Detecção dos extremos da voz em presença de ruído através de uma análise no domínio wavelet

Denilson C. Silva e Fernando G. V. Resende Jr.

***Resumo*— A detecção dos extremos da voz ainda é um grande problema, principalmente em situações de reconhecimento em ambiente impregnado por ruído. Este artigo apresenta um método de detecção de extremos baseado na concentração da energia dos coeficientes wavelets de baixa ordem. Com o método proposto a redução média da taxa de erro na detecção de início foi de até $12{,}59\%$, no caso de ruído no interior de carro, e na detecção de fim foi de até $24{,}14\%$, para o caso de falatório, com *SNR* de 0 a 20dB, em relação ao método robusto baseado na energia e na distância euclidiana cepstral.**

***Palavras-Chave*— Detecção de extremos, Ruído, Wavelets, Processamento robusto.**

***Abstract*— The endpoint detection of speech is still a big problem in situations of speech recognition in noisy environments. This article presents a method for endpoint detection based on energy concentration of the low order wavelet coefficients. With the proposed method the average reduction of the error rate of the begin detection was up to $12{,}59\%$, in corrupted signal by interior car noise, and of the end was up to $24{,}14\%$, in corrupted signal by babble noise, with *SNR* from 0 to 20dB, compared to robust endpoint detection based on energy and cepstral euclidean distance.**

***Keywords*— Endpoint detection, Noise, Wavelets, Robust processing.**

I. INTRODUÇÃO

Em diversas aplicações de processamento de sinais da fala, a determinação dos extremos das locuções é necessária. Os métodos tradicionais de detecção de extremos baseados na energia e na taxa de cruzamentos por zero funcionam muito bem com a fala limpa [1]. Quando temos locuções com fricativas, por exemplo, a detecção dos extremos fica bastante comprometida se o processo de delimitação ocorrer em ambiente ruidoso.

Vários trabalhos têm buscado solucionar a questão da detecção de extremos em condições de ambientes impregnados por ruído [2], [3]. Em [2] temos um algoritmo para detecção dos extremos de palavras isoladas, baseado na teoria de fractais, com excelentes resultados. Em [3], temos um método de detecção de extremos, baseado na energia e na distância euclidiana cepstral, com bons resultados no reconhecimento, mas estes resultados são obtidos comparativamente a um detector de extremos não-robusto, que normalmente apresenta baixa eficiência na maioria dos ambientes impregnados por ruído.

Neste artigo é proposto um método para detecção dos extremos em locuções para reconhecimento de voz baseado na concentração da energia dos coeficientes wavelets de baixa ordem. A proposta visa explorar as propriedades da transformada wavelet sobre o sinal de voz, gerando um conjunto de coeficientes com a maior parte da sua energia concentrada em um número pequeno de dimensões wavelets, com baixa ordem e maior amplitude quando comparados com os coeficientes do ruído.

Denilson C. Silva, M.Sc. e Fernando G. V. Resende Jr., Ph.D., Programa de Engenharia Elétrica, Universidade Federal do Rio de Janeiro, COPPE, Rio de Janeiro, Brasil, E-mails: denilson@lps.ufrj.br, gil@lps.ufrj.br.

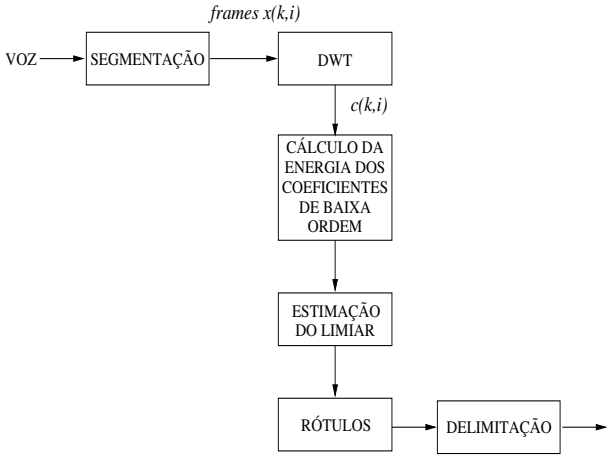

Fig. 1. Método de detecção proposto baseado na concentração da energia dos coeficientes wavelets de baixa ordem.

O método proposto, como representado na Figura 1, resultou em uma redução média da taxa de erro em até $12{,}59\%$ na detecção de início, no caso de ruído no interior de carro, e em até $24{,}14\%$ na detecção de fim, no caso de falatório, comparativamente ao método robusto descrito em [3]. A *SNR* (Relação Sinal-Ruído - *Signal-to-Noise Ratio*) do sinal em análise varia de 0 e 20dB.

Este artigo está organizado da seguinte forma. Na Seção II, apresentamos o método proposto de detecção de extremos das locuções e as considerações para a análise no domínio wavelet. Na Seção III descrevemos a base de dados utilizada, tanto de voz como de ruído. Na Seção IV mostramos os resultados dos testes realizados e as comparações com o método robusto de [3]. Finalmente, na Seção V apresentamos as conclusões e os trabalhos futuros.

II. MÉTODO PROPOSTO DE DETECÇÃO DOS EXTREMOS

A. Análise no domínio wavelet

Em 1995, D. L. Donoho [4] introduziu um método baseado em wavelets para promover "denoising" em sinais de voz contaminados por ruído aditivo gaussiano branco. Donoho procurou explorar as propriedades da transformada wavelet, que baseada na geração de um conjunto de filtros por translações e dilatações de uma wavelet-mãe, possui a vantagem de

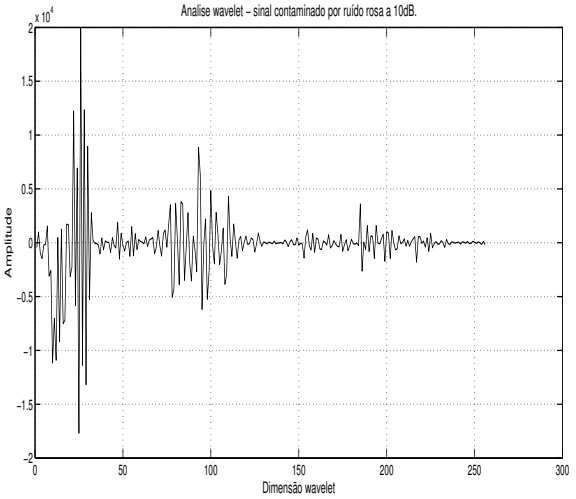

Fig. 2. Concentração da maior parte da energia nos coeficientes de baixa ordem.

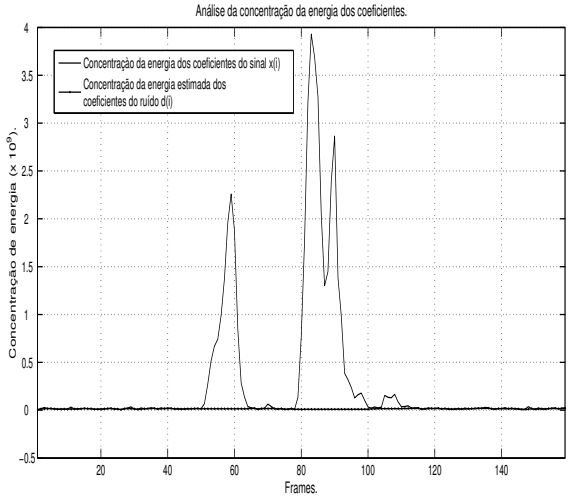

Fig. 3. Concentração da energia quadro-a-quadro.

utilizar janelas temporais de tamanho variável para bandas de freqüência distintas.

Seja o sinal ruidoso $x(i)$ uma composição do sinal de voz limpo $s(i)$ e do ruído $d(i)$:

$$x(i) = s(i) + d(i) \tag{1}$$

O sinal $x(i)$ é dividido em K quadros com tamanho N e superposição de L amostras:

$$x(k,i) = x(k(N-L)+i) \tag{2}$$

onde $0 \leq k \leq K-1$, $0 \leq i \leq N-1$ e N deve ser uma potência de 2. Neste trabalho, os quadros tiveram N fixado em 256 e L em 128 (50%).

Sobre cada quadro $x(k,i)$ é aplicada a DWT (Transformada de Wavelet Discreta - *Discrete Wavelet Transform*), gerando um vetor contendo os coeficientes wavelets $c(k,i)$. Baseado na propriedade da transformada wavelet, os coeficientes de baixa ordem da voz possuem uma amplitude maior comparativamente aos coeficientes do ruído, desde que as condições de contaminação não sejam extremas, impossibilitando o processamento. Com essa observação, basta estabelecer um número reduzido de coeficientes para o cálculo da energia no quadro.

Na Figura 2 temos um quadro de uma locução contaminada por ruído rosa a 10dB, num trecho do sinal onde temos $s(i)+d(i)$. Podemos observar como os coeficientes de ordem mais baixa (aproximadamente até a ordem 40), possuem boa parte da energia concentrada.

Definimos, então, como $\mathcal{E}_x(k)$ a energia dos *J* primeiros coeficientes no *k*-ésimo quadro do sinal ruidoso $x(k,i)$ no domínio wavelet:

$$\mathcal{E}_x(k) = \sum_{j=0}^{J-1} [c(k,j)]^2 \tag{3}$$

onde foi atribuído a J, no cálculo da energia no domínio wavelet, o valor de 32 coeficientes.

B. Estimação do limiar

Um limiar de decisão é estabelecido para avaliarmos os quadros que podem conter fala corrompida ou apenas ruído.

Definimos, também $\mathcal{E}_d(k)$ como a estimativa da energia do ruído no *k*-ésimo quadro, no domínio wavelet, realizada recursivamente através de um filtro IIR (Resposta ao Impulso Infinita - *Infinite Impulse Response*) [5]:

$$\mathcal{E}_d(k) = \alpha(k)\mathcal{E}_d(k-1) + (1-\alpha(k))\mathcal{E}_x(k) \tag{4}$$

O parâmetro $\alpha(k)$ é o elemento que conduz a atualização da estimativa em direção à concentração de energia do sinal ruidoso no quadro k ($\mathcal{E}_x(k)$) ou em direção à estimativa da concentração de energia do ruído no quadro anterior ($\mathcal{E}_d(k-1)$):

$$\alpha(k) = 1 - min(1, \Xi(k)^{-Q}) \tag{5}$$

onde usamos $Q = 5$ e a *SNR* relativa no *k*-ésimo quadro, $\Xi(k)$, é:

$$\Xi(k) = \frac{\mathcal{E}_x(k)}{\frac{1}{M}\sum_{n=0}^{M-1}\mathcal{E}_x(n)} \tag{6}$$

sendo M um número de quadros dentro do intervalo inicial da locução, sem a presença do sinal de voz $s(i)$. Neste trabalho utilizamos $M = 5$.

$$label(k) \quad = \quad \begin{cases} 1, & \text{se } \mathcal{E}_x(k) > \mathcal{E}_d(k) \\ 0, & \text{se } \mathcal{E}_x(k) \leq \mathcal{E}_d(k) \end{cases} \tag{7}$$

Um rótulo, $label(k)$, é estabelecido segundo uma relação da concentração de energia dos coeficientes de baixa ordem, conforme (7). Como a concentração de energia dos coeficientes dos quadros com voz atinge patamares bem maiores do que os quadros que possuem apenas ruído, o filtro IIR traça uma boa estimativa da energia do ruído, e os quadros que estiverem acima desta estimativa têm uma probabilidade maior de possuírem voz. Isto pode ser observado na Figura 3.

C. Delimitação do sinal

Dentro do critério de rotulação dos quadros, é estabelecido que uma seqüência de rótulos permaneça inalterada por um determinado intervalo de tempo. Isso é considerado para evitar que possíveis variações bruscas na amplitude do ruído, possam marcar equivocadamente os extremos.

Sendo assim, fixamos empiricamente que o primeiro trecho de aproximadamente 130ms onde a relação de energia permanecer com rótulo "1" marca o início da locução. O último trecho neste estado determina o fim da locução.

III. Base de dados

As locuções utilizadas neste artigo foram coletadas da base de dados descrita em [6], onde temos 10 palavras isoladas (ANDA, BAIXO, CIMA, DESLIGA, DIREITA, ESQUERDA, FRENTE, MÃO, OLHA e TRÁS), repetidas por diferentes locutores. A taxa de amostragem é de 11025Hz.

A base de dados de ruídos foi utilizada de [7], com 235s de duração em cada tipo de ruído, taxa de amostragem original de 19980Hz sub-amostrada para 11025Hz. Foram selecionados oito tipos de ruídos: falatório (BABBLE), ruído rosa (PINK), ruído no interior de um carro (VOLVO), ruído branco (WHITE), ruído de fábrica (FACTORY), ruído no canal de HF (HFCHANNEL), ruído na sala de operações de um Destroyer (DESTROYEROPS) e cockpit de caça (BUCCANEER).

As locuções ruidosas foram formadas através da adição dos ruídos selecionados às locuções limpas de acordo com a *SNR* estabelecida, que vai de 0 a 20dB.

IV. Resultados obtidos

A detecção dos extremos pode ter a sua performance avaliada de duas formas: uma delas é comparar os resultados obtidos na detecção com valores de referência obtidos de um recorte manual. A outra é passar as locuções recortadas, após a contaminação, em um sistema de reconhecimento de voz, comparando os seus resultados. Optamos pela primeira forma por termos condições de avaliar de forma direta os resultados, fazendo uma comparação pontual entre os métodos.

Foi realizado um teste funcional com sinais de voz contaminados pelos ruídos selecionados com uma *SNR* que vai de 0 a 20dB. O procedimento foi fixar uma *SNR* e o ruído, observando o percentual de erro na detecção ao longo dos vários valores de *SNR*, tanto para detecção de início como para detecção de fim. Para as 121 locuções inseridas no detector, calculamos o erro percentual na detecção de início (ε_i) e o erro percentual na detecção de fim (ε_f) em relação ao recorte manual previamente realizado, da seguinte forma:

$$\varepsilon_i = \frac{|I - e_i|}{F - I} \times 100\% \quad (8)$$

$$\varepsilon_f = \frac{|F - e_f|}{F - I} \times 100\% \quad (9)$$

onde I e F são, respectivamente, os pontos de início e fim do recorte manual, e_i e e_f são, respectivamente, os pontos detectados.

Na Figura 4 apresentamos uma locução referente à palavra "esquerda" na sua versão limpa (em **a**), contaminada por ruído rosa a 10dB (em **b**) e recortada pelo método proposto neste artigo (em **c**). Três regiões do sinal foram definidas e consideramos para análise a versão ruidosa de (**b**):

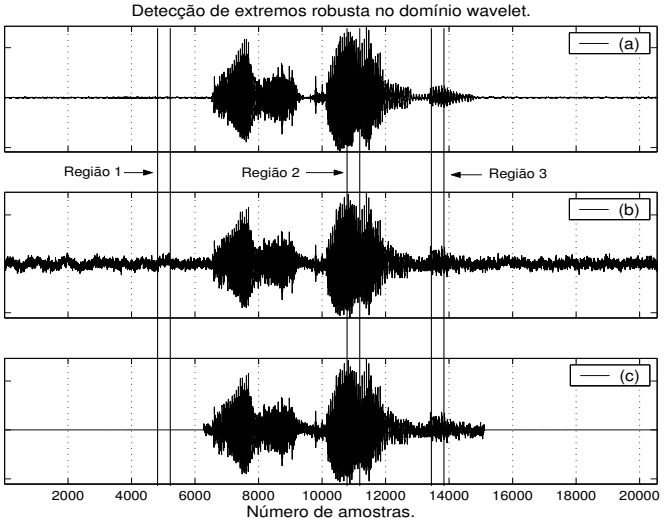

Fig. 4. Detecção dos extremos, com o método proposto, em uma locução referente à palavra "esquerda" contaminada por ruído rosa a 10dB. Em (**a**) temos o sinal limpo, em (**b**), o sinal contaminado por ruído rosa com *SNR* de 10dB e em (**c**) temos o sinal recortado pelo método proposto. As três regiões são definidas nas mesmas posições temporais de todas as versões do sinal.

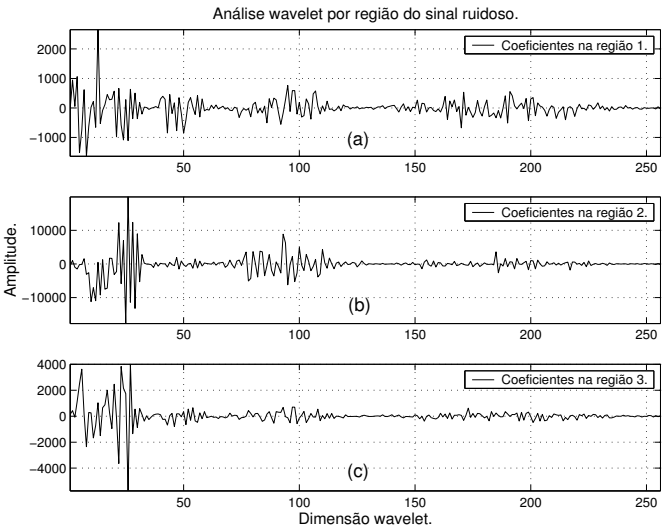

Fig. 5. Variabilidade da amplitude dos coeficientes wavelets de baixa ordem na região 1 (superior), região 2 (centro) e região 3 (inferior), definidas na Figura 4.

- Região 1 - onde somente temos ruído ($x(i) = d(i)$);
- Região 2 - onde temos voz, a alta energia, misturada com ruído ($x(i) = s(i) + d(i)$);
- Região 3 - onde temos voz, a baixa energia, misturada com ruído ($x(i) = s(i) + d(i)$).

Na Figura 5 analisamos a concentração de energia dos coeficientes wavelets das regiões mostradas no sinal ruidoso da Figura 4 (em **b**). Nela temos os coeficientes da região 1 com amplitudes máximas entre -1500 e 2500, da região 2, com os máximos entre -18000 e 20000 e, por fim, a região 3

TABELA I

TABELA COM OS PERCENTUAIS MÉDIOS DE REDUÇÃO DA TAXA DE ERRO OBTIDOS COM O DETECTOR ROBUSTO EM LOCUÇÕES CONTAMINADAS POR RUÍDO COM *SNR* VARIANDO ENTRE 0 E 20dB.

Redução média do erro de detecção		
	Detecção de início	Detecção de fim
BABBLE	11,68%	24,14%
PINK	1,44%	−0,83%
VOLVO	12,59%	10,57%
WHITE	2,49%	5,20%
FACTORY	4,60%	18,27%
HFCHANNEL	2,73%	4,92%
DESTROYEROPS	−0,60%	0,64%
BUCCANEER	5,40%	3,55%

apresenta os máximos entre -6000 e 4000. Podemos observar que por menor que seja a concentração de energia no domínio do tempo (região 3), no domínio wavelet essa concentração consegue se destacar mesmo em presença de ruído, o que favorece a boa performance do detector. Na Figura 3, onde temos uma análise da concentração da energia em cada quadro, referente à Figura 4 (em **b**), tanto do sinal $x(i)$, como do ruído $d(i)$, estimado por $\mathcal{E}_d(k)$, observamos o destaque da concentração de energia na região 3, importante para o processo de delimitação em presença de sinais de voz com baixa energia.

Para os oito tipos de ruídos, a redução média da taxa de erro na detecção, comparativamente ao método robusto de [3], é mostrada na Tabela I. Os valores calculados são uma média das taxas de erro de detecção de início e fim, para cada tipo de ruído, com os cinco valores de *SNR* simulados neste trabalho. Podemos observar que houve uma redução considerável em praticamente todos os ambientes ruidosos simulados.

Os resultados individuais, que foram alcançados nos testes, são comparados com o método robusto de [3] e mostrados no gráfico da Figura 6. É notado que a melhor performance do detector ocorre nos casos de contaminação por falatório, onde a delimitação global (início e fim) apresentou melhores índices de redução do erro de detecção. Para o caso de ruído rosa e ruído na sala de operações do Destroyer praticamente não houve redução do erro de detecção, apresentando, portanto, performance semelhante ao detector apresentado em [3].

V. CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho foi proposto um método para detecção robusta de extremos baseado na concentração da energia dos coeficientes wavelets de baixa ordem. Com este método buscamos explorar a propriedade da transformada wavelet que viabiliza a concentração da energia do sinal de voz em um número pequeno de coeficientes, com amplitude bem maior que os coeficientes do ruído. Os resultados de um teste com 121 locuções contaminadas pelos oito tipos de ruídos com *SNR* de 0 a 20dB foi apresentado. A redução média da taxa de erro no recorte em relação ao método robusto de [3], foi de até 12,59% na detecção de início, no caso de ruído no interior de carro, e de até 24,14% na detecção de fim, no caso de falatório. Nas situações de contaminação por ruído branco, ruído de fábrica, ruído no canal de HF e cockpit de caça, os resultados

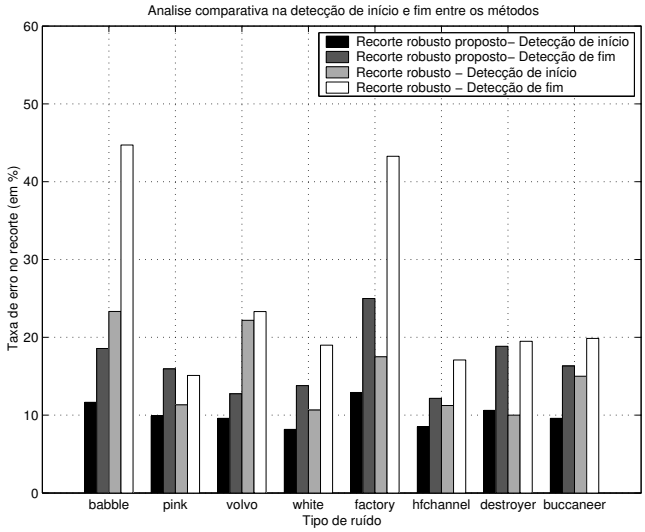

Fig. 6. Resultado comparativo na detecção pelos métodos proposto e robusto de [3]

ainda foram melhores, porém mais modestos. Para o caso de ruído rosa e ruído na sala de operações do Destroyer, os resultados praticamente permaneceram inalterados em relação ao método robusto de [3].

Como trabalho futuro, pretendemos avaliar a eficiência do detector de extremos em um SRRV (Sistema de Reconhecimento Robusto de Voz), com diversas técnicas de redução de ruído inseridas, inclusive em tarefas de processamento on-line.

REFERÊNCIAS

[1] Jia-Lin Shen, Jeih-Weih Hung and Lin-Shan Lee, *Robust entropy-based endpoint detection for speech recognition in noisy environments*, In International Conference on Spoken Language Processing, November, 1998.

[2] L. Gu, J. Gao and J. G. Harris, *Endpoint detection in noisy environment using a poincaré recurrence metric*, In IEEE International Conference on Acoustics, Speech, and Signal Processing, April, 2003.

[3] S. E. Bou-Ghazale and K. Assaleh, *A robust endpoint detection of speech for noisy environment with application to automatic speech recognition*, In IEEE International Conference on Acoustics, Speech, and Signal Processing, v.4, p.3808-3811, May, 2002.

[4] D. L. Donoho, *Denoising by soft thresholding*, IEEE Trans. in Information Theory, v.41, no.3, p.613-627, May, 1995.

[5] L. Lin, W. H. Holmes and E. Ambikairajah, *Subband noise estimation for enhancement using a perceptual Wiener filter*, In IEEE International Conference on Acoustics, Speech, and Signal Processing, v.I, p.80-83, April, 2003.

[6] R. Teruszkin et al., *Biblioteca orientada a objeto para reconhecimento de voz e aplicação em controle de robô*, In Congresso Brasileiro de Automática, Setembro, 2002.

[7] Rice University, *Signal Processing Information Base (SPIB)*. http://spib.rice.edu/spib/select_noise.html.