Um Método Simples e Eficiente para Detecção de Atividade de Voz Usando a Transformada Wavelet

Marco A. Q. Duarte
Curso de Matemática, UUC, UEMS
79540-000, Cassilândia, MS
E-mail: marco@uems.br

Jozué Vieira Filho
Departamento de Engenharia Elétrica – FEIS, UNESP
15385-000, Ilha Solteira, SP
E-mail: jozue@dee.feis.unesp.br

Francisco V. Alvarado
Departamento de Matemática – FEIS, UNESP
15385-000, Ilha Solteira, SP
E-mail: villa@mat.feis.unesp.br

Resumo: *Uma etapa importante no processamento de sinais de voz é a identificação dos trechos do sinal para os quais há atividade de voz ou não, pois a maioria dos métodos de processamento de sinais de voz faz estimativas do sinal nos trechos de silêncio e depois as usa para o processamento de todo o sinal. Por isso, neste trabalho, um detector de atividade de voz baseado na transformada wavelet discreta é apresentado. A análise do sinal é feita através de um janelamento adequado e a identificação das janelas de silêncio e de voz é realizada medindo-se o desvio padrão. Na metodologia proposta, o desvio padrão em cada janela é comparado com a média dos desvios das n primeiras janelas, geralmente os primeiros 200ms que são de silêncio. Se este desvio for menor que a média então a janela será considerada de silêncio, caso contrário ela será considerada uma janela de voz.*

I. INTRODUÇÃO

O processamento de sinais de voz é uma das aplicações mais importantes da área de processamento de sinais na atualidade, pois com o avanço das telecomunicações, cada vez mais surge a necessidade de métodos que trabalhem compactação, codificação, reconhecimento de padrões e redução de ruído em sinais de voz [3]. Uma etapa importante para o processamento de um sinal de voz é a detecção de atividade de voz (VAD) ([4], [5], [6], [9], [10] e [11]), isto é, identificar os trechos do sinal para os quais há fala (voz) ou silêncio (pausa), pois é nas regiões de silêncio que o ruído presente no sinal é mais bem estimado [3]. Os métodos de VAD, geralmente, são métodos estatísticos, ou seja, fazem a detecção de pausa baseados nas propriedades estatísticas do sinal, considerando a variação do ruído, quando há ruído, nos trechos de silêncio e atualizando os parâmetros a cada novo trecho de silêncio [6]. Nestes métodos, o sinal é janelado e a cada janela é dado um valor que a classifica como janela de silêncio ou de voz, geralmente 0 para silêncio e 1 para voz, criando assim um vetor binário. Em [4] é apresentado um método de detecção de atividade de voz baseado na aplicação de um teste de hipóteses bayesiana em amostras descorrelacionadas do sinal. O método proposto em [5] faz detecção de pausa em sinais de banda larga ou sinais filtrados por filtros passa alta ou passa baixa através do cálculo de mínimos adaptativos em partes compactadas do sinal. Em [6], o método é baseado na divergência espectral a longo prazo entre o espectro do ruído e a voz presente no sinal. O método proposto em [9] baseia-se em estimativas a priori do ruído. O algoritmo é dividido em duas partes: a primeira faz uma estimativa a priori do ruído e a segunda define se o trecho analisado é de silêncio ou de voz. O método apresentado em [10] faz detecção de atividade de voz baseado na energia das janelas analisadas. Esse método é proposto para a

análise de sinais com nível de ruído abaixo de 5dB. Em [11] é proposto um método para detecção de atividade de voz em dispositivos usados em comunicações móveis. O método usa características de baixa frequência do sinal para tomar uma decisão lógica baseada numa máquina de estados finitos (sequência finita de operações onde um estado depende do outro). Neste caso, a máquina de estados finitos consiste em quatro estados: (1) estado estacionário do ruído; (2) estado não estacionário de ruído para o sinal de voz; (3) estado estacionário do sinal de voz e (4) estado não estacionário do sinal de voz para o ruído.

Nos últimos anos, a transformada wavelet tem sido explorada de forma intensa no processamento de sinais de voz e algumas das aplicações importantes, dentre várias, são: redução de ruído, codificação de voz e reconhecimento automático de fala. Seguindo essa tendência, neste trabalho o algoritmo proposto é baseado na transformada wavelet discreta (DWT) de Daubechies ([1],[8]) e usa o desvio padrão calculado em cada janela do sinal no domínio wavelet para classificá-la como janela de silêncio ou de voz. O desvio de cada nova janela é comparado com a média dos desvios das janelas de silêncio anteriores. Se este for maior que a média, a janela é classificada como janela de voz; caso contrário, ela é classificada como janela de silêncio. O trabalho está organizado da seguinte forma: na seção II é apresentada uma síntese da transformada wavelet discreta; na seção III, mostra-se como o desvio padrão ou perfil do ruído é estimado; na seção IV, o método proposto é apresentado; na seção V, são apresentados a implementação e os resultados gráficos para três sinais de voz limpos e contaminados por ruído branco ou colorido e, finalmente, na seção VI são apresentadas as conclusões a respeito de tal método.

II. TRANSFORMADA WAVELET DISCRETA

A Transformada Wavelet pode ser vista como uma técnica de janelamento instantânea, o que corresponde a uma filtragem ([1], [7]), muito útil no processamento de sinais não estacionários. A Figura 1 apresenta um exemplo da Transformada Wavelet Discreta, (DWT), com um nível de decomposição. O sinal de entrada, denominado de $c_0[n]$, é passado por dois filtros: o primeiro representa um filtro passa baixa, com resposta ao impulso h[n]; e o segundo representa um filtro passa alta, com resposta ao impulso g[n]. Após a filtragem, os sinais passam por uma dizimação no tempo (no caso, uma sub-amostragem de ordem 2), gerando as saídas $c_1[n]$, que representa as componentes de baixa frequência, e $d_1[n]$, que representa as componentes de alta frequência. Do ponto de vista matemático, $c_1[n]$ contém os chamados coeficientes de aproximação e $d_1[n]$ os coeficientes de detalhes [7].

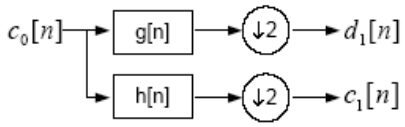


Figura 1. Decomposição em duas faixas

Considerando que o sinal de entrada é real, as equações que expressam as relações entre $c_0[n]$, $c_1[n]$ e $d_1[n]$ são dadas por:

$$c_1[k] = \sum_n \overline{h[n-2k]}\, c_0[n] \quad (1)$$

$$d_1[k] = \sum_n \overline{g[n-2k]}\, c_0[n] \quad (2)$$

A partir da obtenção dos sinais $c_1[n]$ e $d_1[n]$, são feitas as alterações de acordo com a aplicação (compressão, redução de ruído, etc.).

O sinal processado é obtido a partir de uma combinação adequada dos sinais $c_1[n]$ e $d_1[n]$, que é denominada de transformada wavelet inversa [1]. A Figura 2 ilustra o processo de reconstrução. Inicialmente, adiciona-se um zero entre cada par elemento dos sinais $c_1[n]$ e $d_1[n]$. Em seguida,

são efetuadas as convoluções de $c_1[n]$ com h[n] e de $d_1[n]$ com g[n]. Essas duas operações em cascata representam uma sobreamostragem (no caso, de ordem 2).

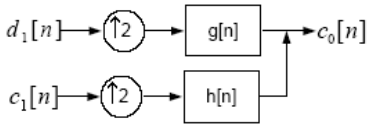


Figura 2. Reconstrução em duas faixas

O sinal reconstruído é a soma dos resultados das convoluções, conforme expresso na equação 3.

$$c_0[n] = \sum_k h[n-2k].c_1[k] + \sum_k g[n-2k].d_1[k] \qquad (3)$$

Para processar um sinal de comprimento N, a decomposição é feita em $L=\log_2 N$ faixas, com L representando o número máximo de faixas de frequências em que o sinal pode ser decomposto, o que equivale a um banco de filtros [7]. A família de wavelets considerada para os testes do algoritmo proposto é a família de wavelets ortonormais de Daubechies e, neste caso, foi usada a wavelet de Daubechies de ordem 10 [1].

III. CÁLCULO DO DESVIO PADRÃO

Com o objetivo de comprimir sinais, em [2] é proposta uma fórmula para o cálculo do perfil de ruído presente num sinal qualquer, que nada mais é que uma aproximação do desvio padrão do sinal. Para um sinal S, de comprimento N**,** no domínio wavelet, esta aproximação é dada pela equação 4.

$$\sigma = mediana(|S|)/0.6745 \qquad (4)$$

Num sinal de voz, o processamento no domínio discreto é feito necessariamente através de um janelamento adequado, o que viabiliza um processamento digital, sendo que todas as janelas têm o mesmo comprimento. Assim, a equação 4 pode ser usada para estimar o perfil do ruído presente em cada janela do sinal; naturalmente, este perfil sofrerá variação de janela para janela [3]. Neste trabalho, como os testes são realizados com sinais limpos e ruidosos, a equação 4 será usada para se estimar o desvio padrão do sinal, no caso de sinal limpo, e o perfil de ruído, no caso de sinal ruidoso.

IV. O MÉTODO PROPOSTO

O algoritmo proposto neste trabalho usa o desvio padrão calculado de acordo com a equação 4. Na Figura 3(a) é apresentado um sinal de voz $v(t)$ limpo, no domínio do tempo, e na Figura 3(b) é apresentada a curva de perfis, ou desvios padrões $\sigma(j)$, das janelas *j* do sinal no domínio wavelet. Nessas figuras é possível notar claramente que o movimento da curva de perfis acompanha o movimento da forma de onda do sinal. Assim, comparando-se a forma de onda do sinal e a curva de perfis, é possível distinguir os trechos de silêncio e os trechos de voz do sinal. Cada ponto da curva de desvios da Figura 3(b) representa o desvio padrão, calculado no domínio wavelet, de uma janela do sinal de voz apresentado na Figura 3(a). Assim, baseado apenas no cálculo do desvio padrão de cada janela no domínio wavelet, é possível criar um detector de silêncio e voz eficiente.

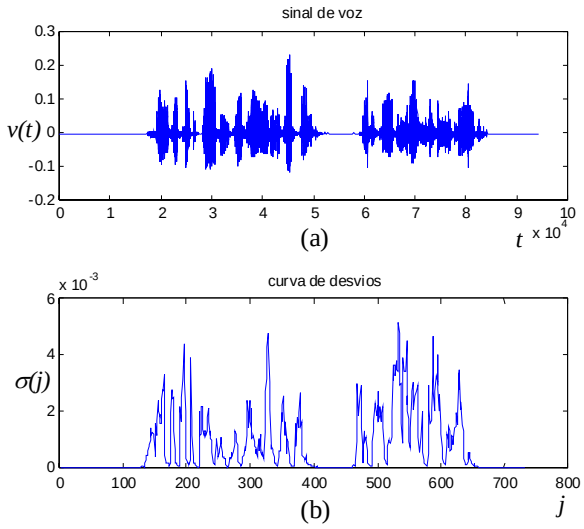


Figura 3. (a) Sinal de Voz – (b) curva de desvios do sinal de voz exposto em (a)

O algoritmo para implementação do detector proposto neste trabalho consiste nos seguintes passos:

1) O sinal de voz é multiplicado por uma janela de 256 pontos (janelamento), que representa um segmento de voz de curta duração (16ms para uma taxa de amostragem de 16 kHz, por exemplo);
2) Aplica-se a DWT (Daubechies 10) ao segmento de sinal janelado e calcula-se o desvio padrão dos coeficientes de detalhes;
3) Repete-se o passo 3 até cobrir um tempo próximo a 200ms, o que irá gerar L janelas analisadas. Nesse tempo assume-se que apenas ruído esteja presente no sinal, que é a realidade das aplicações práticas;
4) Calcula-se a média dos desvios padrões das L janelas analisadas (o número de janelas L depende da taxa de amostragem e da sobreposição entre janelas – para um taxa de amostragem de 16 kHz e uma sobreposição de 50% das amostras entre janelas consecutivas, isto equivale a L = 26);
5) A partir da L-ésima janela considerada, cada vez que um novo desvio padrão é calculado, ele é comparado com a média dos desvios das janelas anteriores. Se este for menor que a média mais um acréscimo, que neste caso é 10% da média dos desvios calculados, esta janela será considerada uma janela de silêncio, caso contrário ela será considerada janela de voz;
6) Se a janela classificada no passo anterior for de silêncio, a média dos desvios é atualizada;
7) Repetem-se os passos 1, 2, 5 e 6 até que seja alcançada a última janela do sinal.

V. IMPLEMENTAÇÃO E RESULTADOS

Nesta seção, são apresentados os resultados da implementação do algoritmo de detecção de atividade de voz proposto neste trabalho. Três trechos de sinais de voz foram analisados: um sinal de voz masculina, em português, amostrado a uma taxa de 16 kHz e com duração de 5,9 segundos (94208 amostras); outro sinal em português, com voz feminina, amostrado a uma taxa de 44,1 kHZ e com duração de 6,2 segundos (273420 amostras); e finalmente um sinal de voz masculina, em inglês, amostrado a uma taxa de 8 kHz e com duração de 2.7 segundos (21801 amostras). Os três sinais foram gravados por falantes nativos e quantizados com 16 bits por amostra. Nos testes, são apresentadas as detecções de silêncio/voz dos três sinais sem qualquer ruído, contaminados por ruído branco e contaminado por ruído colorido. O ruído branco é ruído

artificial, com média zero e variância 1. O ruído colorido considerado aqui é real e se trata de ruído gravado dentro de um automóvel em movimento. Para a implementação computacional foi usado o software MATLAB. O janelamento do sinal foi realizado com uma janela de Hanning de 256 pontos e a sobreposição entre janelas foi de 50% (128 amostras). A sobreposição de 50% evita problemas na aplicação da DWT nas bordas do sinal.

Nas Figuras 4, 6 e 8 são apresentadas as formas de ondas dos três sinais limpos e suas respectivas detecções. Nas Figuras 5, 7 e 9 são apresentados em (a) as formas de ondas dos sinais contaminados por ruído branco com suas respectivas detecções e em (b) os mesmos sinais, porém contaminados por ruído colorido, também com suas respectivas detecções. Nos títulos das Figuras 5, 7 e 9 são informados os níveis de contaminação de cada sinal, com ruído branco e com ruído colorido. Para isto foi usada a relação sinal ruído (SNR), lembrando que quanto menor é a SNR maior é a presença de ruído no sinal [3]. Na Figura 8, o método proposto considera praticamente todo o sinal como tendo apenas atividade de voz (menos as janelas iniciais e finais). A explicação, neste caso, é que a taxa de amostragem de 8 kHz gera poucas janelas (L = 13) para composição da média inicial dos desvios padrões dos coeficientes da DWT dos períodos de silêncio. Isto ocorre também para o sinal com ruído, como observado na Figura 9(b). Na Figura 9(a) também se tem uma detecção atípica. Porém, em todas elas, os trechos de voz foram preservados, ou seja, não foram considerados silêncio. Isto é fundamental para uma aplicação real, pois o ruído pode ser considerado voz, mas o contrário não pode ocorrer [3].

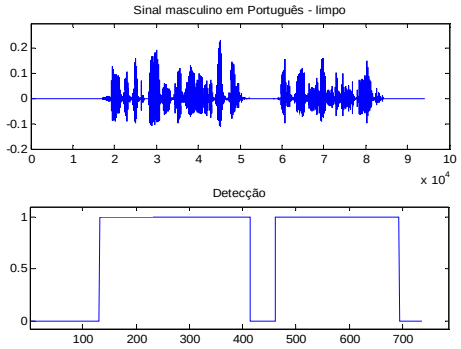


Figura 4. Sinal de voz masculina em Português sem ruído

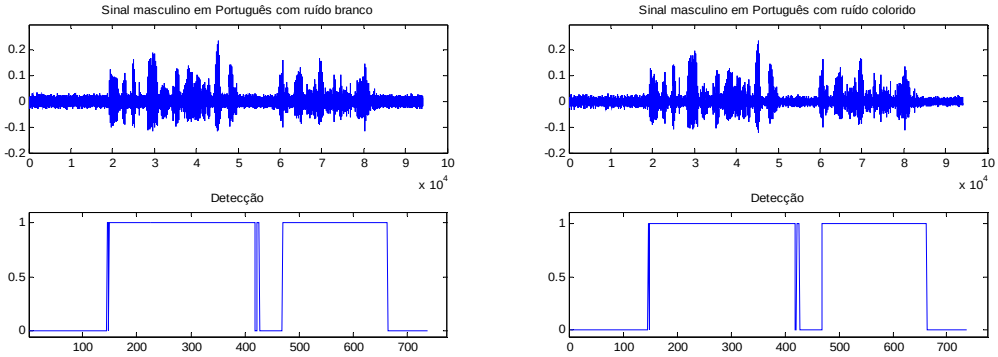


Figura 5. Sinal de voz masculina em Português – (a) ruído branco- SNR 9,643dB e (b) ruído colorido SNR 9.778dB.

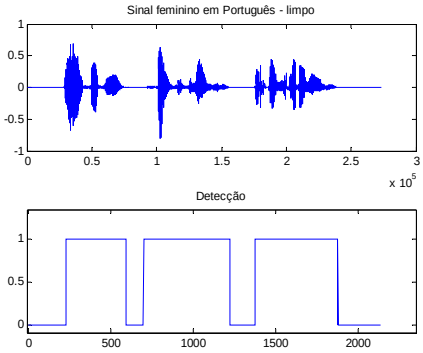
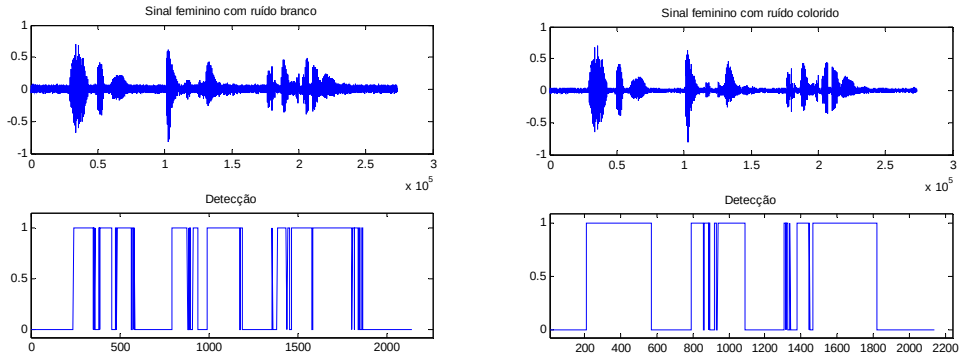


Figura 6. Sinal de voz feminina em Português sem ruído



(a) (b)

Figura 7. Sinal de voz feminina em Português –(a) ruído branco- SNR 8,272dB e (b) ruído colorido - SNR 9,320dB.

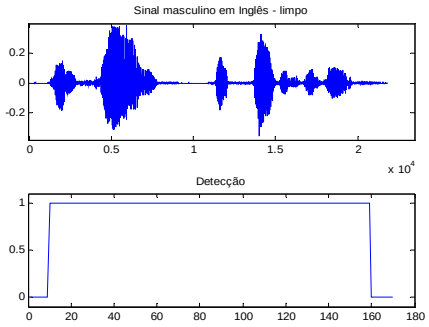
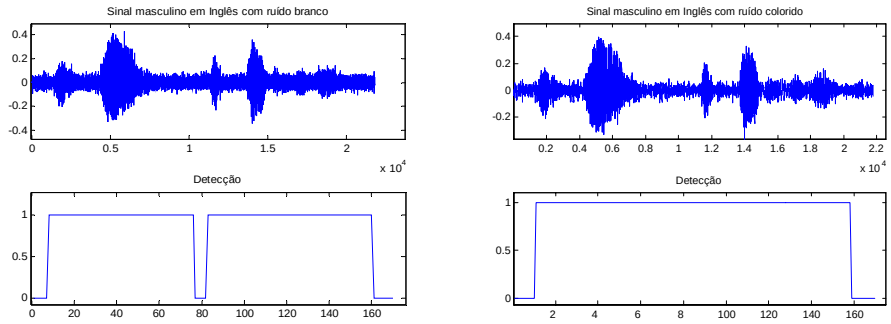


Figura 8. Sinal de voz masculina em Inglês sem ruído



(a) (b)

Figura 9. Detecção do Sinal em Inglês - (a) ruído branco- SNR 9,387dB e (b) ruído colorido – SNR 9,562dB.

VI. CONCLUSÕES

Neste trabalho foi apresentado um detector de atividade de voz baseado na transformada wavelet. O uso da transformada wavelet se justifica pelo fato da mesma evidenciar melhor os detalhes do sinal e pela sua fácil implementação computacional. Para a detecção é calculado o desvio padrão de cada janela do sinal de voz no domínio wavelet e uma média é obtida sempre que o trecho analisado for classificado como de silêncio. Isso cria uma referência para diferenciar-se janela de voz de janela de silêncio. Conforme se pode observar nos resultados visuais apresentados nas Figuras 4 a 9, o detector se mostrou eficiente para sinais sem ruído e para sinais contaminados por ruído branco ou colorido Os testes foram realizados com sinais em português e inglês, com voz masculina e feminina e com diferentes níveis de ruído. Além disso, conforme ilustram as figuras, para nenhum dos sinais testados o detector confundiu voz com silêncio, isto é, os trechos de voz foram sempre classificados como voz. Já o contrário ocorreu, ou seja, o ruído foi detectado como voz, mas isto é facilmente justificável e não causa qualquer prejuízo em aplicações reais.

REFERÊNCIAS

[1] I. Daubechies, “Ten Lectures on Wavelets”, SIAM, Philadelphia,1992.

[2] L. D. Donoho, and I. M. Johnstone, Ideal Spatial Adaptation via Wavelet Shrinkage, Biometrika, Vol. 81, No. 3, pp. 425-455, (1994).

[3] M. A. Q. Duarte, “Redução de Ruído em Sinais de Voz no Domínio Wavelet”, Tese de Doutorado, PPGEE, FEIS-UNESP, 2005.

[4] S. Gazor and W. Shang, “A Soft Voice Activity Detector Based on a Laplacian-Gaussian Model”, IEEE Trans. Speech Audio Processing. vol. 11, pp. 498-505, (2003).

[5] M. Marzinzik, B. Kollmeier, Speech Pause Detection for Noisy Spectrum Estimation by Tracking Power Envelope Dynamics, IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 2, pp. 109-118, (2002).

[6] J. Ramirez, J. C. Segura, C. Benitez, A. De La Torre and A.Rugio, Efficient Voice Activity Detection Algorithms using Lon-term Speech Information, Speech Communication, Vol. 42, pp. 271-287, (2004).

[7] O. Rioul, and M. Vetterli,, Wavelets and Signal Processing, Signal Processing Magazine, vol.8, No.4, pp. 14-38, (1991).

[8] T. K. Sarkar, C. Su, R. Adve, M. S. Palma, L. G. Castilho and R. R. Boix, A Tutorial on Wavelets from an Electrical Engineering Perspective, Part 1: Discrete Wavelet Techniques, IEEE Antennas and Propagation Magazine, Vol. 40, No. 5, (1998).

[9] J. Sohn and W. Sung, “A Voice Activity Detector Employing Soft Decision Based Noise Spectrum Adaptation”, Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’98), vol. 1, pp. 365-368, (1998).

[10] S. G. Tayner and H. Özer, “Voice Activity detection in Nonstationary noise,” IEEE Trans. Speech Audio Processing, vol. 8, pp. 478-482, (2000).

[11] B. Wu, X. Ren, C. Liu, Y. Zhang, A Robust, Real-time, Voice Activity Detection Algorithm for Embedded Mobile Device, International Journal of Speech Technology, Vol. 8, pp. 133-146, (2005).