Comparação de Detectores de Atividade de Voz em Ambiente Ruidoso

Felipe Sander Pereira Clark[1], Mariane Rembold Petraglia[1] e Diego Barreto Haddad[2]

***Resumo*— Neste artigo apresentamos e comparamos três algoritmos de detecção de atividade de voz em sinais ruidosos: o detector linear e adaptativo de energia, o detector linear de energia em sub-bandas e um modelo estatístico de detecção de voz. Esta comparação visa a determinação do algoritmo de melhor desempenho tomando-se como métrica de referência o percentual de falsos positivos e falsos negativos obtido por cada detector ao analisar sinais comprometidos por ruído branco em função da SNR.**

***Palavras-Chave*— Detecção de voz, ruído, baixo SNR.**

***Abstract*— This paper presents and compares three algorithms for the detection of voice activity in noisy signals: the adaptive linear energy based detector, the linear sub-band energy detector and a statistical model based voice activity detector. Our aim is to determine which algorithm performs better in what concerns the false positive and false negative detection percentage, as the SNR is decreased by the introduction of white noise.**

***Keywords*— Voice activity detection, noise, low SNR.**

I. INTRODUÇÃO

Detectores de atividade de voz (do inglês *voice activity detectors - VADs*) são algoritmos capazes de detectar a existência de trechos que contém voz em sinais compostos, alternadamente, por momentos de silêncio e de fala, sejam estes ruidosos ou não. Sua capacidade de otimizar a codificação e a compressão justifica o seu amplo emprego na redução de banda requerida em sistemas VoIP. De maneira simples, se não há informação de interesse (voz) em um sinal, não é necessário transmiti-lo, comprimi-lo ou codificá-lo [1]. Também há aplicação em outros contextos, como em cancelamento de eco e filtragem adaptativa [2] e supressão de ruído [3].

Ao longo das próximas seções apresentaremos três algoritmos e uma comparação de seus desempenhos. Na Seção II apresentamos o detector linear e adaptativo de energia. A Seção III introduz uma generalização deste detector em sub-bandas. Na Seção IV apresentamos um sofisticado modelo estatístico de detecção. Na Seção V, apresentamos testes comparativos de desempenho dos três detectores em contextos ruidosos, usando como métrica objetiva de avaliação o percentual de detecções falso positivas e falso negativas. Finalmente, na Seção VI tecemos nossos comentários finais a respeito desta exposição.

1-2: Universidade Federal do Rio de Janeiro, Dep. de Eletrônica e de Computação - Escola Politécnica e Prog. de Engenharia Elétrica - COPPE, Rio de Janeiro, RJ, CP 68505, Brasil

2: CEFET-RJ Nova Iguaçu, Dep. de Telecomunicações Nova Iguaçu, RJ, CEP 26041-271, Brasil

II. DETECTOR LINEAR E ADAPTATIVO DE ENERGIA

O detector linear e adaptativo de energia (*Adaptive Linear Energy-Based Detector - ALED*) consiste da computação da energia de quadros do sinal segundo

$$E_j = \frac{1}{K} \sum_{i=(j-1)K+1}^{jK} x^2(i), \tag{1}$$

onde E_j é a energia do *j*-ésimo quadro e K é o seu tamanho, seguido da comparação destas energias a limiares.

Para ambientes cuja energia do ruído de fundo é conhecida e invariante, estima-se a presença de voz em um dado quadro quando sua energia for maior que o limiar estabelecido pelo ruído. Todavia, não raro a energia do ruído de fundo apresenta significativas variações temporais, sendo preciso renovar o limiar de detecção.

Neste âmbito, propõe-se a regra de atualização [3]

$$L_{novo} = (1-p)L_{atual} + pE_{silêncio}, \tag{2}$$

onde L_{novo} representa o novo limiar a ser calculado, L_{atual} indica o valor de limiar mais recentemente empregado, $E_{silêncio}$ é a energia calculada para o último quadro em que não se detectou voz e $0 < p < 1$.

O papel do parâmetro p é fundamental para o funcionamento do *ALED*, sendo o critério de escolha do seu valor calcado na variação da estatística de segunda ordem do conjunto das energias calculadas para os m quadros mais recentes onde não houve a detecção de voz. Em outras palavras, dada uma memória de m componentes, denotada por M_m, preenchida pelos valores das energias dos últimos m quadros silenciosos, assim que não se detecta voz em um novo quadro, a energia mais antiga armazenada nesta memória é descartada e substituída pela atual, sendo, posteriormente, computada a variância deste novo conjunto conforme o seguinte cálculo:

$$\sigma^2 = \text{var}(M_m), \tag{3}$$

onde var$(\cdot)$ é a variância dos m componentes da memória.

Uma vez conhecido o valor de σ^2_{novo} - após a troca do último elemento da memória - comparamos este valor àquele obtido antes da substituição (σ^2_{prev}), escolhendo o valor de p segundo a Tabela I, proposta em [1], de onde notamos, pela substituição dos valores sugeridos na Equação (2), que quanto maior é a variação de ξ, maior peso é dado para a contribuição da energia do último quadro silencioso ($E_{silêncio}$). Trata-se de um proceder coerente, na medida em que um maior valor de

TABELA I: Escolha de p.

$\frac{\sigma^2_{novo}}{\sigma^2_{prev}} = \xi$	p
$\xi \geq 1,25$	$0,25$
$1,25 > \xi \geq 1,10$	$0,20$
$1,10 > \xi \geq 1,00$	$0,15$
$1,00 > \xi$	$0,10$

ξ indica maior energia no quadro atual quando comparado a quadros anteriores, ou seja, a energia do ruído de fundo está aumentando e o limiar de detecção precisa acompanhá-la o mais rapidamente possível.

É fácil perceber que os valores iniciais que preenchem M_m são fundamentais para o bom funcionamento do método *ALED*. Portanto, costuma-se assumir que os m quadros iniciais do sinal são desprovidos de fenômenos vocais, armazenando-se suas energias diretamente em M_m e utiliza-se a média destas energias (escalada por um fator $k > 1$, entendido como margem de erro) como limiar inicial.

III. DETECTOR LINEAR DE ENERGIA EM SUB-BANDAS

O detector linear de energia em sub-bandas (*Linear Sub-Band Energy Detector - LSED*) é um processo no domínio da frequência que trabalha as sub-bandas de sinais com voz de modo idêntico ao *ALED*. Neste sistema, após a divisão do sinal $x(i)$ em quadros $q_j = x(i)|_{(j-i)K+1}^{jK}$, onde j é o índice do quadro e K é o seu tamanho em amostras, computam-se suas DCTs, conforme explicitado na Equação (4), dividindo-as em sub-bandas[1] [1].

$$F(q_j) = DCT(q_j). \tag{4}$$

Para um sinal amostrado na taxa de 8000 amostras/seg, adotam-se bandas de largura 1 kHz: 0-1 kHz, 1-2 kHz, 2-3 kHz e 3-4 kHz, sendo suas energias computadas segundo

$$E_j = \sum_n |(F_{j,n})|^2, \tag{5}$$

onde E_j indica a energia da j-ésima sub-banda, cujas n raias são denotadas por $F_{j,n}$, que será comparada ao limiar $L_{n_{atual}}$.

Finalmente, aplica-se a lógica combinacional apresentada na Figura 1 para a tomada de decisão sobre existência de voz no quadro presente.

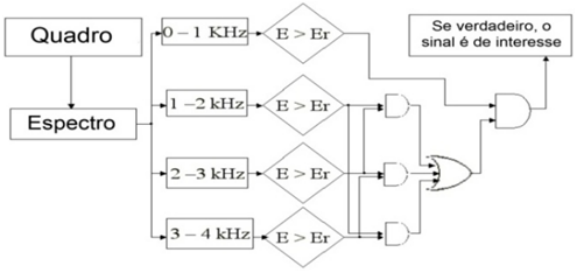


Fig. 1: Diagrama lógico do *LSED* - Adaptado de [1].

[1]Emprega-se a DCT pois esta transformada é computacionalmente menos custosa que a DFT e evita o surgimento de valores complexos.

É interessante notar que, justamente pelo fato de a maior parte da energia da voz estar concentrada na faixa 0-1 kHz, esta é dominante no processo *LSED*. Assim, torna-se fundamental a superação da energia de limiar nesta sub-banda e em qualquer outro par de sub-bandas para que seja indicada a atividade vocal.

Por outro lado, quando um quadro é entendido como silencioso, a energia das sub-bandas é armazenada em memórias respectivas e adota-se o processo descrito pelas Equações (2)-(3) para a atualização dos limiares de cada sub-banda.

IV. MODELO ESTATÍSTICO DE DETECÇÃO DE VOZ

O detector de voz proposto em [4] e aperfeiçoado em [5] (*Statistical Model-Based Voice Activity Detector - SMBVAD*) sugere que é possível determinar com acurácia as regiões com voz de um sinal se as estatísticas do ruído tiverem menor variabilidade que as estatísticas da voz que se deseja localizar. Esta hipótese enseja o emprego de uma métrica estatística para a determinação de um limiar de decisão, o qual distingue trechos nos quais a voz deve estar presente daqueles onde apenas ocorre o ruído de fundo. Para tal, utiliza-se para a estimativa estatística um critério de máximo verossimilhança, adotando-se como ponto de partida a suposição de que podemos conhecer as estatísticas do ruído de fundo através de aferição por um estimador estatístico.

O modelo empregado assume, adicionalmente, que, em cada quadro analisado, podemos enquadrar tanto o ruído como a voz em modelos estatísticos gaussianos e independentes entre si, ou, de maneira sintética,

$$\begin{cases} H_0 : \text{sem voz} \rightarrow Y = N \\ H_1 : \text{com voz} \rightarrow Y = N + S \end{cases}, \tag{6}$$

sendo S (voz), N (ruído) e Y DFTs de dimensão L cujos k-ésimos elementos representamos por S_k, N_k e Y_k, respectivamente. Isto posto, as funções densidade de probabilidade (PDFs, do inglês *probability density functions*) condicionadas por H_0 e H_1 são dadas por

$$\begin{aligned} p(Y_k|H_{0,k}) &= \tfrac{1}{\pi\lambda_{N,k}} e^{-\frac{|Y_k|^2}{\lambda_{N,k}}} \\ p(Y_k|H_{1,k}) &= \tfrac{1}{\pi(\lambda_{N,k}+\lambda_{S,k})} e^{-\frac{|Y_k|^2}{\lambda_{N,k}+\lambda_{S,k}}} \end{aligned}, \tag{7}$$

onde λ_N e λ_S representam, respectivamente, a variância do espectro do ruído e da voz.

Conhecidas estas PDFs, pode-se definir o estimador de máxima verossimilhança supracitado, responsável por determinar a existência ou não de voz em quadros do sinal, como

$$\Lambda_k = \frac{p(Y_k|H_{1,k})}{p(Y_k|H_{0,k})} = \frac{1}{1+\xi_k} e^{\frac{(1+\gamma_k)\xi_k}{(1+\xi_k)}}, \tag{8}$$

onde $\gamma_k = \frac{|Y_k|^2}{\lambda_{N,k-1}}$, $\xi_k = \frac{\lambda_{S,k}}{\lambda_{N,k}}$ e Λ_k é o estimador de máximo verossimilhança que buscamos para cada componente frequencial.

Cabe destacar que estas equações ainda são indefinidas, pois é preciso conhecer a variância λ_S do espectro da voz para calcularmos ξ_k. Ademais, também carecemos de um mecanismo para o cálculo de $\lambda_{N,k}$ para tornarmos definido γ_k.

Portanto, apresentamos em seguida o mecanismo de estimação de $\lambda_{N,k}$ e de ξ_k sem o conhecimento prévio de λ_S.

O cálculo de $\lambda_{N,k}$ tem como ponto inicial a probabilidade de ausência de voz em um quadro de sinal, o que pode ser computado por intermédio do teorema de Bayes pela equação

$$p(H_{0,k}|Y_k) = \frac{1}{1 + \frac{1-p(H_{0,k})}{p(H_{0,k})}\psi_k}, \quad (9)$$

onde ψ será definido na Equação (15) e a probabilidade *a priori* $p(H_{0,k})$ de ausência de fala é estimada de maneira adaptativa [6] aplicando-se

$$p(H_{0,k}) = \frac{1}{1 + e^{-\xi_k} I_0(2\sqrt{\gamma_k \xi_k})}, \quad (10)$$

em que I_0 representa a função de Bessel modificada de ordem zero.

O espectro de potência do ruído de fundo é, então, estimado por

$$\mathrm{E}\left(\left|N_k^{(n)}\right|^2 \left|Y_k^{(n)}\right.\right) = p\left(H_{0,k}\left|Y_k^{(n)}\right.\right)\left|Y_k^{(n)}\right|^2 + \left(1 - p\left(H_{0,k}\left|Y_k^{(n)}\right.\right)\right)\lambda_{N,k}^{(n-1)}, \quad (11)$$

de onde deduzimos a fórmula para a obtenção de γ_k:

$$\lambda_{N,k}^{(n)} = \eta\lambda_{N,k}^{(n-1)} + (1-\eta)\mathrm{E}\left(\left|N_k^{(n)}\right|^2 \left|Y_k^{(n)}\right.\right), \quad (12)$$

onde η atua como fator de controle de aprendizado do sistema e $\mathrm{E}(\cdot)$ é o operador de média estatística.

Finalmente, resta-nos o cálculo de ξ_k sem o conhecimento direto de $\lambda_{S,k}$. Em [5] sugere-se o método *decision-direct* dado por

$$\xi_k = \alpha\frac{|\hat{S}_k^{(n-1)}|^2}{\lambda_{N,k}^{(}n-1)} + (1-\alpha)\max(\gamma_k^{(n)}, 0), \quad (13)$$

como solução para este problema, sendo α uma constante de ponderação e $|\hat{S}_k|$ a amplitude espectral da parcela de voz, estimada utilizando a técnica de minimização do erro quadrado médio proposta em [7], conforme a equação:

$$\hat{S}_k = \frac{\xi_k}{1+\xi_k} Y_k. \quad (14)$$

Uma vez computadas estas grandezas, torna-se possível o cálculo de Λ_k para cada janela de sinal analisada. Dado que esta decisão deve ser tomada a cada quadro (e não a cada raia da DFT), substitui-se Λ_k por sua média geométrica em relação a k antes da tomada de decisão [5]. Com o intuito de contornar eventuais detecções incorretas nas regiões de ataque e decaimento do sinal de voz - fenômeno introduzido pela ação do termo atrasado $\lambda_{N,k}^{(n-1)}$ - [5], adiciona-se memória ao sistema, de tal modo que a verossimilhança de um quadro passe a depender do resultado do quadro anterior, isto é,

$$\psi_k^{(n)} = e^{\iota \log \psi_k^{(n-1)} + (1-\iota)\log \Lambda_k^{(n)}}, \quad (15)$$

onde ι é um fator de suavização cujo valor deve ser escolhido no intervalo $[0, 1)$ e, uma vez aplicada esta relação, a média geométrica passa a ser calculada sobre $\psi_k^{(n)}$, permitindo a obtenção da suavidade pretendida. Este resultado é, então, comparado ao limiar de detecção estatístico determinado heuristicamente.

V. COMPARAÇÃO DE DESEMPENHO

Os testes comparativos do *ALED*, *LSED* e *SMBVAD* foram realizados sobre um mesmo sinal de duração de 33 seg, amostrado na taxa de 8000 amostras/seg e resolução de 16 bits, em que se alternam momentos nos quais há presença de voz e outros silenciosos, gravado em ambiente livre de ruído. Para a condução dos testes, ruído branco gaussiano de média nula foi adicionado artificialmente, de modo que compusemos quatro cenários distintos, representados nas Figuras 2(a)-2(d) com, respectivamente, 10 dB SNR, 3 dB SNR, -3 dB SNR e -10 dB SNR, onde apresentamos o perfil de detecção alcançado por cada algoritmo. Quando a linha de detecção está em nível lógico baixo (ND) não há detecção. Caso contrário, ela está em nível lógico alto (D), assinalando a detecção. Além dos resultados dos 3 detectores de voz discutidos, apresentamos o perfil de detecção ideal (REF), obtido por inspeção visual da forma de onda previamente à adição de ruído ao sinal.

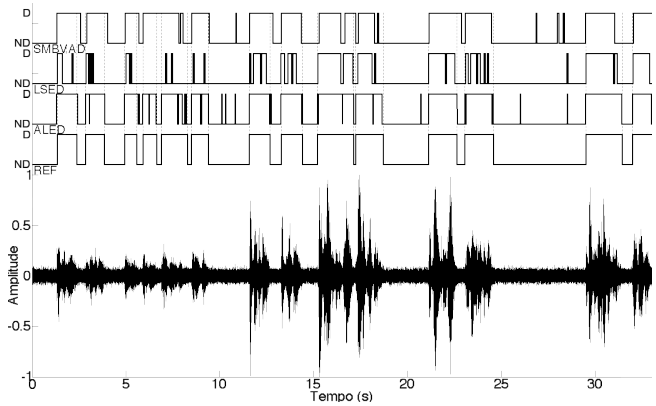
Ademais, as Figuras 3(a)-3(b) apresentam em forma gráfica a taxa de detecção falso positiva e falso negativa dos sistemas apresentados quando confrontados com a detecção ideal nos contextos ruidosos. Cabe enfatizar que a detecção obtida por inspeção da forma de onda não considera pausas entre palavras e sílabas, o que, seguramente, adiciona uma margem de erro aos percentuais de falso negativo apresentados, que estimamos ser de até -5%.

Cabe destacar que para cada teste buscamos o melhor ajuste dos parâmetros de cada um dos detectores de voz de modo a obter a melhor relação entre a taxa de falsos negativos e falsos positivos. Os resultados exibidos nas figuras supracitadas são aqueles que obtiveram maior êxito sob este aspecto, sendo os parâmetros de ajuste dos *VADs* apresentados na Tabela II, onde M_m denota o número de posições disponíveis na memória utilizada pelo *ALED* e pelo *LSED*, T_q denota o tamanho do quadro de análise empregado pelos algoritmos que nos propusemos a comparar e L_e representa o limiar estatístico utilizado pelo *SMBVAD*.

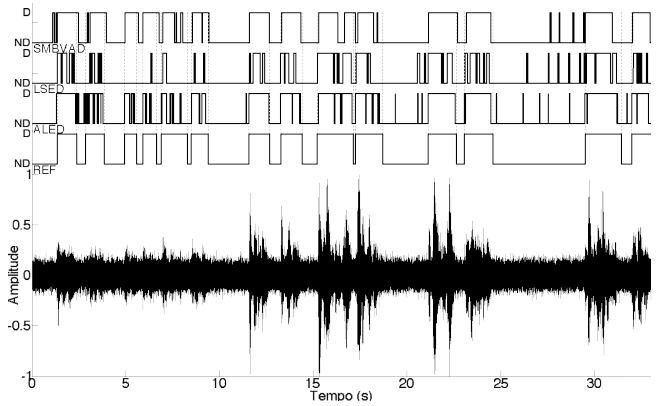
A inspeção da Figura 2 nos permite identificar o modelo estatístico de detecção de voz como superior, mesmo nos contextos com SNRs inferiores sob o ponto de vista da taxa de detecções falso negativas e com os parâmetros usados. Segundo esperado, é clara a tendência crescente desta taxa, independentemente do método empregado, à medida que se reduz a SNR.

Sob o ponto de vista de detecções falso positivas, embora o *SMBVAD* apresente o pior resultado na maioria dos casos, cabe destacar que para um contexto de comunicação, detecções falso positivas são consideravelmente menos críticas do que as falso negativas, uma vez que não comprometem a compreensão da conversação.

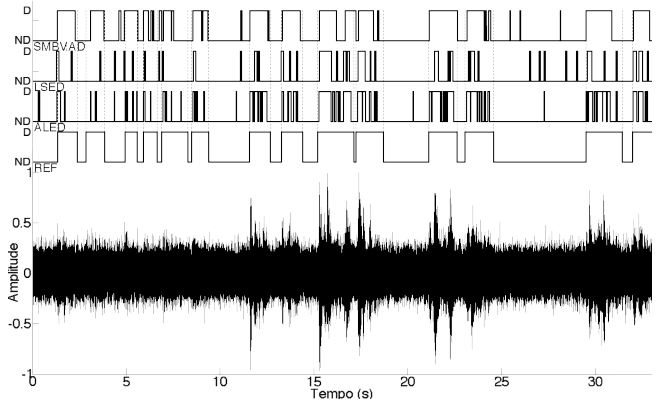
Destaca-se da avaliação da Figura 3(a) o *ALED* como método mais robusto contra falsos positivos, o que pode ser



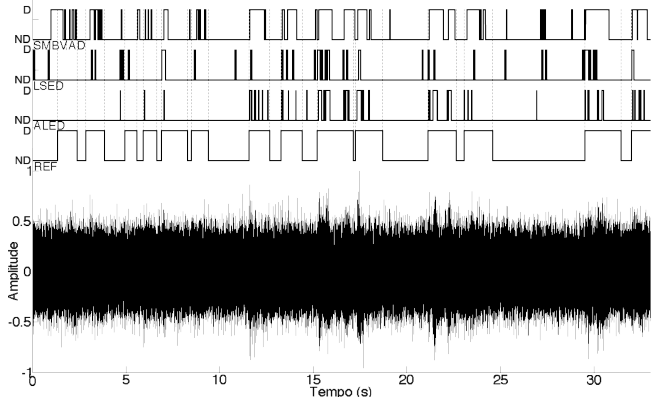
(a) 10 dB SNR.



(b) 3 dB SNR.

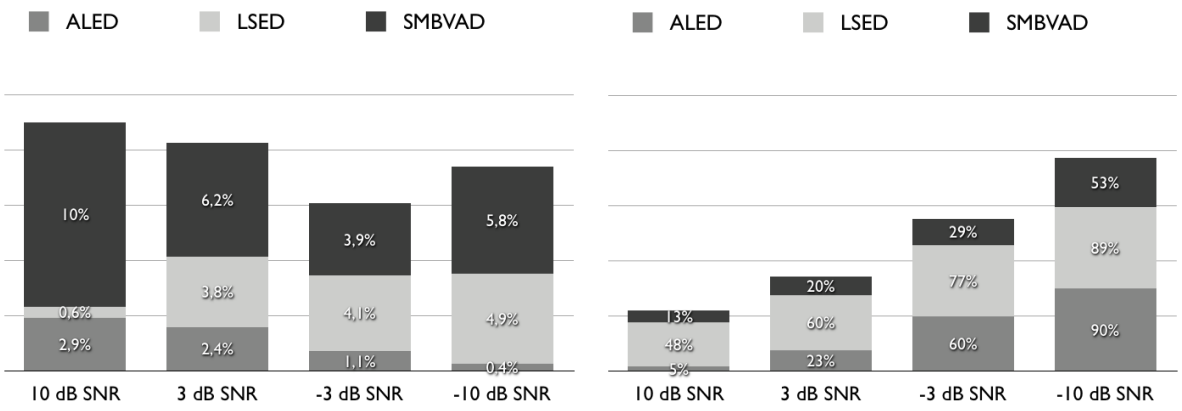


(c) -3 dB SNR.



(d) -10 dB SNR.

Fig. 2: Resultados dos diferentes sistemas com degradação progressiva da SNR.



(a) Falso positivo. (b) Falso negativo.

Fig. 3: Taxas de falsos positivos e falsos negativos.

TABELA II: Parâmetros de teste.

VAD/SNR	10 dB	3 dB	-3 dB	-10 dB
ALED(Tq; k; Mm)	160; 1,50; 32	160; 1,40; 128	160; 1,30; 128	160; 1,30; 128
LSED(T_q; k; Mm)	1024; 1,04; 32	512; 1,06; 128	512; 1,06; 128	512; 1,06; 128
SMBVAD(T_q; η; α; ι; L_e)	128; 0,86; 0,98; 0,8; 1,030	128; 0,85; 0,98; 0,8; 1,015	128; 0,85; 0,98; 0,8; 1,012	128; 0,85; 0,98; 0,8; 1,011

entendido quando nos lembramos de que o critério de decisão deste detector é apenas a energia do ruído de fundo no domínio do tempo, que é praticamente constante nos testes apresentados.

Esperar-se-ia, dada a semelhança entre os algoritmos, que o *LSED* apresentasse comportamento semelhante. Porém, ao contrário do que ocorre no *ALED*, a decomposição em sub-bandas acrescentada na estrutura do *LSED* acaba tornando-se uma desvantagem quando o sinal é comprometido por ruído branco, uma vez que este tem energia igualmente distribuída por todas as frequências, fazendo com que a robustez adicionada ao processo pela decomposição frequencial do sinal seja minimizada.

Apresentadas estas características, discutimos a taxa de falsos positivos apresentada pelo *SMBVAD*. Constata-se que, para SNR elevada, este mecanismo apresenta maior taxa do que para SNRs inferiores. Atribuímos esta característica ao fato de a base deste detector ser a suposição de pequena variação das estatísticas do ruído de fundo, sendo mais fácil angariá-las e estimá-las quando a sua energia é maior, ou seja, conforme ela cresce, mais fácil é esta identificação.

Além dessas avaliações, é pertinente que se pondere sobre a complexidade computacional dos algoritmos apresentados. A operação mais custosa tanto para o *ALED* quanto para o *LSED* é o somatório de *K* termos quadráticos - Equação (1) e Equação (5), respectivamente - havendo, adicionalmente, a computação da *DCT* no segundo algoritmo, tornando-o marginalmente mais complexo. Quanto ao *SMBVAD*, além da computação da *FFT* e somatório de termos quadráticos (Equação (11)), é preciso calcular a função de Bessel, exponencial, logaritmo e realizar a média geométrica de *K* termos, o que torna esse o algoritmo mais complexo dentre os três que foram comparados.

Finalmente, ponderamos que todos os sistemas propostos, embora dependentes de diversos parâmetros (apresentados resumidamente na Tabela II), apresentam grande estabilidade em relação aos seus valores ótimos. Constata-se, na maioria dos casos, uma razoável insensibilidade aos parâmetros ótimos quando a SNR varia, ocorrendo a maior variação na passagem de 10 dB para 3 dB SNR.

VI. CONCLUSÃO

Neste artigo apresentamos a formulação teórica e um comparativo entre os desempenhos de três detectores de atividade de voz em sinais: o detector linear e adaptativo de energia, o detector linear de energia em sub-bandas e o modelo estatístico de detecção de voz. Os testes apresentados demonstraram a maior robustez do último método em ambiente ruidoso e a tendência à menor taxa de falsos positivos apresentada pelo primeiro (empregamos ruído branco nas simulações, mas, possivelmente, ruídos distintos implicariam diferentes conclusões). O método em sub-bandas, por outro lado, demonstrou-se o menos eficaz, fato atribuído ao tipo de ruído (branco e gaussiano) que compromete o sinal.

REFERÊNCIAS

[1] R. Venkatesha Prasad, A. Sangwan, H.S. Jamadagni, M.C. Chiranth, R. Sah, and V. Gaurav. Comparison of voice activity detection algorithms for voip. In *Proc. 2002 Seventh Int. Symp. on Computers and Communications*, pages 530 – 535, Jul. 2002.

[2] Felipe Sander Pereira Clark. Cancelamento de eco acústico e separação cega de fontes aplicados à telefonia viva-voz, Dec. 2010.

[3] Petr Pollák, Pavel Sovka, and Jan Uhlír. Noise suppression system for a car. In *Proc. 1993 Third European Conf. on Speech, Communication and Technology*, pages 1073 – 1076, Sept. 1993.

[4] Jongseo Sohn, Nam Soo Kin, and Wonyong Sung. A statistical model-based voice activity detection. *Signal Processing Letters, IEEE*, 6(1):1 – 3, Jan. 1999.

[5] Yong Duk Cho and A. Kondoz. Analysis and improvement of a statistical model-based voice activity detector. *Signal Processing Letters, IEEE*, 8(10):276 –278, Oct. 2001.

[6] Ing Yann Soon, Soo Ngee Koh, and Chai Kiat Yeo. Improved noise suppression filter using self-adaptive estimator of probability of speech absence. *Signal Processing*, 75(2):151 – 159, Jun. 1999.

[7] Yariv Ephraim and David Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *Transactions on Acoustics, Speech and Signal Processing, IEEE*, 32(6):1109 – 1121, Dec. 1984.