Verificação Automática de Locutor com Captação de Voz em Ambientes Ruidosos

L. Lima e R. Coelho

***Resumo*— Este artigo apresenta a avaliação do desempenho de um sistema de verificação automática de locutor com locuções submetidas a diferentes ruídos acústicos ou sonoros. O sistema investigado é baseado na característica MEL-cepestro e no classificador GMM (*Gaussian Mixture Models*). Nos experimentos foram considerados os ruídos *F16 Cockpit* e *Factory Floor 1* da base NOISEX-92 com SNR (*Signal Noise Rate*) de 10dB e 4dB. Os resultados reforçam a importância da caracterização de ruídos, principalmente, para emprego em sistemas de acesso baseado em biometria. Uma importante contribuição do trabalho é a demonstração de que o grau de impulsividade do ruído afeta de forma distinta a taxa de erro dos sistemas de verificação de locutor.**

***Palavras-Chave*— verificação de locutor, ruídos acústicos, impulsividade.**

***Abstract*— This article presents the performance evaluation of an automatic speaker verification system with speeches collected under different acoustic noises. The system is based on the MEL-cepestro feature and the GMM (*Gaussian Mixture Models*) classifier. The experiments considered the *F16 Cockpit* and *Factory Floor 1* noises from the NOISEX-92 database and 10dB and 4dB SNR (*Signal Noise Rate*). The results reinforced the importance of the noise characterization mainly for use in access systems based on biometrics. An important contribution of this work is the demonstration that the noise impulsiveness degree affects the accuracy of the verification systems.**

***Keywords*— speaker verification, noise, acoustic, impulsiveness.**

I. INTRODUÇÃO

A expansão do uso das comunicações eletrônicas torna o acesso seguro um requisito primordial para sistemas públicos e privados. Neste cenário, os sistemas de autenticação baseados em biometria, tornam-se bastante promissores. As soluções biométricas [1] baseiam-se no reconhecimento de padrões considerando características humanas tais como: a impressão digital, a íris, a face e a voz. Portanto, é fundamental a pesquisa de acesso seguro a sistemas de comunicação baseado em autenticação biométrica.

O sinal acústico resultante do sistema de produção da fala é portador de informações como o pensamento e a identidade do locutor, além de condições físicas e emocionais. Um importante desafio para sistemas de acesso biométrico é a captação de voz em ambientes ruidosos acústicos.

Este trabalho avalia o desempenho de um sistema de verificação automática de locutor em ambientes ruidosos. Os sinais de voz foram submetidos, através do método de fusão, aos ruídos sonoros FC (*F16 Cockpit*) e FF (*Factory Floor 1*) da base NOISEX-92 [2]. O sistema estudado baseia-se na característica MEL-cepestro [3] e no classificador GMM (*Gaussian Mixture Models*) [4]. Estes foram escolhidos por apresentarem as melhores taxas de reconhecimento de locutor entre os sistemas propostos na literatura.

Leonardo Lima é bolsista do programa PIBITI/CNPq. Os autores são do Instituto Militar de Engenharia (IME), Laboratório de Comunicações e Sistemas Ópticos (LaRSO). E-mails: {leolima,coelho}@ime.eb.br.

II. VERIFICAÇÃO AUTOMÁTICA DE LOCUTOR

Um sistema completo de RAL (reconhecimento automático de locutor) é geralmente realizado em duas fases: treinamento e teste. Cada fase engloba, basicamente, três etapas: aquisição e pré-processamento do sinal de voz, extração de características ou atributos da voz e classificação do locutor. Na primeira etapa, é feita a conversão analógico-digital e a preparação para o RAL. O sinal de voz é dividido em janelas de 25 ms. Em seguida, esse sinal passa por filtros e estimadores que irão efetuar a extração de características, representadas por uma matriz de coeficientes. Na etapa de extração, deve-se escolher características que tenham alto poder discriminatório, grande variabilidade entre locutores e pequena variabilidade para um mesmo locutor.

Neste trabalho a verificação é avaliada como tarefa da classificação do locutor. Na verificação, o locutor se declara e o sistema terá que autenticar a identidade clamada. No processo, podem ocorrer dois erros: a falsa aceitação e a falsa rejeição. No primeiro caso, o locutor de teste é aceito, apesar de não possuir a identidade declarada. Na falsa rejeição, o locutor é rejeitado pelo sistema, mesmo sendo o verdadeiro dono da voz. A base de voz elaborada neste trabalho é constituída de 10 locuções. O limiar da verificação foi modelado pelo *Universal Background Model* (UBM) [5] e é composto de 20 locuções não pertencentes à base de voz. O resultado da verificação de locutor foi obtido através do limiar baseado no UBM e calculando as probabilidades de falsa aceitação e falsa rejeição.

A. A característica MEL-cepestro

A característica MEL-cepestro [3] é um atributo fisiológico que baseia-se na percepção não-linear do ouvido humano. Os pontos do espectro auditivo são calculados aplicando-se a transformada discreta de Fourier no sinal de voz através de: $A_j(n) = \sum_{f=f_{jl}}^{f_{jh}} W_j(f)S(n,f)$ onde $W_j(f)$ são os pesos atribuídos ao espectro na faixa de banda crítica (f_{jl}, f_{jh}), e $S(n,f)$ é a densidade espectral de potência do sinal de voz no tempo *n* e frequência *f*. Os coeficientes MEL (MEL_i) são extraídos de um banco de filtros triangular de banda crítica e podem ser definidos pela expressão: $MEL_i = \sum_{k=1}^{N} X_k \cos[i(k-1/2)\pi/N],\ i = 1,2,\ldots,M$

onde M é o número de coeficientes mel-cepestrais, e X_k, $k = 1, 2, \ldots, N$, representa a energia logarítmica do k-ésimo filtro e *N* é o número de filtros do banco de filtros. Por fim, a matriz de coeficientes MEL-cepestro resultante, é entregue ao classificador GMM para realização da verificação.

B. O classificador GMM

O modelo GMM (λ) [4] é definido por uma soma ponderada de M funções densidades de probabilidade (fdp) Gaussianas: $p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x})$, em que $\vec{x}$ é um vetor aleatório de dimensão L, $b_i(\vec{x})$ são as fdps e p_i é a ponderação das misturas, onde $i = 1, \ldots, M$. Cada função Gaussiana de dimensão L é da forma: $b_i(\vec{x}) = \frac{e^{-\frac{1}{2}(\vec{x}-\vec{\mu_i})' K_i^{-1} (\vec{x}-\vec{\mu_i})}}{(2\pi)^{\frac{N}{2}} \sqrt{|K_i|}}$, com vetor média $\vec{\mu_i}$ e matriz covariância K_i, onde $|.|$ indica determinante. As ponderações das misturas devem satisfazer à condição $\sum_{i=1}^{M} p_i = 1$. Os parâmetros do modelo do locutor são dados por: $\lambda = \{p_i, \vec{\mu_i}, K_i\}$, $i = 1, \ldots, M$. Para uma sequência de *T* vetores de treinamento $X = \{\vec{x_1}, \vec{x_2}, ..., \vec{x_T}\}$, o valor da log-verossimilhança normalizada é dada por: $\log p(\vec{X}|\lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(\vec{x_t}|\lambda)$.

III. Ruídos Acústicos ou Sonoros

Os ruídos sonoros são, geralmente, caracterizados na literatura como brancos e Gaussianos. No entanto, estudos atuais [6] demonstram a presença de impulsividade em diversas áreas da ciência, inclusive em ruídos.

Os ruídos caracterizados como impulsivos tem geralmente representação atribuida as funções não-Gaussianas estáveis [6] e possuem um grau de impulsividade definido por ($0 \leq \alpha \leq 2$). Sua principal particularidade, quando comparada com a distribuição Gaussiana ($\alpha = 2$), é a de que a distribuição possui cauda ($P[X > x]$) com decaimento lento ou não-Exponencial. Neste trabalho, o método proposto por McCulloch [7] foi adotado na estimação do grau de impulsividade dos ruídos.

IV. Experimentos e Resultados

A base de voz desenvolvida para este trabalho é constituída de 10 locutores (6 masculinos e 4 femininos), alunos do IME (Instituto Militar de Engenharia), que leram um texto duas vezes, sendo uma para treinamento e outra para teste. Um microfone dinâmico supercardióide *Shure BETA 58A* e um *headphone RH-2005 Roland* foram utilizados durante as gravações, ambos conectados a um amplificador *Fast Track M-Audio*. O UBM usado como modelo do locutor intruso, foi construído com trechos de voz de 20 locutores não pertencentes à base usada para experimentos de teste. Nos experimentos, foram realizados 2066 testes com diferentes tempos de duração (25s, 10s, 5s, 2s e 1s). A duração média das locuções de treinamento é de 21s e das locuções de teste é de 23s. Neste trabalho, foram utilizadas matrizes de atributos com 15 coeficientes MEL e GMM com 32 Gaussianas.

A base de ruídos utilizada foi a NOISEX-92 [2], que apresenta 15 diferentes ruídos. Para este estudo, foram selecionados os ruídos: *F-16 Cockpit* e *Factory Floor 1* e valores de SNR 10dB e 4dB. Os graus de impulsividade desses ruídos, estimados pelo método McCulloch [7], foram: $\alpha = 2.0$ para FC e $\alpha = 1.47$ para FF. Observa-se que o ruído FF é muito impulsivo e que o FC aproxima-se da impulsividade de uma distribuição Gaussiana.

A Tab. 1 apresenta os resultados das taxas de erro obtidos para verificação de locutor considerando locução limpa e com a fusão de ruídos.

TABELA I
TAXA DE ERRO DA VERIFICAÇÃO

Taxa de Erro						
	SNR	Duração do Teste				
		25s	10s	5s	2s	1s
Locução Limpa		2,50%	3,75%	6,38%	9,57%	11,80%
F16 Cockpit Noise	10	40,00%	46,67%	48,00%	49,55%	50,24%
$\alpha = 2.0$	4	45,00%	50,00%	51,00%	53,64%	55,24%
Factory Noise	10	50,00%	53,57%	55,21%	58,33%	59,52%
$\alpha = 1.47$	4	55,00%	55,36%	58,33%	58,80%	60,05%

Os resultados mostram que os ruídos acarretaram em diferentes impactos na taxa de erros de verificação. Para um mesmo valor de SNR houve um aumento na taxa de erro de mais de 5% em todos os casos chegando a 10% para os trechos de 25s. Nota-se que a relação sinal ruído também teve grande influência nos resultados de taxa de erro. Observe ainda, que a taxa de erro cresceu com o aumento do grau de impulsividade, ou seja, de $\alpha = 2.0$ (FC, pouco impulsivo) para $\alpha = 1.47$ (FF, muito impulsivo). Isso, portanto, ressalta a importância da caracterização do ruído para os estudos em verificação de locutor.

V. Conclusões

Neste trabalho apresentou-se a avaliação do sistema de verificação de locutor com locuções submetidas a diferentes ruídos. Nos experimentos foram consideradas locuções submetidas a diferentes ruídos acústicos e valores de SNR. Os resultados mostraram a influência do ambiente de captação de voz na taxa de erros da verificação de locutor. Também foi demostrado que o grau de impulsividade dos ruídos é um atributo promissor na caracterização de ruídos sonoros ou acústicos.

Referências

[1] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[2] H. Steeneken, *NOISEX-92 - Noise Database*. 1992.

[3] S. Imai, "Cepstral analisys synthesis on the mel frequency scale," *IEEE International Conference on ICASSP' 83*, vol. 8, pp. 93–96, April 1983.

[4] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech, and Audio Processing*, vol. 3, pp. 72–83, January 1995.

[5] E. Hosftetter, R. Rose, and D. Reynolds, "Integrated models of signals and background with application to speaker identification in noise," *IEEE Transactions on Speech, and Audio Processing*, vol. 2, pp. 245–267, April 1994.

[6] C. Nikias and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications*. New York: Wiley, 1995.

[7] H. McCulloch, "Simple consistent estimators of stable distribution parameters," *Communications in Statistics*, vol. 15, no. 4, pp. 1109–1136, 1998.