Ajuste de Misturas Gaussianas utilizando Algoritmo de Maximização da Esperança e Critério de Comprimento de Descrição Mínimo para Modelagem de Tráfego VoIP

Cheila Mendes de Oliveira
Escola de Engenharia Elétrica e de Computação - UFG
Av. Universitária S.N. – Setor Leste Universitário – Goiânia – Goiás
Cheila@eee.ufg.br

Flávio Henrique Teles Vieira
Escola de Engenharia Elétrica e de Computação - UFG
Av. Universitária S.N. – Setor Leste Universitário – Goiânia – Goiás
flavio@eee.ufg.br

Marcos Antônio de Sousa
Escola de Engenharia Elétrica e de Computação - UFG
Av. Universitária S.N. – Setor Leste Universitário – Goiânia – Goiás
marcos@eee.ufg.br

Marco Aurélio de Farias Borges
Escola de Engenharia Elétrica e de Computação - UFG
Av. Universitária S.N. – Setor Leste Universitário – Goiânia – Goiás
marcoafborges@gmail.com

RESUMO

O aumento do número de aplicações que geram tráfego VoIP (Voice over IP) requer que modelos adequados para este tipo de tráfego sejam empregados para um melhor dimensionamento dos recursos da rede. Neste artigo, propomos a aplicação de Misturas Gaussianas ajustadas por um algoritmo adaptativo EM (Expectation-Maximization) com seleção de classes utilizando o princípio do Comprimento de Descrição Mínimo (MDL-Minimum Description Length) para descrever o tráfego VoIP. Dados reais de VoIP foram coletados e séries sintéticas foram geradas de acordo com esses dados. O algoritmo proposto é comparado com o algoritmo adaptativo denominado Misturas Adaptativas em termos do desempenho em modelar as séries de tráfego VoIP consideradas.

PALAVRAS CHAVES. VoIP, Tráfego, Misturas.

Área principal: PO em Telecomunicações e Sistemas de Informação

ABSTRACT

The growing number of applications that generate traffic VoIP (Voice over IP) requires that suitable models for this kind of traffic be used in order to better estimate the required network resources. In this paper, we propose the application of Gaussian mixtures adjusted by an adaptive algorithm EM (Expectation-Maximization) with selection of classes using the principle of Minimum Description Length (MDL Minimum Description Length-) to describe the VoIP traffic. Real VoIP data were collected and synthetic traces were generated according to them. The proposed algorithm is compared to the adaptive algorithm called Adaptive Mixtures in terms of performance in modeling the series of VoIP traffic considered.

KEYWORDS. VoIP, Traffic, Mixtures

Main area: Operations Research in Telecommunications and Information Systems

1. Introdução

O crescente aumento do acesso à Internet impulsionou o processo de integração da transmissão das diversas mídias. A competitividade de preços, imposta por grandes indústrias, para equipamentos específicos para VoIP, o aumento da qualidade dos serviços em conjunto com a grande utilização da rede Internet ou de outras redes baseadas em IP e a criação de padrões relacionados ao VoIP permitiram sua entrada definitiva no mercado. Na década de 90, o surgimento do software comercial Internet Phone criado pela empresa *VocalTec Communications* possibilitou o empacotamento de amostras de voz para transmiti-las via rede IP (Colcher et. al., 2005).

O conhecimento das características do tráfego VoIP torna-se crucial, especialmente quando os prestadores de serviços se deparam com a escassez de recursos de rede e são obrigados a alocar seus recursos da forma mais eficiente possível (Hassan et. al., 2006).

Neste trabalho, propomos um algoritmo adaptativo de ajuste de parâmetros para o modelo Misturas Gaussianas baseado no algoritmo de estimação de maximização da esperança (*EM-Expectation-Maximization*), (Martinez at. al., 2008) e no conceito de Comprimento de Descrição Mínimo *(MDL-Minimum Description Length*) (Bouman et. al., 2005) para representar eficientemente o tráfego VoIP.

A principal motivação para estudarmos a aplicação de Misturas Gaussianas para modelar o tráfego VoIP foi a constatação de que suas densidades de probabilidade poderiam ser representadas pela soma de distribuições Gaussianas. Além disso, as Misturas Gaussianas proporcionam uma modelagem precisa para vários tipos de fenômenos cujas densidades de probabilidade podem ser contínuas ou discretas (Martinez at. al., 2008).

Para avaliar o algoritmo proposto aplicado a Misturas Gaussianas, inicialmente foram feitas coletas de dados de tráfego de pacotes em diversas chamadas VoIP. Em seguida, analisou-se o desempenho de modelagem para Misturas Gaussianas variando o tipo de algoritmo adaptativo de estimação do número de Gaussianas para o modelo. Utilizando o modelo obtido, dados sintéticos foram gerados e a partir dos mesmos a média, a variância e o desvio-padrão foram calculados para comparação com os valores dos dados de tráfego reais.

Este artigo está dividido em 5 seções. A seção 2 introduz os conceitos básicos sobre *Misturas Gaussianas*. A seção 3 expõe brevemente o conceito de Comprimento de Descrição Mínima (MDL) e o algoritmo proposto. A seção 4 traz a descrição dos dados coletados e os principais resultados para as análises realizadas. Finalmente, a seção 5 expõe as conclusões obtidas neste trabalho.

2. Misturas Gaussianas

A utilização generalizada das misturas gaussianas é devido ao fato de que uma distribuição gaussiana tem uma representação simples e concisa que requer apenas dois parâmetros, a média μ e variância σ^2 (Zhuang et. al., 1996). A Mistura Gaussiana é uma forma de modelagem de séries temporais que consiste no agrupamento de densidades de probabilidades gaussianas com características individuais. Tem como características desejáveis para um modelo, necessitar de pouco volume de informações armazenadas e de pouca carga computacional para se realizar as estimativas de densidade de probabilidade, principalmente quando se trabalha com conjunto de dados de *n* dimensões.

A Mistura Gaussiana parte do princípio que a função de densidade de probabilidade *f(x)* pode ser modelada como a soma ponderada de um certo número *c* de densidade de probabilidade, com $c<<n$. O caso mais comum é a Mistura Gaussiana univariada que é dada pela Equação 01 (Martinez et. al., 2008):

$$f(x) = \sum_{i=1}^{c} \rho_i \delta(x; \theta_i) \qquad (01)$$

onde ρ_i representa o peso ou o coeficiente da mistura para *i*-ésimo termo e $\delta(x;\theta_i)$ representa uma densidade de probabilidade, com os parâmetros μ e σ^2 representados pelo vetor θ_i e *x* é o ponto da variável que se deseja estimar a densidade. Para a Mistura Gaussiana multivariada pode-se usar a seguinte equação (Martinez et. al., 2008):

$$\hat{f}_{FM}(x) = \sum_{i=1}^{c} \hat{\rho}_i \, \phi(x; \hat{\mu}_i, \hat{\sum} i) \qquad (02)$$

onde *x* é um vetor *n*-dimensional dos valores da série que se deseja estimar a densidade; $\hat{\mu}_{\hat{i}}$ é um vetor *n*-dimensional de médias e $\hat{\sum}_i$ é uma matriz de covariância *n* x *n* de *x*.

A condição de que a soma dos coeficientes ρ_i da mistura deve ser igual à unidade tem que ser obedecida tanto para uma mistura univariada quando para uma multivariada.

Os parâmetros da função de densidade de probabilidade de uma mistura gaussiana são o número de Gaussianas (*c*), seu fator de ponderação ρ_i e os parâmetros μ (média) e Σ (matriz de covariância) de cada função de densidade Gaussiana (Resch, 2010). Geralmente o algoritmo EM é utilizado para estimar estes parâmetros.

O algoritmo EM (*Expectation-Maximization*) executa um cálculo iterativo de estimadores de máxima verossimilhança quando as observações podem ser vistas como dados incompletos (Picard, 2007). É de grande importância iniciar os parâmetros do algoritmo EM com valores adequados, pois o algoritmo encontra um valor ótimo local e não um valor ótimo global e a convergência do algoritmo também depende destes valores iniciais (Resch, 2010). O algoritmo EM possui duas limitações básicas: a taxa de convergência pode ser lenta e, como muitos procedimentos iterativos, é sensível à etapa de inicialização (Picard, 2007). Existem métodos menos sensíveis à etapa de inicialização como a Mistura Adaptativa que será abordada na próxima subseção.

2.1 – Misturas Adaptativas (MA)

A Mistura Adaptativa é um método para se estimar adaptativamente os parâmetros de uma Mistura Gaussiana, assim como o seu número de Gaussianas. A idéia básica na abordagem da mistura adaptativa é obter para cada ponto (amostra) dos dados (série temporal) à distância a partir desta observação a cada componente de densidade no modelo. Se a distância entre a observação e cada componente (termo) da densidade de probabilidade é maior do que algum limiar *(t_c)* de todos os termos, então um termo novo é criado. Se a distância for inferior ao limiar *(t_c)* de todos os termos, as estimativas de parâmetros são atualizadas com base nas equações EM recursivas (Martinez et. al., 2008). Se $t_c=1$ for usado, então um novo termo é criado quando uma nova observação é maior do que um desvio-padrão distante da média de cada termo. Para $t_c=4$, um novo termo será criado para uma observação que é, pelo menos, dois desvios-padrão de distância a partir dos termos existentes (Martinez et. al., 2008). Esta distância é dada pela Distância de Mahalanobis (distância de cada elemento até o elemento central da série). A distância de Mahalanobis pode ser definida como uma medida de similaridade.

A distância de Mahalanobis ao quadrado entre a nova observação $x^{(n+1)}$ e o i-ésimo termo é dado por

$$MD_i^2\left(x^{n+1}\right) = \left(x^{n+1} - \hat{\mu}_i^{\,n}\right)^T \left(\hat{\Sigma}_i^{\,n}\right)^{-1} \left(x^{n+1} - \hat{\mu}_i^{\,n}\right) \qquad (03)$$

A abordagem das Misturas Adaptativas pode ser resumida da seguinte forma, (Martinez et. al., 2008):

Algoritmo 01 – MA

1. Inicializar o algoritmo das misturas adaptativas usando o primeiro ponto do dados $x^{(1)}$:

$$\hat{\mu}_1^{(1)} = x^{(1)}, \hat{\rho}_1^{(1)} = 1 \; e \; \hat{\Sigma}_1^{(1)} = I$$

onde *I* denota a matriz identidade. No caso univariado, a variância do termo inicial é a unidade.

2. Para um novo ponto de dados $x^{(n+1)}$, calcular a distância de Mahalanobis ao quadrado como na Equação (03).

3. Se $\min_i \left\{ MD_i^2 \left(x^{n+1} \right) \right\} > t_c$, criar um novo termo utilizando as Equações (04), (05) e (06). Aumentar o número de termos N em uma unidade.

$$\hat{\mu}_{N+1}^{(n+1)} = x^{(n+1)} \tag{04}$$

$$\hat{\rho}_{N+1}^{(n+1)} = \frac{1}{n+1} \tag{05}$$

$$\hat{\Sigma}_{N+1}^{(n+1)} = \Im\left(\hat{\Sigma}_i \right) \tag{06}$$

onde $\Im\left(\hat{\Sigma}_i \right)$ é a média ponderada das probabilidades a posteriori calculadas usando as Equações (07) e (08).

$$\hat{\tau}_i^{(n+1)} = \frac{\hat{\rho}_i^{n} \, \phi\left(x^{(n+1)}; \hat{\mu}_i^{(n)}, \hat{\sum}_i^{(n)} \right)}{\hat{f}^{(n)}\left(x^{(n+1)} \right)}; i = 1, \ldots, c \tag{07}$$

$$\hat{f}^{(n)}\left(x^{(n+1)} \right) = \sum_{i=1}^{c} \hat{\rho}_i \phi\left(x^{(n+1)}; \hat{\mu}_i^{(n)}, \hat{\sum}_i^{(n)} \right) \tag{08}$$

onde $\hat{\tau}_i^{(n+1)}$ representa a probabilidade estimada a posteriori de que a nova observação $x^{(n+1)}$ pertença ao i-ésimo termo e o subescrito *(n)* denota os valores dos parâmetros estimados com base em *n* observações anteriores. O denominador é a estimativa de densidade da mistura gaussiana para a nova observação usando a mistura dos n pontos anteriores.

4. Se a distância mínima ao quadrado é inferior ao limiar de criação t_c, atualizar os termos existentes usando as Equações (09), (10) e (11).

$$\hat{\rho}_i^{(n+1)} = \hat{\rho}_i^{(n)} + \frac{1}{n}\left(\hat{\tau}_i^{(n+1)} - \hat{\rho}_i^{(n)} \right) \tag{09}$$

$$\hat{\mu}_i^{(n+1)} = \hat{\mu}_i^{(n)} + \frac{\hat{\tau}_i^{(n+1)}}{n \hat{\rho}_i^{(n)}}\left(x^{(n+1)} - \hat{\mu}_i^{(n)} \right) \tag{10}$$

$$\hat{\Sigma}_i^{(n+1)} = \hat{\Sigma}_i^{(n)} + \frac{\hat{\tau}_i^{(n+1)}}{n \hat{\rho}_i^{(n)}}\left[\left(x^{(n+1)} - \hat{\mu}_i^{(n)} \right)\left(x^{(n+1)} - \hat{\mu}_i^{(n)} \right)^T - \hat{\Sigma}_i^{(n)} \right] \tag{11}$$

5. Repetir os passos 2 à 4, utilizando todos os pontos do vetor de amostras.

Na prática, o método de misturas adaptativas é utilizado para obter os valores iniciais para os parâmetros, bem como uma estimativa do número de termos necessários para o modelo de densidade. Esta abordagem apresenta dois fatores que devem ser considerados: a complexidade do modelo que as vezes pode apresentar um número de termos maior do que o necessário e a função de densidade de probabilidade estimada depende da ordem na qual os

dados são apresentados ao algoritmo. Maiores detalhes podem ser encontrados em Martinez, et. al., (2008).

3. Algoritmo de Maximização da Esperança com Seleção de Classes utilizando Comprimento de Descrição Mínimo (MDL)

A abordagem MDL *(Minimum Description Length)* baseia-se no fato de que se existe alguma regularidade nos dados, esta pode ser usada para comprimir os mesmos. Quanto mais os dados forem compactados menos classes e consequentemente menos parâmetros serão necessários para caracterizar as mesmas, (Peter, 2005). O critério MDL tenta obter o número de subclasses que melhor descreve os dados da amostra. Propomos então a partir dos dados compactados, utilizar o algoritmo EM para estimar os parâmetros das subclasses formadas pelos dados da amostra e, dessa forma reduzir o número de Gaussianas na Mistura Gaussiana. As etapas relacionadas abaixo resumem a abordagem proposta de utilização da maximização da esperança em conjunto com o critério MDL (Bouman, et. al. 2005).

Algoritmo 02 - MDL

1. Inicializar a classe com um grande número de subclasses, *Ko*.

2. Inicializar $\theta^{(1)}$ (onde $\theta^{(1)} = (\pi_k^{(1)}, \mu_k^{(1)}, R_k^{(1)})$) usando (12), (13) e (14).

$$\pi_k^{(1)} = \frac{1}{K_0} \tag{12}$$

$$\mu_k^{(1)} = y_n \ \ onde \ \ n = \lfloor (k-1)(N-1)/(K_0 - 1) \rfloor + 1 \tag{13}$$

$$R_k^{(1)} = \frac{1}{N} \sum_{n=1}^{N} y_n y_n^t \tag{14}$$

onde y_n é uma amostra da série que se deseja modelar usando a Mistura Gaussiana; N é o número de elementos da série e $\lfloor . \rfloor$ é uma função piso com arredondamento para baixo onde o resultado é o maior inteiro menor ou igual ao argumento da função.

3. Aplicar o algoritmo iterativo EM - Equação (15) - até a mudança de MDL (K; θ) dado pela Equação (16), seja menor que ε .

$$\left(\pi^{(i+1)}, \mu^{(i+1)}, R^{(i+1)}\right) = \arg_{(\pi,\mu,R)\in\Omega^{(K)}} \max Q\left(\theta; \theta^{(i)}\right) = \left(\overline{\pi}, \overline{\mu}, \overline{R}\right) \tag{15}$$

onde

$$Q\left(\theta, \theta^{(i)}\right) = E\left[\log \rho_{x,y}\left(y, X \mid \theta\right) \mid Y = y, \theta^{(i)}\right] - \frac{1}{2} L \log(NM)$$

$$MDL(K, \theta) = -\log p_y(y \mid K, \theta) + \frac{1}{2} L \log(NM) \tag{16}$$

e L é um termo de penalidade e M é a dimensão da série temporal que se deseja modelar.

4. Armazenar o parâmetro $\theta^{(K, i_{final})}$ e o valor MDL (K, $\theta^{(K, i_{final})}$).

5. Se o número de subclasses K é maior que 1, aplicar a Equação (17) para reduzir o número de subclasses, conjunto $K \leftarrow K - 1$, e voltar para o passo 3.

$$\left(l^*, m^*\right) = \arg_{(l,m)} \min d(l, m) \tag{17}$$

onde l e m são exemplos de subclasses.

6. Escolher o valor de K^* e dos parâmetros $\theta^{(K^*, i_{final})}$ que minimizam o valor do MDL.

Na etapa 3, o valor de ε é escolhido para ser (Bouman et. al., 2005):

$$\varepsilon = \frac{1}{100}\left(1 + M + \frac{(M+1)M}{2}\right)\log(NM)$$

Uma vez obtidos K^*e $\theta^{(K^*, i_{final})}$ é possível utilizar as seguintes equações para calcular a densidade de probabilidade da Mistura e posteriormente gerar uma série sintética.

$$\rho_{y_n|x_n}(y_n|k,\theta) = \frac{1}{(2\pi)^{M/2}}|R_k|^{-1/2}\exp\left\{-\frac{1}{2}(y_n-\mu_k)^t R_k^{-1}(y_n-\mu_k)\right\} \quad (18)$$

$$\rho_{y_n}(y_n|k,\theta) = \sum \rho_{y_n|x_n}(y_n|k,\theta)\pi_k \quad (19)$$

4. Avaliação dos Algoritmos Adaptativos

4.1 - Dados coletados

Os dados foram coletados de chamadas realizadas com um software (Skype) que utiliza a comunicação VoIP entre dois terminais fim-a-fim. Para mostrar os resultados de modelagem neste artigo, foram escolhidas 3 séries de tráfego (chamadas) que representam a maioria dos comportamentos do tráfego VoIP das séries coletadas. A Tabela 01 mostra o comportamento estatístico dos dados coletados para três destas chamadas realizadas.

Chamada	*Nº. de Pacotes (Amostras)*	*Média da Amostra*	*Variância da Amostra*
01	15.337	97,6224	313,5448
02	41.240	455,8242	338890
03	7.400	566,7468	375860

Tabela 01 – Resumo dos Dados Coletados nas Chamadas VoIP.

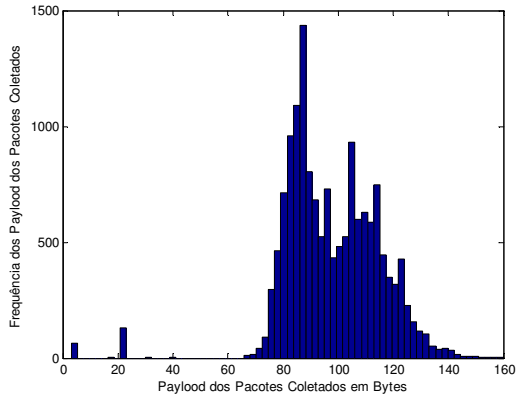

Figura 01 – Histograma Chamada 01.

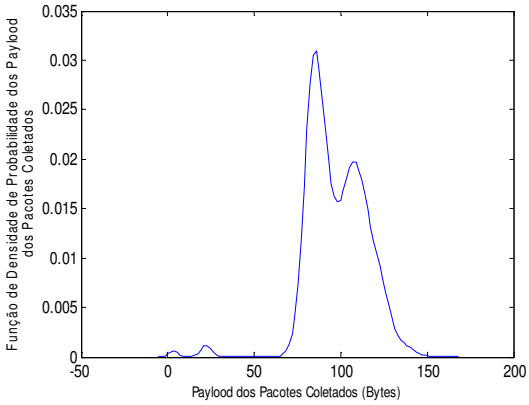

Figura 02 – Densidade de Probabilidade da Chamada 01

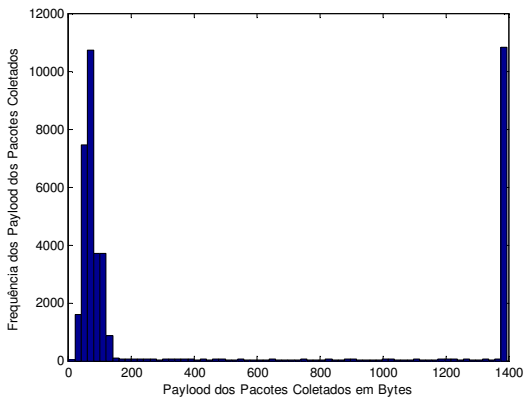

Figura 03 – Histograma Chamada 02.

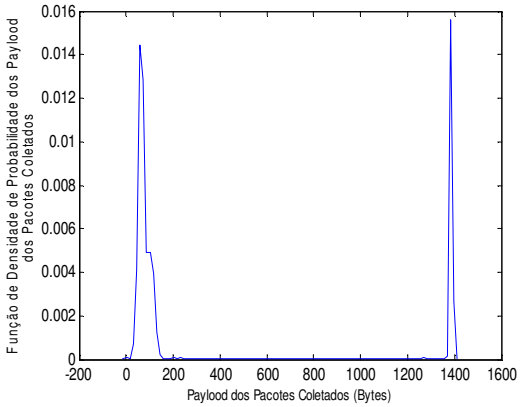

Figura 04 – Densidade de Probabilidade da Chamada 02

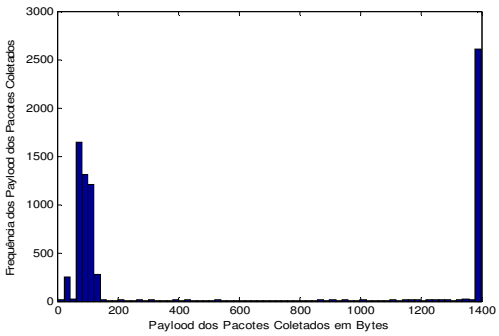

Figura 05 – Histograma Chamada 03.

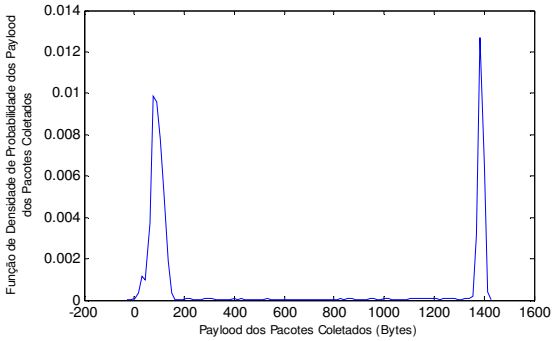

Figura 06 – Densidade de Probabilidade da Chamada 03

As figuras de 01 a 06 apresentam os histogramas e as curvas de densidades de probabilidade para os valores de *payload* (número de bytes de dados de usuário) nas três chamadas. É possível observar que na chamada 01 existe uma regularidade entre os valores dos *payload*, com uma pequena variação na quantidade de bytes para o *payload*. Este comportamento no tamanho dos pacotes já não se repete para as chamadas 02 e 03. A Tabela 02 ilustra este comportamento no tamanho do *payload* para os dez primeiros valores coletados nas três séries VoIP.

Chamada 01	***Chamada 02***	***Chamada 03***
109	49	72
97	80	78
106	51	79
97	76	69
97	40	81
110	47	82
107	79	66
105	1387	82
100	1158	90
92	56	1387

Tabela 02 – Dez Primeiros Valores de *Payload* dos Pacotes das Chamadas VoIP.

4.2 - Análise Realizada utilizando os Métodos MA e MDL

Para encontrar o número adequado de gaussianas para as Misturas Gaussianas foram utilizadas abordagens adaptativas: Misturas Adaptativas e a proposta (MDL-EM). Para realizar as comparações, depois de gerada a amostra sintética, também, foi calculado a média, a variância e o desvio-padrão de ambas as séries.

Chamada 01:

As Tabelas 03 e 04 mostram os erros percentuais calculados para as duas abordagens para a chamada 01. Comparou-se também as funções de densidade das séries geradas com as das séries reais. A Figura 07 mostra a densidade de probabilidade de cada série gerada com os algoritmos adaptativos para a chamada 01.

	Série Coletada	***Série Sintética com 14 Gaussianas***	***Erro(%)***
Média	97,6224	97,6497	0,0280
Variância	313,5448	313,5803	0,0113
Desvio Padrão	17,7072	17,7082	0,0056

Tabela 03 - Parâmetros Estatísticos da Chamada 01 utilizando MA.

	Série Coletada	***Série Sintética com 03 Gaussianas***	***Erro(%)***
Média	97,6224	97,6385	0,0166
Variância	313,5448	311,8561	0,5386
Desvio Padrão	17,7072	17,6594	0,2694

Tabela 04 - Parâmetros Estatísticos da Chamada 01 utilizando MDL-EM.

Comparando os valores mostrados nas Tabelas 03 e 04, nota-se que apesar do algoritmo MA apresentar menores valores de erro percentual para o desvio padrão e conseqüentemente para variância, os valores de erro percentual apresentados para média é maior que o valor apresentado pelo algoritmo MDL-EM. Entretanto, o algoritmo MDL-EM atinge este desempenho utilizando um menor número de gaussianas.

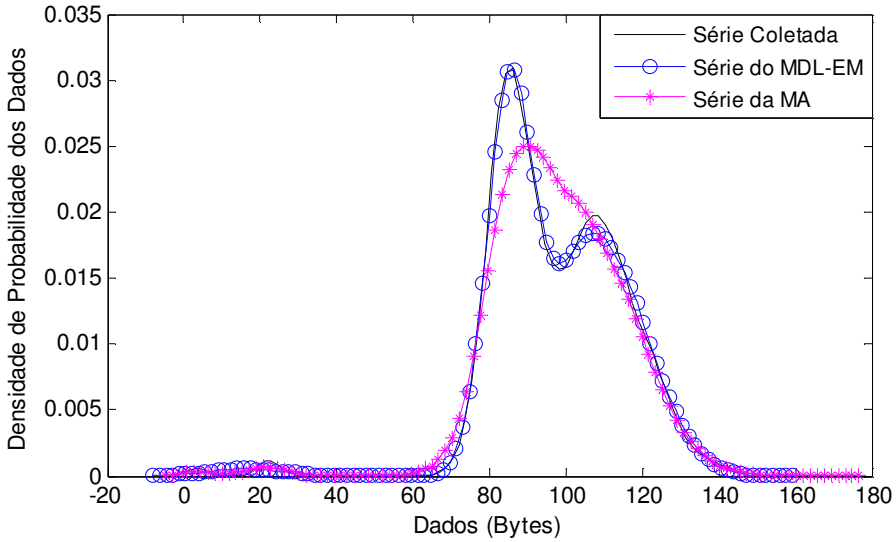

Figura 07 – Densidade de Probabilidade das Séries Sintéticas Geradas pelo MDL-EM e MA e da Série Coletada para a chamada 01

Observa-se na Figura 07, que a curva de densidade de probabilidade da série gerada pelo algoritmo MDL-EM aproximadamente sobrepõe a curva de densidade da série coletada enquanto que a curva de densidade da série gerada pelo algoritmo MA apresentou uma discrepância significativa.

Chamada 02:

Para a chamada 02 utilizou os mesmos procedimentos realizados para a chamada 01, utilizou os algoritmos MA e MDL-EM, com base nas séries VoIP coletadas, para gerar as séries sintéticas para posterior calculo de suas estatísticas. As Tabelas 05 e 06 mostram os erros percentuais calculados para as abordagens MA e MDL-EM. A Figura 08 mostra as curvas de densidade de probabilidade de cada série gerada com os algoritmos adaptativos.

	Série Coletada	***Série Sintética com 10 Gaussianas***	***Erro(%)***
Média	455,8242	456,3812	0,1222
Variância	338890	331780	2,0980
Desvio Padrão	582,1422	576,0041	1,0544

Tabela 05 - Parâmetros Estatísticos da Chamada 02 utilizando MA.

	Série Coletada	***Série Sintética com 09 Gaussianas***	***Erro(%)***
Média	455,8242	455,7431	0,0178
Variância	338890	338650	0,0700
Desvio Padrão	582,1422	581,9382	0,0350

Tabela 06 - Parâmetros Estatísticos da Chamada 02 utilizando MDL-EM

Observando os valores apresentados pelas Tabelas 05 e 06, nota-se que os menores erros percentuais para os parâmetros estatísticos foram obtidos para a série sintética gerada pelo algoritmo MDL-EM. Pode-se dizer então que a série gerada pelo algoritmo MDL-EM se aproximou mais da série VoIP coletada.

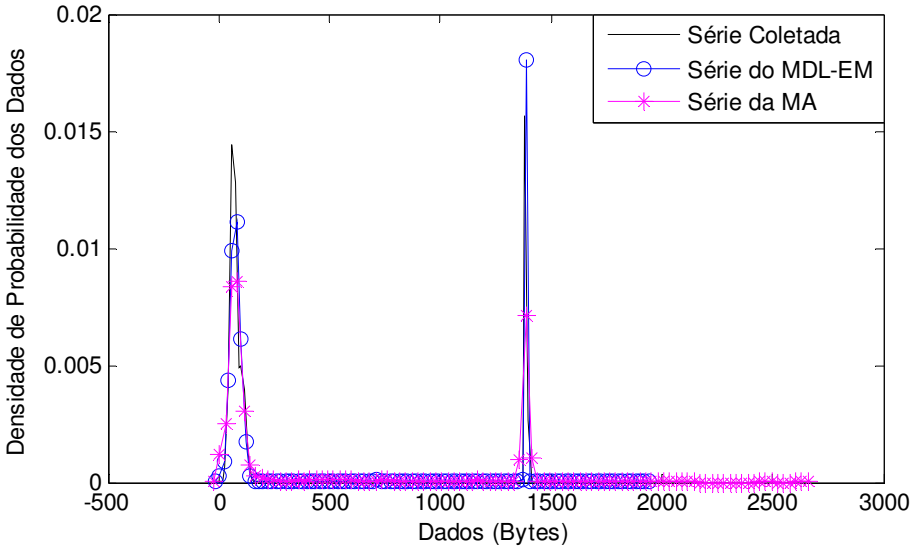

Figura 08 – Densidade de Probabilidade das Séries Sintéticas Geradas pelo MDL-EM e MA e da Série Coletada para a chamada 02

Pode-se observar na Figura 08 que a curva de densidade de probabilidade da série gerada pelo algoritmo MA apresenta picos mais distantes da curva de densidade da série coletada. Ainda na Figura 08, também, pode-se observar que a densidade da série gerada pelo algoritmo MDL-EM acompanha de forma mais precisa a curva de densidade da série coletada do que a densidade da série gerada pelo algoritmo MA.

Chamada 03:

Para a chamada 03 foram mantidos os mesmos procedimentos realizados para as chamadas 01 e 02. As Tabelas 07 e 08 mostram os erros percentuais calculados para as abordagens MA e MDL-EM. A Figura 09 mostra as curvas de densidade de cada série gerada com os algoritmos adaptativos.

	Série Coletada	***Série Sintética com 10 Gaussianas***	***Erro(%)***
Média	566,7468	568,0899	0,2370
Variância	375860	375950	0,0238
Desvio Padrão	613,0720	613,1451	0,0119

Tabela 07 - Parâmetros Estatísticos da Chamada 03 utilizando MA.

	Série Coletada	***Série Sintética com 08 Gaussianas***	***Erro(%)***
Média	566,7468	566,5941	0,0269
Variância	375860	375770	0,0220
Desvio Padrão	613,0720	613,0046	0,0110

Tabela 08 - Parâmetros Estatísticos da Chamada 03 utilizando MDL-EM

Na chamada 03, a série sintética gerada pelo algoritmo MDL-EM, também, apresentou os menores erros percentuais comparados com os erros percentuais calculados para a série sintética gerada pelo algoritmo MA, como mostra as Tabelas 07 e 08. Também para a chamada 03, a série sintética gerada pelo algoritmo MDL-EM possui características mais próximas das características da série VoIP coletada.

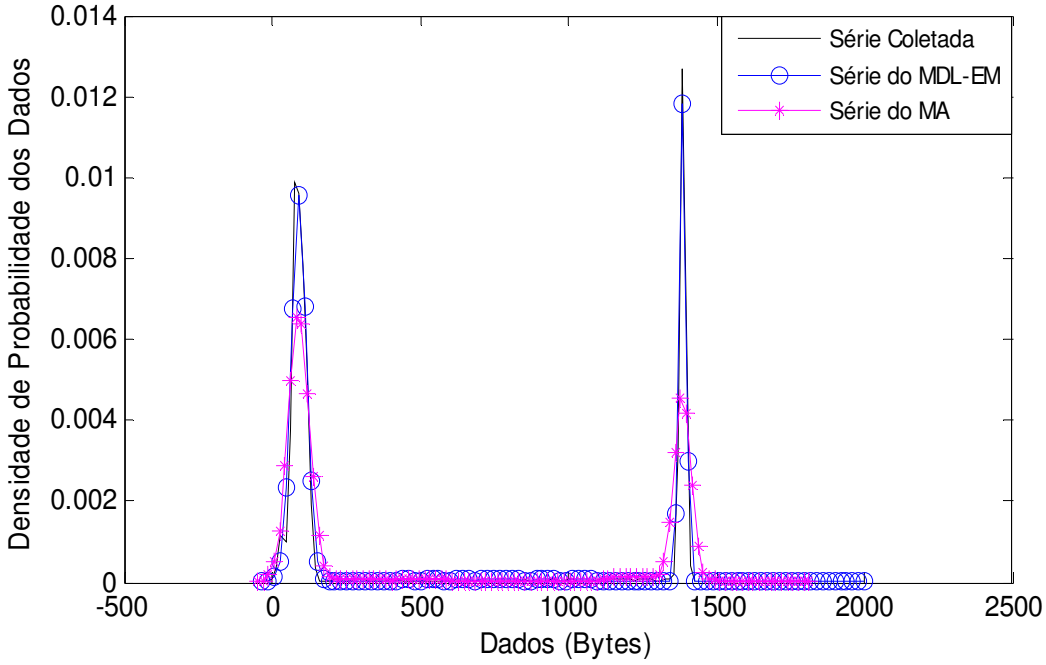

Figura 09 – Densidade de Probabilidade das Séries Sintéticas Geradas pelo MDL-EM e MA e da Série Coletada para a chamada 03.

Observa-se na Figura 09 que a curva de densidade de probabilidade da série gerada pelo algoritmo MDL-EM se aproxima mais da curva de densidade da série coletada do que a curva de densidade da série gerada pelo algoritmo MA. Os picos da curva de densidade da série do MDL-EM se aproxima mais que os picos da curva de densidade da série coletada.

4.3 - Taxa de Perda

Uma análise do comportamento de fila no *buffer* também foi realizada para comparar as séries sintéticas geradas com o da série VoIP original coletada. Para isso, considerou-se nas simulações, um enlace com um servidor e um *buffe*r sendo alimentado pelas séries. A simulação computacional foi realizada mantendo-se o tamanho do *buffer* fixo em 100 bytes e variando a capacidade do enlace em 45, 60, 70, 85 e 100 bytes/s. As Tabelas 09, 10 e 11 mostram os valores obtidos para a taxa de perda de pacotes para as séries coletada e sintéticas.

Tamanho do Buffer(Bytes) / Capacidade de Enlace(Bytes/s)	*Série Coletada Taxa de Perda(%)*	*Série MA Taxa de Perda(%)*	*Série MDL Taxa de Perda(%)*
100 / 45	53,90	53,94	53,90
100 / 60	38,61	38,53	38,60
100 / 70	28,49	28,31	28,44
100 / 85	13,77	13,06	13,18
100 / 100	04,53	00,39	00,47

Tabela 09 – Comportamento de Fila das Séries Coletada, do MA e MDL-EM para a Chamada 01

Tamanho do Buffer(Bytes) / Capacidade de Enlace(Bytes/s)	*Série Coletada Taxa de Perda(%)*	*Série MA Taxa de Perda(%)*	*Série MDL Taxa de Perda(%)*
100 / 200	73,24	69,81	70,31
100 / 500	53,09	49,15	50,09
100 / 900	27,56	23,81	24,61
100 / 1.000	21,40	17,99	18,63
100 / 1.300	03,48	01,40	04,52

Tabela 10 – Comportamento de Fila das Séries Coletada, do MA e MDL-EM para a Chamada 02

Tamanho do Buffer(Bytes) / Capacidade de Enlace(Bytes/s)	*Série Coletada Taxa de Perda(%)*	*Série MA Taxa de Perda(%)*	*Série MDL Taxa de Perda(%)*
100 / 200	74,98	71,82	72,17
100 / 500	54,65	51,55	52,11
100 / 900	28,58	26,12	26,33
100 / 1.000	22,24	19,49	19,78
100 / 1.300	03,81	01,76	04,24

Tabela 11 - Comportamento de Fila das Séries Coletada, do MA e MDL-EM para a Chamada 03

Os valores da capacidade de enlace utilizados na chamada 01 são menores do que para as outras chamadas devido ao fato que os valores de capacidade de enlace iguais a 200, 500, 900, 1.000 e 1.300 geram taxas de perdas iguais a zero para esta chamada. A não ser por este critério, os valores para o tamanho do *buffer* e para capacidade de enlace utilizados nas simulações foram escolhidos aleatoriamente. Observando as tabelas acima nota-se que à medida que a capacidade de enlace aumenta a taxa de perda diminui. Observa-se ainda, que para as três chamadas, a série gerada pelo algoritmo MDL-EM apresentou os valores da taxa de perda mais próximos da taxa de perda da série VoIP coletada, ou seja, o comportamento de fila da série sintética gerada pelo algoritmo MDL-EM é o mais parecido com a série VoIP coletas.

A relação entre a taxa de perda de pacotes e o tamanho do buffer para duas chamadas pode ser averiguada nas figuras 10 e 11.

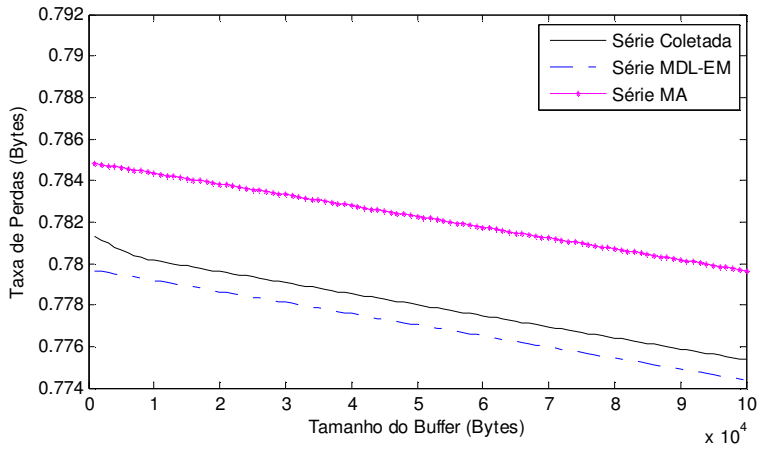

Figura 10 – Gráfico da Relação entre a Taxa de Perda e o Tamanho do Buffer para Série Coletada e Séries do MDL-EM e MA na Chamada 02

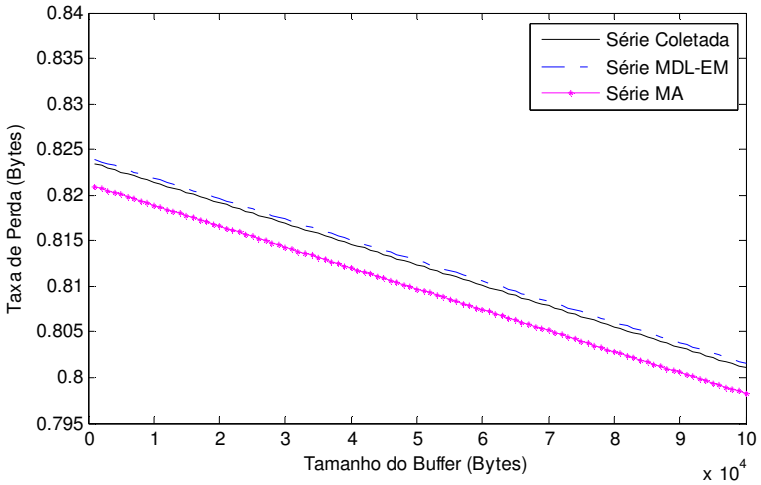

Figura 11 – Gráfico da Relação entre a Taxa de Perda e o Tamanho do Buffer para Série Coletada e Séries do MDL-EM e MA na Chamada 03

As Figuras 10 e 11 corroboram os resultados apresentados nas tabelas 09, 10 e 11 no sentido de que as séries geradas pelo algoritmo MDL-EM também apresentam uma relação entre a taxa de perda e o tamanho do *buffer* similar as das séries VoIP coletadas. Observa-se também que o tamanho do *buffer* influencia no valor da taxa de perda. Um aumento do tamanho do *buffer* proporciona uma queda na taxa de perda.

5. Conclusões

A geração de séries a partir de Misturas Gaussianas depende da estimação da densidade de probabilidade para este modelo. As dificuldades na obtenção da densidade de probabilidade para Misturas Gaussianas é encontrar os parâmetros do modelo e um número de gaussianas que seja eficaz na geração da série sintética. Este artigo apresentou abordagens adaptativas para encontrar o número adequado de gaussianas. Observou-se que é possível obter uma série sintética com características semelhantes das amostras coletadas utilizando misturas gaussianas.

A abordagem proposta MDL-EM, de modo geral, apresentou menores valores para os erros percentuais para as médias e variâncias das séries nas três chamadas VoIP do que o algoritmo MA. O algoritmo MDL-EM, também, apresentou menor número de gaussianas, conseqüentemente necessita de menor volume de informação a ser armazenado e de menor processamento computacional para estimar a densidade do que o algoritmo MA. Avaliando os resultados obtidos, pode-se dizer que a série gerada pelo algoritmo MDL-EM possui em geral características mais próximas das características das séries de VoIP coletadas do que as séries geradas através do algoritmo MA. Outro fato, também observado, é que com a utilização do algoritmo MDL-EM não é necessário determinar o quanto de regularidade possui a série para posteriormente escolher o método a ser usado. O tempo gasto na estimativa dos parâmetros do modelo apresentado pela abordagem MDL-EM é maior que o tempo gasto apresentado pela abordagem MA, mas a diferença entre o tempo computacional dos algoritmos não sobressai às vantagens que o MDL-EM apresenta. Por fim, na análise de taxa de perda em um enlace de comunicação com *buffer* finito, a série gerada pelo algoritmo MDL-EM também apresentou um comportamento mais próximo do das séries coletadas do que o algoritmo MA.

Referências

Bouman, C. A., Shapiro, M., Cook, G. W., Atkins, C. B., Cheng, H., Dy, J.G., Borman, S., (2005), *CLUSTER: An Unsupervised Algorithm for Modeling Gaussian Mixtures*, School of Electrical Engineering, Purdue University, West Lafayerette.

Colcher, S., Gomes, A. T. A., Silva, A. O., Souza, G. L. F. e Soares, L. F. S., (2005), *VoIP – Voz sobre IP*, 4 ed., Editora Elsevier, Rio de Janeiro.

Hassan, H., Garcia, J. M., Bockstal, C., (2006), *Aggregate Traffic Models for VoIP Applications*, LAAS-CNRS, 7 av du colonel ROCHE, 31077 Toulouse cedex4, FRANCE.

Martinez, W. L., Martinez, A. R., (2008), *Computational Statistics Handbook with MatLab*, Chapman & Hall/CRC, Boca Raton, Florida.

Peter G., (2005), *A Tutorial Introduction to the Minimum Description Length Principle, Centrum voor Wiskunde en Informatica*, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands, http://homepages.cwi.nl/~pdg/ftp/mdlintro.pdf.

Picard, F., (2007), *An introduction to mixture models, Laboratoire Estatistique et Génome*, UMR CNRS 8071 – INRA 1152 – Univ. d'Evry, France.

Resch, B. (2010), Mixtures of Gaussians – *A Tutorial for the Course Computational Intelligence*. Signal Processing and Speech Communication Laboratory. http://www.igi.tugraz.at/lehre/CI/tutorials/ MixtGaussian/MixtGaussian.html, Inffeldgasse 16c, last updated: 24. Jun.

Zhuang, X., Huang, Y., Palaniappan, K. end Zhao, Y., (1996), Gaussian Mixture Density Modeling, Decomposition, and Applications, *IEEE TRANSACTIONS ON IMAGE PROCESSING*, VOL. 5, NO. 9.