Um Ambiente para Processamento Digital de Sinais Aplicado à Comunicação Vocal Homem-Máquina

Márcio G. Passos e Patric L. Silva
marcio_passos@yahoo.com.br e lacouth@gmail.com

Silvana Luciene do N. Cunha Costa [1]
silvanacunhacosta@gmail.com

Benedito G. Aguiar Neto e Joseana M. Fechine
bganeto@cct.ufcg.edu.br e joseana@dsc.ufcg.edu.br

Resumo: *Este trabalho trata do estudo e implementação de técnicas de processamento digital de sinais que são utilizadas em sistemas de resposta vocal como: sistemas de reconhecimento de voz, reconhecimento de locutor e sistemas de síntese de voz. Foi utilizada a linguagem de programação C na elaboração de funções que caracterizam os sinais de voz. Foram implementados algoritmos para conversão amostras-texto, detecção de início e fim, pré-ênfase, janelamento de Hamming e extração de parâmetros temporais. Estes parâmetros incluem energia, taxa de cruzamento por zero, número total de picos, diferença no número de picos e coeficiente de correlação normalizado. Foram propostos algoritmos diferenciados para a taxa de cruzamento por zero e para o detector de início e fim. Com o intuito de criar um ambiente didático, construiu-se uma interface gráfica amigável utilizando uma plataforma de desenvolvimento orientada a eventos.*

Palavras Chave: *processamento de sinais, comunicação vocal homem-máquina.*

1. Introdução

Dentre as várias áreas que compõem o campo da comunicação por voz, a área da comunicação vocal homem-máquina é uma das mais interessantes e estimulantes. O desejo, bem como a necessidade das pessoas se comunicarem com as máquinas da maneira mais natural de comunicação – a voz humana – tem dado grande impulso ao crescimento desta área (RABINER; SHAFER, 1978).

Por não requererem nem as mãos nem os olhos do usuário para a sua operação, os sistemas de entrada vocal podem ser utilizados em diversas aplicações, como por exemplo: controle de tráfego aéreo, auxílio a portadores de deficiência física, controle de qualidade e inspeção e controle de acesso a ambientes restritos.

Dos sistemas de entrada vocal hoje disponíveis, destacam-se os sistemas de reconhecimento automático de voz (RAV) e os sistemas de reconhecimento automático de locutor (RAL). Nas aplicações RAV e RAL é necessária uma preparação ou pré-processamento dos sinais da voz. As técnicas de pré-processamento permitem a extração de características que realmente merecem destaque, pois atuam no sentido de fornecer não somente a informação de interesse ao processamento de determinada amostra de som, mas também ocasionar uma redução considerável na quantidade de informações a serem processadas. Tais informações serão responsáveis pela produção de padrões entre determinada referência registrada (PETRY et al, 2000).

Este trabalho apresenta os resultados obtidos utilizando-se técnicas para processamento digital de sinais da fala. A primeira seção tratará do processo de aquisição, gravação e digitalização da voz. Em seguida, é apresentado um algoritmo diferenciado para detecção de início e fim de palavras, bem como as técnicas de pré-ênfase, segmentação, janelamento das amostras e extração de alguns parâmetros temporais do sinal da fala.

Nomenclatura

a = constante, igual a 0,95.
y = sinal pré-enfatizado
x = sinal amostrado
M = número total de amostras
s = amostra após janelamento
L = tamanho (tempo) da janela de *Hamming*

N_a = tamanho (amostras) da janela de *Hamming*
Eseg = energia segmental
COR = coeficiente de autocorrelação normalizado
PNEG = picos negativos
PPOS = picos positivos
NTP = número total de picos
DPN = diferença do número de picos
TCZ = taxa de cruzamento por zero

Índices
n relativos à amostra

2. Processamento do Sinal de Voz

A – Aquisição do sinal

A aquisição dos sinais de voz é realizada inicialmente utilizando-se um microfone. Este converte as variações que a fala causa na pressão do ar em variações de tensão elétrica. O próximo passo do sistema é a amostragem e digitalização das variações de tensão. Geralmente a passagem do sinal de voz da forma analógica para a digital é feita utilizando a modulação por codificação de pulsos (PCM – *pulse code modulation*). Com a finalidade de ser manipulado por um sistema digital, o sinal de voz é representado por uma seqüência de pulsos binários, codificados com uma quantidade de bits proporcional a qualidade e fidelidade desejadas. Para sinais de voz, esta codificação é feita geralmente com 8 ou 16 bits.

Neste trabalho, para a aquisição dos sinais de voz utilizou-se um microcomputador PC com placa de som e microfone comum. O software usado para gravação do som em mídias digitais foi o *GoldWave*© versão 4.26, e o formato de gravação escolhido foi o padrão *WAV*. Este formato de gravação é um dos mais utilizados para este tipo de aplicação, e contém um cabeçalho de 44 bytes com informações sobre o próprio arquivo.

Com a obtenção do arquivo *WAV,* contendo o sinal de voz na forma digital, pode-se agora manipulá-lo no ambiente de processamento digital de sinais implementado. Como passo inicial, foi implementado um algoritmo que extrai as informações do cabeçalho do arquivo *WAV*. Estas informações incluem número de amostras, freqüência de amostragem, tipo de modulação e número de bits por amostra. Foi elaborado um algoritmo para converter arquivos *WAV* para o modo texto. Para manter a compatibilidade, as amostras são multiplicadas por constantes de valores previamente estabelecidos, a fim de que o software *GoldWave*© possa fazer a reprodução audível dos arquivos de voz também no modo texto.

B – Detecção de início e fim das palavras

Durante o processo de gravação da voz, inevitavelmente, parte do tempo alocado para a elocução é ocupado com silêncio ou ruído ambiente. Assim, quando uma palavra é gravada, as amostras sem informação útil podem ser descartadas de maneira segura. A separação das amostras representativas de voz das amostras de silêncio é chamada de detecção de início e fim de palavra.

Em sistemas de palavras isoladas, a detecção de início e fim é fundamental por duas razões principais (COSTA, 1994):

1. A classificação correta da palavra é criticamente dependente da precisão dessa detecção.
2. Os cálculos necessários para o processamento do sinal de voz são minimizados quando o início e o fim são localizados com precisão.

O algoritmo para detecção de início e fim, proposto neste trabalho, é mostrado na Fig. (1).

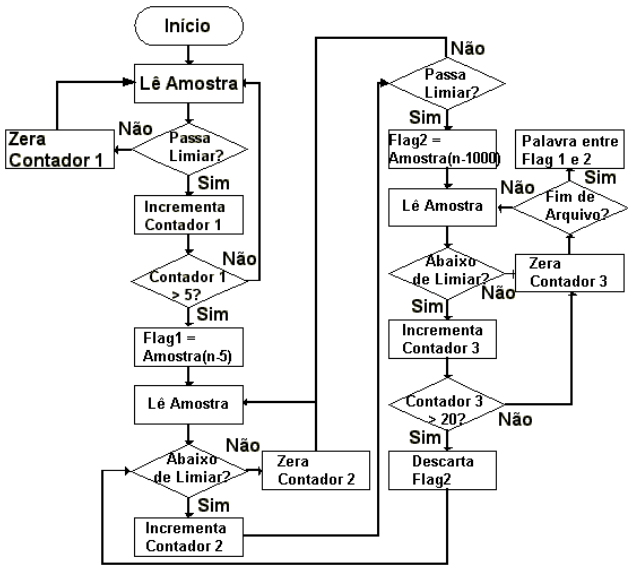

Figura 1. Fluxograma do algoritmo para detecção de início e fim de palavras.

Este algoritmo consiste na leitura ordenada de amostras individuais na busca de um grupo de 5 amostras consecutivas que ultrapassam um limiar pré-determinado. Encontrado este grupo, indica-se que o início da palavra é a primeira amostra deste grupo. Após a determinação do início da palavra, o algoritmo passa a buscar o fim desta.

Esta busca é feita através da análise ordenada de amostras, de forma que se 1000 dessas amostras consecutivamente estiverem abaixo de um limiar

previamente estabelecido, é então delimitado provisoriamente o fim da palavra.

Mesmo após essas detecções, o algoritmo continua a ler as amostras à procura de um novo grupo, desta vez com 20 amostras seguidas, que estejam acima de um certo limiar. Caso seja encontrado este último grupo, é reiniciada a busca pelo fim da palavra. Isto evita que o algoritmo venha a identificar incorretamente o fim de elocuções que possuem intervalos curtos de silêncio entre fonemas.

De posse das amostras que compõem o sinal, é criado um novo arquivo de extensão *WAV* que possui a palavra delimitada. As Fig. (2) e (3) ilustram a forma de onda da elocução "esquerda", codificada com 8 bits e taxa de amostragem de 11025 Hz, antes e depois da detecção de início e fim respectivamente.

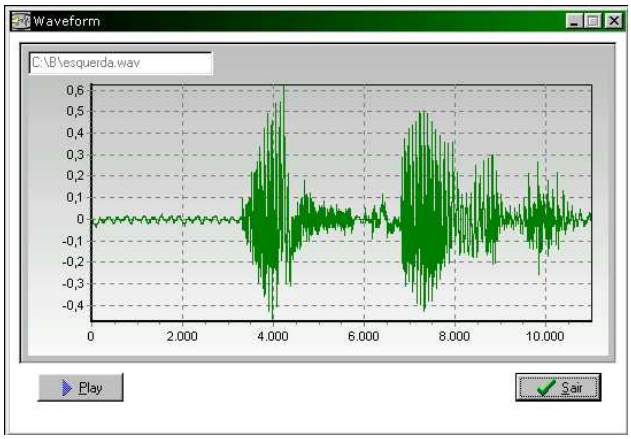

Figura 2. Forma de onda da palavra "esquerda" com codificação em 8 bits e taxa de amostragem de 11025 Hz.

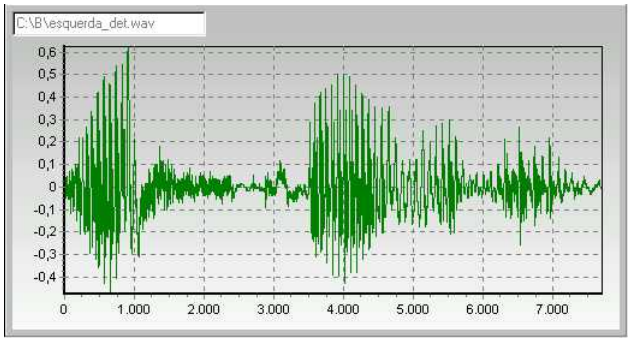

Figura 3. Forma de onda da palavra "esquerda" com início e fim detectados .

C – Pré-Ênfase

O sinal de voz apresenta baixas amplitudes nas altas freqüências o que as torna especialmente vulneráveis ao ruído. Tais freqüências são responsáveis pela geração dos sons surdos (COSTA, 1994).

A pré-ênfase objetiva eliminar uma tendência espectral de aproximadamente –6dB/oitava na fala irradiada dos lábios. Essa distorção espectral não traz informação adicional e pode ser eliminada através de um filtro, que proporcione um ganho de +6dB/oitava, fazendo com que o espectro se nivele. Em um sistema digital a pré-ênfase pode ser implementada como um circuito analógico, precedendo o amostrador, ou diretamente na informação digital através de um filtro do tipo de resposta ao impulso finito (*finite impulse response*) FIR de primeira ordem (PETRY et al, 2000). A Eq. (1) descreve o processo de pré-ênfase realizado neste trabalho:

$$y(n) = x(n) - ax(n-1) \ , \ com\ 1 \le n < M \qquad (1)$$

em que:

$y(n)$ = sinal pré-enfatizado
$x(n)$ = sinal amostrado
M = número de amostras
a = constante, neste caso, usou-se $a = 0{,}95$

A Fig. (4) ilustra o processo de pré-ênfase, com a elocução "esquerda".

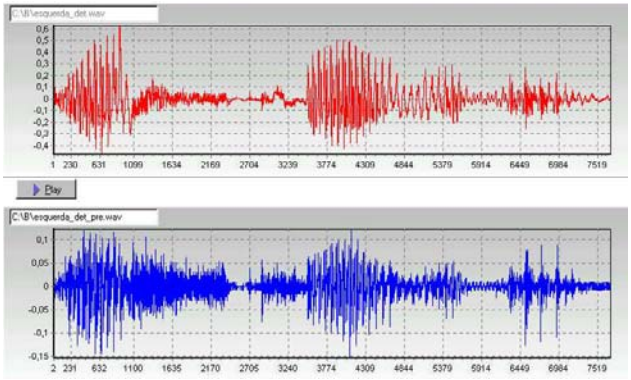

Figura 4. Forma de onda da palavra "esquerda" – original e pré-enfatizada.

D – Segmentação e Janelamento

Em processamento digital de sinais de voz, é necessário trabalhar com segmentos ou *frames* do sinal. Estes segmentos são da ordem de milisegundos, assumindo que nestes pequenos intervalos os sinais podem ser considerados razoavelmente estacionários. Foi definido um frame de voz como sendo o produto de uma janela discreta $w(n)$ de tamanho L, pela seqüência de voz pré-enfatizada (PETRY et al, 2000).

Neste trabalho, optou-se por utilizar a janela de *Hamming*. Este tipo de janela apresenta boas características espectrais bem como atenua a

transição entre quadros adjacentes. O ambiente de processamento de sinais permite que o usuário estabeleça o tamanho das janelas, em milisegundos, de acordo com a sua necessidade. As janelas geralmente são sobrepostas entre si, para que haja uma variação gradual dos parâmetros entre elas. Foi utilizada uma sobreposição fixa entre janelas de 50%. A representação matemática do janelamento de *Hamming* é descrita na Eq. (2).

$$s(n) = \begin{cases} 0 & n < 0 \\ 0{,}54 - 0{,}46\cos\left(\dfrac{2\pi n}{330-1}\right) & 0 \le n < L \\ 0 & n \ge L \end{cases} \quad (2)$$

3. Extração de Parâmetros Temporais do Sinal de Voz

Para aplicações em RAV e RAL, é necessária a extração de informações úteis sobre o sinal da voz. Para se obterem tais informações, foram utilizadas técnicas baseadas no domínio do tempo, pois apresentam baixo custo computacional e produzem informações úteis acerca do sinal processado. Segue-se uma explanação sobre cada um dos parâmetros obtidos pelo ambiente de processamento de sinais.

A – Energia

A energia segmental, E_{seg}, é utilizada para diferenciação do silêncio, sons surdos, sons sonoros e fricativos. Este parâmetro é obtido simplesmente somando-se os quadrados das amplitudes das N_a amostras da janela em análise. A energia por segmento para sinais estacionários é dada pela Eq. (3) (RABINER; SHAFER, 1978).

A Fig. (5) mostra a variação da energia ao longo da elocução "esquerda".

$$Eseg = \sum_{n=0}^{Na-1} [s(n)]^2 \quad (3)$$

B – Taxa de Cruzamento por Zero

As aplicações em que se utilizam métodos de análise no domínio do tempo, a Taxa de Cruzamento por Zero (TCZ) é um parâmetro usado na detecção de blocos com sons surdos (ex. consoante "s"), sonoros (ex. vogal "a") e consoantes fricativas (ex. consoante "f") (RABINER; SHAFER, 1978).

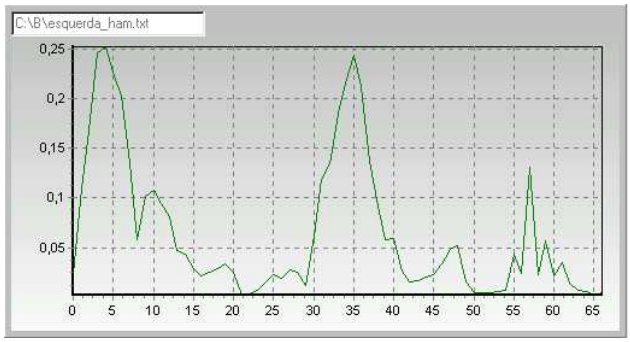

Figura 5. Energia segmental da palavra "esquerda".

Usualmente este parâmetro é definido por Eq. (4) e (5):

$$TCZ = \frac{1}{2}\sum_{n=1}^{Na-1} |\operatorname{sgn}[s(n)] - \operatorname{sgn}[s(n-1)]| \quad (4)$$

em que:

$$\operatorname{sgn}[s(n)] = \begin{cases} 1 & , \quad se\ s(n) \ge 0 \\ -1 & , \quad se\ s(n) < 0 \end{cases} \quad (5)$$

Em janelas que apresentam elevado número de amostras de valor nulo, esta definição não corresponde à realidade, pois, na transição de uma amostra negativa para uma amostra nula será contado um cruzamento por zero mesmo que a próxima amostra também seja negativa.

Neste trabalho, é proposto um algoritmo que forneça uma maior precisão no resultado final da Taxa de Cruzamento por Zero. Em síntese, o algoritmo ignora as amostras nulas e incrementa a TCZ, apenas, quando houver realmente uma inversão dos sinais das amostras consecutivas não nulas. A Fig. (6) ilustra a variação da Taxa de Cruzamento por Zero ao longo da palavra "esquerda" e a Fig. (7) o fluxograma do algoritmo em questão.

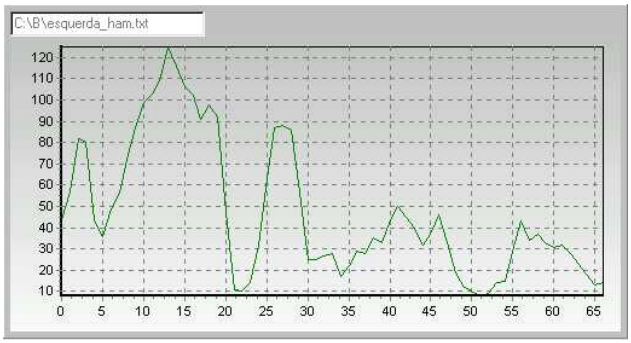

Figura 6. Variação da TCZ ao longo da palavra "esquerda".

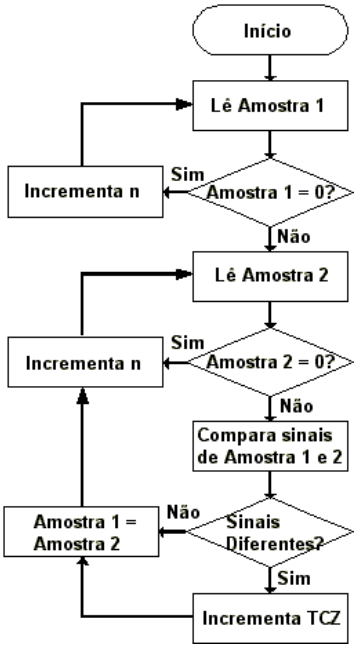

Figura 7. Fluxograma do algoritmo implementado para cálculo da TCZ.

C– Número Total de Picos e Diferença entre o Número de Picos

O sinal de voz apresenta trechos que se repetem quase periodicamente e trechos basicamente aleatórios, sem nenhuma periodicidade. Em sistemas que trabalham com reconhecimento ou síntese de voz, a detecção de diferentes modos de excitação permite a classificação dos sinais de voz em: sons sonoros, sons surdos e sons explosivos.

O Número Total de Picos (NTP) é um parâmetro que auxilia a detecção de fricativos surdos de pequena intensidade como o /f/. A Diferença entre o Número de Picos (DPN) ajuda o reconhecimento de sons fricativos sonoros que podem ser facilmente confundidos com vogais de pequena intensidade (VIEIRA, 1989).

O algoritmo a seguir mostra o procedimento de cálculo do NTP e DPN. As variáveis Picos Positivos (PPOS) e Picos Negativos (PNEG) correspondem ao número de picos da parte positiva e da parte negativa do sinal, respectivamente.

PPOS = 0
PNEG = 0

para [i = 1; i < N; i = i + 1]
  se [($s_n \geq 0$) e ($s_n \geq s_{n-1}$) e ($s_n > s_{n+1}$)]
    PPOS = PPOS + 1;

  se [($s_n < 0$) e ($s_n \leq s_{n-1}$) e ($s_n < s_{n+1}$)]
    PNEG = PNEG + 1;

NTP = PPOS + PNEG;
DPN = PPOS – PNEG;

A Fig. (8) ilustra o Número Total de Picos da elocução "esquerda" e a Fig. (9) ilustra a variação da Diferença entre o Número de Picos na mesma elocução.

Figura 8. Variação do parâmetro NTP na palavra "esquerda".

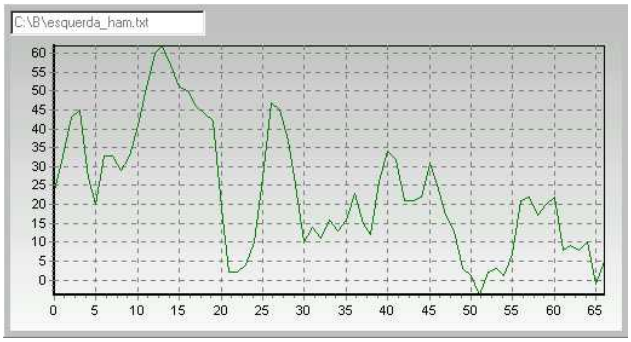

Figura 9. Variação do parâmetro DPN na palavra "esquerda".

D– Coeficiente de Autocorrelação Normalizado

Este parâmetro tem bastante utilidade na distinção de sons surdos e sonoros. Esse coeficiente tem valores próximos a unidade para sons sonoros, por serem sinais que possuem alta concentração de energia. Logo, para sons com baixa concentração de energia como os sons surdos este parâmetro aproxima-se de zero (LIMA, 1994).

O valor do coeficiente de autocorrelação é determinado pela Eq. (6):

$$COR = \frac{\sum_{n=1}^{N} [s_n \cdot s_{n-1}]}{\sqrt{\left[\left(\sum_{n=1}^{N} s_n^2\right) \cdot \left(\sum_{n=0}^{N-1} s_n^2\right)\right]}} \qquad (6)$$

A Fig. (10) ilustra a variação do Coeficiente de Correlação ao longo dos blocos da palavra "esquerda".

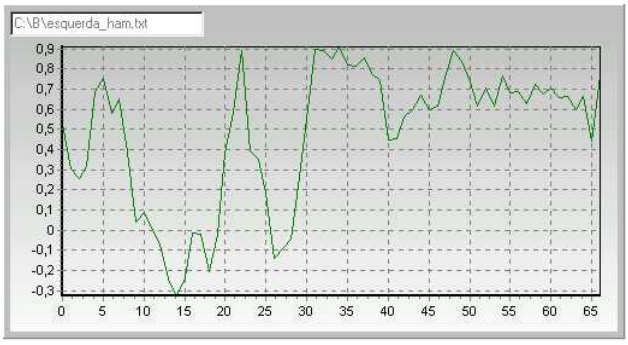

Figura 10. Variação do parâmetro COR ao longo da palavra "esquerda".

4. Interface Gráfica

De forma a tornar-se mais compreensíveis as etapas do processo de extração de características a partir de um arquivo *WAV* , foi implementada uma interface gráfica, através do ambiente de desenvolvimento Borland C++ Builder, amigável e intuitiva que contém todos os algoritmos, anteriormente, discutidos. As figuras contendo os resultados obtidos neste trabalho foram retiradas do *software* implementado. A Fig. (11) apresenta a interface inicial do programa.

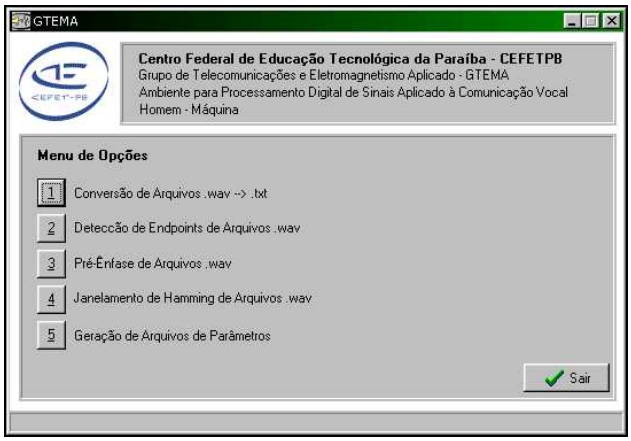

Figura 11. Interface inicial do ambiente de processamento digital de voz.

5. Conclusões

Neste trabalho foram implementadas rotinas que podem servir de base para qualquer sistema de reconhecimento de voz, de locutor ou sistemas de síntese de voz. O programa final é flexível, pois permite que se trabalhe com arquivos *WAV* de 8 ou 16 bits, além de executar passo-a-passo o processamento do sinal da voz o que facilita a compreensão de cada procedimento.

Na fase de detecção de início/fim de palavras foi proposto um algoritmo que demonstrou bons resultados em ambiente de laboratório, mas que deve ainda ser otimizado.

A extração de parâmetros que o programa executa retorna variáveis expressivas para caracterização do sinal da fala. Neste estágio foi implementado um algoritmo diferenciado, do geralmente utilizado, para a contagem da Taxa de Cruzamento pelo Zero, que obteve resultados excelentes na caracterização de sinais de 8 bits.

Outros passos devem ser dados para implementação de um sistema de reconhecimento de voz ou de locutor como a determinação dos coeficientes LPC, Mel, Mel-Cepstrais entre outros. As técnicas de parametrização dos modelos como Modelos de Markov Escondidos (HMM), Redes Neurais Artificiais ou técnicas híbridas podem ser utilizadas. No entanto, o trabalho até aqui realizado serve como base para um sistema que use qualquer dessas técnicas.

Pretende-se ainda, em etapas posteriores, avaliar o desempenho dos algoritmos aqui propostos com os algoritmos existentes na literatura.

6. Referências

COSTA, W. C. da A. **Reconhecimento de Fala Utilizando Modelos de Markov Escondidos (HMM's) de Densidades Contínuas**. 1994. Dissertação (Mestrado em Engenharia Elétrica) – Departamento de Engenharia Elétrica, UFPB, Campina Grande.

LIMA, A. B. O. **Sistema de Resposta Vocal – VOCODER LPC**. 1994. Relatório Técnico – Departamento de Engenharia Elétrica, UFPB, Campina Grande.

PETRY, A..; ZANUZ, A.; BARONE, D. A. C. **Utilização de técnicas de processamento digital de sinais para identificação automática de pessoas pela voz**. 2000. Relatório Técnico – UFRGS, Porto Alegre.

VIEIRA, M. N. **Módulo Frontal para um Sistema de Reconhecimento Automático de Voz**. 1989. Dissertação (Mestrado em Engenharia Elétrica) – Departamento de Engenharia Elétrica, UNICAMP, Campinas.

RABINER, L. R., SHAFER, R. W., **Digital Processing of Speech Signals,** Prentice Hall, 1978.

Responsabilidade de autoria

As informações contidas neste artigo são de inteira responsabilidade de seus autores. As opiniões nele emitidas não representam, necessariamente, pontos de vista da Instituição e/ou do Conselho Editorial.