Reconhecimento de voz utilizando Wavelet e Classificador Neural

Oséas Pereira Rocha
CEFETCAMPOS – Centro Federal de Educação Tecnológica de Campos
Rua Dr Siqueira 273, Bairro Dom Bosco, Campos dos Goytacazes, RJ, Brasil
oseasrocha@cefetcampos.br

Antonio Carlos Gay Thomé
UFRJ – Universidade Federal do Rio de Janeiro
IM – Instituto de Matemática / NCE – Núcleo de Computação Eletrônica
Caixa Postal 2324 – Ilha do Fundão, Rio de Janeiro, RJ, Brasil
thomé@nce.ufrj.br

Sandro Reis Rocha Barros
CEFETCAMPOS – Centro Federal de Educação Tecnológica de Campos
Rua Dr Siqueira 273, Bairro Dom Bosco, Campos dos Goytacazes, RJ, Brasil
sandro@cefetcampos.br

Resumo

Este trabalho consiste na exploração das sub-bandas de detalhe, obtidas pela Análise Wavelet Multi-resolução, como descritores do sinal de voz para fins de reconhecimento. Com base nesta análise, desenvolveu-se um algoritmo alternativo para determinação dos pontos extremos, o qual mostrou-se robusto a diversos tipos de ruído ambiental e concebeu-se um modelo de classificador neural para o reconhecimento de comandos isolados de voz. Os testes foram realizados no modo independente de locutor, atingindo um desempenho de 93,22% de acertos e 99,37% de acertos no modo dependente de locutor.

1. Introdução

O paradigma do reconhecimento de voz tem experimentado diversas técnicas para extração de características ao longo dos anos. Diversos algoritmos tem sido testados na busca de um desempenho cada vez melhor. Uma técnica proposta mais recentemente para utilização em processamentos de sinais (imagens e sons) e que tem se mostrado, em diversos casos, superior as técnicas tradicionais é a Análise Wavelet [1]. Muitas análises, que antes eram executadas com base na Transformada de Fourier, hoje são executadas por Transformada Wavelet obtendo melhores resultados. Inserido neste contexto, o trabalho descrito neste artigo apresenta um relatório de experiências realizadas com objetivo de conceber um sistema de reconhecimento de palavras isoladas utilizando como descritores, as magnitudes de tempo curto, extraídas das sub-bandas de detalhe obtidas pela Análise Wavelet Multi-resolução sobre o sinal de voz.

2. Análise Wavelet Multi-resolução

Para executar a Análise *Wavelet Multi-resolução*, o sinal a ser analisado é introduzido em dois filtros complementares – passa alta e passa baixa, originando dois novos sinais. O sinal "A" (aproximação) contendo as componentes de baixa freqüência e o sinal "D" (detalhe) com as componentes de alta freqüência do sinal original. Este procedimento está ilustrado na Figura 2-1.

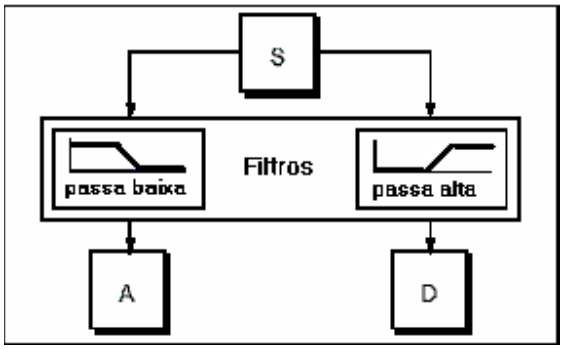

FIGURA 2-1 - Decomposição do sinal original em aproximação e detalhe.

Cada aproximação resultante é submetida a um novo processo de filtragem resultando em um novo nível de aproximação e detalhe. O processo pode ser

repetido até o nível desejado de decomposição, conforme ilustra a figura 2-2.

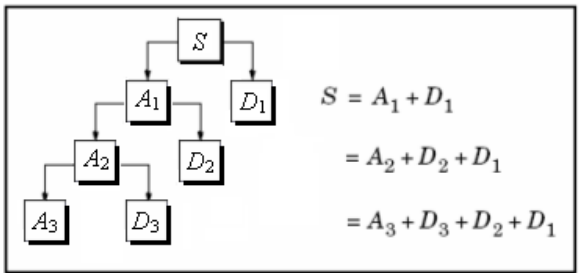

FIGURA 2-2 - Decomposição do sinal original em três níveis.

A freqüência central referente a cada nível pode ser obtida pela equação 2-1 [Misit at al 2002].

$$F_n = \frac{F_c}{2^n . \Delta} \qquad (2\text{-}1)$$

onde:

- F_n representa a freqüência central para a sub-banda do nível *n* (em Hz).
- Δ representa o período de amostragem (em segundos) utilizado na captura do sinal;
- F_c representa a freqüência central associada a função geradora da Wavelet Mãe.

Para este trabalho, foi utilizada a Wavelet Mãe Coif5 da família Coiflet cuja freqüência central associada à função é de 0.6897 Hz [Misiti at al 2002]. Para captura do sinal de voz foi utilizada uma freqüência de amostragem de 11025 Hz. A Tabela 2-1 mostra os valores de freqüência calculados para cada sub-banda nos seis níveis de decomposição utilizados.

TABELA 2-1 – Freqüência central de cada sub-banda na decomposição do sinal.

Nível de de-composição	Freqüência cen-tral em Hz
D1	3800
D2	1900
D3	950
D4	475
D5	237
D6	118

3. A base de dados

O vocabulário utilizado neste trabalho consiste em um conjunto de 21 palavras que sugerem comandos para aplicações em aparelhos eletrodomésticos. A base de dados contém um total de 8148 locuções sendo 2730 de vozes femininas e 3318 de vozes masculinas que foram utilizadas para treinamento e teste do sistema no modo independente de locutor e mais 2100 de voz masculina que foi pronunciada por um único locutor, do sexo masculino, para treinar e testar o sistema no modo dependente de locutor. O grupo de locutores apresentava faixa etária entre 16 e 50 anos.

As sessões de gravações foram realizadas em ambiente, que embora não tivesse a preparação acústica de um estúdio, apresentava baixo ruído de fundo.

Cada locutor gravou cinco repetições de cada comando pronunciando-o de forma natural. Nenhuma locução gravada foi eliminada da base de dados, mesmo apresentando eventuais problemas de dicção.

Para este trabalho foi utilizado um microfone do tipo condensador, omni-direcional, resposta em freqüência entre 50 Hz e 16KHz da marca GOLDSHIP. É um microfone de uso popular, encontrado na maioria das lojas de material de informática, para aplicações multimídia.

4. Implementação do reconhecedor de voz

O sistema reconhecedor de voz, proposto em [6], consiste basicamente de um Classificador Neural do tipo MLP cuja arquitetura está mostrada na Figura 4-1. O algoritmo utilizado para treinamento da rede foi o *Resilient Propagation*.

O processo utilizado para construção dos padrões de entrada para o classificador está esquematizado na Figura 4-3. O pré-processamento do sinal de voz consiste em eliminar o nível DC e normalizar o sinal entre -1 e 1. Considerando S o sinal de voz capturado, a remoção do nível DC é realizada conforme a equação 4-1. e a normalização é feita conforme equação 4-2.

$$S = S - mean(S) \qquad (4\text{ -}1)$$

$$S = S / \max(abs(S)) \qquad (4\text{-}2)$$

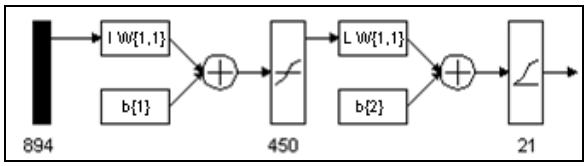

FIGURA 4-1 – Arquitetura do Classificador Neural.

A extração de característica é realizada em três etapas: (1) a decomposição do sinal através da Análise Wavelet, (2) determinação dos pontos extremos (3) o janelamento dentro de cada sub-banda e, finalmente, (4) o cálculo da magnitude média do sinal em cada janela de cada sub-banda, cujos resultados serão os elementos que representarão a locução.

A decomposição do sinal resulta em seis sub-bandas (D1, D2,...,D6), obtidas pela reconstrução do sinal a partir dos coeficientes de detalhe (CD1, CD2,..., CD6). As freqüências centrais de cada sub-banda, mostradas na tabela 2-1, obedecem a uma escala logarítmica, favorecendo a aproximação com a escala de sensibilidade auditiva proposta por Stevens e Volkman [3].

O procedimento para se obter os coeficientes de wavelet pode ser executado em cinco passos [Misiti at al 2002] os quais são descritos a seguir e visualizado na Figura 4-2.

1. Emparelhar a wavelet à uma seção localizada no inicio do sinal a ser analisado.

2. Calcular o valor de C, o qual representa a correlação entre a wavelet e a seção do sinal sob análise. Quanto maior o valor de C maior é a similaridade entre a seção do sinal e a wavelet.

3. Deslocar a wavelet para a direita e repetir os passos 1 e 2 até que todo o sinal seja percorrido.

4.. Alterar a escala ("alongar a wavelet") e repetir os passos 1, 2 e 3. Este procedimento permite correlacionar a wavelet a componentes de freqüência mais baixa.

5. Repetir os passos de 1 até 4 para todas as escalas.

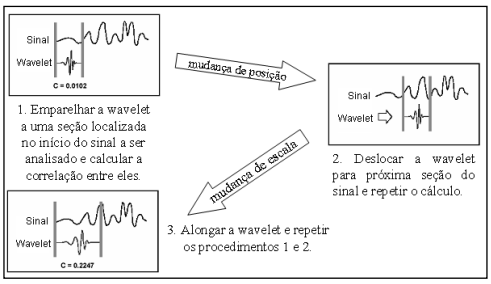

FIGURA 4-2 Procedimento para obtenção dos coeficientes.

Para determinar os pontos extremos foi utilizado o algoritmo proposto por Rocha e Thomé [4][6]. Em cada sub-banda calcula-se a magnitude média do sinal em janelas contendo 100 amostras (aproximadamente 10ms de sinal). Assumindo que os 100ms iniciais e os 100ms finais da locução contém apenas amostras de ruído de fundo (sem sinal de voz), estas regiões são utilizadas para extrair dados estatísticos do ruído atual. Calcula-se a média e o desvio padrão da magnitude do ruído em cada sub-banda e, com base nestas medidas, se estabelece um intervalo de confiança para o ruído. A procura do ponto inicial é realizada varrendo-se cada sub-banda a partir do início da locução. Se a magnitude média da referida sub-banda ultrapassar o limiar (3db) por um período de pelo menos 40 ms (4 janelas consecutivas), tem-se identificado um candidato potencial a ponto inicial. O resultado desta busca produz um vetor contendo seis candidatos. O ponto inicial escolhido será o que estiver mais próximo do início da locução. A procura do ponto final é realizada de forma semelhante, porém iniciando a busca a partir do final da locução em direção ao início e o ponto escolhido será o mais próximo do término do sinal.

A Figura 4-4 mostra a evolução da magnitude média em cada sub-banda. A locução traz a voz de um menino de sete anos pronunciando a palavra "cafofo" em um ambiente com ruído de fundo produzido por um secador de cabelos. Pode-se observar a detecção do ponto inicial realizada pela sub-banda D3 e o ponto final detectado pela sub-banda D4. Neste exemplo pode-se notar a adaptação do algoritmo ao ruído de fundo no sentido de que haverá sempre uma banda de freqüência que sofrerá maior variação na presença do sinal de voz, abrangendo desde as baixas freqüências das vogais até as altas freqüências dos fricativos. A aplicação deste algoritmo resulta na eliminação das regiões de silêncio ou ruído de fundo que precedem e sucedem o sinal de voz recuperando apenas a parte de cada sub-banda (D1', D2',...D6'), compreendida entre o início e o fim da locução.

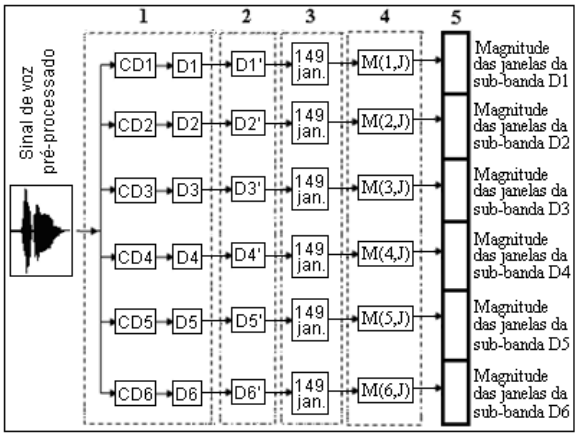

FIGURA 4-3 Etapas da extração de características para a construção do padrão que representará a locução: 1- decomposição do sinal; 2- determinação dos pontos extremos; 3- janelamento; 4- cálculo da magnitude de cada janela; 5- construção do padrão.

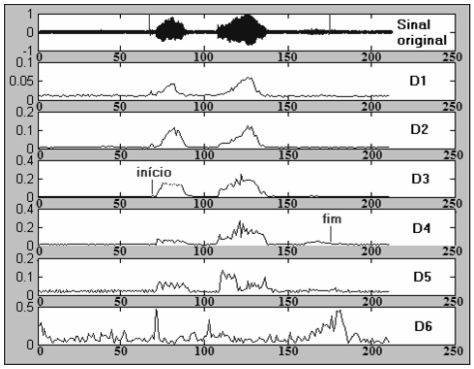

FIGURA 4-4 – Detecção dos pontos extremos pela evolução da magnitude nas seis sub-bandas de detalhe.

A Figura 4-5 traz exemplos dos testes realizados [6] comparando o algoritmo proposto com o algoritmo tradicional (baseado na energia e taxa de cruzamentos por zero). A parte superior das figuras exibe o gráfico gerado pelo algoritmo tradicional enquanto o inferior mostra o comportamento do algoritmo proposto. As barras verticais marcam o início e o fim da locução, detectados pelos respectivos algoritmos.

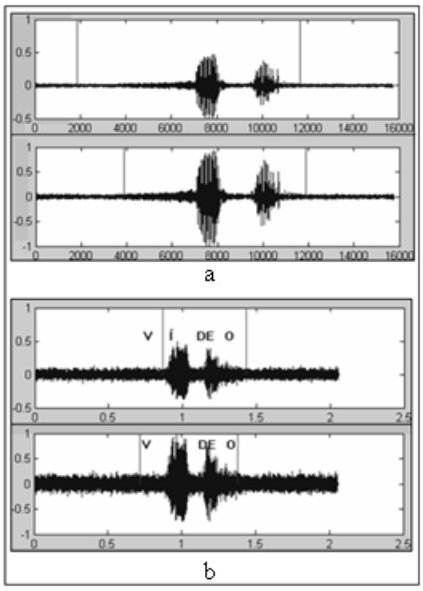

FIGURA 4-5 Detecção dos pontos extremos das locuções das palavras "siga" (a) e da palavra "vídeo" (b) em ambientes com baixa relação sinal/ruído.

A vantagem do algoritmo tradicional é o pequeno esforço computacional exigido tornando-se mais rápido que o algoritmo proposto. Porém em ambientes com baixa relação sinal/ruído o algoritmo proposto mostrou superioridade.

Para atender as exigências do Classificador Neural utilizado, onde os padrões de entrada devem possuir o mesmo comprimento, uma segunda etapa de janelamento foi realizada. Desta vez utilizou-se janelamento dinâmico, com comprimento variando entre 10ms e 40ms, respeitando as características de estacionaridade da voz [3].

Sobre as janelas resultantes é executado o cálculo da magnitude, originando um vetor de 149 elementos para cada sub-banda.

A concatenação destes vetores dá origem ao padrão que representa a locução.

5. Treinamentos e testes do classificador

Além dos testes realizados com objetivo de determinar o desempenho do sistema nas categorias dependente e independente de locutor, foram realizados testes com objetivo de determinar a relevância de cada sub-banda para este fim.

5.1. Desempenho independente de locutor

Neste caso, as locuções utilizadas para treinamento e testes foram fornecidas por locutores distintos, nenhum locutor que contribuiu para o conjunto de treinamento do sistema participou do teste. Foram utilizadas 1281 locuções masculinas e 1050 femininas resultando em 2331 padrões.

A rede convergiu rapidamente para um erro = 10^{-4} tendo decorrido menos de 150 épocas. Conforme pode-se observar em um dos gráficos traçados durante o treinamento, mostrado na Figura 5-1.

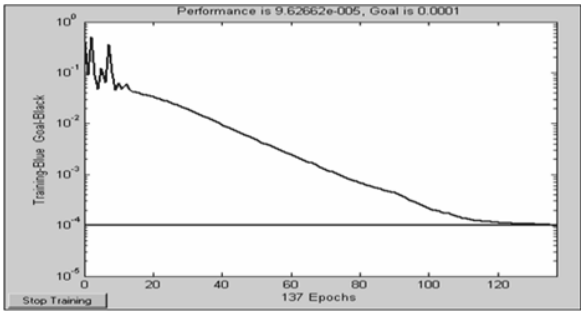

FIGURA 5-1- Comportamento do erro médio quadrático obtido durante um treinamento do sistema para categoria independente de locutor.

O teste foi realizado com um conjunto de padrões com a mesma dimensão do conjunto de treinamento. O resultado obtido foi um desempenho de 93,22% de acertos, conforme detalhado na Figura 5-2.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21
P1- Video	91	0	0	0	0	0	0	0	0	0	6	0	0	0	0	2	5	1	1	0	0
P2- Pare	0	111	0	1	1	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
P3- Grave	0	0	104	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
P4-Pausa	0	0	0	109	5	3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
P5- Avance	0	0	0	1	99	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	1
P6- Volte	1	0	1	0	1	103	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P7- Inicie	0	0	0	0	0	0	106	2	0	0	0	0	0	0	1	2	0	0	0	0	0
P8- TV	0	0	0	0	0	0	1	104	0	0	0	0	0	0	0	4	0	0	0	6	0
P9- Globo	2	0	5	0	0	1	0	0	111	0	2	0	0	0	0	0	2	0	1	0	0
P10- Manchete	0	0	0	0	0	0	0	0	0	107	0	0	0	0	0	0	1	0	0	0	2
P11- Fita	2	0	0	0	0	0	1	1	0	1	94	0	0	0	0	0	0	0	10	0	0
P12- Acenda	0	0	0	0	1	0	0	0	0	0	0	107	2	0	0	0	0	0	0	0	1
P13- Apague	0	0	0	0	4	1	0	0	0	0	0	1	107	0	0	0	0	0	0	0	0
P14- Ventilador	2	0	1	0	0	0	0	0	0	0	3	0	0	103	0	0	0	0	0	1	1
P15- Exaustor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	110	0	0	0	0	0	1
P16- Desligue	2	0	0	0	0	0	2	0	0	2	0	0	0	0	0	98	1	0	0	2	0
P17- Ligue	6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	102	1	0	0	1
P18- Liga	1	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	108	1	2	0
P19- Siga	3	0	0	0	0	0	1	0	0	1	3	0	0	3	0	0	0	0	97	1	0
P20- Desliga	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	3	0	0	0	98	0
P21- Ejete	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	104
% de acertos	82%	100%	94%	98%	89%	93%	95%	94%	100%	96%	85%	96%	96%	93%	99%	88%	92%	97%	87%	88%	94%
Média geral	93,22%																				

FIGURA 5-2 - Matriz de confusão para avaliação do sistema proposto na categoria "independente de locutor".

5.2. Desempenho dependente de locutor

Neste caso utilizou-se um total de 2100 padrões extraídos de locuções pronunciadas por um mesmo locutor.

Usando esses 2100 padrões, construiu-se um conjunto de treinamento com 1470 padrões e os 630 restantes, utilizou-se para teste.

Nesta categoria, o desempenho do sistema alcançou 99,37% de acertos.

5.3. Relevância das sub-bandas

A finalidade deste teste foi investigar a contribuição das informações geradas por cada sub-banda na tarefa de reconhecimento de voz para o sistema proposto. Os testes foram realizados mantendo a configuração da RNA. Em cada teste realizado foi utilizado um sub-conjunto de descritores formado por parte do total de descritores extraído da locução. Como exemplo, no teste que exclui as informações da sub-banda D6 foram excluídos os 149 elementos correspondentes a esta sub-banda. A Figura 5-3 exibe um exemplo do esquema utilizado para a construção dos padrões.

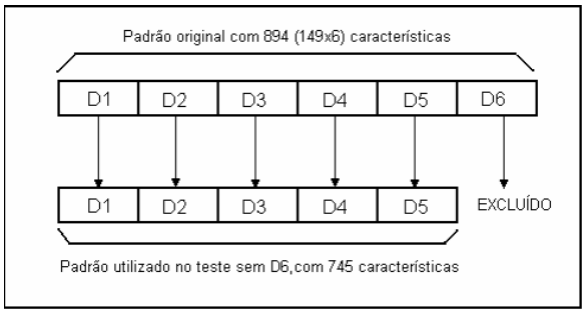

FIGURA 5-3 - Construção dos padrões para treinamento e teste do sistema para investigar a relevância de cada sub-banda.

A Figura 5-4 traz uma visualização gráfica dos resultados obtidos com a remoção de uma sub-banda de cada vez. e a Tabela 5-1 traz um relatório do desempenho obtido em cada teste.

Tanto no gráfico quanto na tabela observa-se uma grande importância atribuída a sub-banda D3. Em contra-partida, pode-se inferir que a sub-banda D6 contribui de forma negativa, dificultando o aprendizado e diminuindo o desempenho do sistema.

Retornando à Tabela 2-1, pode-se especular que a freqüência da sub-banda D6 está dentro da faixa de valores característicos da freqüência fundamental para locutor masculino, sendo assim, além de não contribuir para caracterizar o que está sendo dito [3], os seus valores atuam como ruído para a rede neural dificultando o treinamento, uma vez que os descritores gerados pela maioria das locuções femininas não apresentam esta componente. Por outro lado a sub-banda D3 está localizada na região dos primeiros formantes, contribuindo de forma significativa para a discriminação dos sons (principalmente vocálicos).

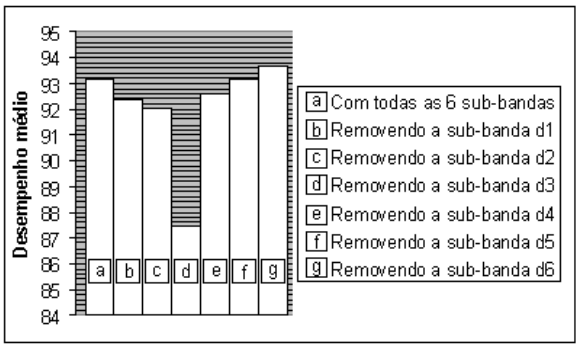

FIGURA 5-4 Efeito da remoção de uma sub-banda no desempenho do sistema

TABELA 5-1 - Desempenho do sistema proposto para cada modalidade de teste na avaliação da relevância de cada sub-banda.

Teste	Desempenho % de acertos
Removendo a sub-banda D1	92,36
Removendo a sub-banda D2	92,06
Removendo a sub-banda D3	**87,47**
Removendo a sub-banda D4	92,62
Removendo a sub-banda D5	93,18
Removendo a sub-banda D6	93,69
Somente com a sub-banda D1	59,25
Somente com a sub-banda D2	64,52
Somente com a sub-banda D3	**76,58**
Somente com a sub-banda D4	71,82
Somente com a sub-banda D5	65,68
Somente com a sub-banda D6	38,05
Com as sub-bandas D1, D2 e D3	90,56
Com as sub-bandas D1, D2, D3 e D4	93,00

6. Conclusão

A Transformada Wavelet em Multi-Resolução mostrou ser uma ferramenta capaz de produzir informações que caracterizam o sinal de voz para fins de reconhecimento. O desempenho de 93,22% de acertos no modo independente de locutor pode ser considerado um bom resultado diante da seguinte realidade: as locuções foram pronunciadas por pessoas de diferentes regiões do país e neste caso os diferentes sotaques acabam por influenciar de forma negativa para reconhecimento, outro caso é o fato de que uma grande parte dos locutores tinha dificuldades em pronunciar as palavras de forma clara, o que é de grande importância para sistemas de reconhecimento que utiliza a palavra como unidade fonética [3].

O algoritmo proposto para detectar os pontos extremos mostrou-se eficiente e robusto ao ruído durante os testes realizados. Esta proposta exigiu maior carga de processamento que o sistema baseado na energia e taxa de cruzamentos por zero, porém esta desvantagem é reduzida quando a etapa de reconhecimento é efetuada com base na Transformada Wavelet, pois uma grande parte dos resultados deste processamento é reaproveitada.

Atualmente, muitos pesquisadores [5] estão desenvolvendo algoritmos para melhorar o sinal de voz utilizando filtros baseados na análise wavelet podendo contribuir significativamente para implementação de sistemas de reconhecimento de voz em ambientes com baixa relação sinal / ruído.

7. Referências Bibliográficas

[1] Graps, A.L.; "An Introduction to Wavelets", IEEE Computational Sciences and Engineering, Volume 2, número 2, verão de 1995, pp 50-61.

[2] Misiti, M., Misiti, Y., Oppenheim, G. & Poggi, J. M., "Wavele Toolbox for use with Matlab", Guia do usuário, Volume 2, 2002.

[3] Diniz, Suelaine S., *Uso de Técnicas Neurais para o Reconhecimento de Comandos a Voz*. Dissertação de Mestrado, Instituto Militar de Engenharia, Rio de Janeiro. 1997.

[4] Rocha, O. P., Thomé, A. C. G., "Utilizando Análise Wavelet Multi-resolução na detecção dos pontos extremos de uma locução em ambiente com baixa relação sinal/ruído". *Revista Militar de Ciência e Tecnologia*, volume XI – 1º Quadrimestre de 2004.

[5] Lu, Ching-ta, Wang, Hsiao-Chuan, "Enhancemente of single channel speech based on masking property and wavelet transform". *Speech Communication* 41 (2003) 409-427. 2003.

[6] Rocha, Oséas P., *Utilização da Transformada Wavelet Multi-Resolução para o Reconhecimento de Comandos a Voz*. Dissertação de Mestrado, Universidade Candido Mendes Campos, Campos dos Goytacazes, 2004.