UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

DESENVOLVIMENTO DE UM PROTÓTIPO PARA RECONHECIMENTO DE VOZ

TRABALHO DE GRADUAÇÃO

Rodrigo Dias Flores

Santa Maria, RS, Brasil

2009

DESENVOLVIMENTO DE UM PROTÓTIPO PARA RECONHECIMENTO DE VOZ

por

Rodrigo Dias Flores

Trabalho de Graduação apresentado ao Curso de Ciência da Computação
da Universidade Federal de Santa Maria (UFSM, RS), como requisito
parcial para a obtenção do grau de
Bacharel em Ciência da Computação

Orientador: Prof. Natanael Rodrigues Gomes

Trabalho de Graduação N. 285

Santa Maria, RS, Brasil

2009

Universidade Federal de Santa Maria
Centro de Tecnologia
Curso de Ciência da Computação

A Comissão Examinadora, abaixo assinada,
aprova o Trabalho de Graduação

DESENVOLVIMENTO DE UM PROTÓTIPO PARA RECONHECIMENTO DE VOZ

elaborado por
Rodrigo Dias Flores

como requisito parcial para obtenção do grau de
Bacharel em Ciência da Computação

COMISSÃO EXAMINADORA:

Prof. Natanael Rodrigues Gomes
(Presidente/Orientador)

Prof. Raul Ceretta Nunes (UFSM)

Prof. Cesar Tadeu Pozzer (UFSM)

Santa Maria, 17 de Julho de 2009.

“O verdadeiro artista é aquele que idealiza o real e realiza o ideal.”
— AUTOR DESCONHECIDO

AGRADECIMENTOS

Agradeço sinceramente a todos que contribuíram, de forma direta ou indireta, para a realização deste trabalho. Em primeiro lugar, à Deus, pela oportunidade de estar aqui trabalhando e aprendendo. Agradeço ao meu Anjo da Guarda que está sempre me guiando e inspirando sempre que necessário. Aos meus notáveis, exemplares e amados pais, Ednezer Rodrigues Flores e Cecilia Dias Flores, agradeço a educação, atenção e amor que recebo e sempre recebi. À toda minha família, em especial à Nena – minha segunda mãe –, à Vó Nilza – a terceira mãe –, ao saudoso Vô Rezende, à Tia Liane e ao Tio Fernando que me acolheram durante a faculdade. Agradeço profundamente à minha querida namorada Simone, pela paciência, compreensão, dedicação, carinho e atenção dedicados a mim. As felicidades e bons momentos que passamos são indispensáveis e muito importantes. Agradeço igualmente aos seus pais, à Lia e ao Danilo, por toda a receptividade, apoio e incentivo, não esquecendo dos incontáveis almoços, jantas e lanchinhos. Ao meu orientador, professor Natanael Gomes, pela amizade, confiança e conselhos fundamentais para a realização desse trabalho. Ao professor Raul Ceretta pela amizade, conselhos e oportunidades dadas ao longo da faculdade, ao professor João Baptista, pela oportunidade de trabalhar e aprender durante esses anos no GMICRO. Não esquecendo da Janice, por sua dedicação e paciência para explicar os trâmites do curso. À UFSM e ao programa de intercâmbio que me possibilitou um incomparável aprendizado de vida na Argentina, sem contar as amizades e passeios magníficos. Agradeço à todos meus colegas, Alexandro, Chico Linux, Dalmazo, Gustavo, Henrique, Jesse, Jesus, João, Suzana, pelo companheirismo, risadas e diversões que passamos juntos. E não esquecendo dos amigos que cederam as amostras de suas vozes.

RESUMO

Trabalho de Graduação
Curso de Ciência da Computação
Universidade Federal de Santa Maria

DESENVOLVIMENTO DE UM PROTÓTIPO PARA RECONHECIMENTO DE VOZ

Autor: Rodrigo Dias Flores
Orientador: Prof. Natanael Rodrigues Gomes
Local e data da defesa: Santa Maria, 17 de Julho de 2009.

A ação de reconhecer uma pessoa pela sua voz, através de uma máquina, é conhecida como reconhecimento automático de locutor. Tal técnica configura em um problema complexo quando consideramos algoritmos que exigem resposta rápida de processamento e as diferentes áreas do conhecimento envolvidas na sua implementação, tais como, processamento de sinais, reconhecimento de padrões, física acústica, linguística, matemática, ciência da computação, entre outras. Este trabalho tem por objetivo investigar sistemas de reconhecimento de voz, mais especificamente sistemas de reconhecimento automático de locutor, buscando como resultado da pesquisa o desenvolvimento de um protótipo. O protótipo implementado, baseado na identificação de locutores de maneira independente de texto, utiliza a extração dos Coeficientes Mel-Cepstrais (MFCC) para obtenção de parâmetros característicos da voz e o algoritmo de Quantização Vetorial LBG (Linde-Buzo-Grey) como técnica de reconhecimento baseada na comparação de padrões. Todo o sistema foi implementado na linguagem ANSI C rodando sobre um PC. A possibilidade de utilização de sistemas de reconhecimento de locutor abrange diversos tipos de aplicações práticas tais como: operações bancárias, controle de acesso à áreas restritas, controle de acesso em softwares, assinatura eletrônica, etc.

Palavras-chave: Reconhecimento de locutor; coeficientes mel-cepstrais; quantização vetorial; algoritmo LBG.

ABSTRACT

Trabalho de Graduação
Curso de Ciência da Computação
Universidade Federal de Santa Maria

DEVELOPMENT OF A PROTOTYPE FOR VOICE RECOGNITION

Author: Rodrigo Dias Flores
Advisor: Prof. Natanael Rodrigues Gomes

The effect of recognizing a person by his voice, through a machine, is known as automatic speaker recognition. This technique sets up a complex problem when considering algorithms that demand a fast processing and the different fields of knowledge, like signal processing, pattern recognition, acoustical physics, linguistics, mathematics, computer science, and others. This work has the purpose of investigating voice recognition systems, specificaly, automatic speaker recognition systems obtaining, as a research result, the development of a prototype. The prototype, based on speakers identification in an text independent classification, uses the Mel-Frequency Cepstrum Coefficients (MFCC) extraction to obtain the characteristic voice parameters and the LBG (Linde-Buzo-Grey) Vector Quantization algorithm as a technique for recognition based on pattern comparison. The whole system was implemented in ANSI C language running on a PC. The speaker recognition systems can be used in many types of practical applications such as banking, access control to restricted areas, access control in software, electronic signature, etc.

Keywords: Speaker recognition; mel-frequency cepstrum coefficients; LBG vector quantization algorithm.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DCT	Discrete Cosine Transform
DSP	Digital Signal Processor
DTW	Dynamic Time Warping
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
IFFT	Inverse Fast Fourier Transform
LBG	Linde-Buzo-Grey
MFCC	Mel-Frequency Cepstrum Coefficients
QV	Quantização Vetorial
RNA	Redes Neurais Artificiais
WAV	Waveform audio format

SUMÁRIO

1 INTRODUÇÃO

O nível de desenvolvimento tecnológico atual nos possibilita a construção de sistemas computacionais cada vez mais complexos, principalmente aqueles envolvendo processamento digital de sinais. Estes sistemas requerem recursos que apresentam alto desempenho e maiores velocidades de processamento. Desse modo, outros campos de atuação surgem para profissionais especializados no tratamento de sinais, como o campo de Processamento Digital da Fala. Essa área compreende algumas sub-áreas bem definidas, como por exemplo, a de Reconhecimento Automático de Locutor.

Uma característica interessante deste campo, é sua natureza interdisciplinar. Diferentes áreas do conhecimento necessitam serem estudadas, tais como, processamento de sinais, reconhecimento de padrões, física acústica, linguística, matemática, inteligência artificial, entre outras. Devido à essa interdisciplinaridade, faz-se necessária a interação com especialistas de tais áreas, a fim de se obter resultados satisfatórios.

Reconhecimento de locutor consiste em identificar uma pessoa através da análise de uma amostra de sua voz. Esta tarefa pode ser realizada de forma automática através da utilização dos recursos computacionais disponíveis atualmente. A maioria dos sistemas de reconhecimento de locutor utilizam-se de técnicas baseadas em comparações de padrões como parte do seu processo de treinamento e reconhecimento. O algoritmo LBG (Linde-Buzo-Grey) para Quantização Vetorial (LINDE; BUZO; GRAY, 1980) é um exemplo de técnica baseada na comparação de padrões na qual pode obter-se excelentes resultados e precisões na identificação de locutores (HE; LIU; PALM, 1999) (SOONG et al., 1987).

A possibilidade de utilização de sistemas de reconhecimento de locutor abrange diversos tipos de aplicações práticas, tais como: operações bancárias, controle de acesso à áreas restritas, controle de acesso em softwares, assinatura eletrônica, etc.

1.1 Classificação do Problema

O reconhecimento de locutor consiste em identificar uma pessoa através da análise de uma amostra de sua voz. Esta tarefa pode ser realizada de forma automática através da utilização de recursos computacionais. De acordo com o objetivo desejado, o reconhecimento automático de locutor é dividido em:

- Verificação: tem por objetivo determinar, automaticamente, se a identidade de um pretenso locutor é verdadeira ou não. Isto pode acontecer quando uma pessoa digita um código e logo em seguida fala uma frase. Trata, portanto, de uma tarefa mais simples, pois compara um padrão teste com um padrão de referência envolvendo uma decisão binária, ou seja, a resposta será sim ou não.

- Identificação: visa determinar qual, dentre os locutores conhecidos, pronunciou a amostra de voz que está sendo avaliada. É usado quando se deseja reconhecer uma pessoa dentre um conjunto de várias outras pessoas, mas sem inicialmente fornecer qualquer informação, ou código da pessoa que se deseja identificar. Ela escolhe dentre um conjunto de N locutores qual deles o padrão em teste melhor se aproxima. Desde que N comparações são necessárias, a taxa de erro no sistema de Identificação pode ser mais alta do que no sistema de Verificação.

Os sistemas de reconhecimento de locutor podem ser ainda divididos quanto ao texto em:

- Dependente de Texto: requer que o locutor forneça elocuções de sentenças ou palavras-chave com o mesmo texto para ambos treinamento e reconhecimento.

- Independente de Texto: não requer que um texto específico seja falado.

Este trabalho tem por objetivo investigar sistemas de reconhecimento de voz, mais especificamente sistemas de reconhecimento automático de locutor, obtendo como resultado da pesquisa um protótipo de identificação de locutores independente de texto através do algoritmo LBG para Quantização Vetorial.

1.2 Organização do Trabalho

O texto está organizado do seguinte modo: no segundo capítulo são apresentados alguns conceitos fundamentais do reconhecimento de locutor, explicando um modelo de

produção da fala humana, e alguns tipos de técnicas utilizadas no reconhecimento de locutor. O terceiro capítulo descreve o projeto do sistema, expondo as etapas e cálculos relacionados à implementação, desde a aquisição da voz até o reconhecimento final. O quarto capítulo aborda a implementação e desenvolvimento, apresentando detalhes acerca da programação realizada. Em seguida, o capítulo cinco traz os resultados obtidos através dos testes realizados sobre as amostras de voz adquiridas, com informações acerca do desempenho do protótipo. Por fim, o sexto capítulo apresenta as conclusões e sugestões de trabalhos futuros.

2 CONCEITOS FUNDAMENTAIS

Neste capítulo são apresentados os conceitos fundamentais para a modelagem e reconhecimento da fala. O aparelho fonador do ser humano é descrito e seu modelo matemático é apresentado. A partir desse modelo matemático são descritos tipos de classificadores de voz utilizados para reconhecer o interlocutor ou a palavra emitida.

2.1 Modelo da Produção da Fala

O ser humano possui um aparelho fonador que é capaz de produzir uma variada quantidade de sons, com diferentes entonações. Essa característica é proveniente da grande flexibilidade do trato vocal que, de acordo com seu posicionamento e movimentação, permite a geração de vários tipos de sons. O trato vocal é o conjunto de órgãos responsáveis pela produção da voz no qual assemelha-se a um tubo, que se inicia na glote, passa pela faringe, pela cavidade oral, e termina nos lábios, com um comprimento médio de 17cm em um homem adulto, como mostra a figura 2.1.

Num primeiro momento, constata-se que os sons produzidos pelo aparelho fonador humano são divididos em apenas duas classes principais: sonoros e não sonoros ou surdos. Os sons sonoros são produzidos pela passagem de ar através da glote, com a tensão das cordas vocais ajustada de maneira a vibrar numa certa frequência, chamada de *pitch*, gerando pulsos quase periódicos de ar que excitam o trato vocal. Os não sonoros são gerados através da produção de uma região de contrição, geralmente a boca, através da qual o ar é forçado a passar numa velocidade suficiente para produzir turbulência. Diante de tais características, pode-se identificar alguns elementos do modelo de produção de fala: o trato vocal, que age como um tubo ressonante e a excitação ocasionada pelo fluxo de ar oriundo dos pulmões.

Como o trato vocal se assemelha a um conjunto de tubos, apresenta ressonâncias em

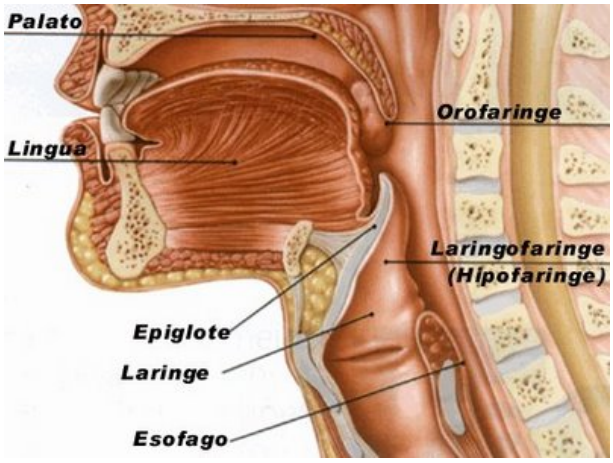

Figura 2.1: Órgãos do aparelho fonador humano

frequências determinadas pelas relações entre as dimensões das diversas seções transversais tomadas ao longo de sua extensão. Tais frequências de ressonância são chamadas de formantes, e possibilitam a diferenciação dos sons produzidos, de acordo com o posicionamento da língua e da boca.

A excitação de trechos não sonoros apresenta características de um sinal ruidoso, composto de ruído branco, de envoltória espectral praticamente plana (SAMUDRAVIJAYA, 2003). Ao atravessar o trato vocal, o espectro da excitação é moldado de acordo com as ressonâncias, resultando no sinal de fala conhecido.

O processo de produção da voz é modelado analiticamente como mostra a figura 2.2. Há um chaveamento entre as fontes de excitação periódica e ruído branco, que produzem os sons sonoros e não-sonoros, respectivamente (PETRY, 2002). Os ganhos para cada um destes sons correspondem à amplitude do sinal gerado. Esse modelo foi desenvolvido através do estudo da produção do sinal de voz, e representa uma simplificação do processo.

2.2 Tipos de técnicas de reconhecimento

Existem três tipos de técnicas de reconhecimento utilizadas no processamento da fala: reconhecimento por comparação de padrões, reconhecimento na análise fonético-acústica e reconhecimento empregando inteligência artificial (MARTINS, 1997).

No reconhecimento por comparação de padrões, o sinal de fala é comparado com padrões previamente armazenados e o padrão mais parecido com o sinal de entrada é escolhido.

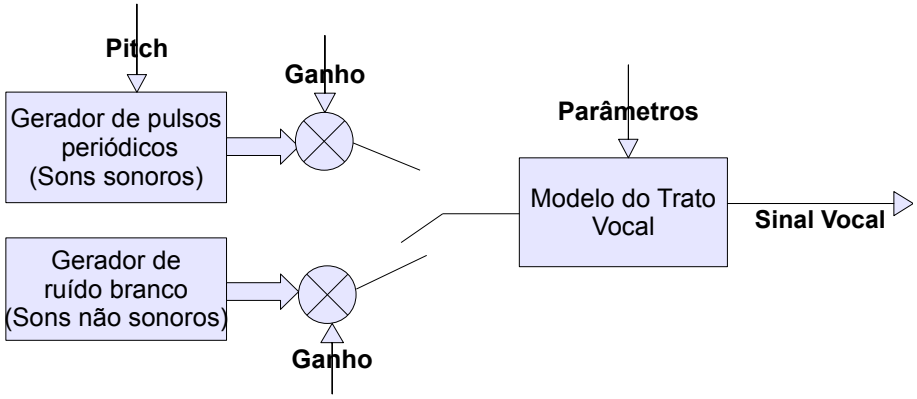

Figura 2.2: Modelo da produção da fala

A abordagem fonético-acústica consiste na detecção sequencial de sons e classes de sons observando as características acústicas do sinal de voz, aplicando as relações conhecidas entre estas característica e os símbolos fonéticos. O método possui dois passos: no primeiro, o sinal é segmentado em regiões discretas onde as propriedades acústicas sejam representativas de um ou mais fonemas, etiquetando-se cada segmento com o fonema ou fonemas prováveis. No segundo passo tenta-se determinar a palavra ou sequência de palavras válidas, a partir da informação do primeiro passo, levando-se em conta a consistência com o vocabulário de reconhecimento, a sintaxe da língua, etc. Alguns exemplos de propriedades acústicas são: localização dos formantes, frequência de *pitch*, etc.

As técnicas que utilizam inteligência artificial para o reconhecimento de voz podem ser consideradas como uma forma híbrida das outras duas abordagens anteriores. Esse método procura mecanizar a função de reconhecimento, incorporando um sistema especialista semelhante ao modo que uma pessoa pensa, analisa e finalmente faz a decisão nas características acústicas medidas para o reconhecimento. As Redes Neurais Artificiais são amplamente usadas para esta finalidade (RABINER; JUANG, 1993).

2.2.1 Reconhecimento por comparação de padrões

Consiste na ideia de treinamento do sistema para o reconhecimento de determinados padrões. São duas fases distintas que esse método utiliza, a saber, treinamento e reconhecimento. A fase de treinamento é responsável pela identificação das informações a serem tomadas como padrões de referência. Se suficientes versões de um padrão a ser reconhecido estiverem incluídas em um conjunto de treinamento, entregue ao algoritmo, a função de treinamento deve ser hábil para, adequadamente, caracterizar as propriedades

acústicas do sinal. Este tipo de caracterização da voz por meio do treinamento é chamado de classificação de padrões porque a máquina aprende quais propriedades acústicas são confiáveis ao longo de todo o processo.

Na fase de reconhecimento, compara-se o sinal desconhecido com os padrões de referência e calcula-se uma medida de similaridade. O padrão que mais se aproximar do sinal desconhecido é escolhido como o padrão reconhecido. Independentemente do tipo de padrão a ser reconhecido é necessário a apresentação de uma quantidade suficiente de material para um bom treinamento do sistema.

O método de reconhecimento por comparação de padrões é relativamente simples de usar e fácil de compreender, existindo uma justificativa matemática forte para muitos dos procedimentos usados no treinamento e no reconhecimento. A figura 2.3 apresenta o diagrama de um sistema que utiliza o método de reconhecimento por comparação de padrões (RUNSTEIN, 1998) (RABINER; JUANG, 1993).

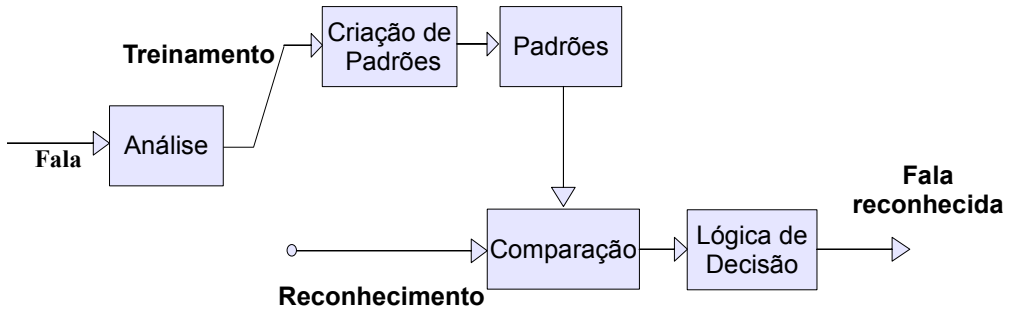

Figura 2.3: Diagrama de um sistema de reconhecimento de voz baseado na comparação de padrões

No primeiro bloco, chamado Análise, o sinal de voz é analisado obtendo-se um conjunto de parâmetros ou atributos que representam as locuções gravadas. Durante a fase de treinamento do sistema, o bloco Criação de Padrões gera padrões, os quais posteriormente, serão utilizados no reconhecimento. Durante a fase de reconhecimento do sistema, o bloco Comparação tem a função de comparar o padrão desconhecido com os padrões existentes, gerando um valor de similaridade para cada comparação.

2.2.2 Estado da Arte

Durante muito tempo vem sendo realizadas pesquisas em reconhecimento de locutor e algumas delas com bons níveis de eficácia. Muitas técnicas têm sido propostas, dentre elas

destacam-se *Dynamic Time Warping* (DTW), *Hidden Markov Models* (HMM), *Gaussian Mixture Models* (GMM), Redes Neurais Artificiais (RNA) e Quantização Vetorial (QV).

Atualmente, o método mais eficaz para o reconhecimento depende principalmente da modalidade de texto associada ao problema, ou seja, dependentes ou independentes do texto (MAFRA, 2002).

Os HMM têm demonstrado os melhores resultados em aplicações que têm dependência de texto, com grande capacidade de modelagem das dependências temporais associadas aos sinais de fala (CHOU; JUANG, 2003).

Os GMM são modelos estatísticos onde as probabilidades de ocorrência dos vetores de atributos para cada locutor são modeladas como combinações ponderadas de variáveis aleatórias vetoriais com Gaussianas. Eles apresentam resultados satisfatórios em aplicações independentes de texto (CARICATI; WEIGANG, 2001).

As RNA são modelos conexionistas não lineares, com grande capacidade de reconhecimento e classificação de padrões. Os melhores resultados são conseguidos pelo uso de arquiteturas baseadas em quantização vetorial, para aplicações independentes de texto (MAFRA, 2002).

Outros estudos mostram que o uso da técnica de QV para aplicações não críticas obtêm resultados satisfatórios (HE; LIU; PALM, 1999).

A técnica empregada para reconhecimento de voz neste trabalho é a QV. Esta técnica foi escolhida em razão de sua facilidade de implementação, de seu índice satisfatório de reconhecimento e do tempo disponível para implementação do sistema.

3 PROJETO DO PROTÓTIPO

Neste capítulo, é apresentado o projeto do protótipo, introduzindo as etapas que constituem o reconhecimento de locutor bem como detalhes importantes para a compreensão do sistema proposto neste trabalho.

3.1 Aquisição da Fala

A primeira etapa do protótipo consiste em realizar a aquisição do sinal da fala através de um transdutor, normalmente um microfone, cuja função é converter as ondas sonoras do locutor em sinais elétricos de tensão analógica. A partir disso, é realizada a digitalização do sinal da fala, que engloba as seguintes operações (PETRY, 2002), conforme a figura 3.1:

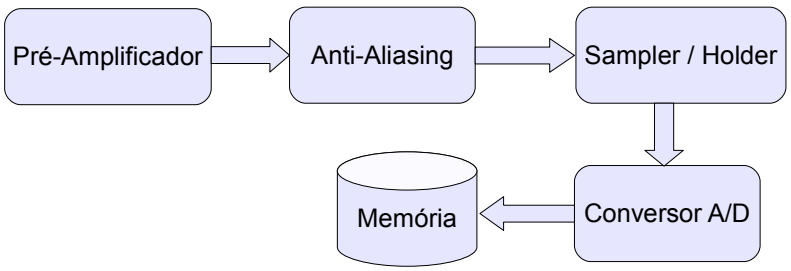

Figura 3.1: Diagrama da aquisição da fala

- **Pré-amplificação**: é realizada uma filtragem analógica de ganho positivo no sinal de voz.
- **Filtro Anti-Anliasing**: filtragem analógica cujo objetivo é eliminar frequências altas, evitando a interferência no espectro relevante durante a análise espectral. Tal interferência produz um ruído chamado *alias*.

- **Sampler/Holder**: ocorre a amostragem do sinal em intervalos de tempo dados pela a frequência de amostragem.

- **Conversão A/D**: recebe os sinais amostrados e os quantiza em uma determinada resolução. Geralmente utiliza-se 8 ou 16 bits para representar cada amostra do sinal.

- **Memória**: uma vez digitalizados, os sinais podem ser armazenados numa memória no formato de arquivos de sons *.wav* (The Canonical WAVE PCM soundfile format, 2009), guardando referência ao locutor e à locução.

3.2 Extração dos Atributos

Qualquer que seja o método utilizado para reconhecer a fala (comparação de padrões, fonético-acústica ou inteligência artificial), existe uma etapa comum a todos os métodos, que é a análise do sinal de fala de forma a extrair informações relevantes para o seu reconhecimento. Tal análise tem como objetivo obter uma representação paramétrica do sinal, que reduza redundâncias, mantendo informações estatísticas suficientes para o reconhecimento. Do ponto de vista do reconhecedor (protótipo), estas representações são os atributos do sinal de voz, que constituem o objeto de reconhecimento.

Diversas representações paramétricas já foram experimentadas em sistemas de reconhecimento de voz e locutor, sendo as que apresentam os melhores resultados, na maioria dos casos, são os Coeficientes Mel-Cepstrais e seus derivados (MAFRA, 2002).

3.2.1 Coeficientes Mel-Cepstrais

Os Coeficientes Mel-Cepstrais (MFCC) surgiram devido aos estudos na área de psicoacústica – estudo da percepção auditiva humana – mostrando que a percepção humana das frequências de tons puros – compostos por uma única frequência – ou de sinais de voz, não segue uma escala linear. Isto estimulou a ideia de serem definidas frequências subjetivas de tons puros, da seguinte forma: para cada tom, medido em Hertz, define-se um tom subjetivo medido em uma escala que se chama escala Mel. Portanto, o Mel é uma unidade de medida de frequência percebida de um tom.

As etapas do cálculo dos Coeficientes Mel-Cepstrais são mostradas no diagrama da figura 3.2 (SIGURDSON; PETERSEN; LEHN-SCHIOLER, 2006). Assim, os Coeficientes Mel-Cepstrais são definidos como a Transformada Inversa de Fourier do logaritmo

do espectro de um sinal aplicado a um banco de filtros digitais triangulares do tipo passa-banda espaçados segundo a escala mel (PETRY; ZANUZ; BARONE, 2000).

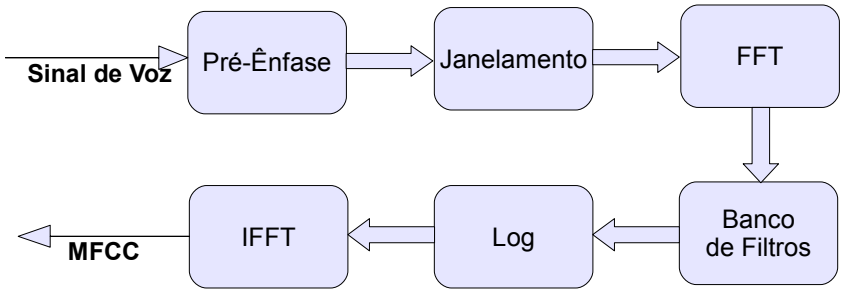

Figura 3.2: Diagrama do cálculo dos Coeficientes Mel-Cepstrais

3.2.1.1 Pré-ênfase

Observa-se que, para sinais de voz, a energia carregada pelas altas frequências é pequena quando comparada com as baixas frequências. A pré-ênfase das frequências altas é necessária para que se obtenha amplitudes mais homogêneas das frequências formantes, pois informações importantes sobre a locução também estão presentes nas altas frequências (MARTINS, 1997). Isto pode ser feito através de um filtro digital cuja função de transferência no domínio z é:

$$H(z) = 1 - a.z^{-1}, \;\; 0 \leq a \leq 1 \tag{3.1}$$

sendo a o parâmetro responsável pela pré-ênfase, da ordem de 0,95. No domínio do tempo o filtro é implementado da seguinte maneira:

$$X' = \{x'(n)\}, 1 \leq n \leq N \tag{3.2}$$

$$x'(n) = x(n) - ax(n-1) \tag{3.3}$$

onde x representa o sinal de voz. O efeito da pré-ênfase pode ser observado comparando o espectro do sinal original e o pós filtrado nas figuras 3.3 e 3.4 respectivamente.

3.2.1.2 Janelamento

O sinal de voz varia lentamente ao longo do tempo (chamado quase estacionário). Quando examinado em um curto período do tempo (janela entre 10 e 30ms), apresenta

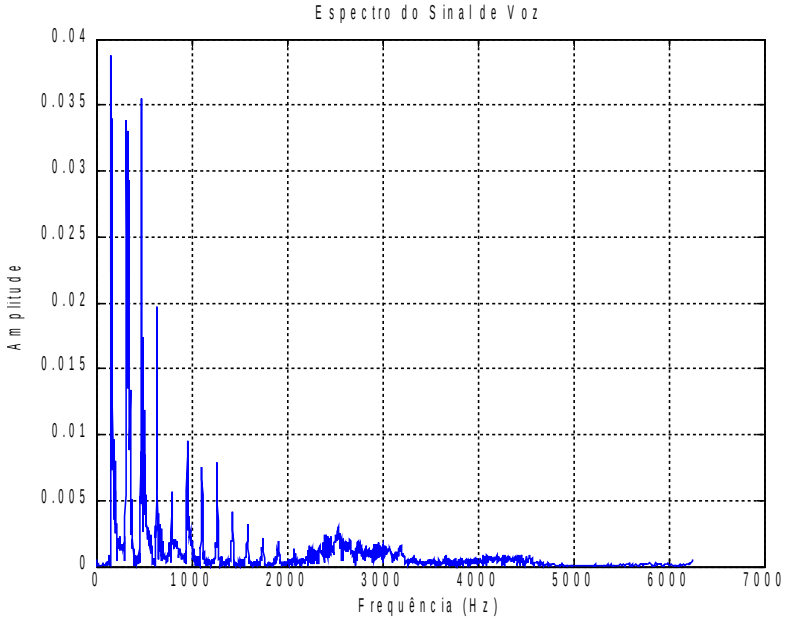

Figura 3.3: Espectro de uma amostra de voz antes da pré-ênfase

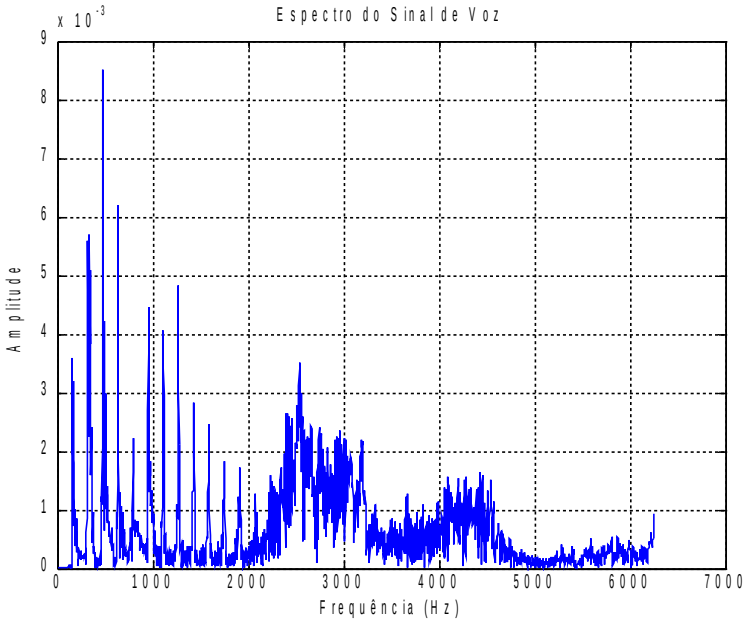

Figura 3.4: Espectro de uma amostra de voz após a pré-ênfase

características estacionárias (PEGORARO, 2000). Entretanto, quando observado num período longo do tempo, a característica do sinal muda, apresentando múltiplas variações espectrais dos diferentes sons pronunciados.

No gráfico da figura 3.5 pode-se observar uma característica quase estacionária da voz, tornando-se possível a análise *short term* do mesmo.

O janelamento de um sinal consiste em dividi-lo em janelas (*frames*) de largura N sendo uma potência de dois e sobrepostos segundo um fator M de deslocamento. Na

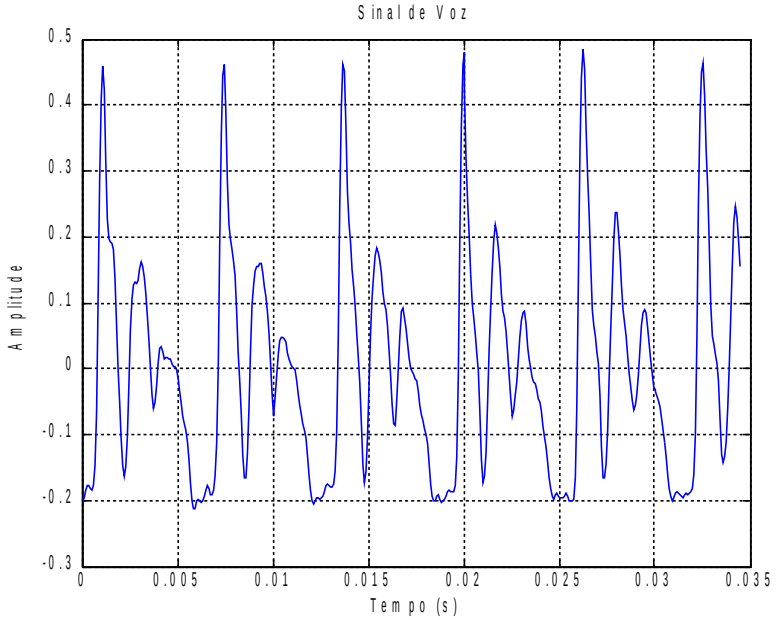

Figura 3.5: Exemplo de um sinal de voz

figura 3.6 as janelas estão sobrepostas em 66,6% (2/3).

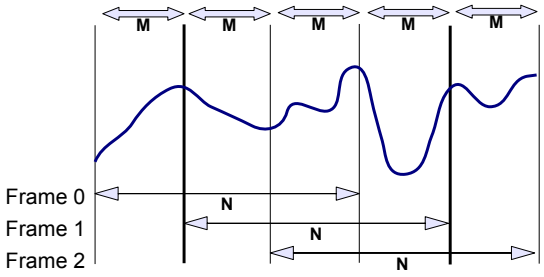

Figura 3.6: Divisão em *frames* do sinal de voz

Os diferentes tipos de janela são apresentados, sendo a *Hamming* e *Hanning* as mais usadas para reconhecimento de locutor.

Tabela 3.1: Diferentes tipos de janelas

Nome	***Equação***
Barllet	$1 - \left\lvert \frac{n-\frac{1}{2}N}{\frac{1}{2}N} \right\rvert, \quad 0 \leq n \leq N-1$
Blackman	$0,42 - 0,5\cos\left(\frac{2\pi n}{N-1}\right) + 0,08\cos\left(\frac{4\pi n}{N-1}\right), \quad 0 \leq n \leq N-1$
Hamming	$0,54 - 0,46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1$
Hanning	$0,5 - 0,5\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1$
Welch	$1 - \left(\frac{n-\frac{1}{2}N}{\frac{1}{2}N}\right)^2, \quad 0 \leq n \leq N-1$

Cada tipo de janela, apresentada no gráfico da figura 3.7, tem um efeito final diferente quando aplicada ao sinal. Se não for utilizada qualquer janela para a análise, melhor dizendo, se for aplicada uma janela retangular, poderá ocorrer o efeito do fenômeno de *leakage* – introdução de ruídos em frequência devidos à segmentação (Wikipedia – Leakage, 2009).

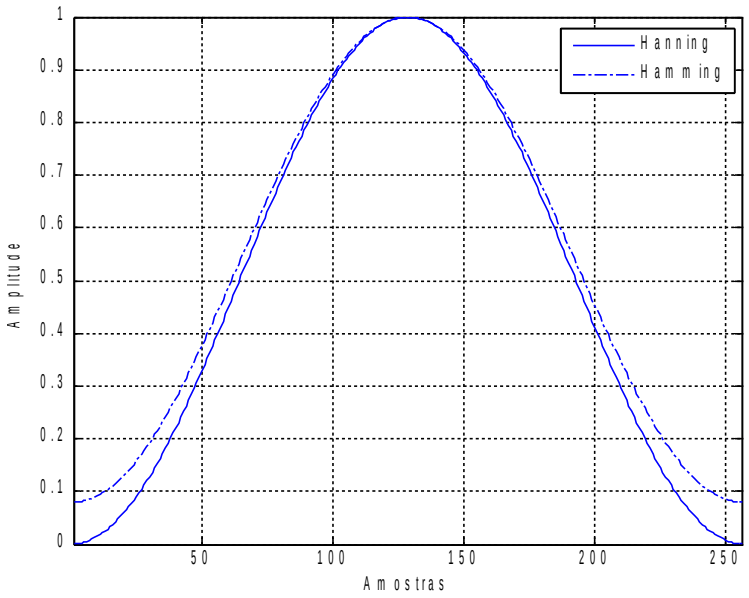

Figura 3.7: Janelas no domínio do tempo com 256 pontos

Após o processo de janelamento é iniciada a análise Fourier de cada segmento ou *frame* através da FFT (COOLEY; TUKEY, 1965). A análise de Fourier permite obter, em cada segmento analisado, a faixa de frequências que compõe o sinal de voz naquele segmento. A FFT é definida no conjunto de N amostras x_n, como segue na equação 3.4:

$$X_n = \sum_{x=0}^{N-1} x_n e^{-j2\pi kn/N}, \quad k = 0, 1, 2, ..., N-1 \tag{3.4}$$

Geralmente, o resultado X_n é um conjunto de números complexos cujas magnitudes representam a intensidade de cada componente de frequência (valores absolutos) do sinal de voz no segmento analisado.

3.2.1.3 Banco de Filtros Triangulares

A aplicação de um banco de filtros, com configuração apresentada na figura 3.8, é realizada após o cálculo da FFT. Cada filtro triangular está centrado nas frequências da escala Mel de modo a abranger o espectro da voz humana a ser analisada. Geralmente,

utilizam-se 20 filtros, espaçados aproximadamente em 150 mels, com largura de banda em torno de 300 mels.

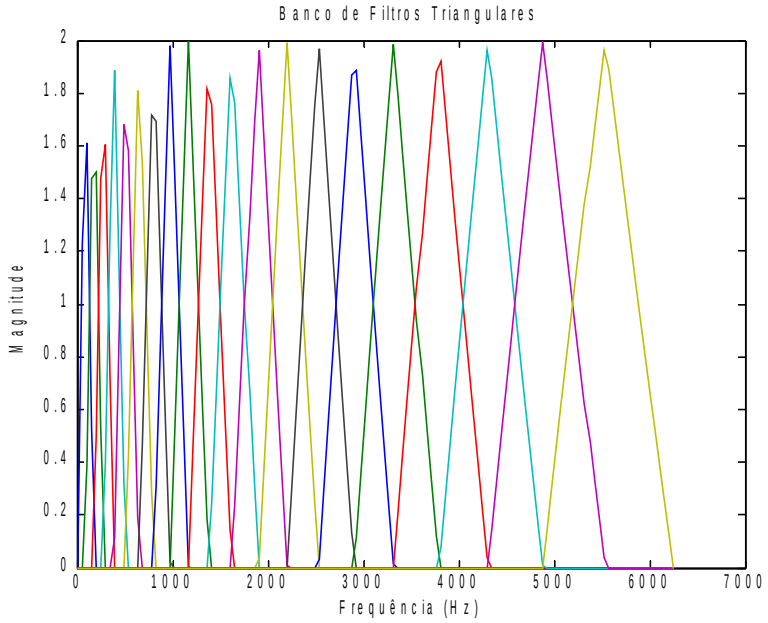

Figura 3.8: Banco de filtros triangulares

3.2.1.4 Função Logarítmica

Segundo o modelo de produção da fala, o sinal de voz $f(t)$ pode ser definido como o resultado da operação de convolução do sinal de excitação $u(t)$ com a resposta impulsiva do trato vocal $h(t)$ (PETRY; ZANUZ; BARONE, 2000).

$$f(t) = u(t) \otimes h(t) \tag{3.5}$$

A convolução é definida como o produto do cálculo da integral de duas funções produzindo uma terceira que é tipicamente vista como uma combinação dessas funções. A operação de convolução de duas funções no domínio do tempo corresponde a uma multiplicação, dessas funções, no domínio da frequência. Aplicando-se a função logarítmica nos dois sinais multiplicados, transforma-se tal multiplicação em sobreposição (soma) desses sinais, facilitando assim, a separação das duas partes:

$$\log(F\{f(t)\}) = \log(F\{u(t)\}) + \log(F\{h(t)\}) \tag{3.6}$$

onde $F\{x\}$ representa a aplicação da FFT.

3.2.1.5 Transformada Inversa de Fourier (IFFT)

Como a maior parte da energia espectral do sinal de excitação $u(t)$ e a resposta impulsiva do trato vocal $h(t)$ ocupam bandas espectrais diferentes, é possível utilizar as informações de apenas um deles. Sabe-se que o trato vocal varia mais lentamente que a excitação. Aplicando-se a transformada inversa de Fourier (equação 3.7) nesse sinal, volta-se para o domínio do tempo obtendo o cepstrum ou Coeficientes Cepstrais. Os primeiros coeficientes contêm informações relativas ao trato vocal, que está intimamente relacionada ao locutor.

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j2\pi kn/N}, \;\; n = 0, 1, 2, ..., N-1 \tag{3.7}$$

Desse modo, concluída a aplicação da transformada inversa, obtém-se um conjunto de coeficientes para cada janela do sinal da fala inicialmente obtido. Cada conjunto é conhecido como vetor de treinamento. Assim, os MFCC são constituídos de uma coleção de vetores de treinamento, os quais são utilizados no algoritmo de treinamento do sistema.

3.3 Treinamento

O processo de extração dos atributos da fala gera um conjunto de vetores os quais caracterizam um indivíduo. O treinamento do sistema é realizado utilizando a Quantização Vetorial através do algoritmo LBG.

3.3.1 Quantização Vetorial

A Quantização Vetorial (MAKHOUL; ROUCOS; GISH, 1985) é um processo de mapear vetores de um grande espaço vetorial para um número finito de regiões nesse espaço. Cada região é chamada de *cluster* e pode ser representada por um centroide chamado *codeword*. O conjunto de *codewords* é chamado *codebook*.

Após o processo de extração dos atributos da voz é construído um *codebook* específico para cada locutor usando os vetores de treinamento (Coeficientes Mel-Cepstrais referentes a cada *frame*). O algoritmo escolhido para tal tarefa chama-se LBG, que irá agrupar o conjunto dos coeficientes em um conjunto K de *codebooks*. A figura 3.9 mostra o diagrama do algoritmo (Digital Signal Processing Mini-Project, 2009):

1. Calcula o primeiro centroide do *codebook*; este é o centroide de todo o conjunto de vetores de treinamento;

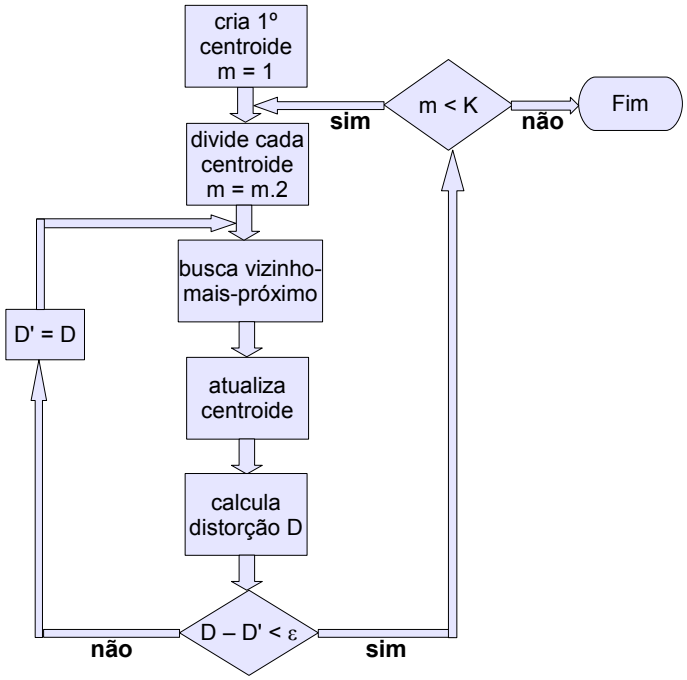

Figura 3.9: Diagrama de funcionamento do algoritmo LBG

2. Dobra o tamanho do *codebook* dividindo cada centroide segundo a regra:

$$y_n^+ = y_n(1+\epsilon) \tag{3.8}$$

$$y_n^- = y_n(1-\epsilon) \tag{3.9}$$

onde ϵ é um parâmetro de divisão $0,01 \leq \epsilon \leq 0,05$;

3. Busca de vizinho mais próximo: para cada vetor de treinamento, encontra-se o *codeword* no *codebook* que é mais perto (em termos de medida de similaridade), e atribuir na célula correspondente (associada com o codeword mais perto);

4. Atualização do centroide: atualizar o centroide em cada célula, a partir dos vetores de treinamento designados para aquela célula;

5. Iteração 1: repete os passos 3 e 4 até a distância (distorção) média ser menor que um limiar;

6. Iteração 2: repete os passos 2, 3 e 4 até o tamanho K do *codebook*.

3.4 Reconhecimento

A etapa de reconhecimento envolve basicamente uma busca por todo o *codebook* para encontrar o locutor que é mais próximo. Para isso, se faz necessária a extração MFCC da

locução a ser reconhecida, que são chamados, agora, de vetores de atributos.

O reconhecimento de um indivíduo consiste em comparar os vetores de atributos com os vetores de treinamento do codebook para encontrar o mais próximo. O processo de reconhecimento é composto das seguintes etapas:

1. Calcula-se os MFCC da amostra de voz a ser reconhecida, gerando assim, os vetores de atributos a serem reconhecidos;

2. Busca do centroide mais próximo: encontrar no *codebook* o centroide que é mais perto (em termos de medida de similaridade) do vetor a ser reconhecido;

3. Classificar o vetor a ser reconhecido no centroide mais próximo.

Pode-se notar que quanto maior o número de locutores registrados no sistema, maior será o número de comparações para o reconhecimento, e maior será o tempo de processamento. A figura 3.10 mostra uma ilustração do processo de criação do *codebook*.

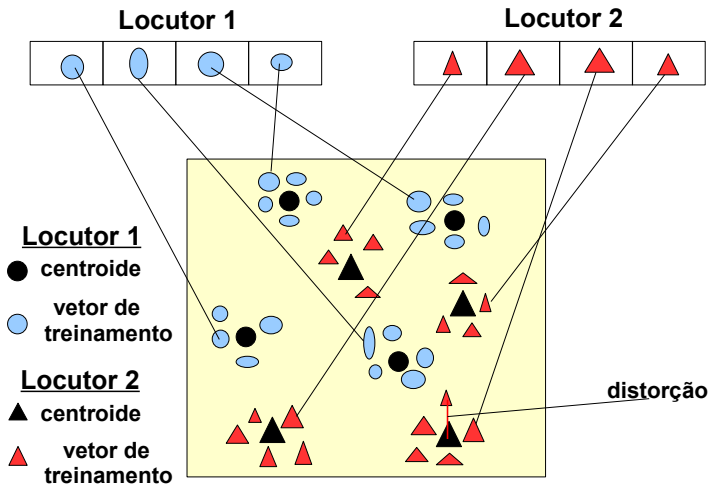

Figura 3.10: Diagrama ilustrativo da formação do *codebook* de QV. Cada locutor é diferenciado do outro baseado na localização dos centroides

4 IMPLEMENTAÇÃO E DESENVOLVIMENTO

Optou-se para o desenvolvimento do protótipo a programação de um sistema de identificação para o reconhecimento automático de locutor em modo independente de texto. O banco de vozes utilizado foi composto por um grupo fechado de 10 locutores distintos com a finalidade de avaliar o desempenho das técnicas selecionadas.

4.1 Banco de Vozes

Inicialmente, para obter-se um bom desempenho, é necessário um banco de vozes variado, que possua locuções de pessoas de gêneros e idades distintas, a fim de evitar resultados atribuídos ao acaso. Sendo assim, é necessário que cada locutor forneça um bom número de repetições de sua locução para que o sistema armazene seu padrão vocal e utilize algumas gravações para os testes. O local de gravação das amostras de voz deve ser o mais silencioso possível para evitar ruídos externos que podem prejudicar o reconhecimento posterior.

Cada locutor forneceu 7 gravações de frases balanceadas para o treinamento do sistema, e 1 frase para o reconhecimento. Cinco, dentre os 10 locutores, forneceram duas palavras distintas para a realização do reconhecimento a partir de palavras isoladas. As frases balanceadas, extraídas do trabalho de (ZANUZ et al., 2004), são caracterizadas por um conjunto de informações fonéticas e fonológicas variadas, de modo a cobrir o máximo de sons da língua utilizada, neste caso, a Língua Portuguesa. Geralmente, em grandes sistemas, faz-se a construção de um *Corpus*, que é um vasto conjunto de frases balanceadas. As tabelas 4.1, 4.2 e 4.3 apresentam maiores detalhes sobre as locuções.

Tabela 4.1: Frases balanceadas usadas no treinamento (BAL)

Número	***Frase***
1	O presidente da República faz advertência ao ministro da justiça.
2	O jovem atacante convenceu fácil na partida contra o México.
3	O Zorro é outro dos filmes muito procurados nas locadoras atualmente.
4	A primeira maior guerra de todas foi entre o bem e o mal, o céu e a terra.
5	É melhor engomar os lençóis azuis debaixo do sol.
6	Eu prefiro ser essa metamorfose ambulante.
7	Do que ter aquela velha opinião formada sobre tudo.

Tabela 4.2: Frase utilizada no reconhecimento de locutor

Número	***Frase***
8	Meu nome é <nome completo> e tenho <idade> anos.

Tabela 4.3: Palavras utilizadas no reconhecimento de locutor

Número	***Palavra***
1	Chuveiro
2	Tartaruga

4.2 Ambiente de Gravação

A gravação foi executada em uma sala pequena e silenciosa. Foi usado um microfone profissional conectado a um computador, e o software CoolEdit Pro 2.0 (CoolEdit Pro 2.0, 2009) para a gravação e criação dos arquivos de áudio. Após cada gravação, foi aplicado um ganho de 26dB em cada locução, de modo a realizar a pré-amplificação. Não foi executado qualquer processamento para eliminação de ruídos ou silêncios sobre as amostras. A tabela 4.4 apresenta maiores detalhes da aquisição das vozes.

4.3 Conjunto de Locutores

O conjunto de locutores (tabela 4.5) foi composto por 10 pessoas, todas brasileiras de diferentes estados do país, na faixa etária entre 10 anos e 49 anos.

Tabela 4.4: Dados de gravação

Descrição	***Utilizado***
Microfone	SM-58 P4 LeSon
Taxa de Amostragem	12,5KHz
Resolução	16 bits/amostra
Modo de Gravação	Mono
Formato	*.wav*
Locutores Femininos	4
Locutores Masculinos	6

Tabela 4.5: Composição do conjunto de locutores

Locutor	***Nome***	***Sexo***	***Idade***	***Estado***
1	André	M	10	Rio Grande do Sul
2	Egon	M	25	Rio Grande do Sul
3	Lia	F	47	Rio Grande do Sul
4	Rodrigo	M	22	Rio Grande do Sul
5	Simone	F	23	Rio Grande do Sul
6	Danilo	M	49	Rio Grande do Sul
7	Santos	F	26	Rio de Janeiro
8	Sabatini	M	21	São Paulo
9	Barbosa	F	44	São Paulo
10	Teixera	M	26	Tocantins

4.4 Característica do Sistema

A linguagem de programação foi o ANSI C, devido a possível utilização na programação de um DSP, ideia inicialmente concebida mas abandonada momentaneamente devido à complexidade de implementação no intervalo de tempo disponível para este trabalho. Utilizou-se o ambiente de programação Dev-C++ v4.9.9.2 para Windows rodando num PC com processador Intel Core 2 Duo 2.20GHz, memória RAM de 2GB DDR2, placa de som integrada SIGAMTEL STAC 92XX C-Major HD Audio.

Como ilustrado na tabela 4.4, as vozes foram coletadas numa taxa de amostragem de 12,5KHz, com resolução de 16 bits/amostra e armazenadas em arquivos no formato *.wav*. Foi utilizado na pré-ênfase o valor de a=0,95. Após isso, as amostras foram divididas em janelas de 20,48ms (256 amostras) para a aplicação da janela de *Hamming*.

Após a etapa de janelamento, inicia-se a análise Fourier dos segmentos através da FFT, – código extraído de (ORFANIDIS, 1996) – seguido da aplicação dos filtros triangulares

e da função logarítmica. Às janelas, a partir disso, são aplicadas o algoritmo da IFFT para a conclusão da extração dos MFCC. Porém, as janelas, antes desse último procedimento, contêm apenas valores reais, anulando a parte imaginária da IFFT. A partir disso, optou-se pela aplicação da DCT (*Discrete Cosine Transform*) (Wikipedia – Discrete Cosine Transform, 2009) no lugar da IFFT, devido à sua simplicidade de implementação.

O protótipo foi implementado de modo a gerar 1 *codebook* por locução, sendo cada um deste com 16 centroides. Portanto, o número total de *codebooks* gerados para 10 locutores na fase de treinamento do sistema foi de 70. O algoritmo LBG utilizou o valor de 1% para o limiar de distorção. Este limiar define a distância euclidiana mínima necessária para a associação dos vetores de treinamento aos centroides. Em outras palavras, as etapas de busca do vizinho mais próximo e de atualização dos centroides, do LBG, são repetidas até que a distância média dos vetores de treinamento em relação ao centroide seja menor que 1%.

Tabela 4.6: Dados do protótipo

Descrição	***Utilizado***
Coeficiente de Pré-ênfase	0,95
Extrator de Atributos	MFCC
Número de MFCC	20/janela
Tamanho da janela (N)	256 amostras (20,48ms)
Tipo da Janela	*Hamming*
Técnica de Classificação	Algoritmo LBG de QV
Número de Centroides	16/*codebook*
Medida de Distorção	Distância Euclidiana
Limiar de Distorção	1%

5 TESTES E RESULTADOS

Este capítulo apresenta os testes e resultados obtidos a partir das fases de treinamento e reconhecimento do protótipo. De um modo geral, as análises dos testes foram divididas entre as fases de treinamento e reconhecimento. No treinamento, utilizou-se 7 frases balanceadas, enquanto que no reconhecimento, uma frase e duas palavras isoladas e distintas.

5.1 Dados dos Testes do Treinamento

O primeiro teste realizado foi o treinamento do protótipo a partir das 7 frases mostradas na tabela 4.1 para um número de 10 locutores, segundo a configuração mostrada na tabela 4.6. A tabela 5.1 apresenta o número de janelas totais obtidas por locutor e as durações totais das gravações, utilizando o fator M de deslocamento igual à 85. O fator de deslocamento diz respeito à parcela de sobreposição dos *frames* obtidos na etapa de janelamento (figura 3.6).

Tabela 5.1: Número de janelas e duração das gravações por locutor utilizando M=85

Locutor	***Número de Janelas***	***Duração (s)***
1	3494	23,875
2	3574	25,393
3	3242	24,177
4	3586	22,035
5	3716	23,869
6	4591	31,335
7	2823	19,299
8	4264	29,107
9	3157	24,212
10	4087	27,905

Após a análise das janelas obtidas, fez-se o teste do tempo de treinamento total do

protótipo, que consistiu em alterar o fator M de deslocamento de maneira a sobrepor as janelas nas porcentagens de 75%, 66,6%, 50%. A tabela 5.2 mostra o comportamento dessa alteração.

Tabela 5.2: Variações no número de janelas e no tempo de treinamento de acordo com valores de M para 10 locutores

Fator de Deslocamento (M)	***Número de Janelas***	***Tempo de Treinamento***
64 amostras (5,12ms)	49045	104s
85 amostras (6,8ms)	36534	76,5s
128 amostras (10,24ms)	23832	48,8s

A medição de tempo foi realizada diversas vezes através da biblioteca *time.h* da linguagem C, gerando uma média aritmética dos tempos. Porém, esta medida apesar de sua simples obtenção, não é de muita representabilidade quanto aos elementos envolvidos na execução, que podem variar muito devido ao processador, memória, sistema operacional, etc.

5.2 Dados dos Testes do Reconhecimento

Para o reconhecimento, foram realizados dois tipos de testes: teste com frase e teste com palavras isoladas. O primeiro deles utilizou a frase da tabela 4.2, comparando o número de locutores registrados com as porcentagens de acertos de identificação segundo diferentes valores do fator M de deslocamento, como mostra a tabela 5.3. Inicialmente, treinou-se o protótipo com dois locutores, acrescentando um a um até o último.

Tabela 5.3: Porcentagens de identificações corretas no reconhecimento com frase

Nº de Locutores Registrados	***M = 64***	***M = 85***	***M = 128***
2	100%	100%	100%
3	100%	100%	100%
4	100%	100%	100%
5	100%	100%	100%
6	100%	100%	100%
7	100%	100%	100%
8	100%	100%	100%
9	100%	100%	100%
10	90%	90%	90%

Tabela 5.4: Tempo de Reconhecimento total na identificação de 10 locutores de acordo com valores de M

Nº Locutores	***M = 64***	***M = 85***	***M = 128***
10	30,420s	30,529s	30,342s

O segundo teste, realizado com palavras, treinou o protótipo com os 10 locutores a partir das 7 frases balanceadas. Para o reconhecimento, entretanto, apenas 5 locutores participaram. As palavras ocupadas são apresentadas na tabela 4.3, e os valores obtidos, na tabela 5.5.

Tabela 5.5: Porcentagens de identificações corretas no reconhecimento com palavras de 5 locutores com M=85, dentre 10 locutores registrados

Nº Loc. Participantes	***Palavra***	***Porcentagem de acertos***	***Tempo de Rec.***
5	Chuveiro	40%	3,651s
5	Tartaruga	80%	3,510s

5.3 Análise dos Resultados

Nos testes realizados para o treinamento (tabela 5.1) pôde-se observar que, ao alterar o valor do fator M de deslocamento, o tempo de treinamento também se alterou. Isto pode ser claramente explicado devido à variação do número de janelas para análise, pois para cada *frame* gerado é necessária a análise espectral através dos MFCC (ver tabela 5.2).

No primeiro teste feito para o reconhecimento, observou-se que, apesar da variação do fator M, o protótipo não mostrou nenhuma mudança na porcentagem de acertos, mesmo com o aumento de janelas para análise. O segundo teste buscou conferir o desempenho do protótipo utilizando palavras isoladas para o reconhecimento, pois elas carregam poucas informações acerca de cada locutor. Obteve-se 2 acertos, dentre 5 locutores, para a palavra *Chuveiro* e 3 acertos para a palavra *Tartaruga*. Tais resultados mostram que o desempenho do reconhecimento depende da palavra pronunciada e da quantidade de informações fonéticas fornecidas pelo locutor.

6 CONCLUSÕES

Tendo como um dos principais objetivos aplicar os conhecimentos obtidos no curso de Ciência da Computação aliado à outras áreas do conhecimento, este trabalho apresentou o desenvolvimento do protótipo de um sistema de reconhecimento automático de locutor em modo independente de texto, utilizando amostras de vozes de pessoas para testes. Foi descrita a modelagem da produção da voz humana a fim de apresentar os tipos de sons produzidos e suas características. Em seguida, foi executada a aquisição das locuções a partir de 10 pessoas, sendo 4 do sexo feminino e 6 do sexo masculino, envolvendo um conjunto de 7 frases balanceadas para compor o banco de vozes, uma frase e duas palavras isoladas para testes de reconhecimento. O sistema se baseou na extração dos Coeficientes Mel-Cepstrais para caracterizar cada locutor e adotou o algoritmo LBG de Quantização Vetorial para a classificação dos padrões, gerando 1 *codebook* para cada amostra de voz, sendo cada um com 16 centroides.

O número de locutores registrados no sistema é de fundamental importância para a medida da performance em sistemas de reconhecimento de locutor. Observou-se que a porcentagem de acertos diminuiu de acordo com o aumento de pessoas. Entretanto, a variação do fator M de deslocamento não interfiriu no resultado final. O tempo de treinamento e reconhecimento estão ligados diretamente com a duração da locução e, consequentemente, com o número de janelas para análise.

O desempenho geral do protótipo foi satisfatório para o número de locuções e locutores utilizados. Entretanto, para uma maior abrangência de testes necessita-se de maiores repetições, pessoas, frases, e até mesmo a consideração da presença de ruídos externos.

Com esse trabalho, pôde-se adquirir muitos conhecimentos teóricos e práticos da área de Reconhecimento de Voz, destacando a importância e necessidade da interdisciplinaridade para o desenvolvimento de sistemas com maiores desempenhos.

6.1 Trabalhos futuros

O trabalho apresentado serviu como iniciação ao tema de Reconhecimento de Voz, introduzindo conceitos, técnicas e mostrando a estrutura básica de um sistema. Pensando em um trabalho futuro, seria pertinente para a melhoria do desempenho do sistema:

- o aumento no número de frases balanceadas para treinamento;
- o processo de aquisição da fala em tempo real sem a utilização de arquivos *.wav*;
- a aplicação de filtros para remoção de ruídos externos;
- a otimização dos algoritmos utilizados, reduzindo a utilização dos recursos de processamento.

Aliado à esses tópicos, sugere-se a implementação do sistema num DSP para uso embarcado.

REFERÊNCIAS

CARICATI, A. M.; WEIGANG, L. Reconhecimento de Locutores em Língua Portuguesa com Modelos de Redes Neurais e Gaussianos. **V Congresso Brasileiro de Redes Neurais**, UNB, p.25–30, 2001.

CHOU, W.; JUANG, B. H. Pattern Recognition in Speech and Language Processing. **Ed. CRC Press.**, [S.l.], p.162–240, 2003.

CoolEdit Pro 2.0. Disponível em: http://www.adobe.com, último acesso em junho de 2009.

COOLEY, W. J.; TUKEY, J. W. An algorithm for the machine computation of the complex Fourier series. **Mathematical Computation, Vol. 19**, [S.l.], p.297–301, 1965.

Digital Signal Processing Mini-Project. Disponível em: http://www.ifp.illinois.edu/∼mminhdo/teaching/speaker_recognition, último acesso em junho de 2009.

HE, J.; LIU, L.; PALM, G. A Discriminative Training Algorithm for VQ-based Speaker Identification. **IEEE Transactions on Speech and Audio Processing, Vol. 7, Nº 3**, [S.l.], p.353–356, 1999.

LINDE, Y.; BUZO, A.; GRAY, R. M. An Algorithm for Vector Quantizer Design. **IEEE Transactions on Communications, Vol. com-28, Nº 1**, [S.l.], 1980.

MAFRA, A. T. **Reconhecimento Automático de Locutor em Modo Independente de Texto por Self-Organizing Maps**. Dissertação de Mestrado em Engenharia Mecatrônica, USP, 2002.

MAKHOUL, J.; ROUCOS, S.; GISH, H. Vector Quantization in Speech Coding. **IEEE, Vol. 73, Nº 11**, [S.l.], p.1551–1587, 1985.

MARTINS, J. A. **Avaliação de Diferentes Técnicas para reconhecimento de fala**. Tese de Doutorado em Engenharia Elétrica, UNICAMP, 1997.

ORFANIDIS, S. J. **Introduction to Signal Processing.** [S.l.]: Prentice Hall, 1996. 513–532p.

PEGORARO, T. F. **Algoritmos Robustos de Reconhecimento de Voz Aplicados a Verificação de Locutor**. Dissertação de Mestrado em Engenharia Elétrica, UNICAMP, 2000.

PETRY, A. **Reconhecimento Automático de Locutor Utilizando medidas de invariantes dinâmicas não-lineares**. Tese de Doutorado em Ciência da Computação, UFRGS, 2002.

PETRY, A.; ZANUZ, A.; BARONE, D. A. C. Reconhecimento Automático de Pessoas pela Voz através de Técnicas de Processamento Digital de Sinais. **III Workshop em Internet, Linux e Aplicações.**, UFRGS, 2000.

RABINER, R. L.; JUANG, B.-H. **Fundamentals of Speech Recognition**. [S.l.]: Prentice-Hall, 1993.

RUNSTEIN, O. F. **Sistema de Reconhecimento de Fala Baseado em Redes Neurais Artificiais**. Tese de Doutorado em Engenharia Elétrica, UNICAMP, 1998.

SAMUDRAVIJAYA, K. Speech and Speaker Recognition: a tutorial. **Int. Workshop on Tech. Development in Indian Languages**, Kolkata, 2003.

SIGURDSON, S.; PETERSEN, K. B.; LEHN-SCHIOLER, T. Mel frequency cepstral coefficients: an evaluation of robustness of mp3 encoded music. **ISMIR**, Vicoria, Canadá, 2006.

SOONG, K. F.; ROSENBERG, E. A.; JUANG, B.-H.; RABINER, R. L. A Vector Quantization Approach to Speaker Recognition. **AT&T Technical Journal, Vol. 66**, [S.l.], p.14–26, 1987.

The Canonical WAVE PCM soundfile format. Disponível em: http://ccrma.stanford.edu/courses/422/projects/WaveFormat, último acesso em junho de 2009.

Wikipedia – Discrete Cosine Transform. Disponível em: http://en.wikipedia.org/wiki/Discrete_cosine_transform, último acesso em junho de 2009.

Wikipedia – Leakage. Disponível em: http://en.wikipedia.org/wiki/Leakage, último acesso em junho de 2009.

ZANUZ, A.; SCHRAMM, M. C.; FREITAS, L. F. R.; BARONE, D. A. C. Uma Base de Dados Vocais para a Língua Portuguesa Falada no Brasil. **VII Seminário de Iniciação Cientifíca**, ULBRA - Guaíba, 2004.