Virginio Velloso Freire

Desenvolvimento de um Sistema de Reconhecimento Pessoal de Fala Baseado em Características Prosódicas do Sinal de Voz

Campina Grande, PB

7 de outubro de 2014

Virginio Velloso Freire

Desenvolvimento de um Sistema de Reconhecimento Pessoal de Fala Baseado em Características Prosódicas do Sinal de Voz

Projeto de Conclusão de curso apresentado ao Curso de Graduação em Engenharia Elétrica da Universidade Federal de Campina Grande, em cumprimento às exigências para obtenção do Grau de Engenheiro Eletricista.

Universidade do Federal de Campina Grande – UFCG

Engenharia Elétrica

Programa de Graduação

Orientador: Marcelo Sampaio de Alencar

Campina Grande, PB

7 de outubro de 2014

Virginio Velloso Freire
Desenvolvimento de um Sistema de Reconhecimento Pessoal de Fala Baseado em Características Prosódicas do Sinal de Voz/ Virginio Velloso Freire. – Campina Grande, PB, 7 de outubro de 2014-
74 p.

Orientador: Marcelo Sampaio de Alencar

Monografia – Universidade do Federal de Campina Grande – UFCG
Engenharia Elétrica
Programa de Graduação, 7 de outubro de 2014.

1. Reconhecedor fonético. 2. Reconhecedor de fala. 3. Segmentador fonético. 4. Codificador de voz. I. Marcelo Sampaio de Alencar. II. Universidade Federal de Campina Grande. III. Curso de Engenharia Elétrica. IV. Desenvolvimento de um Sistema de Reconhecimento Pessoal de Fala Baseado em Características Prosódicas do Sinal de Voz

Virginio Velloso Freire

Desenvolvimento de um Sistema de Reconhecimento Pessoal de Fala Baseado em Características Prosódicas do Sinal de Voz

Projeto de Conclusão de curso apresentado ao Curso de Graduação em Engenharia Elétrica da Universidade Federal de Campina Grande, em cumprimento às exigências para obtenção do Grau de Engenheiro Eletricista.

Trabalho aprovado. Campina Grande, PB, 7 de outubro de 2014:

Marcelo Sampaio de Alencar
Orientador

Professor
Convidado

Campina Grande, PB
7 de outubro de 2014

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.

Agradecimentos

Agradeço primeiramente à minha família, meus amigos e minha namorada que me apoiram nos momentos mais difíceis que esse curso me proporcionou.

Agradecimentos especias são direcionados ao Departamento de Engenharia Elétrica da Universidade Federal de Campina Grande, em especial ao Instituto de Estudos Avançados em Comunicações e ao meu orientador Marcelo Sampaio de Alencar, por ter me ajudado a realizar grande parte dos meus trabalhos científicos.

Resumo

Este Trabalho apresenta o desenvolvimento de um método diferenciado para o reconhecimento pessoal de fala que tem como principal característica o uso de informações prosódicas do sinal, tais como energi, *pitch* e autocorrelação, para reconhecer fonemas.

Sua implementação está dividida na criação de um sistema de segmentação fonética e o sistema de reconhecimento fonético em si. O segmentador é usado principalmente como entrada do reconhecedor de fala, e por isso seu desenvolvimento é importante, pois é responsável por separar todos os fonemas encontrados em uma frase em questão. Esse sistema não depende de locutor e não necessita de fase de treinamento, ao contrário dos que se encontram no mercado.

O reconhecedor de fala aqui proposto é responsável por identificar qual frase, palavra ou fonema foi pronunciado pelo orador sob teste. Por fim, o trabalho explicita resultados de ambos os sistemas e os compara com dados atuais, com a finalidade de poder melhorar os programas desenvolvidos com a adição de métodos mais recentes.

Palavras-chaves: Reconhecimento de fala. Reconhecimento fonético. Segmentador fonético. Codificador de voz.

Abstract

This work presents the development of a differentiated method for personal speech recognition, whose main purpose is the use of prosodic information of the signal, such as energy, textit pitch and autocorrelation to recognize phonemes.

Its implementation is divided in creating a phonetic segmentation system and the phonetic recognition system itself. The segmenter is mainly used as input to the speech recognizer, and therefore its development is important because it is responsible for separating all phonemes found in a phrase. This system is not dependent upon the speaker and does not require training phase, unlike those already on the market.

The speech recognizer proposed here is responsible for identifying which phrase, word or phoneme was pronounced by the speaker under test. Finally, the work explains results of both systems and compares them with current data, in order to be able to improve the programs developed with the addition of recent methods.

Key-words: Speech recognition. Fonetic recognition. Fonetic segmentation. Voice Coder.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ASR	Automatic Speech Recognition
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
HMM	Hidden Markov Models
STT	Speech To Text

Lista de símbolos

E	Energia de um sinal discreto no tempo
$x(n)$	Sinal discreto no tempo
m	Variável discreta
n	Variável discreta
$w()$	Função Janela
$h()$	Função janela ao quadrado
E_n	Energia de tempo curto
$X(\omega)$	Transformada de Fourier de um sinal discreto no tempo
R	Parte real da equação
I	Parte imaginária da equação
N	Número de pontos
W_N^{kn}	Fator de fase
y	Espectro normalizado
x	Espectro não normalizado
r_{xy}	Correlação cruzada entre dois sinais
r_x	Autocorrelação
$f(x)$	Função de frequência do espectro
f_s	Frequência de amostragem

Sumário

1 Introdução

1.1 Considerações iniciais

O reconhecimento de fala é utilizado em tarefas diárias, tais como ligação por voz, controle de aparelhos domésticos (automação residencial) e processamento de fala para texto,como processadores de palavras ou emails. Essas tarefas tornam-se cada vez mais importantes e facilitam o trabalho de seus usuários. Em ciência da computação e engenharia elétrica, o reconhecimento de fala é a conversão de palavras faladas em texto, e também é conhecido como reconhecimento automático de fala (*Automatic Speech Recognition* – ASR), reconhecimento de fala por computador, ou apenas fala para texto (*Speech to Text* – STT).

Um sistema de reconhecimento de voz tem como objetivo determinar qual palavra, frase ou sentença foi pronunciada. Esses sistemas possuem diversas aplicações, tais como, atendimento automático, interfaces para computadores pessoais, controle de equipamentos, sistemas de codificação de voz.

O reconhecedor de fala pode ser classificado como dependente ou independente de locutor. No primeiro caso, ele é caracterizado por possuir informações prévias da voz dos usuários que deverão utilizar o sistema. Por outro lado, o reconhecedor independente de locutor é implementado de forma a reconhecer com uma boa taxa de acerto a fala pronunciada por qualquer locutor.

Além disso, os reconhecedores de fala podem ser classificados como irrestrito, no qual são capazes de reconhecer qualquer sentença pronunciada, ou restrito, quando estão programados a reconhecer sentenças pré-definidas.

Este trabalho de conclusão de curso tem como objetivo o desenvolvimento de um sistema de reconhecimento de fala dependente de locutor. Desta forma, cada usuário deverá pronunciar sentenças pré-estabelecidas, que serão segmentadas por meio de um segmentador fonético e, em seguida, amostras de cada fonema serão armazenados, porém dada a complexidade do reconhecimento fonético, primeiramente o reconhecimento será feito em palavras. O reconhecimento de fala se dará por meio da comparação de características prosódicas da fala pronunciada com os fonemas e frases armazenadas em um banco de dados.

1.2 Motivação e objetivos

Este trabalho foi desenvolvido para servir como base de um codificador de voz paramétrico pessoal de baixa taxa a ser utilizado principalmente em sistemas de comunicações móveis celulares, já que a taxa de transmissão necessária atualmente é de cerca de 64 k bits/s e o codificador de voz a ser desenvolvido pode diminuir tal taxa para valores proxímos a 150 *bits* por segundo. Uma diminuição de aproximadamente 4250% em relação ao atual, o que pode fazer com que mais celulares possam usar um canal de mesma largura de banda, aumentando o número de possíveis usuários e baixando os custos da telefonia móvel celular em geral.

1.2.1 Objetivos Específicos

- Estudar os sistemas de reconhecimento de fala disponíveis;
- Elaborar bancos de unidades acústicas por meio da segmentação de frases pré-selecionadas em palavras, fonemas, sílabas e encontros vocálicos;
- Desenvolver o sistema de reconhecimento automático de fala utilizando principalmente a correlação entre o banco de unidades fonéticas e a locução sob teste;
- Verificar a eficiência do sistema de reconhecimento de fala produzido a partir das medidas de qualidade nos testes objetivos e subjetivos.

1.3 Estrutura do trabalho

O atual capítulo trata de fazer uma breve introdução ao leitor do sistema de reconhecimento pessoal de fala explicando os usos, métodos usados na atualidade, resultados obtidos por tais métodos, conceitos de reconhecedores de fala e sua importância, bem como os atributos necessários ao desenvolvimento de um sitema do tipo. Os objetivos principais do trabalho também são explicados sucintamente.

O Capítulo 2 busca realizar um breve resumo teórico com os principais assuntos envolvidos para a completa realização do programa. Serão descritos métodos para encontrar a DFT e FFT, assim como será apresentada a teoria por trás da correlação cruzada e autocorrelação de sinais de voz.

A segmentação fonética e o programa desenvolvido para isso neste trabalho será alvo da discussão do Capítulo 3, onde o sistema será descrito e desenvolvido por meio da utlização de duas etapas principais, são elas a identificação de regiões audíveis e a identificação de novas fronteiras dentro das regiões audíveis.

O sistema de reconhecimento pessoal de fala, que será o objetivo principal deste texto, é desenvolvido utilizando o algoritmo da correlação cruzada no quarto capítulo. Juntamente com isso, será explicado o motivo pelo qual o sistema é teoricamente possível de se implementar.

Os resultados serão finalmente apresentados no Capítulo 5, porém antes será necessário mostrar todos os passos realizados para programar o sistema e os resultados obtidos para a segmentação fonética. Feito isso os resultados do reconhecimento de fala serão divididos em reconhecimento de palavras e reconhecimento de fonemas. Os fonemas ainda estão divididos em apenas vogais e todos os fonemas. Uma pequena avaliação qualitativa dos resultados do sistema desenvolvido será realizada ao final deste capítulo.

Por fim, serão realizadas as considerações finais, explicitando de maneira clara as contirbuições dadas pelo trabalho e mostrando as dificuldades e as formas de superá-las. Ao fim do Capítulo 6 serão propostos trabalhos futuros que foram pensados com base nos resultados obtidos.

2 Desenvolvimento teórico

Essa parte introduz algumas definições e informações que virão a ser úteis no decorrer do trabalho. Será apresentada uma base para a compreensão de teorias que incluem energia de tempo curto, DFT (*Discrete Fourier Transform* - Transformada discreta de Fourier), FFT (*Fast Fourier Transform* - Transformada Rápidade Fourier), normalização do espectro e algoritmos como os da correlação cruzada e autocorrelação.

2.1 Energia de tempo curto

A voz audível apresenta amplitude maior que a voz inaudível ou o silêncio (ruído de fundo). A energia de curta duração de um sinal provê uma representação conveniente que reflete as variações de amplitude do sinal.

A energia de um sinal discreto no tempo pode ser definida como

$$E = \sum_{m=-\infty}^{\infty} x^2[m], \tag{2.1}$$

logo, a energia de tempo curto poderá ser dada por

$$E_n = \sum_{m=-\infty}^{\infty} [x(m) \cdot w(n-m)]^2, \tag{2.2}$$

em que w é a função janela. Ainda, pode-se reescrever a equação acima como

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m) \cdot h(n-m), \tag{2.3}$$

na qual $h(n) = w^2(n)$.

É possível perceber que se a janela for muito longa e constante em amplitude, E_n variará muito pouco em relação ao tempo. Essa janela seria o equivalente a um filtro passa-baixa de banda (muito) estreita, porém uma janela estreita demais não consegue produzir uma função suave de energia. Mas se a janela for da ordem de vários picos do sinal, E_n não irá refletir as variações do sinal. Este conflito é de grande importância na representação em tempo curto de sinais de voz.

Desta maneira, o tamanho da janela varia desde 20 amostras, para uma voz aguda de mulher, a 250 amostras para uma voz grave de homem. Na prática, para uma frequência de amostragem do sinal de 10 k Hz, deve-se utilizar uma janela da ordem de $100 < N < 200$ amostras, ou seja, uma janela de duração entre 10 e 20 milissegundos.

A maior significância de E_n está em conseguir destinguir entre segmentos com voz audível e inaudível, de modo que os valores de E_n são significativamente maiores para sinas audíveis, sendo útil para determinar o momento em que um sinal audível torna-se inaudível e vice-versa. Ainda por cima, se o sinal for de boa qualidade, é possível distinguir a voz do silêncio.

2.2 DFT

A DFT nada mais é que apenas um tipo de transformada de Fourier para tempo discreto *x(n)* e não de um sinal continuo *x(t)*. A equação de Fourier torna-se então:

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n)e^{j\omega n}. \tag{2.4}$$

A função principal dessa transformada é a de transformar a variável discreta n na variável ω, o que significa transformar sinais do domínio do tempo para o domínio da frequência.

Assumindo que o sinal de voz gravado *x(n)* é um vetor que consiste em valores complexos, como $x(n) = R + I$, na qual R é a parte real do valor e I a imaginária. Como:

$$e^{j\omega n} = cos(\omega n) + j \cdot sen(\omega n), \tag{2.5}$$

então:

$$x(n)\cdot e^{j\omega n} = (R+I)\cdot[\cos(\omega n)+j\cdot sen(\omega n)] = R\cdot\cos(\omega n)+R\cdot j\cdot sen(\omega n)+I\cdot\cos(\omega n)+I\cdot j\cdot sen(\omega n). \tag{2.6}$$

Rearranjando as partes real e imaginaria da equação, é possível obter:

$$x(n) \cdot e^{j\omega n} = [R \cdot \cos(\omega n) + I \cdot \cos(\omega n)] + [R \cdot j \cdot sen(\omega n) + I \cdot j \cdot sen(\omega n)]. \tag{2.7}$$

Substituindo na equação (2.4):

$$x(\omega) = \sum[R \cdot \cos(\omega n) + I \cdot \cos(\omega n)] + \sum j[R \cdot sen(\omega n) + I \cdot sen(\omega n)]. \tag{2.8}$$

Como, geralmente, apenas o valor real do sinal $x(n)$ é usado, a parte imaginária da equação é $I = 0$. Daí, a transformada de Fourier é:

$$X(\omega) = \sum_{n=-\infty}^{\infty} [R \cdot \cos(\omega n)] + \sum_{n=-\infty}^{\infty} [j \cdot R \cdot sen(\omega n)]. \quad (2.9)$$

A análise acima são os passos gerais para se construir a transformada de Fourier programando o fator de frequência de computação que consiste na parte real e imaginária com a magnitude do sinal.

2.3 FFT

A transformada rápida de Fourier ainda é essencialmente a DFT utilizada para transformar um sinal de tempo discreto do domínio do tempo para o domínio da frequência. A diferença se dá devido a FFT ser mais rápida e mais eficiente computacionalmente falando. Existem várias maneiras de se aumentar a eficiência computacional da DFT, mas o algoritmo mais usado para o cálculo da FFT é o *Radix-2 FFT Algorithm* (PROAKIS; MANOLAKIS,)

Como a FFT ainda é o cálculo da DFT, então é conveniente investigar a FFT considerando primeiramente a equação da DFT com N pontos, sabendo que $W_N^{kn} = e^{j\omega_k n}$:

$$X(\omega) = \sum_{n=0}^{N-1} x(n) W_N^{kn}, k = 0, 1, 2, ..., N-1. \quad (2.10)$$

Inicialmente, separa-se $x(n)$ em duas partes: $x(\acute{i}mpar) = x(2m+1)$ e $x(par) = x(2m)$, onde $m = 0, 1, 2, ..., (N/2) - 1$. Então a equação da DFT com N pontos também se torna duas partes para cada $N/2$ pontos:

$$\begin{aligned} X(\omega) &= \sum_{n=0}^{N-1} x(n) W_N^{kn} = \sum_{n=0}^{N/2-1} x(2m) W_N^{2mk} + \sum_{n=0}^{N/2-1} x(2m+1) W_N^{(2m+1)k} \\ &= \sum_{n=0}^{N/2-1} x(2m) W_N^{2mk} + W_N^k \sum_{n=0}^{N/2-1} x(2m+1) W_N^{2mk}, \end{aligned} \quad (2.11)$$

onde $m = 0, 1, 2, ..., (N/2) - 1$. Sabendo que $e^{j(\omega_k + \pi)n} = -e^{j\omega_k n}$

Então quando o fator de fase for deslocado por meio período, o valor do fator de fase não irá mudar, mas o sinal do fator de fase será o oposto. Isso é chamado de propriedade da simetria do fator de fase (PROAKIS; MANOLAKIS,). Como o fator de fase também pode ser expressado por $W_N^{kn} = e^{j\omega_k n}$, então:

$$W_N^{(k+\frac{N}{2})n} = -W_N^{kn} \quad (2.12)$$

e

$$(W_N^{kn})^2 = -W_{N/2}^{kn} = e^{j\frac{4\pi k}{N}n}. \tag{2.13}$$

A equação da DFT com N pontos finalmente se torna:

$$X(k) = \sum_{n=0}^{N/2-1} x_1(m)W_{N/2}^{mk} + W_N^k \sum_{n=0}^{N/2-1} x_2(m)W_{N/2}^{mk}, = X_1(k) + W_N^k X_2(k), k = 0, 1, 2, ...N/2 \tag{2.14}$$

$$X(k + N/2) = X_1(k) - W_N^k X_2(k), k = 0, 1, 2, ...N/2. \tag{2.15}$$

Então a DFT de N pontos é separada em duas DFTs de $N/2$ pontos. Da equação (2.14), $X_1(k)$ possui $(N/2) \cdot (N/2) = (N/2)^2$ multiplicações comlpexas. $W_N^k X_2(k)$ possui $N/2 + (N/2)^2$ multiplicações complexas. Então o número total de multiplicações complexas para $X(k)$ é $2 \cdot (N/2)^2 + N/2 = N^2/2 + N/2$. A equação original da DFT de N pontos possui N^2 multiplicações complexas. Então separar $x(n)$ em duas partes faz com que o número de multiplicações complexas caia de N^2 para $N^2/2 + N/2$, ou seja, esse número foi reduzido pela metade.

Esse é o processo para reduzir o número de cálculos de N pontos para $N/2$ pontos. Logo, se o processo de separar $x_1(m)$ e $x_2(m)$ independentemente em partes pares e ímpares continuar do mesmo modo, os cálculos de $N/2$ pontos serão reduzidos para $N/4$ pontos, e assim sucessivamente. Caso o sinal para a DFT de N pontos seja continuamente separada até que a sequência final seja uma sequência de um ponto, assumindo que existem 2^s pontos que precisam ser calculados para a DFT, então o súmero de separações que podem ser feitas é $s = \log_2(N)$. O número total de multiplicações complexas será aproximadamente reduzido a $(N/2)\log_2(N)$. Para os cálculos de adição, o número será reduzido para $N\log_2(N)$(PROAKIS; MANOLAKIS,). Devido a redução no número de adições e multiplicações, a velocidade de computação da DFT é aumentada. A ideia principal do *Radix-2 FFT* é separar a sequencia de dados antiga em partes pares e ímpares continuamente para reduzir os cálculos originais em aproximadamente a metade.

2.4 Normalização Do Espectro

Quando se fala sobre a frequência do sinal de fala para diferentes palavras, cada uma pussui uma banda de frequência diferente, não só uma única frequência. Na banda de frequência de cada palavra, o espectro ($|X(\omega)|$) ou a potência do espectro ($|X(\omega)|^2$) possui um valor máximo e mínimo. Quando se comparam diferenças entre dois sinais de fala diferentes, é difícil comparar dois espectros em diferentes padrões de medidas. Logo, usar

a normalização do espectro pode fazer com que os padrões de medição sejam os mesmos, tornando a comparação mais simples.

É importante salientar que a normalização pode reduzir o erro quando comparam-se espectros, o que é bom para reconhecimento de fala (BUERA ANTONIO MIGUEL,). Então, antes de analizar a diferença entre espectros o primeiro passo é a normalização dos mesmos utilizando a normalização linear. A equação que rege a normalização linear é a seguinte:

$$y = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{2.16}$$

na qual x_{min} e x_{max} são, respectivamente, os valores mínimo e máximo do sinal x.

Depois da normalização os valores de amplitude do espectro $|X(\omega)|$ estarão no intervalo entre $[0, 1]$. A normalização apenas muda os a faixa de alcance dos valores do espectro, porém não muda a forma da informação contida no proprio espectro.

Usando o *MatLab* é possível visualizar a diferença entre um espectro não normalizado e um normalizado por meio de uma normalização linear. Primeiramente um sinal de voz é gravada, então a FFT do sinal é encontrada. Tomam-se, então, os valores absolutos do espectro da FFT. O gráfico obtido é o encontrado na Figura 1:

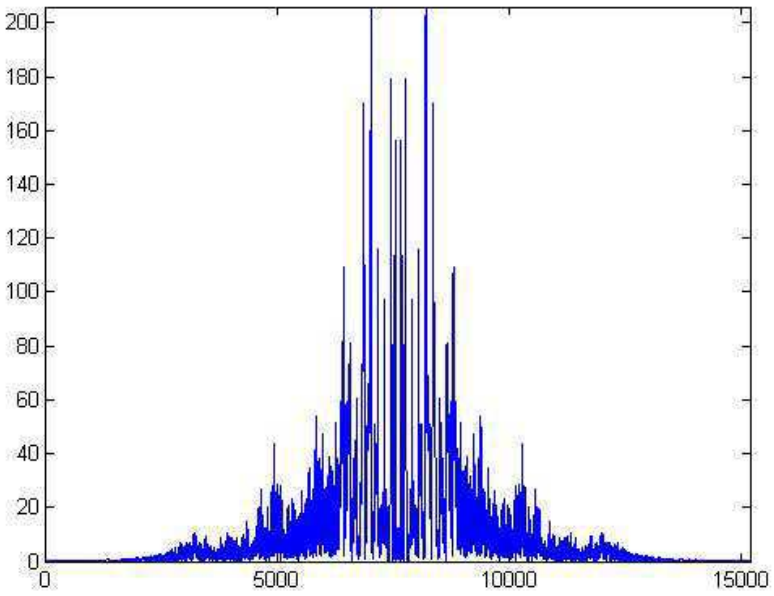

Figura 1 – Valores absolutos da FFT não normalizada da palavra casa

Utilizando a nomalização linear no espectro da Figura 1 encontramos o espectro nomalizado da Figura 2

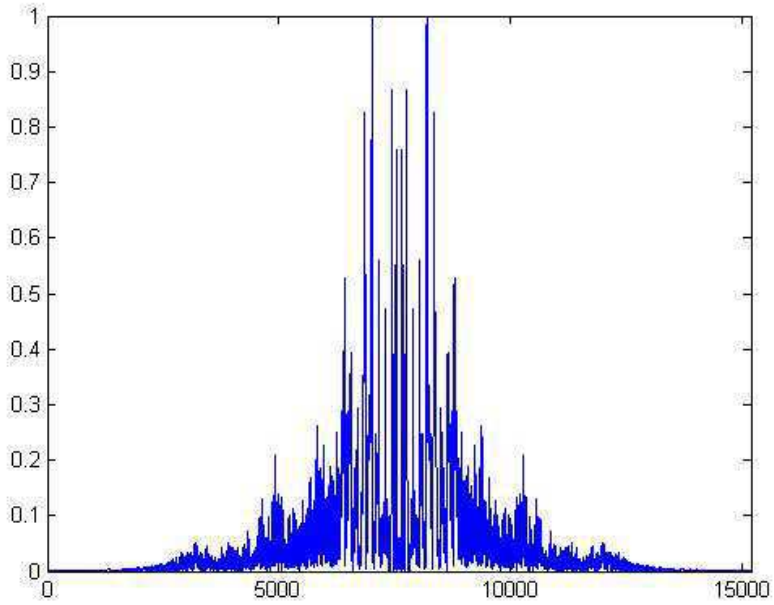

Figura 2 – Valores absolutos da FFT normalizada da palavra casa

Das Figuras 1 e 2 é possível identificar que a única diferença entre os dois espectros é o intervalo dos valores de amplitude, que no primeiro caso variam de $[0, 206]$ e no segundo caso $[0, 1]$. Nenhuma outra informação do espectro é mudada. Depois da normalização dos valores absolutos da FFT, o próximo passo para o reconhecimento de palavras é separá-las de uma frase.

3 Segmentação fonética

Um sistema de segmentação de fala possui como objetivo principal determinar as fronteiras que separam os elementos essenciais, como palavras, sílabas ou fonemas, em uma locução (SELMINI, 2008). Esse sistema pode ser usado nos codificadores de voz fonéticos, assim como em sistemas de síntese e reconhecimento automático de fala (trabalho proposto neste texto).

A segmentação da fala é uma etapa essencial para o desenvolvimento de um sistema de reconhecimento de voz, pois para um reconhecimento de fala adequado é necessário que se extrair de cada fonema informações prosódicas, como duração, energia, frequência fundamental, entre outros, para que seja possível realizar a comparação com os fonemas ou palavras da base de dados. Desta forma, o sistema de segmentação de fala afeta diretamente o desempenho dos reconhecedores fonéticos e de palavras, visto que é necessária uma segmentação robusta para a extração exata dos parâmetros com o mínimo de erro (ROCHA, 2012).

3.1 Descrição Do Sistema

O sistema de segmentação de fala proposto neste texto é classificado como implícito, pois não utiliza dados da transcrição fonética para segmentar a fala, e tem o objetivo de obter os instantes iniciais e finais de cada fonema em uma locução, além de obter as fronteiras entre partes audíveis e inaudíveis (silêncio), separando desse modo palavras e sílabas.

Este sistema não utiliza métodos estatísticos, como os modelos de Markov escondidos (*Hidden Markov Models – HMM*), e sistemas de refinamento em que é necessário o conhecimento prévio da transcrição fonética presente em uma locução, além de outras características, diferentemente dos demais trabalhos encontrados na literatura sobre segmentação de fala, o sistema descrito segmenta a fala em fonemas mediante a observação de uma característica prosódica do sinal de voz, a energia (PARANAGUá, 2012; SELMINI, 2008).

As marcas de segmentação são encontradas por meio de cálculos de valores de energia em intervalos de duração pré-definidos entre regiões de silêncio, no decorrer da forma de onda de uma locução. Como resultado, fornece uma matriz com os valores zeros e uns, por meio da comparação da energia média do intervalo entre regiões inaudíveis e a energia de curta duração a cada passo. A observação desta matriz permite a identificação de regiões com concentração de zeros, devido a energia de curta duração ser menor que a

energia média no intervalo, além de regiões onde a energia de curta duração é maior que a energia média no intervalo, formando, assim, uma concentração de uns na matriz.

O método usado identifica o ponto de transição entre as regiões da matriz que apresentam muitos zeros seguidos por uns e vice versa para encontrar as marcas de segmentação entre dois fonemas adjacentes, uma vez que essa mudança implica fonemas distintos.

3.2 Desenvolvimento Do Sistema

O método desenvolvido para segmentação de fala é essencialmente composto por duas etapas. A descrição de cada uma delas pode ser vista a seguir nas próximas subseções.

3.2.1 Primeira Etapa: Identificação De Regiões Audíveis

A etapa inicial consiste em fazer uma segmentação preliminar com a identificação de regiões audíveis e não audíveis por meio da energia de curta duração. Este parâmetro é utilizado, pois apresenta valores significativamente maiores para regiões audíveis em uma locução, sendo possível distinguir a voz do silêncio.

A intenção deste procedimento é melhorar os resultados da segmentação por meio da obtenção das fronteiras entre silêncio e fonemas e vice-versa, que são mais fáceis de identificar, denominadas fronteiras de referência.

Para a realização desta tarefa, é necessário calcular a energia em toda a forma de onda e perceber quando a mesma se encontra com um valor nulo (ou muito baixo, no caso de ruído), para que então seja possível a delimitação de fronteiras de referência (entre fonemas e silêncio e vice-versa) nesses locais.

De acordo com a literatura, a energia de curta duração é calculada por meio de janelas cujo tamanho deve variar entre 20 amostras para uma voz aguda a 250 amostras para voz grave. Na prática, para uma taxa de amostragem na ordem de 10 k amostras/s, deve-se utilizar uma janela entre 100 e 200 amostras (10 ms $<$ t $<$ 20ms).

A taxa de amostragem das locuções usadas no desenvolvimento do sistema de segmentação é de 22050 k amostras/s. Assim, o cálculo da energia é feito com uma janela com 500 amostras, utilizando um deslocamento de janela de 20 amostras. A Figura 3 ilustra um exemplo de aquisição de fronteiras de referência, localizadas entre fonemas e silêncio e vice-versa.

Figura 3 – Fronteiras que delimitam o silêncio (vermelho) localizadas em uma locução (azul)

3.2.2 Segunda Etapa: Identificação De Novas Fronteiras Dentro Das Regiões Audíveis

O segundo passo consiste em uma nova segmentação feita apenas nas regiões audíveis, delimitadas pelas fronteiras de referência encontradas na primeira etapa, com o objetivo de encontrar novas fronteiras de presentes em cada região audível, identificando, assim, o início e o fim de cada fonema presente na locução.

As novas fronteiras são identificadas mediante a análise da energia da forma de onda do sinal de voz confinado entre as fronteiras de referência (região audível), encontradas na etapa anterior. A energia é calculada para cada intervalo de duração pré-definido (200 amostras). A Figura 4 ilustra o comportamento da energia (vermelho) para uma locução (azul).

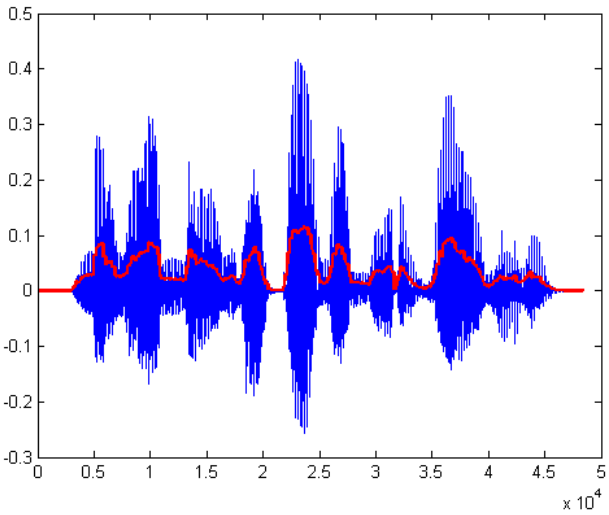

Figura 4 – Curva da energia (vermelho) para a locução (azul)

É facil perceber que as novas fronteiras estão localizadas no início e no fim dos vales (regiões onde a energia cai drasticamente e logo depois sobe) presentes na curva de potência, de acordo com a Figura 4, pois parte das fronteiras entre fonemas está nas regiões em que a energia está crescendo ou decrescendo.

A identificação dos instantes inicial e final da cada um desses vales é realizada por meio de uma codificação dos valores de energia. Primeiramente, a locução a ser segmentada é dividida em quatro regiões com o mesmo número de amostras. Para cada uma destas regiões é encontrado um valor médio de energia, o qual será utilizado como limiar para a codificação, sendo atribuído o código 1 aos valores acima dele e o código 0 para valores de energia abaixo desse limiar.

Uma matriz formada por regiões de zeros e uns é obitida como resultado desse procedimento. Para identificar as fronteiras, uma busca da transição entre as regiões de

uns e zeros é realizada. Dessa forma, novas marcas de segmentação, ou seja, aquelas com maior probabilidade de estarem corretas, são encontradas, uma vez que estão localizada na região de transição entre as regiões da referida matriz, sendo assim, estão posicionadas no início e final dos vales.

4 Algoritmo do sistema

Este capítulo trata do algoritmo desenvolvido, utilizando a correlação cruzada e autocorrelação, para determinar primeiramente a palavra e em seguida o fonema pronunciados pelo orador. O orador deve ser o mesmo que construiu a base de voz, pois o sistema em questão é dependente do locutor, ou seja, possui informações prévias da voz dos usuários que deverão utilizar o sistema.

4.1 O Algoritmo Da Correlação Cruzada

Existe uma quantidade de dados substancial na frequência fundamental da voz (F_0) na fala de oradores que diferem em idade e sexo (TRAUNMULLER,). Para o mesmo orador, as palavras diferentes também possuem diferentes bandas de frequência e formatos de espectro, devido a vibrações diferentes na corda vocal. Essa é a base para o reconhecimento de fala proposto neste trabalho. Para realizar o reconhecimento de fala, é necessário que se façam comparações entre os espectros gravados. algumas palavras serão gravadas como referência, então uma palavra qualquer será pronunciada para que a identificação da mesma seja feita pelo programa. A seguir, serão feitos testes com fonemas de maneira parecida com os testes realizados com palavras. Ao final o sistema julgará qual palavra ou fonema foi pronunciado. O método desenvolvido é realmente útil para determinar o parâmetro de deslocamento (CHEN; GUPTA,). No resto do texto esse parâmetro será chamado por desvio da frequência.

A correlação cruzada entre dois sinais tem como definição a seguinte:

$$r_{XY} = r(m) = \sum_{n=-\infty}^{\infty} x(n)y(n+m), m = 0, \pm1, \pm2, \pm3, ... \tag{4.1}$$

Da equação anterior, a ideia principal do algoritmo para a correlação cruzada pode ser descrito em três passos:

- Passo 1: Fixar um dos dois sinais $x(n)$ e deslocar o outro sinal $y(n)$ da esquerda para direita com algumas unidades de tempo.
- Passo 2: Multiplicar o valor de $x(n)$ pelo sinal deslocado $y(n+m)$ posição por posição.
- Passo 3: Tomar a resultado da soma de todos os resultados das multiplicações $x(n) \cdot y(n+m)$.

Como exemplo, dois sinais digitais $x(n) = [00100]$ e $y(n) = [01000]$, os tamanhos de ambos sinais é $N = 5$. Então a correlação cruzada entre $x(n)$ e $y(n)$ pode ser encontrada como observado nas figuras seguintes

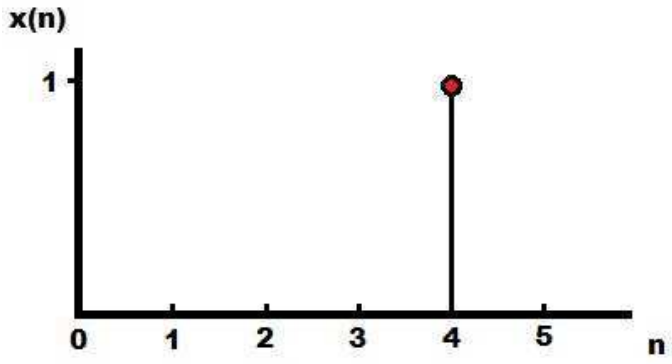

Figura 5 – Sequência de sinal $x(n)$

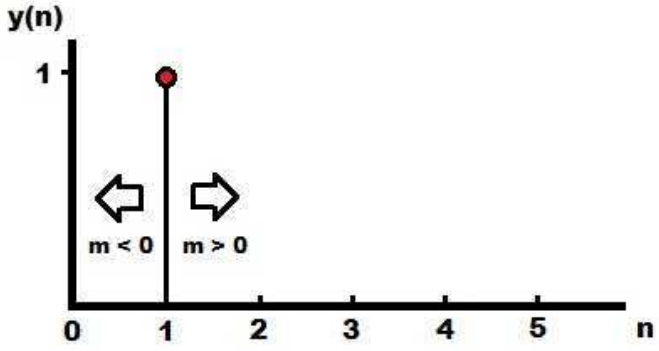

Figura 6 – A sequência de sinal $y(n)$ irá ser deslocada para direita ($m > 0$) e esquerda ($m < 0$) e será multiplicada por $x(n)$

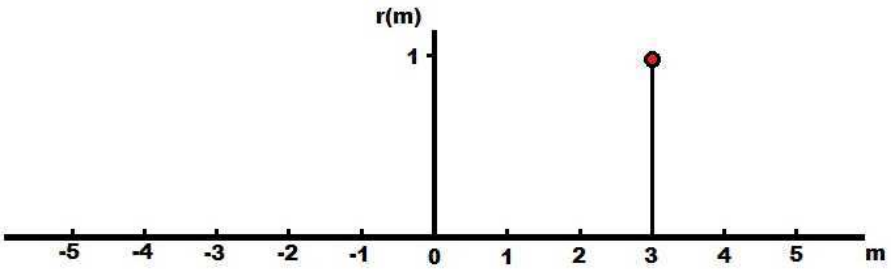

Figura 7 – O resultado da correlação cruzada é o somatório de todas as multiplicações

No exemplo dado, existe um deslocamento discreto de duas unidades de tempo entre os sinais $x(n)$ e $y(n)$. Da Figura 7, a correlação cruzada $r(m)$ possui um resultado com valor não nulo na posição $m = 3$. O eixo m da Figura 7 não é mais o eixo do sinal,

é agora o eixo do deslocamento no tempo. Como o comprimento dos dois sinais $x(n)$ e $y(n)$ são ambos $N = 5$, então o comprimento do eixo do deslocamento no tempo é $2N + 1$. Quando se usa o *MatLab* para fazer cálculos de correlação cruzada, o comprimento da correlação cruzada $r(m)$ continua sendo $2N + 1$, porém a plotagem do gráfico ocorre do zero até $2N$, e não de $-N$ a N, como mostrado no exemplo. Então o zero no eixo do deslocamento no tempo será deslocado para a posição N. Logo, quando dois sinais não possuem deslocamento no tempo entre eles, o máximo valor da sua correlação cruzada se encontrará na posição $m = N$ no software *MatLab*, que é a posição do meio do comprimento total da correlação cruzada.

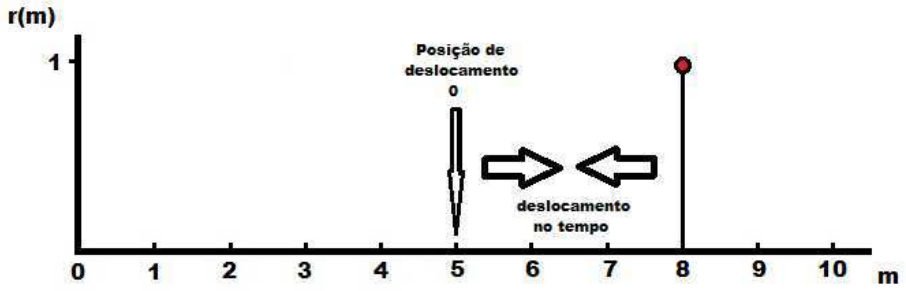

Figura 8 – O resultado da correlação cruzada como deve ser encontrado no *MatLab*

Da Figura 8, o valor máximo da correlação cruzada entre os sinais $x(n)$ e $y(n)$ não se encontra na posição do meio em relação ao comprimento total da correlação cruzada. No exemplo dado, o comprimento de ambos os sinais é $N = 5$, então o comprimento total da correlação cruzada é $2N + 1 = 11$. Então quando os dois sinais não possuem deslocamento no tempo, o valor máximo de sua correlação cruzada deveria ser em $m = 5$, porém, na Figura 8, o valor máximo se encontra na posição $m = 8$, o que significa que os dois sinais originais possuem um deslocamento no tempo de três unidades de tempo entre eles.

Do exemplo, duas informações importantes a respeito da correlação cruzada podem ser dadas. A primeira é que quando os dois sinais originais não possuirem um deslocamento no tempo entre eles, a correlação cruzada entre eles deve ser máxima. A segunda informação importante que pode ser retirada do exemplo anterior é que a diferença entre a posição de valor máximo e a posição do meio da correlação cruzada é o comprimento do deslocamento no tempo entre os dois sinais originais.

Agora assumindo que os dois sinais de voz gravados são exatamente iguais, então seus espectros também serão os mesmos. Então quando se faz a correlação cruzada entre esses dois sinais e o resultado é plotado em um gráfico, o resultado esperado é um gráfico totalmente simétrico em relação a posição do meio (N), de acordo com o algoritmo discutido anteriormente. No entanto, para gravações reais de voz, os espectros dos sinais gravados duas vezes (gravados damesma palavra, por exemplo) não são exatamente os

mesmos, porém devem ser similares, o que implica que o gráfico da correlação cruzada entre eles deve ser aproximadamente simétrico. Esse é o conceito mais importante usado nesse trabalho para o recohecimento de palavras e fonemas a partir de um banco de dados previamente gravado e sinais de teste.

Comparando a simetria da correlação cruzada entre um sinal sob teste e o banco de voz, o sistema é capaz de fazer a decisão sobre qual dos sinais do banco de voz é o mais próximo do sinal sob teste, ou seja, qual sinal do banco de dados possui o espectro mais parecido com o sinal sob teste. Em outras palavras, possivelmente esses sinais são de uma mesma palavra ou fonema.

Como exemplo, foram gravadas duas palavras no banco de dados: "casa"e "teste". Então uma terceira palavra foi inserida para fazer o papel de sinal sob teste. Essa terceira palavra foi uma nova gravação de "casa". Foram então realizados os cálculos da correlação cruzada entre o sinal sob teste e as palavras inseridas no banco de dados. A Figura 9 é o gráfico da correlação cruzada entre a primeira palavra gravada ("casa") e o sinal sob teste ("casa"). Já a Figura 10 explicita o gráfico obtido da correlação cruzada entre as palavras "teste"e "casa".

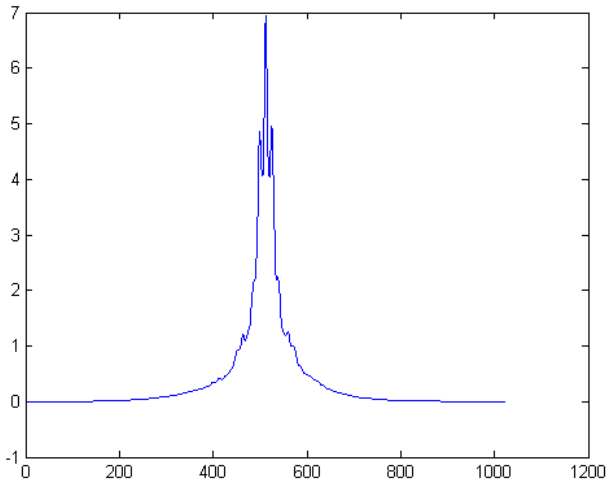

Figura 9 – Correlação cruzada entre duas gravações diferentes da palavra ”*casa*”

É possível perceber apenas observando os gráficos que a correlação cruzada entre duas amostras diferentes de palavras "casa"possui um pico mais centralizado, é mais simétrica e também é mais suave que a correlação cruzada obtida por meio da comparação entre as palavras "teste"e "casa".

Matematicamente falando, seja a função de frequência do espectro $f(x)$, de acordo com a definição da propriedade da simetria axial: para a função $f(x)$, se x_1 e x_3 são simetricos em relação ao eixo x_2, então $f(x_1) = f(x_3)$. Para a comparação no reconhecimento de fala, depois de calcular a correlação cruzada de dois espectros de frequência gravados, existe

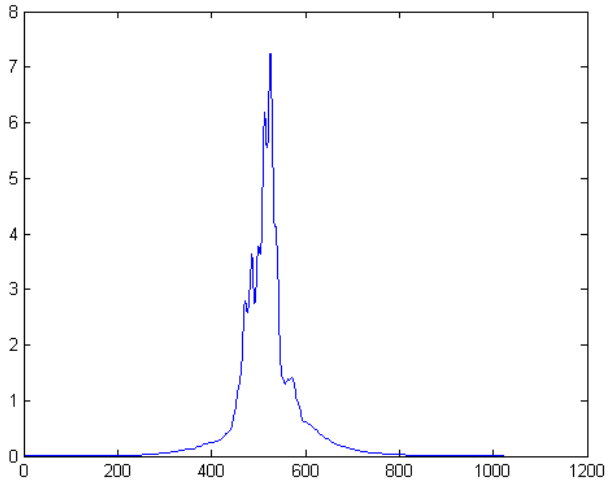

Figura 10 – Correlação cruzada entre as palavras "*teste*" e "*casa*"

uma necessidade de encontrar a posição exata do valor máximo da correlação cruzada e subtrair os valores à direita desse ponto dos valores à esquerda, tomar o valor absoluto dessa diferença e encontrar o erro médio quadrático desse valor. Quanto mais semelhantes forem os dois sinais, mais a correlação cruzada é simétrica, e quanto mais a correlação cruzada for simétrica, menor deve ser o valor do erro médio quadrático. Então, ao comparar esse erro, o sistema pode decidir qual sinal de referência foi gravado no terceiro momento.

4.2 Autocorrelação

A autocorrelação é um caso particular da correlação cruzada, pois pode ser tratada como a correlação cruzada entre sinal gravado com ele mesmo, em vez de usar outro sinal para comparação. Definição essa encontrada no *MatLab*. O algoritmo da autocorrelação mede o quanto o sinal se correlaciona com ele mesmo, ou seja, existe uma dependência temporal entre os valores sucessivos do sinal em questão. Quanto maior a autocorrelação, maior é a dependência da próxima amostra do sinal com a anterior.

A equação para a autocorrelação no caso em questão é:

$$r_X(k) = r_{XX}(k) = \sum_{n=-\infty}^{\infty} x(n)x(n+k) \tag{4.2}$$

Como exemplo, a Figura 11 mostra a autocorrelação do espectro de frequência da palavra "emocionantes"

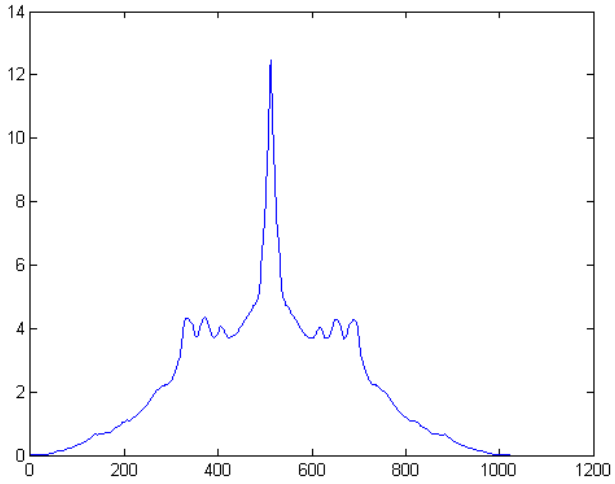

Figura 11 – Autocorrelação do sinal de voz gravado "*emocionantes*"

5 Passos da programação e resultados da simulação

Os passos usados na programação dos sistemas desenvolvidos (segmentação e reconhecimento) serão apresentados e explicados. Os testes do segmentador foram realizados utilizando uma base de voz gravada por um locutor homem paulista, e os resultados foram comparados com a segmentação realizada a mão por um foneticista. Os testes com o reconhecimento de palavras foram feitos com dois locutores: um homem e uma mulher. Já os testes com fonemas, utilizou a mesma base de voz utilizada para o sistema de segmentação previamente citado.

O sistema de segmentação tem como entrada uma locução inteira, ou seja, uma frase completa, pronunciada por um orador em ambiente de ruído quase nulo. Tem como saída as marcas que delimitam inicio e fim de fonemas, chamadas fronteiras.

Para a simulação do sistema de reconhecimento de palavras são gravados primeiramente os sinais de referência que serão armazenados em um banco de dados, utlizando o software *Audacity* em um ambiente pouco ruidoso. Após gravados sinais de referência, os sinais sob teste devem ser gravados do mesmo modo para serem inseridos no programa.

A versão do programa usada no teste com palavras utliza apenas três sinais de referência e um sinal sob teste como entrada, ou seja, o programa decide qual das três palavras o sinal sob teste deve ser. No teste com fonemas, primeiramente serão apresentados os resultados utilizando apenas vogais como sinais de referência e sinais sob teste, e seguida serão introduzidos todos os fonemas encontrados na lingua portuguesa. Todos os fonemas serão retirados do sistema de segmentação desenvolvido como base deste trabalho.

5.1 Resultados Do Sistema De Segmentação

Existem duas formas de avaliar os resultados obtidos de um sistema de segmentação de fala: avaliações subjetivas ou objetivas. A avaliação subjetiva se dá quando os avaliadores escutam os segementos de fala, no caso os fones, obtidos pelas marcas de segmentação e avaliam o quão próximo estão dos reais. No caso da avaliação objetiva, utilizada neste trabalho, as fronteiras de segmentação encontradas no sistema de segmentação automática são comparadas com as marcas de segmentação obtidas de forma manual. Neste caso, o erro entre tais fronteiras não deve ultrapassar 20 ms (SELMINI, 2008; PARK; KIM, 2007; PARANAGUá, 2012; JARIFI; PASTOR; ROSEC, 2008; TOLEDANO; GOMEZ; GRANDE, 2003).

A avaliação objetiva foi escolhida para avaliar o sistema de segmentação proposto, devido a dificuldade em encontrar avaliadores que possam fazer os testes em um ambiente controlado, ou seja, sem ruido ou interrupções. Assim, os valores das fronteiras obtidos pelo sistema de segmentação desenvolvido foram comparados aos resultados obtidos de forma manual. O banco de dados utilizado nos testes, segmentado manualmente, apresentado em (SELMINI, 2008), e cedido por (PARANAGUá, 2012) para utilização neste trabalho, é composto por 200 frases, gravadas por um locutor paulista, do interior do Estado de São Paulo, não abordando, desta forma, os regionalismos do País, assim como os diferentes tipos de pronúcia para algumas locuções.

As locuções foram gravadas a uma taxa de 22,05 k amostras/s e quantizadas com 16 *bits* por amostra. As sentenças têm, em média, três segundos e foram gravadas com o mínimo de ruído possível.

Para efeito da avaliação, as fronteiras entre ditongos não foram consideradas, uma vez que o sistema de segmentação de fala proposto está sendo desenvolvido para ser utilizado em um codificador de voz fonético, já que este tipo de encontro vocálico é considerado pelo codificador como uma única locução, ou seja, um único fonema.

A avaliação objetiva foi realizada considerando 50 locuções selecionadas da base de dados de voz. De acodo com a Tabela 1, que apresenta os resultados da taxa de segmentação, o sistema de segmentação proposto foi capaz de localizar $72,41\%$ das fronteiras de segmentação, sendo $77,6\%$ das marcas de segmentação encontradas com erro menor que 10 ms (221 amostras), enquanto $22,4\%$ das fronteiras obtidas foram conseguidas com um erro entre 10 e 20 ms (221 a 442 amostras).

Pode-se perceber pela Tabela 2, a qual apresenta o número de fronteiras falsas obtidas bem como a quantidade de fronteiras não detectadas em cada locução, que o sistema desenvolvido obteve em média $6,74$ falsas fronteiras e não foi capaz de localizar $7,6$ fronteiras encontradas na segmentação manual.

O algoritmo de segmentação proposto apresenta uma complexidade de algoritmo inferior aos demais sistemas de segmentação encontrados na literatura. O seu desenvolvimento não depende de um banco de fala robusto para treinar modelos estatísticos, como é o caso dos sistemas que utilizam os modelos de Markov escondidos para a segmentação. Além disso, não se faz necessário o prévio conhecimento da transcrição fonética para a realização da segmentação. O sistema proposto neste trabalho obteve resultados competitivos com outros sistemas já desenvolvidos, sem usar um sistema de refinamento dos resultados obtidos.

As locuções utilizadas para os testes serão apresentadas em anexo.

Tabela 1 – Resultados da taxa de segmentação obtidos pelo sistema desenvolvido.

Locução	**Total de Fronteiras**	**10 ms**	**20 ms**	**Taxa de Segmentação (%)**
L001	48	29(76,3%)	9(23,7%)	79,16
L002	22	11(78,6%)	3(21,4%)	63,63
L003	23	11(68,7%)	5(31,3%)	69,56
L004	18	12(80,0%)	3(20,0%)	83,33
L005	41	19(73,0%)	7(27,0%)	63,41
L006	41	24(70,0%)	10(30,0%)	82,92
L007	27	14(77,8%)	4(22,2%)	66,66
L008	34	24(92,3%)	2(7,7%)	76,47
L009	43	28(82,3%)	6(17,7%)	79,06
L010	20	14(73,7%)	5(26,3%)	95,00
L011	21	16(88,9%)	2(11,1%)	85,71
L012	21	10(66,7%)	5(33,3%)	71,42
L013	35	24(88,9%)	3(11,1%)	77,14
L014	29	17(77,3%)	5(22,7%)	75,86
L015	31	17(77,3%)	5(22,2%)	70,96
L016	33	18(66,7%)	9(33,3%)	81,81
L017	37	24(82,7%)	5(17,3%)	78,37
L018	31	18(85,7%)	3(14,3%)	67,74
L019	34	16(69,5%)	7(30,5%)	67,64
L020	25	10(62,5%)	6(37,5%)	64,00
L021	27	18(94,7%)	1(5,3%)	70,37
L022	31	10(62,5%)	6(37,5%)	51,61
L023	28	12(60,0%)	8(40,0%)	71,42
L024	25	13(65,0%)	7(35,0%)	80,00
L025	27	13(81,2%)	3(18,8%)	59,25
L026	30	16(64,0%)	9(36,0%)	83,33
L027	24	12(63,1%)	7(36,9%)	79,16
L028	15	9(81,8%)	2(18,2%)	73,33
L029	17	7(77,8%)	2(22,2%)	52,94
L030	33	17(73,9%)	6(26,1%)	69,69
L031	12	7(87,5%)	1(12,5%)	66,66
L032	25	17(89,5%)	2(10,5%)	76,00
L033	26	13(72,2%)	5(27,8%)	69,23
L034	37	25(83,3%)	5(16,7%)	81,08
L035	30	19(86,4%)	3(13,6%)	73,33
L036	31	22(84,6%)	4(15,4%)	83,87
L037	34	14(66,7%)	7(33,3%)	61,76
L038	23	12(80,0%)	3(20,0%)	65,21
L039	26	17(85,0%)	3(15,0%)	76,92
L040	28	15(71,4%)	6(28,6%)	75,00
L041	30	14(73,6%)	5(26,4%)	63,33
L042	30	17(80,9%)	4(19,1%)	70,00
L043	22	8(88,9%)	1(11,1%)	40,90
L044	14	8(100,0%)	0(0%)	57,14
L045	23	11(73,3%)	4(26,7%)	65,21
L046	28	15(68,2%)	7(31,8%)	78,57
L047	27	22(91,6%)	2(8,4%)	88,88
L048	26	17(80,9%)	4(19,1%)	80,76
L049	36	21(80,7%)	5(19,3%)	72,22
L050	24	15(75,0%)	5(25,0%)	83,33
Média		77,6%	22,4%	72,41

Tabela 2 – Fronteiras falsas e não detectadas obtidas em cada locução

Locução	**Fronteiras Falsas**	**Não Detectadas**
L001	9	10
L002	6	8
L003	9	7
L004	7	3
L005	5	15
L006	15	7
L007	8	9
L008	4	8
L009	4	9
L010	6	1
L011	5	3
L012	6	6
L013	6	8
L014	8	7
L015	6	9
L016	9	6
L017	4	8
L018	6	10
L019	10	11
L020	5	9
L021	8	8
L022	10	15
L023	5	8
L024	7	5
L025	9	11
L026	6	5
L027	7	5
L028	4	4
L029	4	8
L030	5	10
L031	4	4
L032	9	6
L033	9	8
L034	7	7
L035	8	8
L036	3	5
L037	14	13
L038	7	8
L039	6	6
L040	3	7
L041	5	11
L042	11	9
L043	6	13
L044	3	6
L045	8	8
L046	6	6
L047	6	3
L048	6	5
L049	7	10
L050	6	4
Média	6,74	7,6

5.2 Passos Do Programa Da Correlação Cruzada

- Inicializar as variáveis;
- Escolher as palavras ou fonemas do banco de dados que serão usadas para comparação;
- Definir como sinal sob teste a voz gravada de algum dos locutores, para que seja inserida no programa;
- Usar a função *spectrogram* para processar os sinais gravados e ter como retorno uma matriz de sinais;
- Fazer a transposta da matriz de sinais para linhas e colunas, para que seja possível tomar a soma da matriz e ter como retorno um vetor linha para cada resultado de soma de coluna. Esse vetor linha é o espectro de frequência do sinal
- Como descrito em capítulos anteriores, normalizam-se os espectros de frequência utilizando uma normalização linear, para que sinais com diferentes valores absolutos possam ser comparados;
- Realizar a correlação cruzada entre os sinais do banco de dados e o sinal sob teste, obtendo um gráfico para cada sinal do banco de dados relativo a correlação cruzada entre ele e o sinal sob teste;
- É checado o desvio de frequência de cada correlação cruzada. Se a diferença entre os dois menores desvios de frequência for maior que 2, então o sistema decidirá levando em consideração apenas esse desvio. Quanto menor o desvio de frequência, maior a semelhança entre os sinais. Porém se a diferença entre os dois menores desvios de frequência for menor que 2, então o sistema nada pode decidir, passando assim para o próximo passo;
- O sistema vai localizar o ponto central e a partir daí vai calcular quão simétrica é a correlação cruzada. O algoritmo irá fazer a diferença entre os pontos simetricos em relação ao eixo do ponto central e somar seus valores absolutos para retirar, assim, o valor do erro médio quadrático para cada uma das funções de correlação cruzada. O sinal que obtiver o menor erro será decidido pelo sistema como sinal correto.

O programa desenvolvido encontra-se no Anexo A, com partes comentadas para uma melhor compreensão.

5.3 Resultados Da Simulação

Os resultados estão apresentados em duas seções, onde a primeira analisa o desempenho do sistema proposto para reconhecimento de palavras pessoal utilizando amostras

de voz femininas e masculinas. A segunda seção analisa o comportamento da rotina desenvolvida em reconhecimento de fonemas, utilizando um banco de voz previamente segmentado por um foneticista.

Para um melhor entendimento os resultados obtidos pelo programa desenvolvido estão dispostos em formas de tabela e gráficos e serão explicados conforme aparecem no texto.

5.3.1 Resultados Para o Reconhecimento De Palavras

Antes de serem usadas para esse teste, as palavras gravadas passaram pela primeira etapa do sistema de segmentação fonética para que fossem extraídas as partes relativas ao silêncio nas gravações, deixando assim, apenas as palavras em si.

As Tabelas 3 e 4 exibem os resultados alcançados para o reconhecimento de palavras pessoal utilizando sinais de voz femininos gravados com o auxílio do *software Audacity* em um ambiente pouco ruidoso.

A Tabela 3 trata de observar o funcionamento do algoritmo do desvio de frequência, para isso foram propostas gravações de duas palavras com grandes diferenças, ou seja, palavras com sílabas, números de sílabas e fonemas distintos. As palavras escolhidas foram "desligar"e "casa"por não possuírem nenhuma sílaba em comum e por serem utilizadas com frequência em sistemas que necessitam de reconhecimento de voz.

Tabela 3 – Resultados Reconhecimento de palavras para voz feminina entre duas palavras com grandes diferenças.

Nº Teste	**Palavra Pronunciada**	**Deslocamento**		**Erro Médio Quadrático**		**Veredito**
		Desligar	**Casa**	**Desligar**	**Casa**	
01	Desligar	0	0	0,0055	0,8247	Desligar
02	Desligar	0	-8	–	–	Desligar
03	Desligar	0	0	0,0345	0,4730	Desligar
04	Desligar	0	-7	–	–	Desligar
05	Desligar	0	-13	–	–	Desligar
06	Desligar	0	-8	–	–	Desligar
07	Desligar	0	0	0,0184	0,5149	Desligar
08	Desligar	0	-8	–	–	Desligar
09	Desligar	0	-8	–	–	Desligar
10	Desligar	0	0	0,0074	0,4844	Desligar
11	Casa	24	2	–	–	Casa
12	Casa	28	2	–	–	Casa
13	Casa	28	3	–	–	Casa
14	Casa	26	2	–	–	Casa
15	Casa	7	0	–	–	Casa
16	Casa	28	2	–	–	Casa
17	Casa	28	1	–	–	Casa
18	Casa	25	1	–	–	Casa
19	Casa	28	2	–	–	Casa
20	Casa	29	2	–	–	Casa

Observando a Tabela 3 é possível perceber que a taxa de acerto do sistema desenvolvido foi de 100%. Ainda, praticamente todos os acertos foram devido ao algoritmo de desvio de frequência utilizado na rotina, o que mostra que é parte fundamental nesse sistema de reconhecimento de palavras pessoal.

É possível avaliar no gráfico da Figura 12 como se dá o reconhecimento por desvio de frequência. Sabendo que a palavra "desligar"foi pronunciada até o teste de número 10 e nos testes restantes foram pronunciadas palavras "casa", é fácil perceber que nos primeiros 10 testes, a diferença de desvio de frequência entre as duas palavras foi quase sempre maior que 2, tornando a identificação utilizando o algoritmo do desvio de frequência possível de ser realizada. Houveram apenas quatro momentos em que não foi percebido um desvio de freqûencia da palavra pronunciada com a palavra de referência "casa", nos testes de número $1, 3, 7$ e 10, porém o sistema ainda foi capaz de decidir a palavra correta graças ao algoritmo do erro médio quadrático, citado anteriormente. No caso dos testes de número 10 a 20, o desvio de frequência foi mais acentuado e o sistema foi capaz de identificar a palavra pronunciada sem maiores dificuldades, apesar de um desvio de frequência ocorrer mesmo com relação a palavra de referência correta ("casa"). Esse desvio de frequência ocorreu pois a palavra de referência foi escolhida de forma aleatória no banco de dados. Futuramente será possível melhorar o sistema criando um algoritmo que trate de decidir quando uma palavra de referência é melhor que outra.

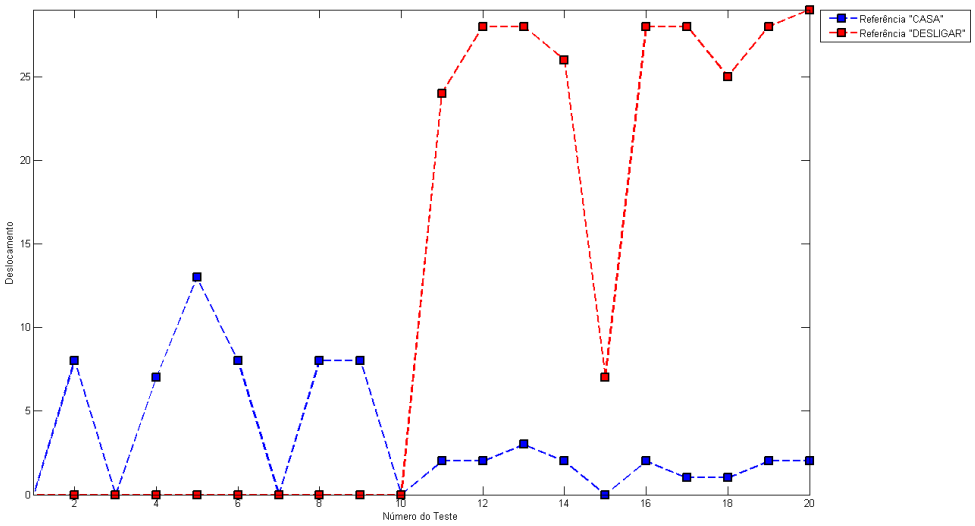

Figura 12 – Desvio em frequência referente a Tabela 3

Por meio da Tabela 4 foram estudadas as reações do sistema a um número maior de palavras com semelhaça de sílabas, número de sílabas e fonemas. As palavras selecionadas foram "ligar", "carro"e "marca"por possuírem, além de sílabas comuns, uma grande utilização em sistemas de reconhecimento de fala. Nesse caso o sistema deve reconhecer as palavras

utilizando o algoritmo do erro médio quadrático na maioria dos casos, já que a semelhança entre as locuções é maior, causando um deslocamento de frequência menor.

Tabela 4 – Resultados Reconhecimento de palavras para voz feminina entre três palavras com semelhança.

Nº Teste	**Palavra Pronunciada**	**Deslocamento**			**Erro Médio Quadrático**			**Veredito**
		Ligar	**Carro**	**Marca**	**Ligar**	**Carro**	**Marca**	
01	Ligar	0	-19	0	0,0517	2,6919	0,2518	Ligar
02	Ligar	0	-16	0	0,0246	1,0446	0,3623	Ligar
03	Ligar	0	-16	0	0,0233	1,4358	0,3366	Ligar
04	Ligar	0	-15	1	0,0197	1,3444	0,2404	Ligar
05	Ligar	0	-19	1	0,0168	2,2466	0,3791	Ligar
06	Ligar	0	-20	1	0.0604	1,5059	0,5059	Ligar
07	Ligar	0	-19	0	0,0458	4,1443	0,2361	Ligar
08	Ligar	0	-20	1	0,0639	1,6061	0,5428	Ligar
09	Ligar	0	-16	0	0,0358	0,8342	0,3509	Ligar
10	Ligar	0	-15	0	0,0264	1,5812	0,1801	Ligar
11	Carro	17	1	4	–	–	–	Carro
12	Carro	16	-1	1	1,2507	0,0258	0,1976	Carro
13	Carro	17	0	2	–	–	–	Carro
14	Carro	17	1	0	1,1794	0,0443	0,3014	Carro
15	Carro	17	0	3	–	–	–	Carro
16	Carro	20	0	2	–	–	–	Carro
17	Carro	20	0	2	–	–	–	Carro
18	Carro	19	0	2	–	–	–	Carro
19	Carro	17	0	-1	2,0277	0,0689	0,3205	Carro
20	Carro	19	0	-1	2,8391	0,0472	0,4047	Carro
21	Marca	-1	0	0	0,3121	0,4351	0,0523	Marca
22	Marca	0	-1	0	0,3503	0,2314	0,0209	Marca
23	Marca	-1	-1	-1	0,3590	0,2608	0,0394	Marca
24	Marca	-1	-1	0	0,2137	0,5300	0,1111	Marca
25	Marca	0	0	0	0,3922	0,0789	0,0687	Marca
26	Marca	0	0	0	0,3276	0,2106	0,0262	Marca
27	Marca	-1	0	-1	0,4479	0,2116	0,0756	Marca
28	Marca	0	0	0	0,2591	0,1046	0,0437	Marca
29	Marca	0	0	0	0,2515	0,1046	0,0437	Marca
30	Marca	-1	1	0	0,4301	0,2391	0,0459	Marca

Visualizando a Tabela 4, as palavras "ligar"e "marca"foram sempre reconhecidas por meio do erro médio quadrático, porém a locução "carro"foi reconhecida na maioria das vezes pelo algoritmo anterior do desvio da frequência, isso mostra que a semelhança entre a primeira sílaba de "carro"(CA) e a segunda sílaba de "marca"(CA) foram provavelmente pronunciadas de maneira distinta pelo orador em questão.

Ainda a respeito da Tabela 4, o desempenho do sistema foi de 100%, e é possível perceber que o erro médio quadrático entre a palavra de pronunciada e a locução de referência correta foi em sua maior parte cerca de 5 a 10 vezes menor que o erro entre a palavra pronunciada e as locuções de referência incorretas. O gráfico disponível na Figura 13 pode simplificar o entendimento mostrando que o erro médio quadrático da palavra de referência a ser tomada como correta deve ser o mais baixo entre os três, sendo assim, a

palavra a ser escolhida como correta entre os testes de número 1 e 10 é "ligar", assim como entre os testes 11 e 20 é "carro"e entre os testes 21 e 30 a locução a ser adotada como certa seria "marca".

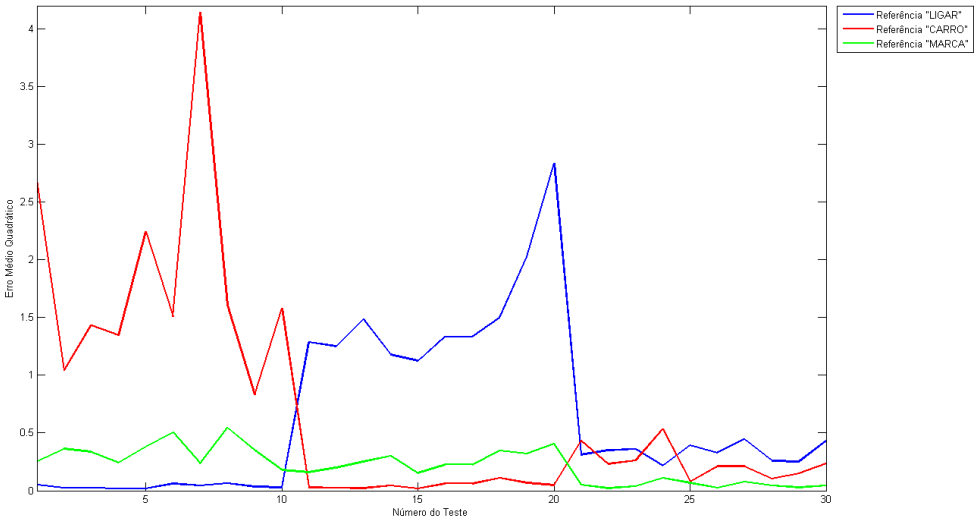

Figura 13 – Erro médio quadrático referente a Tabela 4

É disponibilizado, também, o gráfico do desvio de frequência na Figura 14 para que sejam observadas as diferêncas entre os devios. A diferença do deslocamento em frequência entre as palavras "ligar"e "carro"é visível, sendo assim, o sistema poderia ser aprimorado futuramente, eliminando a palavra com maior desvio de frequência do cálculo do erro médio quadrático, diminuindo os gastos computacionais.

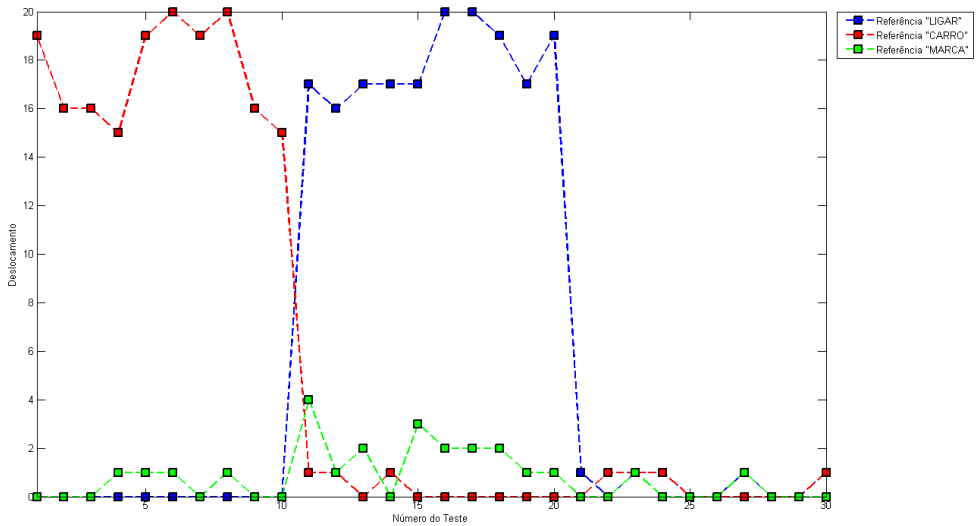

Figura 14 – Desvio em frequência referente a Tabela 4

Do mesmo modo que foram realizados os experimentos com a base de voz feminina, foram conduzidos experimentos com uma base de voz masculina para comparação de resultados entre os mesmos, além de determinar se o sistema é valido para os dois tipos de voz, mais graves (masculino) e mais agúdas (feminino)

A Tabela 5 foi criada a partir do mesmo tipo de testes utilizado para a construção da Tabela 3, ou seja, trata de observar o funcionamento do algoritmo do desvio de frequência, então foram propostas gravações de duas palavras com sílabas, números de sílabas e fonemas distintos. As palavras escolhidas foram as mesmas dos testes com a voz feminina, "desligar"e "casa"para que pudesse ser feita uma comparação mais precisa entre o resultado de ambos os sinais de voz gravados por um orador feminino e outro masculino.

Tabela 5 – Resultados reconhecimento de palavras para voz masculina entre duas palavras com grandes diferenças.

Nº Teste	**Palavra Pronunciada**	**Deslocamento**		**Erro Médio Quadrático**		**Veredito**
		Desligar	**Casa**	**Desligar**	**Casa**	
01	Desligar	0	-11	–	–	Desligar
02	Desligar	0	-10	–	–	Desligar
03	Desligar	0	-10	–	–	Desligar
04	Desligar	0	-11	–	–	Desligar
05	Desligar	0	-10	–	–	Desligar
06	Desligar	0	-10	–	–	Desligar
07	Desligar	0	-10	–	–	Desligar
08	Desligar	0	-10	–	–	Desligar
09	Desligar	0	-10	–	–	Desligar
10	Desligar	0	-11	–	–	Desligar
11	Casa	10	0	–	–	Casa
12	Casa	11	0	–	–	Casa
13	Casa	12	0	–	–	Casa
14	Casa	11	0	–	–	Casa
15	Casa	12	0	–	–	Casa
16	Casa	11	0	–	–	Casa
17	Casa	12	0	–	–	Casa
18	Casa	11	0	–	–	Casa
19	Casa	12	0	–	–	Casa
20	Casa	12	0	–	–	Casa

Na Tabela 5 observa-se uma discrepância constante em relação aos desvios de frequência das duas palavras quando comparado com o caso feminino, provavelmente devido ao homem produzir sons mais graves, ou seja de menor frequência que as mulheres, qualquer diferência de frequência se torna mai sutil, portanto mais constante. Qualquer variação na voz aguda feminina pode produzir variações altas, tanto para cima quanto para baixo do desvio de frequência.

Os dados da Tabela 5 indicam que o desempenho do algoritmo desenvolvido para a detecção do desvio de frequência é de 100% de acerto, pelo menos nesse tipo de teste. E todos os acertos foram devido a uma diferença entre o desvio de frequência em relação

a palavra de referência correta, o que mostra que esse deve ser um método eficiente de identificar palavras bastante diferentes.

O gráfico da Figura 15 é semelhante ao gráfico, anteriormente citado, da Figura 12, porém é de uma visualização mais fácil, podendo ser identificado de maneira rápida onde ocorre a transição das locuções pronunciadas pelo orador. Essa mudança na fala se dá quando há o cruzamento do gráfico entre os testes de número 10 e 11. Entre os testes de número 1 ao 10 o orador pronunciou a palavra "desligar", já nos testes entre 11 e 20 o locutor falou "casa". A diferença de desvio de frequência entre a palavra pronunciada com a locução de referência correta e a palavra pronunciada com a locução de referência incorreta é sempre maior que 2, ou seja, o sistema é capaz de identificar perfeitamente qual das palavras esta sendo dita pelo orador.

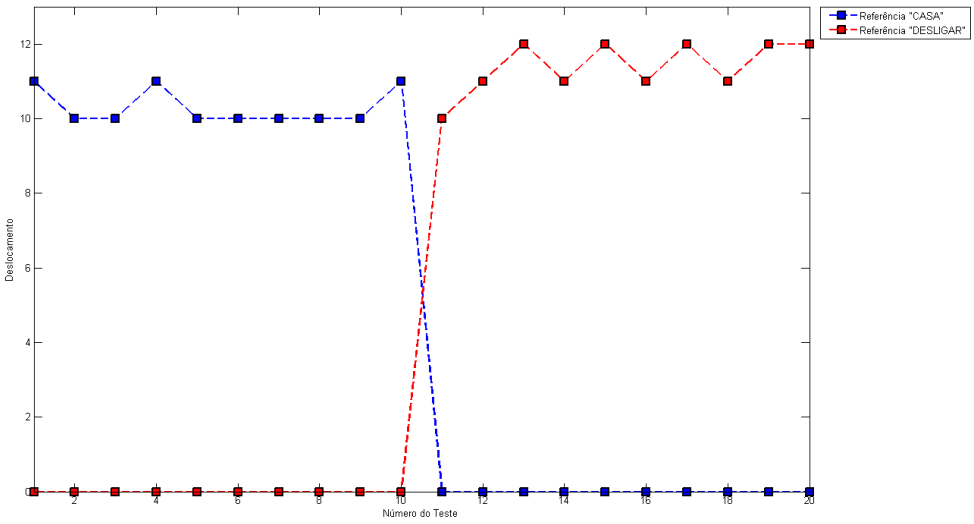

Figura 15 – Desvio em frequência referente a Tabela 5

Já a Tabela 6 é referente ao teste de palavras semelhantes utilizando a voz masculina como banco de dados, análogo à Tabela 4, onde as locuções gravadas pelo orador são "ligar", "carro"e "marca", com o propósito de verificar o desempenho do algoritmo do erro médio quadrático na voz masculina.

A taxa de acerto do sistema, como em todos os outros resultados foi de 100%, porém dessa vez o sistema calculou erros médios quadráticos mais próximos uns dos outros, o que poderá levar a erros em casos extremos. Isso se deve além da frequência mais baixa da voz masculina, a amplitude e sobretudo o timbre da voz do orador, que é um timbre rouco, pois, segundo (PONTES VANESSA P. VIEIRA, 2002), em faixas de frequência normal para vozes masculinas o ruído de vozes roucas é maior que o ruído de outros timbres de voz.

Os gráficos das Figuras 16 e 17 complementam a informação disposta na Tabela 6,

Tabela 6 – Resultados Reconhecimento de palavras para voz feminina entre três palavras com semelhança.

Nº Teste	**Palavra Pronunciada**	**Deslocamento**			**Erro Médio Quadrático**			**Veredito**
		Ligar	**Carro**	**Marca**	**Ligar**	**Carro**	**Marca**	
01	Ligar	0	0	-15	0,0340	0,3996	0,6396	Ligar
02	Ligar	0	0	-15	0,0133	0,3738	1,0401	Ligar
03	Ligar	0	0	0	0,0024	0,2979	0,5978	Ligar
04	Ligar	0	0	0	0,0350	0,3606	0,5549	Ligar
05	Ligar	0	0	-15	0,0550	0,5544	0,5998	Ligar
06	Ligar	0	0	0	0,0058	0,2443	0,4644	Ligar
07	Ligar	0	0	0	0,0085	0,1799	0,3823	Ligar
08	Ligar	0	0	0	0,0190	0,2444	0,5527	Ligar
09	Ligar	0	0	-15	0,0273	0,4098	0,8552	Ligar
10	Ligar	0	0	0	0,0192	0,1850	0,4415	Ligar
11	Carro	0	0	0	0,3340	0,0192	0,1170	Carro
12	Carro	0	0	0	0,2264	0,1672	0,3482	Carro
13	Carro	0	0	0	0,3697	0,0350	0,1317	Carro
14	Carro	0	0	0	0,3585	0,0221	0,1067	Carro
15	Carro	0	0	0	0,2631	0,0198	0,1267	Carro
16	Carro	0	0	0	0,3790	0,0258	0,0705	Carro
17	Carro	0	0	0	0,3292	0,0207	0,0860	Carro
18	Carro	0	0	0	0,2785	0,1743	0,3518	Carro
19	Carro	0	0	0	0,2234	0,2161	0,3978	Carro
20	Carro	0	0	0	0,3619	0,0177	0,0913	Carro
21	Marca	14	0	0	1,1538	0,2871	0,0420	Marca
22	Marca	14	0	0	1,0052	0,2199	0,0299	Marca
23	Marca	0	0	0	0,4897	0,0578	0,0531	Marca
24	Marca	14	-1	0	1,4542	0,2567	0,0039	Marca
25	Marca	0	-1	0	0,5988	0,1669	0,0116	Marca
26	Marca	0	-1	0	0,5433	0,2165	0,0186	Marca
27	Marca	0	-1	0	0,5993	0,1945	0,0260	Marca
28	Marca	0	-1	0	0,4494	0,0738	0,0699	Marca
29	Marca	14	0	0	0,9240	0,1017	0,0366	Marca
30	Marca	0	-1	0	0,4896	0,0958	0,0338	Marca

mostrando que é possível diminuir o custo computacional do sistema retirando, sempre que o método do desvio da frequência detectar uma diferença entre os desvios de 2 ou mais, um dos sinais de referência (Figura 16). A Figura 17 ainda mostra que a palavra de referência que está possui um erro menor, ou seja, está na parte mais inferior do gráfico, deve ser a escolhida como resultado do reconhecimento de palavras.

O sistema desenvolvido foi testado em diversas formas e com locutores de ambos os sexos, sendo assim, pode-se considerá-lo um sistema robusto, já que não apresentou nenhum erro no que se trata de reconhecimento pessoal de fala. Outra grande vantagem é a simplicidade do sistema, além do pequeno custo computacional necessário para aplicá-lo. A rotina utilizada, ainda pode ser melhorada usando o algoritmo do desvio de frequência para retirar algumas das palavras que possuem uma diferença de desvio de mais de 2 do menor desvio encontrado, diminuindo ainda mais o custo computacional.

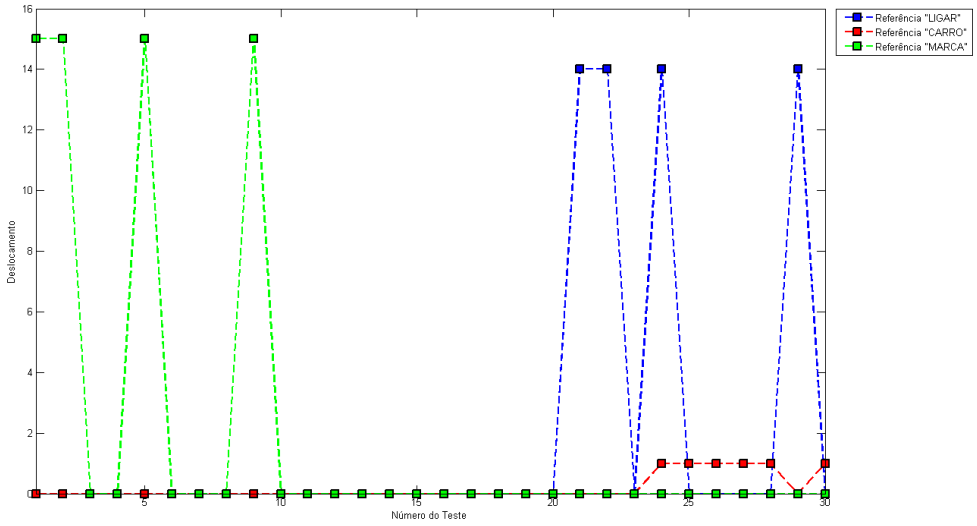

Figura 16 – Desvio em frequência referente a Tabela 6

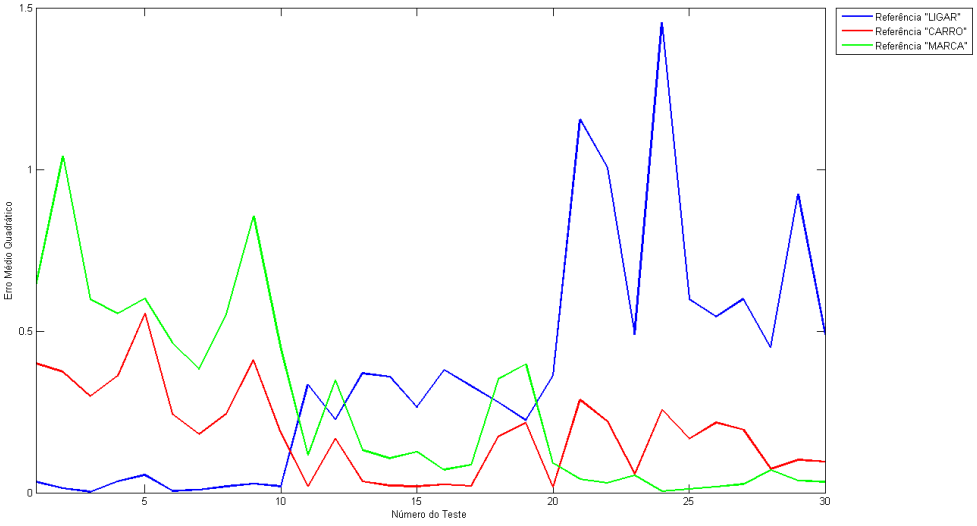

Figura 17 – Erro médio quadrático referente a Tabela 6

5.3.2 Resultados Para o Reconhecimento De Fonemas

Os fonemas usados para os testes das vogais foram obtidos a partir do sistema de segmentação de fala proposto neste trabalho, porém, como o sistema de segmenteção não apresenta uma taxa de 100% de acerto, o teste com o resto dos fonemas da lingua portuguesa foi realizado levando-se em conta as delimitações dos fonemas fornecidas por um profissional da área de fonética.

O reconhecimento de fonemas é feito com base em um banco de dado de voz gravado de um orador do sexo masculino do interior de São Paulo, e foi segmentado por

um foneticista profissional manualmente. São 200 frases gravadas e as palavras e fonemas a serem testados foram escolhidas aleatoriamente.

5.3.2.1 Apenas Vogais

No total foram utilizadas 55 vogais para obter o resultado do teste. Foram recolhidas 11 amostras de cada uma das 5 vogais que compõem o alfabeto, e uma amostra de cada vogal foi escolhida para ser a referência. Todas essas vogais foram retiradas do banco de voz com auxílio do sistema de segmentação proposto nesse trabalho. Os resultados estão dispostos em forma de tabelas para facilitar o entendimento e a visualização dos mesmos.

A Tabela 7 mostra o resultado obtido com o auxílio do programa desenvolvido para o reconhecimento de fonemas apenas com vogais retiradas de frases gravadas, comparando em qual parte do programa a vogal foi identificada corretamente(algoritmo do desvio de frequência ou algoritmo do erro médio quadrático), e se foi identificada incorretamente, qual das outras vogais foi escolhida para representar o fonema sob teste.

Tabela 7 – Resultados Reconhecimento de vogais para voz masculina.

Vogal	**Nž de amostras**	**Acertos com desvio de freq.**	**Acertos com erro médio**	**Vogal errada**					**Total de Acertos**
				A	**E**	**I**	**O**	**U**	
A	10	8	1	–	0	0	0	1	90%
E	10	0	6	1	–	1	0	2	60%
I	10	1	6	0	3	–	0	0	70%
O	10	4	2	0	0	0	–	4	60%
U	10	0	9	0	0	0	1	–	90%
Total de acertos em 50 testes									74%

É possível perceber que os erros das vogais "E", "I"e "O"estão acima da média. No caso do "E"a explicação mais plausível é devido as grandes variações em sua pronúncia, algumas vezes se fala "É"outras vezes se diz "Ê"e ainda é possivel se falar "E"(como no caso da frase: "*ïsso* ***e*** *aquilo.*"), tornando difícil escolher uma amostra de referência, fazendo com que o erro se torne imprevisível. As vogais "O"e "I"possuem um erro, de certa forma, previsível, como mostra a quinta coluna da Tabela 7, pois em algumas palavras o "O"soa ligeiramente como um "U"(por exemplo na palavra "como", o último "O"possui som de "U", pois está precedido por uma consoante) e o "I"tem som semelhante ao de um "E"(por exemplo nas palavras "boi"e "pães"o "I"e o "E"possuem mesmo som, sendo assim considerados semivogais nessas palavras).

A rotina desenvolvida para o reconhecimento pessoal de fala pode ser considerada robusta no caso de identificação de vogais retiradas de uma frase, pois mesmo que a taxa média de acerto de vogais tenha sido 74%, percebe-se que o erro é dado por sons que se assemelham com aqueles que se pretendia encontrar (No caso do "E"e do "I"quando são semivogais representam o mesmo fonema "Y". O mesmo acontece com o "O"e o "U", quando

semivogais representam o fonema "W"). O sistema também pode ser considerado simples, já que não faz uso da transcrição fonética e não necessita de uma fase de treinamento excessiva, ou seja, a gravação de cada fonema uma única vez é o necessário para fazer com que o programa funcione perfeitamente.

5.3.2.2 Todos Os Fonemas Da Lingua Portuguesa

Os testes a seguir foram realizados com palavras retiradas da base de voz gravada por um locutor paulista, por meio da primeira etapa da rotina desenvolvida para a segmentação fonética que trata de encontrar os momentos de silêncio na frase. Depois de retirar as palavras da frase, os fonemas são delimitados usando o trabalho realizado pelo foneticista. Então o programa terá como entrada a palavra retirada da locução, as fronteiras delimitadas pelo foneticista e o banco de fonemas, formado por fonemas aleatoriamente retirados das locuções gravadas pelo mesmo locutor.

A Tabela 8 foi encontrada depois do teste de 10 palavras diferentes, com 50 fonemas no total, ditas por um mesmo orador. Os fonemas de referência foram retirados aleatoriamente da banco de voz de 200 frases, assim como no teste anterior. Tomou-se nota da palavra sob teste, dos fonemas que foram reproduzidos pelo locutor e dos fonemas encontrados como resultado do sistema desenvolvido, então foi realizado um cálculo simples para descobrir a taxa de acerto de cada palavra. No final, foram contabilizados a quantidade total de fonemas nas palavras e a quantidade total de fonemas corretos descobertos pela rotina em questão, para então verificar a taxa de acerto total que o programa obteve.

Tabela 8 – Resultados Reconhecimento de fonemas para voz masculina.

Palavra	**Fonemas Corretos**	**Fonemas Encontrados**	**Taxa de Acerto**
fez	F Ê Z	Ũ Ê Z	66,7%
aluno	A L U N U	A U B N U	60%
cada	K A D A	K R D A	75%
ela	É L A	É Ũ G	33,3%
escute	Ê S K U T Y	Ê S G Ũ J F	33,3%
coracao	K Ó R A S Ã Ô	G U R A Z B Ô	28,6%
encontros	Ẽ K Õ T R U S	Ẽ G Õ I K Ũ S	42,3%
para	P A R A	N B R A	50%
carla	K A R L A	Ô A T A A	40%
educar	Ê D U K A R	N Y U I A P	33,3%
Taxa de acerto total			48%

Tomando os resultados da Tabela 8 como referência, foi concluído que palavras ditas em começo de frases como "FEZ", "ALUNO"e "CADA"foram melhores aceitas pelo sistema, encontrando taxas de acerto de até 75%, porém palavras ditas principalmente em finais de frase, como "CORAÇÃO", conseguiram uma taxa de acerto bem menor, chegando a $28,6\%$ de fonemas encontrados corretamente. Isso mostra que o efeito de "*fading*"(ou

desvanecimento como conhecido no português), que acontece na voz humana ao final de frases, influecia negativamente na utilização desse sistema.

É possível retirar outra conclusão da Tabela 8. As vogais possuem uma taxa de acerto muito maiores que as consoantes. Isso é devido ao tamanho de amostras médio das consoantes ser muito menor que o das vogais, por exemplo, a quantidade de amostras do fonema "B"é cerca de 200 amostras, caso tenha sido amostrado a uma taxa de 22,05 k amostras/segundo, ou seja, a duração médida desse fonema é de 9,07 m s, enquanto que a quantidade de amostras média da vogal "A"é 1500 amostras ou 68,03 m s se amostrada na mesma taxa, algo em torno de 7,5 vezes a duração do fonema "B", assim a correlação cruzada entre as consoantes sempre são parecidas umas com as outras.

Percebe-se também que ocorrem trocas com frequência entre fonemas que possuem grande semelhança, como "S"e "Z"ou "Ô"e "U", pois as vezes nem mesmo o ouvido humano é capaz de distinguir quando os escuta isoladamente, fazendo com que a correlação cruzada entre eles possua grande semelhança com a autocorrelação dos mesmos, confundindo o programa muitas vezes, já que a medida mais importante é essa.

5.4 Avaliação Geral Do Reconhecimento De Fala

Para que seja possível a implementação de um reconhecedor de fala em um codificador utilizado em sistemas móveis celulares, esse sistema tem de ter requisitos como: taxa de acertos elevada, velocidade de compilação alta e custo computacional baixo.

A taxa de acerto do reconhecimento fonético do sitema desenvolvido não está próxima das taxas de sistemas atuais que utilizam *HMM* e redes neurais, que podem chegar a uma taxa de acerto de $82,9\%$ (COSI; HOSOM,), porém é simples de ser implementada e pode, facilmente ser usada como método de refinamento em sistemas atuais. O sistema desenvolvido, no entanto, possui uma taxa de acerto de 100% para reconhecimento pessoal de palavras, o que é um resultado competitivo com qualquer trabalho encontrado na atualidade, como o trabalho de (PRABHAVALKAR PREETHI JYOTHI; FOSLER-LUSSIER, 2010) mostra que já é possível alcançar uma taxa de acerto de $92,1\%$ independente de orador.

A velocidade de compilação do sistema em questão é satisfatória, porém para ser utilizada em um codificador real, precisa de algum aprimoramento. Uma forma de aprimorar seria passar o programa feito em *MatLab* para a linguagem de programação *C*, reduzindo consideravelmente o tempo necessário para que o programa complete sua execução.

O custo computacional do programa é baixíssimo, não requer treinos excessivos nem necessita de várias utilizações para melhorar o desempenho, o que acontece com os métodos

que usam *HMM* e redes neurais. Além de ser um programa simples de implementar, ele pode ser facilmente modificado para outros proprósitos que sejam necessários ao usuário.

6 Considerações finais e trabalhos futuros

Um sistema de reconhecimento pessoal de fala tem como objetivo principal determinar qual fonema, palavra ou sentença foi pronunciada por um locutor predeterminado, podendo ser utilizado como interface entre os segmentadores e codificadores fonéticos para construir um codificador de voz paramético pessoal.

O reconhecimento de fala permite que várias aplicaçoes sejam possíveis. Desde automação residencial, com o acionamento de aparelhos por comandos de voz, a aplicações de processamento texto para fala.

Este trabalho descreve o desenvolvimento de um reconhecedor pessoal de fala baseado em características prosódicas do sinal de voz. O reconhecedor é do tipo fonético, e foi implementado para a utilização em codificadores de voz a baixa taxa de *bits*, que serão usados em sistemas de telefonia móvel celular.

Para um melhor resultado, foi desenvolvido um sistema de segmentação de fala como trabalho paralelo, para que as locuções da base de voz pudessem ter suas palavras extraídas e seus fonemas delimitados. Esse segmentador de voz fonético poderá também ser incluído em um codificador de voz de baixa taxa.

Para avaliar o desempenho do segmentador foram escolhidas 50 sentenças pronunciadas por um locutar do sexo masculino do interior do estado de São Paulo. A segmentação foi então comparada com aquela desenvolvida por um foneticista profissional. Fronteiras com um erro maior que 20 m s foram consideradas incorretas. As fronteiras corretas foram separadas naquelas com erros menores que 10 m s e outras com erro entre 10 e 20 m s para que a precisão do sitema fosse testada. O sistema se mostrou preciso obtendo $77,6\%$ das fronteiras encontradas dentro da faixa de 10 m s, e se mostrou robusto quando encontro $72,41\%$ de todas as fronteiras sem requerer nenhum tipo de teste ou treino anterior, resultados esses que são competitivos com sistemas atuais que dependem de treino.

O reconhecimento de voz foi avaliado segundo um método objetivo. Primeiramente, o sistema foi analisado do ponto de vista de recohecer palavras completas pronunciadas em uma locução. Cinco palavras foram postas à prova em dois testes para cada tipo de voz (feminina e masculina). No total foram 4 testes e 100 usos do sistema (50 para vozes femininas e 50 para vozes masculinas), onde o resultado obtido foi uma taxa de acerto de 100% para o reconhecimento de palavras pessoal, mostrando robustez aliada a um baixo custo computacional.

Em seguida, foram realizados testes de como o sistema se comportava ao tentar reconhecer fonemas vocálicos obtidos por meio da segmentação fonética de frases pronun-

ciadas por um locutor. A taxa de reconhecimento das vogais foi de 74% mesmo com a presença de semivogais, que fonéticamente respresentam os sons de outra vogal, como no caso das vogais "E"e "I". O sistema se mostrou simples e confiável, e não necessitou de um tempo elevado para compilar, fazendo com que ele seja mais uma opção para codificadores de voz reais.

Por último, a avaliação foi por meio do reconhecimento dos fonemas em geral, ou seja, o sistema teria de reconhecer todos os fonemas presentas na lingua portugesa. O resultado não foi muito satisfatório, porém levando em consideração que o sistema não requer um período de treino elevado nem uma quantidade exaustiva de frases para retirar o banco de voz fonético, provavelmente, se o sistema for usado em conjunto com algum método de refinamento, seus resultados poderão ser competitivos com os sistemas mais atuais que chegam a uma taxa de acerto de $82,9\%$ (COSI; HOSOM,).

Por tanto, o trabalho apresentou uma proposta nova, tanto na área de segmentação de voz como na área de reconhecimento de fala, e seus resultados foram, de maneira geral, satisfatórios, podendo ser melhorados com trabalhos futuros.

6.1 Contribuições

6.1.1 Desenvolvimento De Um Segmentador Fonético

A primeira etapa realizada neste trabalho costituiu no desenvolvimento de um sistema de segmentação de fonemas. Esse sistema faz uso de técnicas relacionadas a energia de curta duração dos sinais de voz.

Consiste em um sistema que independe de locutor e de contexto, pois utiliza apenas o cálculo de informações prosódicas do sinal de voz sob teste, que são características presentes em todos os sinais de voz.

O segmentador fonético foi usado no trabalho principalmente para remover regiões de silêncio presentes antes e depois das palavras gravadas em uma locução, porém foi utilizado, em menor proporção, para identificar a fronteira de certos fonemas usados para compor a base de dados fonética.

6.1.2 Desenvolvimento De Um Reconhecedor De Fala Pessoal

O desenvolvimento de um reconhecedor de fala pessoal é a contribuição mais relevante deste trabalho. Foi projetado para ser utilizado principalmente em sistemas móveis celulares, porém, como reconhecedor de palavras, pode possuir várias outras aplicações, inclusive acionamentos de equipamentos industriais.

O reconhecedor é pessoal e utiliza informações prosódicas do sinal de voz, como a

energia e o pitch, para identificar cada fonema ou palavra presente nas locuções sob teste. um reconhecedor de fala mais completo pode ser formado por um sistema de segmentação fonética acoplado ao reconehcedor de fala em si, dispensando assim a atuação humana sobre o sistema.

O sistema em questão pode ser usado para três finalidades distintas. Cada uma requer um pouco mais de cuidado que a anterior. O reconhecedor pode ser usado como reconhecedor de palavras com uma taxa de acerto de 100% (segundo testes realizados), como um reconhecedor de fonemas vocálicos com uma taxa de acerto de 74% (segunso testes realizados) ou ainda como reconhecedor fonético com uma taxa de acerto de 48% (segundo testes realizados, todos eles não necessitam de treinamento exaustivo, apenas a gravação de cada um dos fonemas que será utilizado já é o bastante.

6.2 Dificuldades

A maior dificuldade do trabalho proposto foi encontrar sujeitos dispostos a gravar uma grande quantidade de frases e palavras, o que foi resolvido graças a uma base de voz cedida por (SELMINI, 2008) e a ajuda de duas pessoas para gravar os testes com palavras. Outro pequeno problema decidir o local das gravações para evitar a adição de ruído aso sinais.

Uma preocupação constante do trabalho foi manter os programas simples, para que sejam facilmente modificados pelo usuário de acordo com as necessidades do mesmo. A escolha de frases e palavras que procurassem representar a maior parcela de fonemas encontrados na lingua portuguêsa falada foi um dos desafios do projeto.

6.3 Trabalhos Futuros

Como o reconhecedor de fala pessoal desenvolvido nesse trabalho tem como objetivo a utilização em um codificador de voz utilizado na telefonia móvel celular, a continuidade futura do projeto deve se dar pela implementação dos sistemas desenvolvidos em um codificador real, para que seja possível a realização de testes, comparando-os com os disponíves no mercado. Para isso, uma série de modificações deve ser realizada nos sistemas propostos:

1. O sistema de segmentação fonética necessita de algum tipo de refinamento, facilmente resolvido empregando técnicas bem desenvolvidas na atualidade, como o uso de redes neurais.

2. A diminuição de falsas fronteiras encontradas nas frases segmentadas pelo sistema desenvolvido poderá ser sanada com a utilização da transcrição fonética da frase sob

teste.

3. Criação de uma etapa de treinamento para o sistema de reconhecimento de fala, onde o reconhecedor terá como entrada varias cópias de cada fonema, e decidirá qual cópia seria a mais apropriada para uso como sinal de referência, aumentado de forma significativa a taxa de acertos, como pode ser percebido nos testes realizados.

4. Implementação de um sistema completamente automático de segmentação e reconhecimento de fala, onde o usuário deverá pronunciar a locução e o sistema automaticamente segmentará a frase em fonemas, passando a informação para o reconhecedor fonético que terá como saída os fonemas pronunciados pelo orador.

5. Como último ponto para o desenvolvimento do codificador, será necessário a criação de um sintetizador por concatenação, para que os fonemas que são encontrados pelo reconhecedor fonético possam ser inseridos novamente em uma mesma frase, sem perda de inteligibilidade.

Após o desenvolvimento de todas as etapas citadas anteriormente, o codificador de voz poderá ser desenvolvido criando um sistema que possa conectar cada uma de suas partes: o segmentador, o reconhecedor e o sintetizador.

Referências

ALENCAR, M. S. *Telefonia Celular Digital.* São Paulo: Editora Érica, 2012.

BUERA ANTONIO MIGUEL, E. L. O. S. A. O. L. Robust speech recognition with on-line unsupervised acoustic feature compensation. *Communication Technologies Group (GTC), 13A*. Citado na página 31.

CHEN, J.; GUPTA, J. Estimation of the shift parameter of headway distributions using crosscorrelation function method. *IEEE Transaction on Audio, Speech, and Language Processing*. Citado na página 37.

COSI, P.; HOSOM, J.-P. High performance "general purpose"phonetic recognition for italian. Citado 2 vezes nas páginas 58 e 62.

HUANG, X. D.; ARIKI, A.; JACK, M. A. *Hidden Markov Models for Speech Recognition.* [S.l.]: Edinburgh University Press, 1990.

JARIFI, S.; PASTOR, D.; ROSEC, O. A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis. *Speech Communication*, v. 50, p. 67–80, 2008. Citado na página 43.

LEVINSON, L. R. R. S. E.; SONDHI, M. M. An Introduction to the Application of the Theory of Probabilist Functions of a Markov Process to Automatic Speech Recognition. *The Bell System Technical Journal*, v. 62, n. 4, p. 1035–1068, April 1983.

NEWELL, A. Speech understanding systems. *Academic Press, N*, New York, 1975.

PARANAGUá, E. D. S. *Segmentação automática do sinal de voz para sistemas de conversão texto-fala*. Dissertação (Tese de Doutorado) — Universidade Federal do Rio de Janeiro, 2012. Citado 3 vezes nas páginas 33, 43 e 44.

PARK, S. S.; KIM, N. S. On using multiple models of automatic speech segmatation. *IEEE Transaction on Audio, Speech, and Language Processing*, v. 15, n. 8, p. 2202–2212, November 2007. Citado na página 43.

PONTES VANESSA P. VIEIRA, M. I. R. G. A. A. I. P. Paulo a. l. Características das vozes roucas, ásperas e normais: análise acústica espectrográfica comparativa. *Revista Brasileira de Otorrinolaringologia*, v. 68, n. 2, p. 182–190, Abril 2002. Citado na página 53.

PRABHAVALKAR PREETHI JYOTHI, W. H. J. M. R.; FOSLER-LUSSIER, E. Investigations into the crandem approach to word recognition. *Human Language Technologies*, p. 725–728, June 2010. Citado na página 58.

PROAKIS, J. G.; MANOLAKIS, D. G. *Digital Signal Processing Principles, Argorithms and Applications*. Upper Saddle River: [s.n.]. Citado 2 vezes nas páginas 29 e 30.

ROCHA, R. B. *Desenvolvimento de Um Codificador de Voz Pessoal de Baixa Taxa Baseado em Modelos de Markov Escondidos*. Dissertação (Tese de Mestrado) — Universidade Federal de Campina Grande, 2012. Citado na página 33.

SELMINI, A. M. *Sistema Baseado em Regras para o Refinamento da Segmentação Automática de Fala*. Dissertação (Tese de Doutorado) — Universidade Estadual de Campinas, 2008. Citado 4 vezes nas páginas 33, 43, 44 e 63.

SILVA, C. P. A. *Um Software de Reconhecimento de Voz para Português Brasileiro.* Dissertação (Dissertação de Mestrado) — Universidade Federal do Pará, Belém, Brasil, 2010.

SILVA, D. D. C. da. *Reconhecimento de Fala Contínua para o Português Brasileiro em Sistemas Embarcados*. Dissertação (Tese de Doutorado) — Universidade Federal de Campina Grande, Campina Grande, Brasil, Dezembro de 2011.

TEVAH, R. T. *Implementação de um Sistema de Reconhecimento de Contínua Com Amplo Vocabulário Para o Português Brasileiro.* Dissertação (Dissertação de Mestrado) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, Junho de 2006.

TOLEDANO, D. T.; GOMEZ, L. A. H.; GRANDE, L. V. Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, v. 11, n. 6, p. 617–325, November 2003. Citado na página 43.

TRAUNMULLER, A. E. H. The frequency range of the voice fundamental in the speech of male and female adults. *IEEE Transaction on Audio, Speech, and Language Processing*, Stockholm, Sweden. Citado na página 37.

Anexos

ANEXO A – Reconhecedor De Fala

```
close all
clear all
clc

n = 1;

tamanho(1) = length(wavread('a1'));  %Sinal de Referencia do fonema A
tamanho(2) = length(wavread('e1'));  %Sinal de Referencia do fonema E
tamanho(3) = length(wavread('i1'));  %Sinal de Referencia do fonema I
tamanho(4) = length(wavread('o1'));  %Sinal de Referencia do fonema O
tamanho(5) = length(wavread('u1'));  %Sinal de Referencia do fonema U


tam = max(tamanho);

[sobteste, fs] = wavread('ateste');  %Sinal sob teste
fonema(:,1) = [wavread('a1'); zeros(tam-length(wavread('a1')),1)];
fonema(:,2) = [wavread('e1'); zeros(tam-length(wavread('e1')),1)];
fonema(:,3) = [wavread('i1'); zeros(tam-length(wavread('i1')),1)];
fonema(:,4) = [wavread('o1'); zeros(tam-length(wavread('o1')),1)];
fonema(:,5) = [wavread('u1'); zeros(tam-length(wavread('u1')),1)];

nfft = min(1023,length(sobteste));
st=specgram(sobteste, nfft, fs, hanning(511),380);
absolutet=transpose(abs(st));
at=sum(absolutet);
at_norm=(at-min(at))/(max(at)-min(at));
Ft=transpose(at_norm);

ok = 0;
while n <= length(tamanho)

s(:,:,n) = specgram(fonema(:,n), nfft, fs, hanning(511),380);

absolute(:,:,n) = transpose(abs(s(:,:,n)));

a(:,n)=sum(absolute(:,:,n));

a_norm(:,n)=(a(:,n)-min(a(:,n)))/(max(a(:,n))-min(a(:,n)));

F(:,n)=transpose(a_norm(:,n));
```

```
[x(:,n), lag] = xcorr(Ft,F(:,n));
[mx, indice(n)] = max(x(:,n));
shift(n) = lag(indice(n));

n = n+1;

end

AUX = sort(abs(shift));
menor_shift = AUX(1);
menor_shift2 = AUX(2);

if abs(menor_shift - menor_shift2)>=2  %Se a diferenca entre os desvios de
    i = find(shift==menor_shift);      %frequencia forem maiores que 2, o
    fprintf('Fonema %g\n', i);         %fonema eh identificado.
    ok = 1;
else                                   %Se a diferenca nao for maior que 2
    n = 1;                             %utiliza-se o algoritmo do erro
    while n <= length(tamanho)         %medio quadretico.

        if indice(n)<length(x(:,n))/2
            q=1:(indice(n)-1);
            p=indice(n)+length(q):-1:indice(n)+1;
            length(p);
            length(q);
            x_left=x(q,n);
            min(x_left);
            x_right=x(p,n);
            min(x_right);
            error(n)= mean((abs(x_right-x_left)).^2);
        else
            q=1+shift(n)*2:indice(n)-1;
            p=length(x(:,n)):-1:indice(n)+1;
            length(q);
            length(p);
            x_left=x(q,n);
            x_right=x(p,n);
            error(n)= mean((abs(x_right-x_left)).^2);
        end

        n = n+1;

    end
end

if ok==0
```

```
    error
    i = find(error==min(error));  %O fonema que obtiver o menor erro eh
    fprintf('Fonema %g\n', i);    %escolhido como correto.
    ok = 0;
end
```

ANEXO B – Locuções Usadas Para Os Testes

01 - Muitas pessoas participam da construção de um texto

02 - Tenho vergonha do meu país

03 - Cada aluno fez a sua avaliação

04 - Escute o seu coração

05 - Um dos encontros mais emocionantes foi o último do ano

06 - Para karla, educar é uma atitude de esperança

07 - Ela afirmou que tinha sido um fato

08 - O jabuti comi muita jabuticaba

09 - O texto da professora trazia bons pensamentos

10 - Minha mãe estava nervosa

11 - O cavalo pulou a cerca

12 - A chave do chaveiro inferrujou

13 - Conhecera grandes pescadores no mar

14 - Ricardo escreve de modo fácil

15 - Fica aqui o convite para sua leitura

16 - A essa altura todos estavam emocionados

17 - Ele gostava mais das músicas animadas

18 - A atriz recebeu o prêmio com entusiasmo

19 - Todos correram para pegar o cachorro

20 - O aparelho de jantar caiu no chão

21 - O sol faz propaganda do verão

22 - O povo não estava feliz com o governo

23 - Levante a vela para limpar a mesa

24 - A chuva derrubou várias casas

25 - O feijão da feijoada estava azedo

26 - O seu primo comeu todas as laranjas

27 - A guerra acabou com fracasso

28 - Tenha uma boa noite

29 - Era uma ótima opção

30 - O calor me deixou de garganta seca

31 - O anjo caiu do céu

32 - Cobras são animais peçonhentos

33 - A gente sempre faz o que pode

34 - Meu computador não conversa com a impressora

35 - Os padres se reuniram na sacristia

36 - Meu gravador de dvd quebrou

37 - Diga sempre a verdade para seus pais

38 - Diga pata bem baixinho

39 - Beba muita água durante o dia

40 - Ganhei um novo aparelho de dvd

41 - Temos um belo presente para você

42 - Bebidas são prejudiciais a saúde

43 - Por favor fume longe de mim

44 - Corra lola corra

45 - É a coisa mais linda do mundo

46 - Bonito é a cidade de bonito

47 - A rússia tem uma cozinha com história

48 - A faculdade abriu concurso

49 - Para ter boas notas é preciso estudar

50 - A criança acredita em fadas