# Estudo da Interferência de Ruídos Impulsivos na Identificação Automática de Locutor

P. Araújo e R. Coelho

***Resumo*— Este artigo apresenta um estudo sobre a influência de ruídos com características impulsivas nos sistemas de reconhecimento automático de locutor (RAL), na tarefa de identificação. O sistema de identificação avaliado é baseado na característica MEL-cepestro e no classificador GMM (*Gaussian Mixture Models*).**

***Palavras-Chave*— reconhecimento de locutor, ruídos, impulsividade**

***Abstract*— This paper presents a study about the influence of impulsive noise on automatic speaker recognition (ASkR) systems, on the identification task. The evaluated identification system is based on MEL-cepstral feature and GMM (Gaussian Mixture Models) classifier.**

***Keywords*— speaker recognition, noise, impulsiveness**

## I. Introdução

Um sinal de voz é portador de informações resultantes de um pensamento do locutor e de algumas de suas características, tais como: idioma, sexo e condições emocionais. O reconhecimento automático de locutor (RAL) consiste em um sistema de análise do sinal acústico da fala que permite a identificação ou verificação do locutor. A utilização de sistemas de RAL, tem ampla aplicação na área de segurança, seja no controle de acesso ou na proteção da integridade das informações.

Um sistema de RAL é, geralmente, composto de três fases: pré-processamento da voz, extração de características e classificação [1] [2]. Na primeira etapa, é feita a aquisição do sinal da voz, a conversão analógico-digital e o janelamento (divisão em quadros) do sinal. Na segunda fase, é realizada a extração dos atributos ou características da locução. A etapa de classificação é responsável por produzir uma representação do locutor, ou seja, o modelo da voz [1]. No sistema RAL, destacam-se ainda duas etapas: treinamento e testes [2]. Na primeira, é obtido o modelo de cada locutor. Na etapa de teste, realiza-se uma tarefa de identificação ou de verificação. Na identificação, a amostra de voz deve ser reconhecida como pertencente a um dos locutores cadastrados. Na verificação, o sinal de voz é apenas aceito ou não como pertencente a um locutor declarado.

Os ruídos considerados nos ambientes de captação das locuções são, geralmente, caracterizados como Gaussianos ou brancos. Diversos estudos, no entanto, demonstram a presença de impulsividade em medidas reais de diversas áreas da ciência, inclusive na caracterização dos ruídos sonoros [3]. Amostras com características impulsivas possuem distribuições com caudas que decaem mais lentamente que uma distribuição Gaussiana.

Este trabalho investiga o desempenho de um sistema de identificação de locutor, independente do texto, com sinais de voz submetidos a ruídos sonoros impulsivos. Sinais e ruídos com natureza impulsiva não podem ser caracterizados por distribuições Gaussianas [4]. Assim, espera-se que essa impulsividade tenha um grande impacto no reconhecimento de locutor. O sistema estudado baseia-se na característica MEL-cepestro [5] e no classificador GMM (*Gaussian Mixture Models*) [6], que apresenta as melhores taxas de reconhecimento entre os propostos na literatura.

Priscilla de Araújo Farias é bolsista do CNPq/PIBITI. E-mail: priscilla@ime.eb.br. Rosângela Coelho. E-mail: coelho@ime.eb.br. Este trabalho é desenvolvido no Laboratório de Comunicações e Sistemas Ópticos (LaRSO), do Instituto Militar de Engenharia (IME)

## II. O sistema RAL avaliado

Os critérios para que uma característica seja interessante para os sistemas de RAL são: alto poder discriminatório, grande variabilidade entre locutores e pequena variabilidade para um mesmo locutor. Esses atributos devem considerar as limitações computacionais e de armazenamento além de se adequarem ao método de classificação em avaliação.

### A. A característica MEL

A característica MEL-cepestro [5] 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 $(S(n,f))$ e atribuindo-se pesos às freqüências através de: $A_j(n) = \sum f = f_{jl}^{f_{jh}} W_j(f) S(n,f)$ em que $W_j$ são os pesos atribuídos ao espectro na faixa da banda crítica $(f_{jl}, f_{jh})$, e $S(n,f)$ é a densidade espectral de potência, obtida da análise do sinal de voz na janela $n$ e freqüência $f$. Os pesos são extraídos do banco de filtros que modelam o ouvido humano.

O efeito do canal, no qual a voz foi captada, é dado pela convolução da resposta ao impulso do canal e o sinal de voz. Aplicando o logarítimo e supondo a janela grande suficiente para considerar que a resposta em freqüência do canal não muda, tem-se: $\log A'_j(n) = \log A_j(n) + \log H(f)$, em que $H(f)$ é a transformada da resposta ao impulso. A constante presente nessa expressão representa o canal, e o método CMS (*Cepstrum Mean Subtraction*) foi proposto para sua extração.

### B. O classificador GMM

O GMM (modelo $\lambda$) [6] 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 $L$ 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 $\sum_{i=1}^{M} p_i = 1$. Assim, os parâmetros do modelo do locutor são dados por: $\lambda = \{p_i, \vec{\mu_i}, K_i\}, \qquad i = 1, \ldots, M$. Para uma seqüência de T vetores de treinamento $X = \{\vec{x_1}, \vec{x_2}, ..., \vec{x_T}\}$ o valor da log-verossimilhança normalizada é dada por: $\log p(\vec{X}|\lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(\vec{x_t}|\lambda)$. Na identificação é escolhido o modelo de locutor que obteve o máximo valor nessa expressão.

### C. Impulsividade

A impulsividade ($0 \leq \alpha \leq 2$) é uma característica atribuida as funções não-Gaussianas mas estáveis [4]. A principal particularidade das distribuições estáveis, quando comparada a distribuição Gaussiana ($\alpha = 2$), é a de que sua fdp possue cauda ($P[X > x]$) com decaimento lento ou não-exponencial. O método de estimação do grau de impulsividade proposto por McCulloch [7] apresenta, segundo a literatura [8], o melhor desempenho até o momento.

## III. Identificação de locutor: Resultados

A base de voz adotada neste trabalho é constituída de 70 locutores, sendo um sub-conjunto da (BaseIME)[1]. As gravações foram realizadas através de um canal telefônico fixo. Dois textos diferentes foram lidos por cada locutor. Para fase de treinamento, foram utilizados trechos de voz de 1 minuto. E, para a fase de teste, as durações foram de 5 e 10 segundos.

A base de ruídos utilizada foi a NOISEX-92 [9], que apresenta 15 diferentes ruídos. Para este estudo, foram selecionados os ruídos: *Jet Cockpit Noise 2* (JCN) e *Speech Babble* (SB). Os graus de impulsividade desses ruídos, estimados pelo método McCulloch, foram: $\alpha = 1.76$ para JCN e $\alpha = 2.0$ para SB. Observa-se que o ruído JCN é muito impulsivo e que o SB aproxima-se da impulsividade de uma distribuição Gaussiana.

Os ruídos foram adicionados a cada uma das locuções utilizando o aplicativo *Audacity*. Inicialmente, realizou-se a fase de treinamento com cada um dos locutores. Em seguida, trechos de 10s e 5s foram submetidos à extração de características. Para os testes, cada matriz característica foi comparada com os 70 modelos gerados retornando o de maior probabilidade.

A Tabela I apresenta o desempenho do sistema de identificação com locuções sem ruídos, com o ruído SB ou o JCN adicionado as locuções.

Ao todo, foram realizados 2190 testes para trechos de 5s e 1076 para 10s. Como visto na Tabela I, ocorre uma acentuada redução no desempenho do sistema RAL. Essa queda obedeceu uma relação com seus graus de impulsividade. Quanto mais impulsivo, maior foi a interferência do ruído no processo de reconhecimento. Para o SB obteve-se uma redução de 81,34% e 81,7% para 10s e 5s, respectivamente. Para o JCN a redução foi de 82,75% para 10s e 83,71% para 5s.

[1]Esta base foi desenvolvida no Departamento Engenharia Elétrica do IME com recursos FAPERJ e Secretaria de Segurança Pública do Rio de Janeiro e está disponível em http://larso.ime.eb.br

TABELA I

DESEMPENHO DO RAL EM AMBIENTES COM RUIDOS IMPULSIVOS

| Duração dos testes | Sem ruído | *Speech Babble* | *Jet cockpit* |
|---|---|---|---|
| 10s | 98,33% | 16,99% | 15,58% |
| 5s | 98,36% | 16,66% | 14,65% |

Em seguida, utilizando o Método de *Matching* e a subtração espectral (CMS) [10], foi extraído o ruído SB de cada uma das 70 locuções. A Tabela II apresenta o desempenho do sistema, antes e após a extração do ruído.

TABELA II

DESEMPENHO DO SISTEMA APÓS A EXTRAÇÃO DO RUIDO

| Duração dos testes | Antes da extração | Depois da extração |
|---|---|---|
| 10s | 16,99% | 20,13% |
| 5s | 16,66% | 19,85% |

Observa-se que o método utilizado para a extração de ruídos, creditou uma melhora no desempenho do sistema de 3,14% para os testes de 10s e, 3,19%, para os de 5s.

## IV. Conclusões

Neste trabalho foram apresentados os resultados de testes realizados no sistema RAL em ambientes ruidosos. Verificou-se uma redução considerável no desempenho desse sistema. Esse fato ocorre devido à soma dos espectros da voz com a do ruído que altera o sinal de voz.

Após a extração de ruído, nota-se uma pequena melhora no desempenho. No entanto, parte da locução também tem seu espectro subtraído, ocasionando perda de informações sobre o locutor. Isso impede uma melhora maior no desempenho.

## Referências


[1] D. O'SHAUGHNESSY, *Speech Communication.* 2000.

[2] J. P. J. CAMPBELL, "Speaker recognition: A tutorial," *Proceedings of IEEE*, vol. 85, pp. 1437–1462, September 1997.

[3] K. C. GEORGIOU G. P., TSAKALIDES P., "Alpha-stables modeling of noise and robust time-delay estimation en the presence of impulsive noise," *IEEE Transactions on Multimedia*, vol. 3, pp. 291–301, September 1999.

[4] M. SHAO and C. L. NIKIAS, "Signal processing with fractional lower order moments: Stable processes and their applications," vol. 81, pp. 986–1010, July 1993.

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

[6] D. A. REYNOLDS and R. C. ROSE, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, January 1995.

[7] J. H. McCULLOCH, "Simple consistent estimators of stable distribuition parameters," *Communication Statistic*, vol. 15, no. 4, pp. 1109–1136, 1986.

[8] S. BATES and S. McLAUGHLIN, "The estimation of stable distribution parameters from teletraffic data," *IEEE Transactions on Signal Processing*, vol. 48, pp. 865–870, March 200.

[9] H. STEENEKEN, *NOISEX-92 - Noise Database.* 1992.

[10] S. F. BOLL, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transaction on Acoustic, Speech and Signal Processing*, vol. ASSP-27, pp. 113–120, April 1979.