knowledge-Based Intelligent Information Engineering Systems (1999).