Identificação da natureza do ruído baseada em HMM

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

***Resumo*— Este artigo apresenta um método eficiente para identificar tipos de ruído em sinais de voz com baixa SNR. De acordo com a natureza do ruído e sua potência, a técnica mais apropriada de tratamento robusto pode ser escolhida. O processo de identificação é realizado através de HMM e os parâmetros extraídos são entropia espectral, taxa de cruzamentos por zero e log-energia de 16 sub-bandas. Com o método proposto, para quatro possíveis tipos de ruído (branco, rosa, falatório e interior de carro), obtivemos 97,22% de correta identificação da natureza do ruído quando misturados com comandos de voz, e 100% quando o sinal é composto apenas por ruído.**

***Palavras-Chave*— Ruído, Processamento Robusto de Voz, HMM, Espectro de Potência.**

***Abstract*— This article presents an efficient method to identify types of noise in speech signals with low SNR. In accordance with the type of noise and its power, the most suitable technique for robust treatment can be selected. The process is carried out by HMM and the extracted parameters are spectral entropy, zero crossing rate and log-energy from 16 sub-bands. With the considered method, for four possible types of noise (white, pink, babble and car interior), we got 97.22% of correct identification of the nature of the noise when mixed with commands of voice, and 100% when the signal is composed only for noise.**

***Keywords*— Noise, Robust Speech Processing, HMM, Power Spectrum.**

I. INTRODUÇÃO

Pesquisas na área de reconhecimento robusto têm sido intensamente realizadas no sentido de viabilizar aplicações práticas de reconhecimento de voz. Em atividades recentes de processamento robusto de voz, estimativas do espectro do ruído são realizadas, a partir das amostras do sinal ruidoso [1], [2].

Neste trabalho, não só a potência do ruído é estimada, mas também a sua natureza, o que permite a escolha da técnica de processamento robusto mais apropriada. Isso pode ser observado em [3], onde temos uma técnica eficiente para alguns tipos de ruídos e ineficiente para outros.

O algoritmo que utilizamos para identificar a natureza dos ruídos é baseado na extração da entropia espectral, taxa de cruzamentos por zero e log-energia de 16 sub-bandas na classificação utilizando modelos de Markov escondidos (hidden Markov models - HMM). Com o método proposto obtivemos 97,22% de acerto na identificação da natureza do ruído em locuções contaminadas e 100% em sinal composto somente por ruído.

Este artigo está organizado da seguinte forma. A Seção II descreve o procedimento usado para estimativa da relação sinal-ruído (signal noise rate - SNR). Na Seção III, apresentamos o método de identificação do tipo de ruído. A Seção IV apresenta a origem dos sinais utilizados, tanto os ruídos como as locuções. Por fim, na Seção V temos os resultados obtidos e na Seção VI, as conclusões.

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

II. ESTIMATIVA DA *SNR*

Em termos práticos, quanto menor for a *SNR*, maior é a necessidade de que um processamento robusto seja realizado. Para a estimativa da *SNR*, uma atualização adaptativa da potência do ruído é realizada, onde um filtro IIR de um pólo é utilizado [1], [2], [4].

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

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

Definimos como $\sigma_x(k)$ a potência do sinal ruidoso no k-ésimo frame.

$$\sigma_x(k) = \sum_{i=0}^{L-1} x^2(kL + i) \tag{2}$$

onde k é o índice do frame e L é o comprimento do frame.

Definindo também $\sigma_d(k)$ como a potência do ruído no k-ésimo frame e assumindo que ela varia de forma mais lenta que a potência do sinal ruidoso, $\sigma_x(k)$, uma estimativa da potência do ruído é realizada, recursivamente, através do filtro IIR:

$$\sigma_d(k) = \alpha(k)\sigma_d(k-1) + (1 - \alpha(k))\sigma_x(k) \tag{3}$$

O parâmetro $\alpha(k)$ é o elemento que conduz a atualização da estimativa em direção ao sinal ruidoso no frame k ou em direção à estimativa da potência do ruído no frame anterior:

$$\alpha(k) = 1 - min(1, RSNR(k)^{-Q}) \tag{4}$$

onde $Q = 5$ e a *SNR* relativa no k-ésimo frame, $RSNR(k)$, é:

$$RSNR(k) = \frac{\sigma_x(k)}{\frac{1}{M}\sum_{n=0}^{M-1}\sigma_x(n)} \tag{5}$$

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

Assim, se um dado frame possui apenas ruído, a sua $RSNR$ fica bem próxima de 1, resultando num $\alpha(k)$ pequeno (em torno de 0). Desta forma, a atualização do ruído segue o sinal ruidoso no frame k. Caso o frame possua voz, a $RSNR$ é bem mais elevada e $\alpha(k)$ fica mais próximo de 1. Neste caso, a estimativa do ruído segue a atualização do frame anterior.

A Figura 1 mostra a estimativa da potência do ruído $d(i)$ a partir do sinal ruidoso $x(i)$, para o caso de uma locução contaminada por ruído rosa com *SNR* de 5dB. Podemos

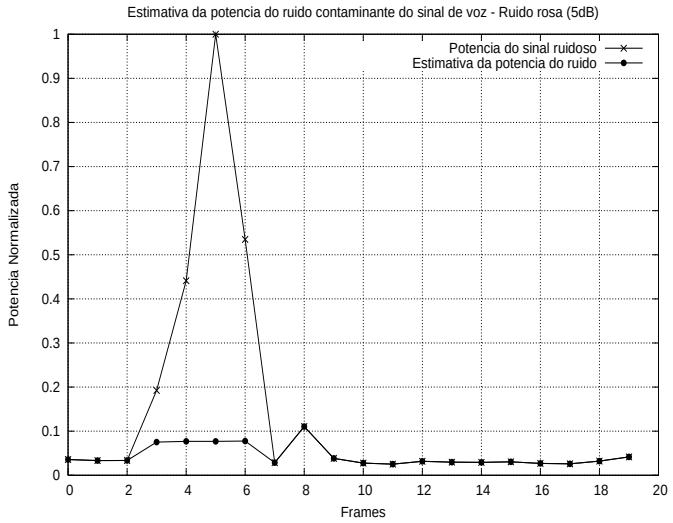

Fig. 1. Estimativa da potência do ruído num sinal contaminado por ruído rosa.

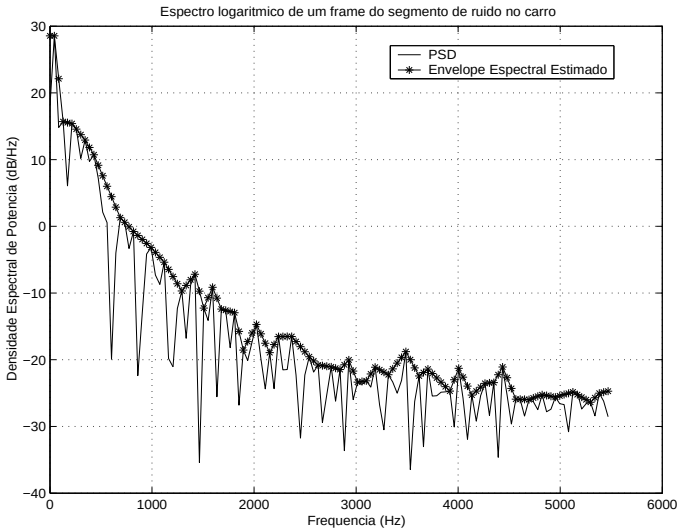

Fig. 2. Envoltória do espectro de um frame com ruído no carro.

observar nesta figura que a estimativa da potência do ruído até o frame dois e após o frame sete, segue perfeitamente a potência do sinal ruidoso $x(i)$, por se tratar de trechos onde a variação de $\sigma_x(k)$ é lenta ao longo dos frames. Fora deste trecho a variação entre frames de $\sigma_x(k)$ é brusca, denotando a presença de $s(i)$ nesta região e a potência do ruído $\sigma_d(k)$ passa a seguir as estimativas anteriores.

Com o conhecimento de $\sigma_x(k)$ e $\sigma_d(k)$, uma estimativa da *SNR* é feita segundo a equação a seguir:

$$SNR = 10log\left(\frac{\sum_{k=th1}^{th2}\sigma_x(k) - \sum_{k=th1}^{th2}\sigma_d(k)}{\sum_{k=th1}^{th2}\sigma_d(k)}\right) \quad (6)$$

onde $th1$ e $th2$ são os limites inferior e superior, respectivamente, do conjunto de frames onde temos variações mais bruscas de σ_x e, conseqüentemente, presença de $s(i)$. Assim, para o cálculo da *SNR*, serão utilizados apenas os k-ésimos frames onde a relação de potências, $\frac{\sigma_x(k)}{\sigma_d(k)}$, ultrapassar um limiar $TH = 2$. No exemplo da Figura 1, $th1 = 2$ e $th2 = 7$.

Com a estimativa da *SNR* do sinal, temos condições de dizer se o nível de ruído é tolerável ou não, dependendo da aplicação do processamento.

III. Identificação do tipo ruído

A. Extração de parâmetros

No processo de identificação do tipo de ruído, para cada frame k são extraídos 18 coeficientes:

- Entropia espectral (01);
- Taxa de cruzamentos por zero (01);
- Log-energia (16);

A entropia espectral, H, é definida na equação a seguir:

$$H = -\sum_{k=0}^{P-1} p_k log(p_k) \quad (7)$$

onde p_k é uma função densidade de probabilidade do espectro na freqüência de índice k, estimada pela normalização sobre todas as componentes de freqüência e P é o número de componentes de freqüência. O espectro $S(f_j)$, para cada freqüência f_j, é obtido em cada frame através da transformada rápida de Fourier (fast Fourier transform - FFT).

$$p_j = \frac{S(f_j)}{\sum_{k=0}^{P-1} S(f_k)}, \text{ para } j = 0, ..., P-1 \quad (8)$$

A entropia espectral está relacionada com a idéia de que um sinal tem tanto mais informação quanto maior for o seu grau de imprevisibilidade [5], [6].

A taxa de cruzamentos por zero, ZCR, é definida na equação a seguir:

$$ZCR = \frac{1}{L-1}\sum_{i=1}^{L-1}\frac{|sgn\{x(i)\} - sgn\{x(i-1)\}|}{2} \quad (9)$$

onde

$$sgn\{x(i)\} \quad = \quad \begin{cases} +1, & x(i) \geq 0 \\ -1, & x(i) < 0 \end{cases}$$

A taxa de cruzamentos por zero, muito utilizada em métodos de detecção de extremos [7], representa do número de vezes que o sinal cruza o eixo onde o valor de $x(i)$ é zero.

Considerando, agora, o espectro do sinal $x(i)$, definimos como $\hat{S}(f)$ uma estimativa da envoltória do espectro logarítmico. Uma divisão do espectro em K sub-bandas com P' freqüências em cada sub-banda é realizada, para a extração de K parâmetros. O k-ésimo parâmetro de log-energia, $LogE(k)$, é definido como a energia contida na k-ésima sub-banda, limitada pela envoltória, conforme a seguir:

$$LogE(k) = \sum_{j=0}^{P'-1}\hat{S}^{(k)}\{F(kP'+j)\} \quad (10)$$

onde $k = 0, ..., K-1$ e $F = \frac{f_s}{N}$, sendo f_s a freqüência de amostragem e N o número de pontos usados na FFT.

Na Figura 2 apresentamos o espectro de potência logarítmico, juntamente com a estimativa da envoltória, para

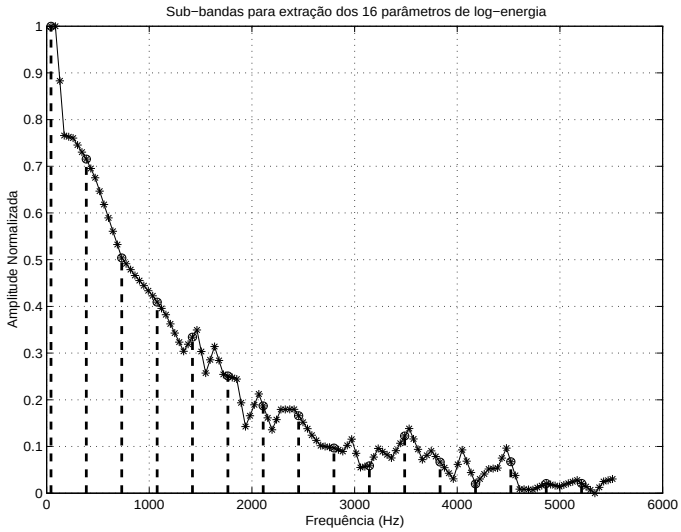

Fig. 3. Sub-bandas para extração de 16 parâmetros de log-energia

o caso de ruído no interior do carro. A envoltória foi estimada através de interpolações entre os picos do espectro logarítmico. Partindo deste espectro, foi feita a divisão em sub-bandas para a extração dos coeficientes. Cada frame foi dividido em 16 sub-bandas, gerando 16 coeficientes de log-enegria, $LogE$. Na Figura 3 temos a concentração de energia em cada uma das 16 sub-bandas, para o caso do ruído no interior do carro, mostrado na Figura 2. Os intervalos das sub-bandas, delimitados pelas linhas tracejadas da Figura 3, são apresentados na Tabela I.

TABELA I

TABELA COM OS INTERVALOS DE FREQÜÊNCIAS PARA EXTRAÇÃO DOS PARÂMETROS DE LOG-ENERGIA

BANDA	FREQÜÊNCIA
1	$0 - 344.5Hz$
2	$344.5 - 689.0Hz$
3	$689 - 1033.5Hz$
4	$1033.5 - 1378Hz$
5	$1378 - 1722.5Hz$
6	$1722.5 - 2067Hz$
7	$2067 - 2411.5Hz$
8	$2411.5 - 2756Hz$
9	$2756 - 3100.5Hz$
10	$3100.5 - 3445Hz$
11	$3445 - 3789.5Hz$
12	$3789.5 - 4134Hz$
13	$4134 - 4478.5Hz$
14	$4478.5 - 4823Hz$
15	$4823 - 5167.5Hz$
16	$5167.5 - 5512Hz$

B. Configuração do HMM

Cada um dos 4 tipos de ruídos têm os seus parâmetros usados no treinamento dos HMM discretos de [8], os quais possuem as seguintes características:

- 50% de superposição;
- 04 modelos;
- 04 estados;
- 128 centróides;
- 18 parâmetros por frame, sendo 01 de entropia espectral, 01 de taxa de cruzamentos por zero e 16 de log-energia espectral;

C. Critérios de decisão

A decisão acerca da natureza do ruído é realizada partindo de locuções contaminadas e com tamanho fixo de dois segundos de duração. Elas são segmentadas em trechos de 100 ms com superposição de 50%. Cada um desses segmentos é submetido ao classificador e a identificação do mesmo é registrada. Em seguida, uma contagem é realizada apenas sobre os segmentos de 100 ms que pertencem à região do sinal $x(i)$, sem a presença de $s(i)$. Para isto, levamos em consideração a potência estimada do ruído no frame k, $\sigma_d(k)$, realizada na Seção II. Os parâmetros $th1$ e $th2$ são os delimitadores das regiões de $x(i)$ com potência mais baixa. Este proced imento é adotado para evitar erros de classificação no caso de ocorrerem locuções muito longas com intervalos de voz muito curtos. No exemplo da Figura 1, os frames submetidos à análise são os que vão até $th1 = 2$ e estão após $th2 = 7$.

IV. BASE DE DADOS

As locuções utilizadas neste artigo foram coletadas da base de dados descrita em [8], onde temos 10 palavras isoladas (ANDA, BAIXO, CIMA, DESLIGA, DIREITA, ESQUERDA, FRENTE, MÃO, OLHA e TRAS), repetidas por diferentes locutores. A taxa de amostragem é de 11025Hz.

A base de dados de ruídos foi utilizada de [9], em formato "wave", onde cada ruído possui 9.0Mb, 235 segundos de duração, taxa de amostragem original de 19980Hz subamostrada para 11025Hz e tomadas da NOISEX-92. Foram selecionados quatro tipos de ruídos:

- Ruído branco (WHITE);
- Ruído rosa (PINK);
- Ruído no interior de um carro (VOLVO); e
- Falatório (BABBLE).

Os dois primeiros foram selecionados por serem tradicionais em tarefas de processamento robusto e os dois últimos por representarem situações reais de operação em condições adversas.

Os ruídos foram segmentados em trechos de 100 ms com superposição de 50% a fim de formar uma sub-base de dados para treinar os HMM. Com cada um dos tipos de ruídos foram obtidos 9178 segmentos com 100 ms de ruído.

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

V. RESULTADOS OBTIDOS

Inicialmente foi feito um treinamento com 500 segmentos de 100 ms de cada um dos quatro tipos de ruídos envolvidos. Em seguida, foram feitos dois testes com o sistema, buscando-se medir sua eficiência na identificação dos tipos de ruídos. O primeiro teste foi realizado com ruído, onde foram inseridos 100 segmentos de cada um dos tipos envolvidos, diferentes

TABELA II

TABELA DE CONFUSÃO DO TESTE COM SEGMENTOS DE RUÍDO

	BABBLE	PINK	VOLVO	WHITE
BABBLE	100%	0%	0%	0%
PINK	0%	100%	0%	0%
VOLVO	0%	0%	100%	0%
WHITE	0%	0%	0%	100%
Taxa de acerto média: 100%				

TABELA III

TABELA COM RESULTADOS DO TESTE COM LOCUÇÕES CONTAMINADAS

	BABBLE	PINK	VOLVO	WHITE
0dB	100%	100%	100%	100%
5dB	100%	100%	100%	100%
10dB	100%	100%	100%	100%
15dB	100%	98.34%	100%	98.34%
20dB	99.17%	85.12%	89.25%	74.38%
Taxa de acerto média: 97.22%				

dos usados no treino, com 100 ms cada. Através de HMM, como mencionado na Sub-seção III-B, foram extraídos 18 parâmetros de cada frame, com 20 ms e superposição de 50%. Os resultados são apresentados na Tabela II.

Em seguida foi realizado um teste funcional com sinais de voz contaminados pelos ruídos selecionados a *SNR*'s que vão de 0dB a 20dB. O procedimento foi fixar uma *SNR* e o ruído, observando o acerto ao longo das várias *SNR*. Foram inseridas 121 locuções no sistema. Após contaminadas pelos ruídos, as locuções foram segmentadas em trechos de 100 ms e analisadas através de HMM. Como mencionado na Sub-seção III-C, os segmentos com relação de potências acima do limiar TH não foram considerados na classificação. Os resultados obtidos são mostrados na Tabela III.

VI. CONCLUSÕES FINAIS E TRABALHOS FUTUROS

Neste trabalho apresentamos um método de identificação da natureza do ruído baseado em HMM para diferentes *SNR*'s. Foram extraídos como parâmetros entropia espectral, taxa de cruzamentos por zero e log-energia de 16 sub-bandas. Os testes foram realizados somente com ruídos e, também, com locuções contaminadas por quatro tipos de ruídos (branco, rosa, falatório e interior do carro). Os resultados obtidos atingiram 100% no teste com ruídos e 97,22% na identificação em locuções contaminadas.

Trabalhos futuros estão sendo desenvolvidos, baseados na proposta aqui apresentada, visando atividades de processamento robusto através de múltiplas bandas, onde, dependendo do tipo de ruído e da *SNR*, podemos conduzir o nosso processamento de forma apropriada.

REFERÊNCIAS

[1] L. Lin, W. H. Holmes and E. Ambikairajah, *Subband noise estimation for enhancement using a perceptual Wiener filter*. In ICASSP, v.I, p.80-83, April 2003.

[2] I. Cohen, *Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging*. In IEEE Trans. on Speech and Audio Processing, v.11, p.466-475, September 2003.

[3] S. Tibrewala and H. Hermansky, *Sub-band based recognition of noisy speech*. In Proc. ICASSP, v.II, p.1255-1258, April 1997.

[4] D. C. Popescu and I. Zeljkovic, *Kalman filtering of colored noise for speech enhancement*. In ICASSP, v.2, p.997-1000, May 1998.

[5] J. G. Proakis, *Digital communications*. McGraw-Hill, New York, Fourth Edition, 2000.

[6] A. Papoulis, *Probability, random variables, and stochastic processes*. McGraw-Hill, New York, Second Edition, 1984.

[7] R. Teruszkin, T. A. Consort and F. G. V. Resende Jr., *Endpoint detection analysis for an implementation of a speech recognition system applied to robot control*. In Proc. SAWCAS, November 2001.

[8] R. Teruszkin, F. G. V. Resende Jr., S. B. Villas-Boas and F. Lizarralde, *Biblioteca orientada a objeto para reconhecimento de voz e aplicação em controle de robô*. In CBA, September 2002.

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