Um Novo Sistema de Reconhecimento Automático de Locutor Independente do Texto Baseado no Parâmetro de Hurst

R. Sant'Ana, R. F. Coelho e A. Alcaim

***Resumo*— Este trabalho propõe um sistema completo de reconhecimento automático de locutor (RAL) independente do texto considerando uma nova característica estatística e um novo classificador. Neste sistema, denominado SR_{Hurst} (*Speaker Recognition with Hurst*), a característica utilizada é a *parâmetro de Hurst* (*pH*) obtida aplicando-se o estimador multi-dimensional baseado em *wavelets* ($Mdim\_wavelets$) às janelas de sinal de voz. O classificador é o fBm (*fractional brownian motion*) multi-dimensional ($Mdim\_fBm$). O modelo de cada locutor é gerados utilizando-se os vetores de valores de Hurst obtidos em cada janela além dos vetores de médias e variâncias. Os resultados demonstraram que o *parâmetro de Hurst* agrega informação nova de locutor e que, o classificador $Mdim\_fBm$ apresenta melhores desempenhos quando comparados a outros esquemas da literatura.**

***Palavras-Chave*— Reconhecimento de locutor, Parâmetro de Hurst, fBm, wavelets**

***Abstract*— This paper proposes a text independent automatic speaker recognition (ASR) based on a new statistical characteristic and a new classifier. In this system, referred to as SR_{Hurst} (Speaker Recognition with Hurst), the characteristic is the Hurst parameter obtained from a wavelet based multi-dimensional estimator ($Mdim\_wavelets$) and the classifier is the multi-dimensional-fBm (fractional brownian motion) or $Mdim\_fBm$. Each speaker model is generated from the vectors of Hurst parameter, means and variances obtained in each window. The results have shown that the Hurst parameter agregates new information from the speaker and that the $Mdim\_fBm$ classifier presents better performances as compared to other schems found in the literature.**

***Keywords*— Speaker Recognition, Hurst Parameter, fBm, wavelets**

I. INTRODUÇÃO

A principal função de um sistema de reconhecimento automático de locutor pode ser caracterizada pelas tarefas de identificação de locutor (Id) e verificação de locutor (Ve). No processo de identificação, a amostra de voz deve ser reconhecida como pertencente a um dos locutores cadastrados. No processo de verificação, um sinal de voz é apenas aceito ou não como pertencente a um locutor declarado. Em ambos os sistemas de reconhecimento, observamos as fases de treinamento e de teste. Os sistemas de RAL também podem ser classificados como dependente ou independente do texto. Neste trabalho, propomos um sistema completo de RAL, independente do texto, considerando uma nova característica e um novo classificador, denominado SR_{Hurst}[1] (*Speaker Recognition with Hurst*).

Um sistema completo de RAL geralmente engloba as fases de aquisição/pré-processamento do sinal de voz, extração de características e classificação (Id,Ve). A FIG. 1 ilustra um diagrama com estas etapas. O SR_{Hurst} foi implementado para, além de atender à crescente necessidade de baixo custo computacional [7], prover um sistema com alta taxa de acertos [9].

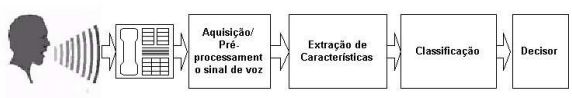


Fig. 1. Diagrama em Blocos de um Sistema de Reconhecimento de Locutor

As informações do sinal de voz de um locutor podem ser classificadas como características denominadas de alto nível e características fisiológicas. As características de alto nível, tais como estado emocional, estão associadas ao comportamento de fala do locutor e são difíceis de avaliar e quantificar. Portanto, características fisiológicas tais como *pitch* (freqüência fundamental de sons sonoros), cepestros (por ex: LPCC, LFCC) e mel-cepestros (MCC) são comumente utilizadas. Nos sistemas RAL, para a etapa de extração de características propomos uma nova característica estatística, denominada *parâmetro de Hurst* (pH). O *parâmetro de Hurst* é bastante utilizado na área de modelagem estocástica de fontes de tráfego[2]. Num estudo preliminar desta proposta, apresentado em [19], foi demonstrada a viabilidade da utilização do *parâmetro de Hurst* num sistema RAL de forma isolada e agregando-o à característica *pitch*.

Para a etapa de classificação, os sistemas atuais de RAL utilizam diversos tipos de classificadores. Alguns classificadores comumente empregados são o GMM [17] , o AR vetorial [11] e a distância Bhattacharyya [13]. Estes classificadores foram desenvolvidos e utilizados na avaliação do desempenho do classificador proposto $Mdim\_fBm$. O $Mdim\_fBm$ modela as características de voz utilizando vetores de *parâmetro de Hurst* além de vetores de médias e variâncias.

Este trabalho está organizado da seguinte forma: A seção II descreve a característica *parâmetro de Hurst* e os métodos

R. Sant'Ana e R. F. Coelho s˜ao do Instituto Militar de Engenharia Rio de Janeiro, Brasil, E-mail:ricksant@epq.ime.eb.br,coelho@ime.eb.br (**Patente Requerida junto ao INPI, No.PI0302179-3**). A. Alcaim ´e do CETUC/PUC-Rio, Rio de Janeiro, Brasil, E-mail: alcaim@cetuc.puc-rio.br

[1]Registro de Pedido de Patente no INPI n^o protocolo 06670/03.07.03

[2]O conhecimento do valor do *par^ametro Hurst* auxilia o desempenho de sistemas de comunicaç ˜oes no dimensionamento de *buffers* e enlaces.

de extração/estimação desta característica. Na seção III são apresentados a definição de movimento browniano fracionário e o classificador proposto $Mdim\_fBm$. Os resultados obtidos para identificação e verificação são apresentados na seção IV. Finalmente, a seção V apresenta as conclusões principais deste trabalho.

II. A CARACTERÍSTICA *Parâmetro de Hurst*

O parâmetro de Hurst (H), também denominado de grau de dependência temporal de um processo estocástico, pode ser definido pela taxa de decaimento da função auto-correlação $\rho(k)$ do processo quando $k \to \infty$. Considerando um sinal de voz representado pelo processo estocástico $X(t)$ com variância finita e função auto-correlação normalizada $\rho(k)$ definida por $\rho(k) = Cov[X(t), X(t+k)]/Var[X(t)], k = 0, 1, 2, \ldots$, onde a função auto-correlação toma valores na faixa $[-1, 1]$ e $\lim_{k\to\infty} \rho(k) = 0$. O comportamento assintótico da função $\rho(k)$ é dado por $\rho(k) \sim H(H-1)\, l(k)\, k^{2(H-1)}$, onde $l(k)$ é uma função de variação lenta no infinito e H $(0 < H < 1)$ é o expoente da função auto-correlação [3].

Ao contrário das características fisiológicas da voz, tais como o mel-cepestro, o *pH* é uma característica estatística. Neste trabalho não se procurou relacionar diretamente o *parâmetro de Hurst* com alguma etapa ou processo da produção de voz pelo ser humano. No entanto, o *parâmetro Hurst* representa o grau de dependência temporal e fornece a tendência dos valores das amplitudes das amostras do sinal de voz, permitindo a previsão do comportamento destas amostras no infinito. Isto o tornou uma característica bastante atraente para a utilização em sistemas RAL.

Em alguns trabalhos, processos estocásticos com presença de dependência de longo alcance ($H > \frac{1}{2}$) são denominados de processos auto-similares ou fractais. No entanto, um processo estocástico só pode ser definido como auto-similar ou fractal se além do grau de dependência, possuir a propriedade de invariância da distribuição para quaisquer incrementos do processo estocástico. Assim, nem todo processo que possui um grau de dependência é auto-similar ou fractal. O principal processo estocástico auto-similar foi proposto por Mandelbrot [2] e é denominado movimento browniano fracionário (fBm - *fractional Brownian motion*). Este processo deriva do movimento browniano puro onde $H = \frac{1}{2}$. Para processos fractais ou auto-similares e, apenas nestes casos, podemos relacionar o *parâmetro de Hurst* com a dimensão fractal (D_h) do processo através da equação $D_h = 2 - H$. Exemplos de estudos sobre a utilização da dimensão fractal para reconhecimento de padrões podem ser vistos em [16] e [20]. Em [18] um estudo sobre discriminação de sons fricativos é feito também utilizando a dimensão fractal. Em [1] um sistema de identificação de locutor baseado em cepestros, é comparado com um sistema que utiliza os cepestros agregados à dimensão fractal utilizando uma base de dados de dígitos. O sistema que incorpora a informação da dimensão fractal apresentou maior taxa de acertos. Os estudos apresentados em [18] e [1] compartilham a hipótese de que o sinal de voz é um sinal fractal.

No presente trabalho, adotamos o *parâmetro de Hurst* como característica da voz. Não presumimos, portanto, que o sinal de voz é um sinal fractal ou auto-similar. Para analisar o impacto da dependência temporal do sinal de voz na taxa de acertos de um sistema de reconhecimento de locutor, é necessário possuir métodos confiáveis para estimar o *parâmetro Hurst*. Diversos métodos para realizar a estimação do *parâmetro de Hurst* têm sido apresentados ([21]) e, amplamente utilizados, na área de pesquisa de modelagem estocástica de tráfego ([8], [14] e [15]). Entretanto, para a escolha do estimador adequado para a tarefa de reconhecimento de locutor, deve-se levar em conta a possibilidade da estimação de H de forma automática e a complexidade computacional do estimador. O estimador Abry-Veitch (AV) [22] baseado em *wavelets* foi selecionado no presente trabalho, por permitir uma estimação automática de H, pela possibilidade da implementação de uma estimação de H em tempo real e pelo baixo custo computacional deste método.

A. O estimador Abry-Veitch (AV)

O estimador AV [22] decompõe as amostras de um processo em seqüências de aproximação (passa-baixa) e detalhe (passa-alta) através da transformada discreta de *wavelet* (DWT - *Discrete Wavelets Transform*)[3].

Estas seqüências de detalhe são obtidas através de filtros digitais especialmente projetados, cujos coeficientes são determinados pela *wavelet* escolhida na estimação. Não há restrições quanto à escolha das *wavelets* a serem utilizadas no estimador AV. Para manter compatibilidade com a proposta original do estimador apresentado em [22], utilizamos os filtros *wavelets* de *Daubechies* [23] com diferentes números de coeficientes.

Logo, partindo de uma amostra original, sucessivas seqüências de aproximação e detalhe são decompostas. Estas seqüências são obtidas aplicando-se filtragem digital em cascata, ou seja, a saída de um estágio de filtragem é novamente aplicada ao estágio de filtragem e assim por diante. Em resumo, o estimador AV pode ser descrito nas seguintes etapas:

1) Decomposição em *wavelets*: A DWT é aplicada nas amostras, gerando as seqüências de detalhe $d(j,k)$ onde j representa a escala de decomposição e k o índice de cada coeficiente gerado no banco de filtro em uma dada escala. A DWT é obtida através de um algoritmo piramidal. Este algoritmo é descrito detalhadamente em [23]. O algoritmo piramidal possui baixa complexidade computacional de $O(n)$[4], onde n é o número de amostras do sinal de voz, permitindo um cálculo rápido das seqüências $d(j,k)$.
2) Estimação da variância dos coeficientes de detalhe: Para cada escala j, obtém-se uma estimativa da variância dos coeficientes $d(j,k)$ denominada μ_j. Como estes coeficientes possuem média nula, tem-se que $\mu_j = \frac{1}{n_j} \sum_k d(j,k)^2$, onde n_j indica o número de coeficientes obtidos na escala j. Pode-se mostrar que, o valor

[3] O estudo das *wavelets* representa uma extensa ´area matem´atica com diversas aplicaç˜oes em processamento de sinais. Um referˆencia mais detalhada sobre este assunto pode ser vista em [23].

[4] É importante observar que a complexidade computacional da transformada r´apida de Fourier (FFT - *fast Fourier transform*), utilizada na obtenç˜ao da caracter´ıstica mel-cepestro, ´e de $O(nlog(n))$.

de $E[\mu_j]$ segue uma lei de potência em j com expoente $\alpha = 2H - 1$, tal que

$$E[\mu_j] = 2^{\alpha j}$$

3) Estimação do parâmetro H: Para estimar o *parâmetro de Hurst*, traça-se o gráfico de $y_j = \log_2(\mu_j)$ versus j, denominado diagrama log-escala. Através de regressão linear ponderada, obtém-se a inclinação α do gráfico e, portanto, a estimativa do parâmetro $H = (1 + \alpha)/2$.

A etapa de regressão linear ponderada deve ser realizada somente nos pontos que apresentarem um alinhamento no diagrama log-escala. Esta região de alinhamento é denominada região de escalamento.

Como mencionado anteriormente, o estimador AV mostrou-se bastante interessante para a aplicação no RAL devido à sua simplicidade de implementação, baixo custo computacional e possibilidade de implementação em tempo real. No entanto, esse estimador gera um único valor de *parâmetro Hurst* para cada sinal de voz. Ou seja, este estimador seria ideal apenas para processos mono-dependentes com a presença de apenas um valor de grau de dependência temporal. Assim, propomos um novo método capaz de estimar mais de um *parâmetro de Hurst* para cada sinal de voz permitindo que, o sistema SR_{Hurst} obtenha bom desempenho em termos de taxa de acertos e custo computacional.

B. Estimador Multi-Wavelet

O estimador *multi-dimensional baseado em wavelets* ($Mdim\_wavelets$), da mesma forma que o estimador AV, decompõe as amostras de um processo em seqüências de aproximação e detalhe através da DWT. De cada seqüência de detalhe $d(j,k)$, gerado no banco de filtro em uma dada escala, estimamos o *parâmetro de Hurst* H_j, utilizando o estimador AV. Assim, pode-se obter um valor Hurst para cada seqüência de detalhe. O conjunto dos H_j forma um vetor de parâmetros de Hurst. A FIG. 2 ilustra um exemplo do estimador $Mdimstwavele$ considerando 3 escalas de decomposição.

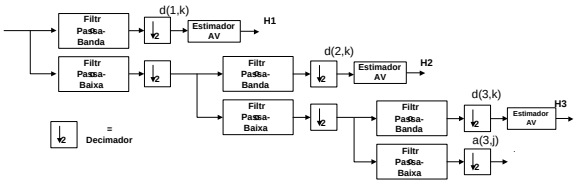


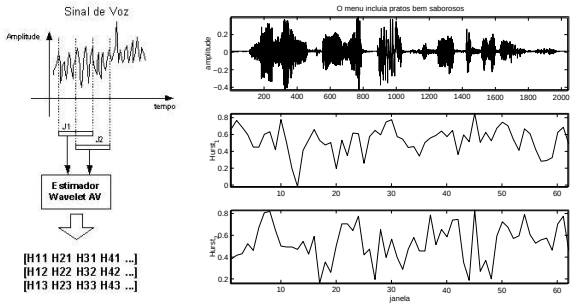
Fig. 2. Estimador $Mdimstwave$ proposto. Como podemos observar, pode-se estimar um valor de Hurst para cada seq¨u^encia de detalhe $d(j,k$.

Nesta implementação desenvolvemos as *wavelets* de *Daubechies* com 4, 6 e 12 coeficientes. O estimador $Mdim\_wavele\ ts$ pode ser descrito em duas etapas principais:

1) Decomposição do sinal de voz em *wavelets*. A DWT é aplicada nas amostras do sinal de voz obtendo-se assim as seqüências de detalhe $d(j,k)$.
2) Aplicação do estimador AV no sinal de voz (H_0) e nas j seqüências de detalhe obtidas na etapa anterior obtendo $(j+1)$ valores de Hurst ou seja, obtemos um vetor $[H_0 H_1 H_2 H_3 ... H_j]$, onde H_0 é o valor do Hurst para o sinal de voz e os H_j são os valores do *parâmetro de Hurst* para cada seqüência de detalhe $d(j,k)$.

C. Extração do Parâmetro Hurst

Para a aplicação no RAL, o *parâmetro de Hurst* foi extraído a partir da divisão do sinal de voz em N janelas (com superposição) aplicando-se em cada janela de voz o estimador proposto $Mdim\_wavelets$. Assim, de cada janela ($n, 1 < n < N$) do sinal de voz, são estimados vários valores de Hurst $[H_{1n} H_{2n} H_{3n} H_{4n} \ ]$. Este processo está ilustrado na FIG. 3 (a). Neste trabalho, aplicamos este processo em um sinal de voz utilizando janelas de 80ms com superposição de 50%, para obter uma matriz de valores H. A FIG.3 (b) mostra os valores de H extraídos para as seqüências de detalhe $d(1,k)$ e $d(2,k)$.



(a) (b)

Fig. 3. (a) Processo Hurst por Janela (b) Exemplo Hurst por Janela aplicado a um sinal de voz para para as seq¨u^encias de detalhe $d(k$ e $d(k$.

No processo de extração do *parâmetro Hurst*, diversas variáveis devem ser ajustadas tais como, número de amostras das janelas, superposição entre as janelas, região de escalamento, o número de coeficientes dos filtros *Daubechies*, entre outras. Em nossa análise, a configuração para obtenção do melhor resultado na taxa de acertos foi a seguinte:

1) Janelas de tamanho igual a 80ms;
2) Número de escalas igual a 6 ;
3) *Wavelets Daubechies* com 12 coeficientes;
4) Região de escalamento de 3 a 5.

III. O Classificador $Mdim\_fBm$

A principal função de um classificador é a de medir a distância entre um determinado modelo de locutor e o modelo utilizando as características do sinal de voz desconhecido ou em análise. Para a etapa de classificação, diversos tipos de classificadores foram propostos na literatura ([6]). Segundo [12] e [4], o classificador *Gaussian Mixtures Models* (GMM)([17]) é considerado como um dos bons referenciais para desempenho. Portanto, consideramos o GMM como um bom parâmetro de comparação adicionando também para um análise mais completa, o modelo autorregressivo (AR vetorial) ([11]) e a distância *Bhattacharyya* (dB)([13]). Estes classificadores

foram desenvolvidos e utilizados na avaliação do desempenho do classificador proposto nesta dissertação $Mdim\_fBm$. O classificador proposto neste trabalho modela as características extraídas a partir de sinais de voz com dependência temporal, podendo ou não ser fractais. Os principais processos estocásticos que exploram a dependência temporal são o fBm, *fractional gaussian motion* (fGn) e os modelos f-ARIMA (*fractional Autoregressive Moving Average*)[5].

A. Movimento Browniano fracionário

O processo estocástico fBm é definido como auto-similar de média nula e variância unitária de parâmetro contínuo t, ou seja, suas características estatísticas se mantem para qualquer escala no tempo. Um processo fBm, $X_H(t)$, possui as seguintes propriedades [2]:

1) $X_H(t)$ possui incrementos estacionários.
2) $X_H(t)$ possui variância de seus incrementos dada pela relação
$$Var[X_H(t_2) - X_H(t_1)] \propto |t_2 - t_1|^{2H} \quad (1)$$
para quaisquer instantes t_1 e t_2.
3) $X_H(t)$ é um processo gaussiano. Assim, para qualquer conjunto $t_1, t_2, \ldots, t_n$, as variáveis aleatórias $X_H(t_1), X_H(t_2), \ldots, X_H(t_n)$, possuem distribuição conjunta gaussiana.
4) $X_H(0) = 0$ e $E[X_H(t)] = 0$ para qualquer instante t.
5) $X_H(t)$ apresenta caminhos amostrais (*sample paths*) contínuos.

Portanto, o fBm, por ser um processo auto-similar de parâmetro H, suas características estatísticas se mantêm para qualquer escala no tempo. Para quaisquer τ e $r > 0$,

$$[X_H(t+\tau) - X_H(t)] \overset{d}{\approx} r^{-H}[X_H(t+r\tau) - X_H(t)] \quad (2)$$

onde r é o fator de escala do processo e $\tau = |t_2 - t_1|$. Pode-se dizer que, $X_H(t)$ é um processo gaussiano e completamente caracterizado por sua média, variância, *parâmetro de Hurst* e função auto-correlação dada por [3]

$$\rho(k) = \frac{1}{2}\sigma^2[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}] \quad (3)$$

B. O classificador $Mdim\_fBm$

Como vimos, o fBm é um processo estocástico mono-fractal, ou seja, utiliza um único valor do *parâmetro de Hurst*. Para a aplicação no RAL, propusemos o classificador fBm multi-dimensional ($Mdim\_fBm$). O procedimento para gerar o modelo $Mdim\_fBm$ de um locutor (fase de treinamento) a partir da matriz de características $MC_{nl,nc}$, sendo nl é o número de linhas da matriz de características e nc é o número de colunas, de voz pode ser descrito em etapas. A FIG. 4 mostra o diagrama dessas etapas do $Mdim\_fBm$.

1) *Pré-processamento*: a matriz de características (n_l x n_c) é dividida em j regiões[6], de tamanho múltiplos de janelas, de forma a obter um conjunto de modelos possíveis que representam o locutor;
2) *Decomposição*: para a obtenção do modelo de n dimensões decompomos cada uma das n_l linhas das regiões, em n escalas *wavelets*.
3) *Estimação de Parâmetros*: de cada dimensão, estimamos a média (μ_n), a variância (σ_n^2) e a característica *parâmetro Hurst* (H_n) das seqüências de detalhes.
4) *Geração dos Processos fBm*: com esses parâmetros geramos os n processos fBm utilizando o algoritmo *Random Midpoint Displacement* (RMD). Portanto, teremos n processos fBm para cada linha de uma determinada região.
5) *Cálculo do Histograma e Geração dos Modelos*: calculamos e armazenamos o histograma de cada processo fBm. O conjunto de todos os n x n_l histogramas será o modelo $Mdim\_fBm$ considerando todas as j regiões.

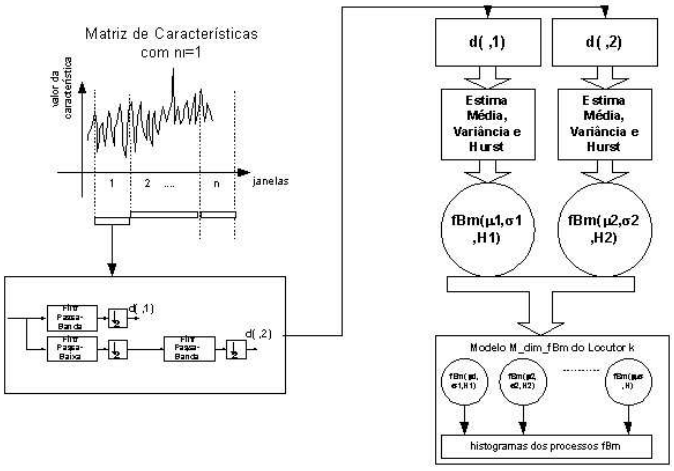


Fig. 4. Exemplo da fase de treinamento do classificador $Mdim\_fBm$ considerando a dimensão igual a 2 ($n = 2$) e o número de linhas igual a 1 ($n_l = 1$).

A FIG. 5 exemplifica a fase de teste de um sistema RAL utilizando o classificador $Mdim\_fBm$. A probabilidade de uma determinada matriz de características de teste MC_t pertencer a um dado locutor k é determinada usando o modelo $Mdim\_fBm$ armazenado deste locutor. Cada linha desta matriz de características de teste é decomposta em n escalas. As seqüências de detalhes são comparadas com os respectivos histogramas e determina-se e acumula-se a probabilidade do valor do coeficiente pertencer ao histograma referente ao locutor k. Quanto maior o resultado final da soma acumulada (S), maior a probabilidade da matriz de características de teste pertencer ao locutor k. Em seguida, pode-se utilizar o resultado numa tarefa de identificação ou verificação.

IV. Resultados

Nesta seção, são apresentados os principais resultados de identificação e verificação dos diferentes sistemas RAL, com o objetivo de comparamos os resultados do sistema SR_{Hurst}

[5]Os modelos fGn e f-ARIMA são mais apropriados para processos estocásticos com $H > \frac{1}{2}$ ou com dependência de longo alcance, enquanto que, o fBm modela processos com qualquer valor de H ($0 < H < 1$).

[6]No GMM isso é similar a definição do número de gaussianas.

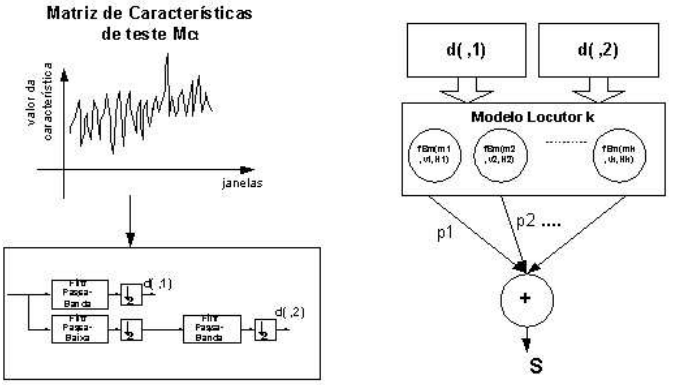


Fig. 5. Exemplo da fase de teste com o classificador $Mdim\_fBm$ considerando a dimensão igual a 2 ($n = 2$) e o número de linhas igual a 1 ($n_l = 1$).

com os demais. A base de dados (BaseIME)[7] utilizada neste trabalho é constituída de 75 locutores (masculinos e femininos) com gravações feitas em canal telefônico e celular, nos testes foram consideradas durações de 20 segundos, 10 segundos e 5 segundos. Para a fase de treinamento, foram utilizados trechos de voz com duração de 1 minuto.

A. Identificação

Os resultados do desempenho de identificação dos sistemas $Mdim\_fBm$, GMM, AR vetorial e dB, são apresentados pela taxa de acertos desses sistemas. Os resultados utilizando a característica *parâmetro de Hurst*, apresentados na seção IV-A.1, atendem à exigência de baixo custo computacional [7]. Na seção IV-A.2, os resultados destes sistemas RAL são apresentados para outras características (mel-cepestro e mel-cepestro agregado ao *pH*) onde não há necessidade de baixo custo computacional [12]. A característica mel-cepestro foi escolhida por apresentar resultados de taxa de acertos promissores [12].

1) Parâmetro Hurst: Um primeiro estudo foi feito para atender a necessidade de sistemas RAL para identificação de locutor que necessitem de baixo custo computacional. Utilizamos sistemas baseados no *parâmetro Hurst* com a forma de extração por janela (seção II-C), já que esta obteve melhores resultados, com os classificadores $Mdim\_fBm$, GMM, AR-vetorial e distância Bhattacharyya (dB). A TAB. I mostra os resultados de sistemas RAL para identificação utilizando esses classificadores para gravações feitas em telefone fixo. A TAB. II apresenta os resultados para gravações feitas em telefone celular.

Como pode-se observar, os melhores resultados foram obtidos para o $Mdim\_fBm$ tanto para gravações em telefone fixo como para gravações em telefone celular. Os classificadores $Mdim\_fBm$ e GMM apresentaram os melhores resultados quando comparados ao AR vetorial e ao dB tanto para as gravações em telefone fixo como gravações em celular.

[7]Esta base foi desenvolvida no Departamento de Engenharia Elétrica do IME dentro de um projeto FAPERJ/Secretaria de Segurança Pública do Rio de Janeiro.

TABELA I

TAXA DE ACERTOS PARA SISTEMA DE IDENTIFICAÇÃO COM GRAVAÇÕES EM TELEFONE FIXO BASEADO NO *parâmetro Hurst*

	$Mdim\_fBm$	GMM	dB	AR
20 s	95,48	95,48	82,20	91,24
10 s	94,22	94,09	72,62	83,39
5 s	89,98	89,69	56,13	64,45

TABELA II

TAXA DE ACERTOS PARA SISTEMA DE IDENTIFICAÇÃO COM GRAVAÇÕES EM CELULAR BASEADO NO *parâmetro Hurst*

	$Mdim\_fBm$	GMM	dB	AR
20 s	87,53	86,85	74,66	79,86
10 s	84,93	84,89	62,29	72,07
5 s	61,43	61,10	44,78	52,41

É importante notar também que, em todos os sistemas apresentados nesta seção, foram utilizados somente 7 valores do *parâmetro Hurst* para cada janela. Vale salientar que, na etapa de extração de características, a estimação do parâmetro H é computacionalmente mais simples (O(n)) que a extração da característica mel-cepestro (custo computacional da FFT é O(nlog(n))). Obtivemos uma menor complexidade computacional também para as fases de treinamento e testes para todos os classificadores se compararmos com os sistemas de reconhecimento convencionais baseados na característica mel-cepestro que normalmente utilizam 15 características por janela.

2) Parâmetro Hurst + Característica mel-cepestro: Neste segundo estudo, utilizamos um sistema de identificação baseado na característica mel-cepestro agregada ao *parâmetro Hurst*. Os resultados de identificação também foram analisados para a característica mel-cepestro isoladamente.

As tabelas III e IV apresentam os resultados desses sistemas RAL para identificação com gravações feitas em telefone fixo para os classificadores $Mdim\_fBm$, GMM, dB e AR vetorial utilizando a característica mel-cepestro e também mel-cepestro agregado ao *parâmetro Hurst*, respectivamente. Comparando estas tabelas, pode-se verificar que, a não ser para os testes de 10s com o classificador AR vetorial, os melhores resultados foram obtidos nos sistemas baseados na característica mel-cepestro agregada ao *parâmetro de Hurst*.

TABELA III

TAXA DE ACERTOS PARA SISTEMA DE IDENTIFICAÇÃO COM GRAVAÇÕES FEITAS EM TELEFONE FIXO BASEADO NA CARACTERÍSTICA MEL-CEPESTRO

	$MdimfBm$	GMM	dB	AR
20 s	98,54	97,95	95,48	96,81
10 s	97,99	97,99	93,38	95,13
5 s	97,59	97,46	84,17	59,47

A TAB. V apresenta os resultados de sistemas baseados na característica mel-cepestro agregada ao parâmetro de Hurst para diversos classificadores com gravações feitas em telefone celular. Comparando os resultados das tabelas IV e V pode-se verificar que os classificadores $Mdim\_fBm$ e GMM são mais robustos aos efeitos do canal celular que o AR vetorial e dB.

TABELA IV

TAXA DE ACERTOS PARA SISTEMA DE IDENTIFICAÇÃO COM GRAVAÇÕES FEITAS EM TELEFONE FIXO BASEADO NA CARACTERÍSTICA MEL-CEPESTRO E NO *parâmetro de Hurst*

	$Mdim\_fBm$	GMM	dB	AR
20 s	98,57	98,40	95,75	97,21
10 s	98,62	98,51	94,61	92,54
5 s	97,91	97,66	87,81	72,83

TABELA V

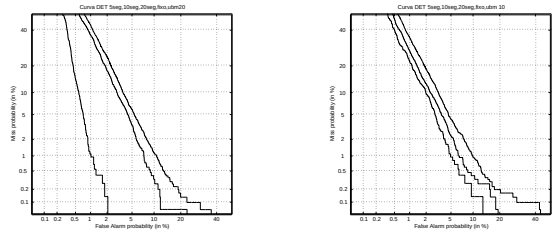
TAXA DE ACERTOS PARA SISTEMA DE IDENTIFICAÇÃO BASEADO NA CARACTERÍSTICA MEL-CEPESTRO E NO *parâmetro de Hurst* COM GRAVAÇÕES FEITAS EM TELEFONE CELULAR

	$Mdim\_fBm$	GMM	dB	AR
20 s	98,19	98,14	88,22	85,75
10 s	92,56	92,03	84,80	74,68
5 s	89,96	89,96	76,09	52,28

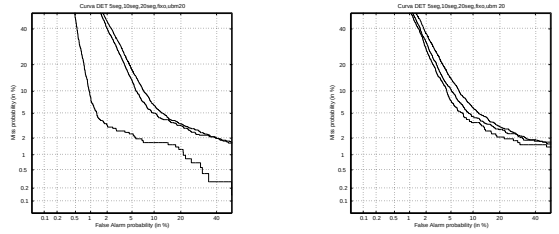
B. Verificação

A apresentação dos resultados dos sistemas RAL para verificação está ilustrada através de curvas DET (*Detection Error Tradeoff*) [10]. Utilizamos como modelos de *background* o UBM (*Universal Background Model*) [5] com 20 locutores não pertencentes aos 75 locutores originais para o GMM e o $Mdim\_fBm$, utilizando gravações em telefone fixo e celular.

1) Parâmetro Hurst: Nesta seção o SR_{Hurst} foi comparado com um sistema também baseado no *parâmetro Hurst* mas utilizando o classificador GMM. A FIG. 6 mostra as curvas DET para um sistema de verificação baseado no *parâmetro de Hurst* utilizando respectivamente o classificador $Mdim\_fBm$ (SR_{Hurst}) e GMM para testes de 20s, 10s e 5s de duração.



(a) $Mdim\_fBm$ fixo (b) GMM fixo



(a) $Mdim\_fBm$ celular (b) GMM celular

Fig. 6. Curva DET para sistema baseado no *parâmetro Hurst* utilizando o classificador $Mdim\_fBm$ e GMM. As curvas, da mais interior para a mais exterior, são respectivamente para testes com 20s, 10s e 5s.

A TAB. VI apresenta a taxa de acertos selecionando-se o ponto de operação da curva DET onde a probabilidade de falso alarme (fa) é igual a probabilidade de falsa rejeição (fr). Novamente, podemos observar um melhor desempenho do classificador $Mdim\_fBm$ em relação ao GMM comprovando a modelagem mais precisa obtida pelo $Mdim\_fBm$.

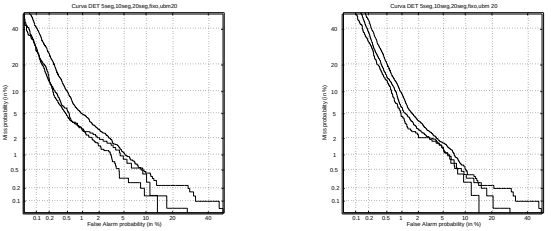
TABELA VI

TAXA DE ACERTOS PARA SISTEMA DE VERIFICAÇÃO BASEADO NO *parâmetro de Hurst* COM GRAVAÇÕES FEITAS EM TELEFONE FIXO

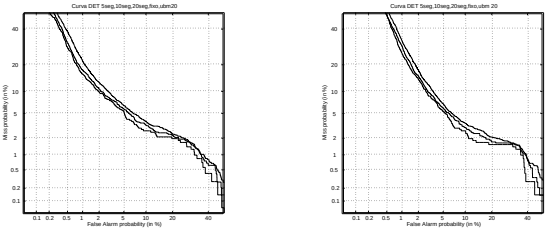
	$MdimfBm$ fixo	GMM fixo	$MdimfBm$ cel	GMM cel
20 s	98,93	96,67	97,12	96,69
10 s	95,66	95,66	92,98	93,01
5 s	94,76	94,37	92,67	92,11

2) Parâmetro Hurst + Característica mel-cepestro: Neste estudo, foram implementados sistemas de verificação que obtivessem melhores taxa de acertos independente do custo computacional. Esses sistemas implementados foram baseados na característica mel-cepestro agregada ao *parâmetro de Hurst* utilizando o $Mdim\_fBm$ e o GMM.

Para comparação, usamos sistemas baseados somente nas características mel-cepestro. A FIG. 7 ilustra as curvas DET para estes sistema baseado nas características mel-cepestro agregada ao *parâmetro Hurst*. A FIG. 8 mostra as mesmas curvas DET para o sistema baseado apenas na característica mel-cepestro.



(a) $Mdim\_fBm$ fixo (b) GMM fixo



(a) $Mdim\_fBm$ celular (b) GMM celular

Fig. 7. Curva DET para sistema baseados na característica mel-cepestro utilizando o classificador $Mdim\_fBm$ e GMM. As curvas, da mais interior para a mais exterior, são respectivamente para testes com 20 s, 10 s e 5 s.

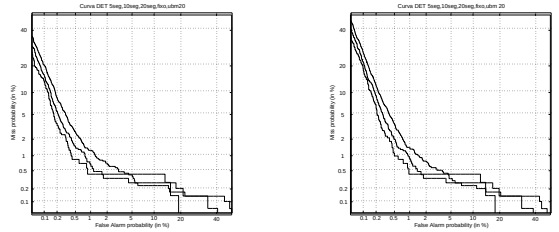
O resumo dos resultados referentes as curvas DET das figuras 7 e 8 está apresentado nas tabelas VII e VIII selecionando o ponto de operação da curva onde $fa = fr$.

TABELA VII

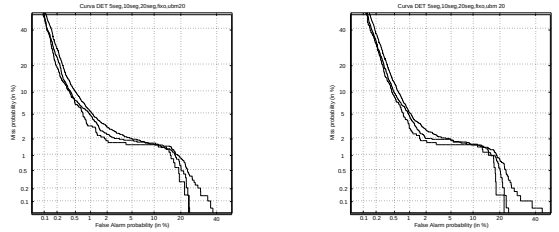
TAXA DE ACERTOS PARA SISTEMA DE VERIFICAÇÃO BASEADO NA CARACTERÍSTICA MEL-CEPESTRO COM GRAVAÇÕES FEITAS EM TELEFONE FIXO E CELULAR

	$MdimfBm$ fixo	GMM fixo	$MdimfBm$ cel	GMM cel
20 s	98,08	98,00	95,06	94,93
10 s	98,31	97,59	94,67	94,63
5 s	97,62	97,27	94,34	94,32

Podemos observar uma melhora bastante acentuada apresentada pelos sistemas baseados na característica mel-cepestro



(a) $Mdim\_fBm$ fixo (b) GMM celular



(a) $Mdim\_fBm$ fixo (b) GMM celular

Fig. 8. Curva DET para sistema baseado na caracter´ıstica mel-cepestro agregada ao *par^ametro de Hurst* utilizando o classificador $Mdim\_fBm$ e GMM. As curvas, da mais interior para a mais exterior, s˜ao respectivamente para testes com 20s, 10s e 5s.

TABELA VIII

TAXA DE ACERTOS PARA SISTEMA DE VERIFICAÇÃO BASEADO NA CARACTERÍSTICA MEL-CEPESTRO AGREGADA AO *par^ametro Hurst* COM GRAVAÇÕES FEITAS EM TELEFONE FIXO E CELULAR

	$Mdim\_fBm$ fixo	GMM fixo	$Mdim\_fBm$ cel	GMM cel
20 s	99,33	99,20	98,21	98,09
10 s	99,15	99,09	97,96	98,01
5 s	98,87	98,81	97,49	97,42

agregada ao *parâmetro Hurst*. É importante também notar a melhora média de 1% em gravações de telefone fixo e de 3% em gravações de telefone celular nas taxas de acerto para a característica mel-cepestro agregada ao *parâmetro de Hurst*. Além disso, o classificador $Mdim\_fBm$ novamente obteve melhores taxa de acertos que o classificador GMM.

V. CONCLUSÕES

Neste artigo foi apresentado um novo sistema de reconhecimento de locutor (RAL) denominado SR_{Hurst}. Este sistema é baseado numa nova característica, o *parâmetro Hurst*, e num novo classificador, o fBm multi-dimensional ($Mdim\_fBm$). Para ambientes que necessitam baixo custo computacional, os sistemas que exploram apenas o *parâmetro de Hurst* mostraram-se bastante atrativos para as fases de extração de características e de classificação. Já para aplicações que não necessitam de baixo custo computacional, o desempenho de sistemas baseados no *parâmetro Hurst*, agregado à característica mel-cepestro, mostrou-se superior a sistemas que exploram apenas as características mel-cepestro ou o *parâmetro de Hurst* individualmente.

O desempenho, em termos de taxa de acertos, do classificador $Mdim\_fBm$ foi comparado com os classificadores GMM, AR vetorial e dB utilizando o *parâmetro de Hurst*, o mel-cepestro e o *parâmetro Hurst* agregado ao mel-cepestro. O $Mdim\_fBm$ apresentou melhores taxa de acertos para todos os cenários examinados. Assim, demonstramos que o $Mdim\_fBm$ faz uma modelagem mais precisa do que o GMM e que, o *parâmetro de Hurst* agrega informação importante aos sistemas baseados na característica mel-cepestro. O SR_{Hurst} foi também testado para o reconhecimento automático de comandos de voz com algumas modificações na forma de apresentação das características de voz ao classificador apresentando resultados próximos a 100% de taxa de acertos para base de comandos utilizada. O SR_{Hurst} é, portanto, uma proposta bastante interessante para as diversas tarefas e aplicações na área de pesquisa em reconhecimento automático de locutor.

REFERÊNCIAS

[1] A. PETRY, D. B. Fractal dimension applied to speaker identification. *Proceedings ICASSP* (2001).

[2] BARNSLEY, M., AND ET AL. *The Science of Fractal Images*. Springer-Verlag New York Inc., USA, 1988.

[3] BERAN, J. *Statistics for Long-Memory Processes.* Chapman & Hall, 1994.

[4] CAMPBELL, J. P. J. Speaker recognition: A tutorial. *Proceedings of the IEEE 85*, 9 (September 1997), 1437–1461.

[5] D.A. REYNOLDS, R.C. ROSE, E. H. Integrated models of signal and background with application to speaker identification in noise. *IEEE Transactions on Speech, and Audio Processing 2*, 2 (April 1994), 245–267.

[6] JAYANT, M. N. Speaker verification: A tutorial. *IEEE Communication Magazine* (January 1990), 42–48.

[7] KUMAGAI, J. Talk to the machine. *IEEE Spectrum 39*, 9 (September 2002), 60–64.

[8] LELAND, W., WILLINGER, W., TAQQU, M., AND WILSON, D. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking 2*, 1 (February 1994), 1–15.

[9] MARTIN, A. F., AND PRZYBOCKI, M. A. The nist speaker recognition evaluations: 1996-2001. *http://www.nist.gov/speech/publications*.

[10] MARTIN, A. F. E. A. The det curve in assessment of detection task performance. *Proceedings of EuroSpeech 97* (1997), 1895–1898.

[11] MONTACIÉ, C., AND LE FLOCH, J. L. Ar-vector models for free-text speaker recognition. *Proceedings of the ICSLP* (1992), 611–614.

[12] NIST. The nist year 2001 speaker recognition evaluation plan. *http://www.nist.gov/speech/publications*.

[13] PETRY, ZANUZ, A. B. Bhattacharyya distance applied to speaker identification. *Proceedings ICASSP* (2000).

[14] PONTES, R., AND COELHO, R. The scaling characteristics of the video traffic and its impact on acceptance regions. *Proceedings of the 17th International Teletraffic Congress 4* (December 2001), 197–210.

[15] PONTES, R., AND COELHO, R. Admission control for video connections traffic streams with scaling characteristics. *Revista da Sociedade Brasileira de Telecomunicaç˜oes (RevSBrT)* (Dezembro 2002), 87–96.

[16] R. ESTELLER, G. VACHTSEVANOS, T. H. Fractal dimensions characterizes seizure onset in epileptic patients. *IEEE Proceedings, ICASSP99 4* (1999), 2343–2346.

[17] REYNOLDS, D. A. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication* (1995).

[18] S. FERNÁNDEZ, S. FEIJÓO, R. B. Fractal characterization of spanish fricatives. *Proceedings of the ICPhS* (1999), 2145–2148.

[19] SILVA, D., AND COELHO, R. Aplicabilidade da an´alise fractal ao reconhecimento de locutor. *XIX Simp´osio Brasileiro de Telecomunicaç˜oes* (setembro 2001).

[20] T. MORIMOTO, T. TAKEUCHI, V. H. Pattern recognition of fruit shape based on the concept of chaos and neural networks. *Computer and Electronics in Agriculture* (2000), 171–186.

[21] TAQQU, M., TEVEROVSKY, V., AND WILLINGER, W. Estimators for long–range dependence: An empirical study. *Fractals 3*, 4 (1995), 785–798.

[22] VEITH, D., AND ABRY, P. A wavelet-based joint estimator of the parameters of long-range dependence. *IEEE Trans. on Information Theory 45*, 3 (1998), 878–897.

[23] VETTERLI, M., AND KOVACEVIC, J. *Wavelets and Subband Coding.* Prentice Hall, 1985.