# CLASSIFICAÇÃO DO GÊNERO MUSICAL UTILIZANDO REDES NEURAIS ARTIFICIAIS

**Elmo BORGES (1), Eduardo SIMAS FILHO (2), Cláudia FARIAS (3), Igor RIBEIRO (4), Diego LOPES (5).**

(1) IFBA, Rua Emídio dos Santos, s/n, (71) 2102-9423, e-mail: elmo.alberto@ifba.edu.br; (2) IFBA, e-mail: esimas@ifba.edu.br; (3) IFBA, e-mail: cfarias@ifba.edu.br; (4) IFBA, e-mail: igorsr@ifba.edu.br; (5) IFBA e-mail: diegobastos@ifba.edu.br;

## RESUMO

A quantidade de arquivos de áudio disponível na internet e em coleções pessoais vem crescendo rapidamente motivado principalmente pelo desenvolvimento de formatos de compactação, como o MP3, o aumento da capacidade e a popularização de dispositivos para transmissão de dados na internet. Numa coleção com grande número de arquivos, sistemas automáticos de classificação e busca baseados no conteúdo do áudio são muito importantes para permitir acesso adequado aos dados desejados, uma vez que os arquivos não carregam explicitamente essa informação, e a busca "manual" em um grande conjunto de sinais pode tornar-se exaustiva. Este trabalho propõe a aplicação de técnicas de processamento de sinais para a extração de características de arquivos de áudio digital visando à classificação em gêneros musicais. Características como os coeficientes mel-cepstrais, a taxa de cruzamento por zero e o histograma rítmico são estimadas e utilizadas para alimentar um classificador neural supervisionado (na arquitetura perceptron de múltiplas camadas). Através da metodologia proposta, elevados índices de acerto na classificação foram obtidos.

**Palavras-chave:** Processamento digital de sinais, extração de características, MFCC, classificação, redes neurais.

## 1. INTRODUÇÃO

O avanço do poder computacional, a disseminação da Internet e o desenvolvimento de novos algoritmos de codificação possibilitaram o crescimento contínuo da produção e do armazenamento de arquivos digitais, sobretudo os de áudio. Neste contexto, a automação de processos para manipulação de grandes conjuntos de dados pode representar considerável economia no tempo de busca. O registro de busca pode conter os mais variados indexadores de classificação: gênero musical, artista, época, instrumentos utilizados, tipo de voz. Uma das classificações de maior utilização e a de gêneros musicais.

Em uma ferramenta automática para classificação em gêneros musicais a etapa inicial é a extração de parâmetros. As características extraídas devem fornecer informações relevantes dos sinais em análise, de forma a tornar possível a classificação dos mesmos quanto a critérios pré-estabelecidos (PEREIRA, 2009).

Nos últimos anos, alguns trabalhos foram desenvolvidos com o objetivo de classificar arquivos de áudio à partir do seu conteúdo. Em (LU *et al*., 2002) e (KIM *et al*., 2005) diferentes algoritmos de processamento de sinais foram utilizados para diferenciar os arquivos entre sinais de música, fala ou som ambiente. Mais especificamente, visando à classificação em gêneros musicais o trabalho de (Tzanetakis e Cook, 2002) utiliza classificadores como GMM (*Gaussian Mixture Model*) e KNN (*k-Nearest Neighbor*) para efetuar a discriminação.

Neste trabalho, realizou-se estudo acerca da extração de características discriminantes de arquivos de áudio, a fim de classificá-los automaticamente, quanto aos diferentes gêneros, com utilização de redes neurais artificiais. Para criar um vetor de classificação foram extraídas características como: textura timbral (concentração da energia nas diferentes faixas de freqüências, coeficientes Mel-Cepstrais (MFCC-*Mel-frenquencey cepstral coeficients*), taxa de cruzamento por zeros (ZCR - *Zero Crossing Rate*) e relacionadas à batida (histograma rítmico).

# 2. FUNDAMENTAÇÃO TEÓRICA

## 2.1 Sistemas de Classificação de Gêneros Musicais

As redes neurais artificiais são um método para solucionar problemas através da simulação do cérebro humano, inclusive em seu comportamento, ou seja, aprendendo, errando e fazendo descobertas. São técnicas computacionais que apresentam um modelo inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência.

As redes neurais possuem nós ou unidades de processamento. Cada unidade possui ligações para outras unidades, nas quais recebem e enviam sinais. Essas unidades são a simulação dos neurônios, recebendo e retransmitindo informações. A propriedade mais importante das redes neurais é a habilidade de aprender a partir de seu ambiente e generalizar para exemplos não vistos no conjunto de treino (HAYKIN, 2008). Isso é feito através de um processo iterativo de ajustes aplicado a seus pesos, o treinamento.

Para a vasta maioria dos problemas práticos um único neurônio não é suficiente, sendo assim utilizam-se neurônios interconectados, sendo que a decisão de como interconectar os neurônios é uma das mais importantes decisões a se tomar em um projeto de uma rede neural artificial. Para permitir a melhoria dos resultados da RNA é conveniente a utilização de camadas intermediárias (ou ocultas).

O sistema de classificação de gêneros musicais utilizando inteligência artificial é composto por dois módulos principais, conforme Figura 1, são eles: extração de características e classificação de gêneros musicais.

A extração de características pode ser efetuada em curtas janelas de tempo ou em janelas maiores, a depender do parâmetro que está sendo estimado.

Por se tratar de um sistema que utiliza algoritmos de classificação, opera em três conjuntos (treinamento, teste e validação) e em dois modos diferentes: treinamento, teste. No modo de treinamento os vetores de características são utilizados com seus respectivos gêneros musicais. O conjunto de validação é utilizado para determinar o momento de parada do treinamento, evitando o treinamento excessivo que pode levar a rede ao erro. No modo de teste o conjunto de teste é usado para realizar a avalição de desempenho do sistema (HAYKIN, 2008).

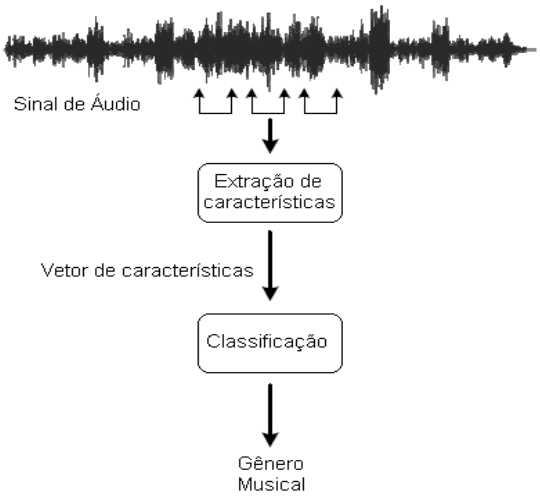


**Figura 1. Visão geral do sistema.**

Em outros trabalhos como o apresentado por Matityaho *et al*.(1995), para obter a classificação de 100% de um conjunto com somente dois gêneros, clássico e pop, a arquitetura da rede necessitou de duas camadas ocultas, com 60 e 20 neurônios cada, o que torna o processamento da rede muito lenta. Já em Tzanetakis *et al*.(2002) que utiliza para classificação padrões estatísticos como o GMM(modelo de mistura Gaussiana) consegui-se a classificação de 61% (tempo não- real) e 44% (tempo real), para um conjunto constituído por dez gêneros musicais.

## 2.2 Extração de Características

### 2.2.1 Concentração do Centróide da Energia em Faixas de Freqüências

A concentração do centróide da energia em faixas de freqüência é uma que analisa a variação da concentração das maiores amplitude do sinal no domínio da frequência, em faixas (bandas) de frequência diferentes, como uma característica para diferenciação de gêneros.

A energia é estimada via Densidade Espectral de Potência (PSD – *Power Espectral Density*) de um sinal x(t) pode ser definida como a transformada de Fourier da função de autocorrelação do sinal.

### 2.2.2 Taxa de Cruzamento por Zeros

A taxa de cruzamento por zeros (ZCR - *Zero Crossing Rate*) é uma técnica comumente utilizada na caracterização de sinais de áudio (LU, *et al*, 2002).

O ZCR é a taxa de mudanças do sinal de positivo para negativo. Descreve a freqüência dominante da música, sendo útil também para encontrar quadros em silêncio. A Eq.01 mostra como o ZCR pode ser obtida a partir de um sinal s(n) (KIM, *et al*, 2005).

$$ZCR = \frac{1}{2}\sum_{n=1}^{N-1}\left|sign(s(n)) - sign(s(n-1))\right|\frac{Fx}{N} \qquad [Eq.01]$$

Onde Fx é a frequência de amostragem e N é o numero de amostra em s(n).

### 2.2.3 Coeficientes Mel-cepstral

Os coeficientes cepstrais de freqüência-mel (MFCC-*Mel-frenquency cepstral coeficients*) são características bastante utilizadas na literatura para a descrição de sinais de áudio, principalmente no processamento da fala (LOGAN, 2000). Os MFCCs são baseados na extração da energia do sinal dentro de críticas bandas de freqüências por meio de uma série de filtros triangulares, cujo centro de freqüências são espaçados de acordo com a escala mel (KIM, *et al*, 2005). As medidas cepstrais, atravéz da escala mel, fornecem informação sobre o formato do espectro, tentando simular a percepção do ouvido humano, que tem resposta não-linear (aproximadamente logarítmica) em função da frequência. A escala mel pode ser mapeada da linear (em Hz) através da Eq. 2 (Pereira, 2009).

$$mel(f) = 1127.\ln\left(1 + \frac{f(Hz)}{700}\right) \qquad [Eq.02]$$

Os coeficientes mel-cespstrais são finalizados com obtenção da Transformada Discreta do Cosseno definida por Eq.03 (KIM, *et al*, 2005).

$$C(i) = \alpha(i)\sum_{x=0}^{N-1} f(x).\cos\left(\frac{(2x+1)i\pi}{2N}\right) \qquad [Eq.03]$$

### 2.2.4 Histograma Rítmico

O histograma rítmico busca encontrar a batida principal da música e seu período em BPM (batidas por minuto). A presença de batidas secundárias relevantes indica que a música tem um conteúdo rítmico mais intenso.

A primeira parte para extração do histograma rítmico consiste na analise do sinal via Transformada Discreta Wavelet. Em musica padrões se repetem no tempo e possuem composição de frequência determinada. A análise com Wavelets pode ser vista como uma decomposição detalhada, onde se busca os componentes mais básicos dos sinais (MALLAT, 1989).

Um modo eficiente de se implementar a TWD (Transformada Wavelet Discreta) foi desenvolvido por Mallat. A análise de multiresolução via algoritmo piramidal de Mallat refere-se ao procedimento de se

obter "aproximações" e "detalhes" de um dado sinal. Uma aproximação é uma representação de baixa frequência do sinal original enquanto que um detalhe representa os componentes de alta frequência (MALLAT, 1989).

A segunda parte é a estimação da envoltória seguindo-se da autocorrelação. As etapas para estimação da envoltória são detalhadas abaixo (TZANETAKIS e COOK, 2002):

Retificação de onda completa:

$$yFull[n] = abs(x[n]) \quad \text{[Eq.04]}$$

Filtragem Passa – baixa:

$$y_{FBP}[n] = (1-\alpha)x[n] + \alpha y[n-1] \quad \text{[Eq.05]}$$

Redução da resolução (subamostragem):

$$y_{down}[n] = x[kn] \quad \text{[Eq.06]}$$

Extração da média:

$$ymed[n] = x[n] - E(x[n]) \quad \text{[Eq.07]}$$

A Autocorrelação é uma ferramenta matemática utilizada para encontrar padrões de repetição, como a presença de um sinal periódico que foi mascarado pelo ruído, ou identificar a freqüência fundamental ausente em um sinal implícito por suas freqüências harmônicas. Considerando um sinal digital x[n], a função de autocorrelação é definida pela Eq.09.

$$yAuto[n] = \frac{1}{N}\sum_n x[n].x[n+k] \quad \text{[Eq.08]}$$

### 2.2.5 Avaliação do treinamento das Redes

Para avaliar o desempenho de cada treinamento foi calculada a taxa de acerto médio definida por:

$$\text{Acerto Médio} = \frac{1}{N}\sum_{i}^{N} \text{Acerto da classe } i \quad \text{[Eq.09]}$$

## 3. METODOLOGIA

Neste trabalho a maioria das características estimadas foi de tempo curto, apenas o histograma rítmico precisa de uma janela de tempo maior para ser extraído.

O processo de amostragem do sinal temporal (em janelas de curto tempo ou janelas de longo tempo) foi realizado a partir de janelas de Hamming (DINIZ *et al*., 2004) com 30% de sobreposição entre janelas adjacentes.

O conjunto de sinais utilizado neste trabalho é composto por 468 amostras (156 para treino, 156 para teste e 156 para validação). Utilizou-se arquivos de músicas de cinco gêneros musicais diferentes: blues, MPB, Reggae, Rock e Samba. Para garantir a variedade das amostras, trechos foram retirados de CD's (compact Disc) e arquivos de áudio MP3. Os arquivos foram armazenador com uma taxa de amostragem de 44100 Hz e codificados com 16 bits.

### 3.1 Concentração do Centroide da Energia em faixas de Freqüências

Foram utilizadas bandas de freqüências dividindo o espectro em baixas (0 a 600 Hz), médias (600 a 2400 Hz) e altas freqüências (2400 a 41000 Hz) A energia em cada faixa é obtida a partir do seguinte procedimento (ver Figura 2): primeiramente, o sinal é dividido em janelas de 30 ms, em seguida filtrado usando filtros FIR (*Finite impulse response*) (Diniz, et al., 2004) passa-baixa (0-600Hz) passa-faixa (600 – 2.400Hz) e passa-alta (2.400 – 41.000Hz). A potência do sinal é extraída das faixas de freqüências através da aplicação de algoritmos de obtenção da PSD. Por fim, a potência em cada janela é somada e normalizada.

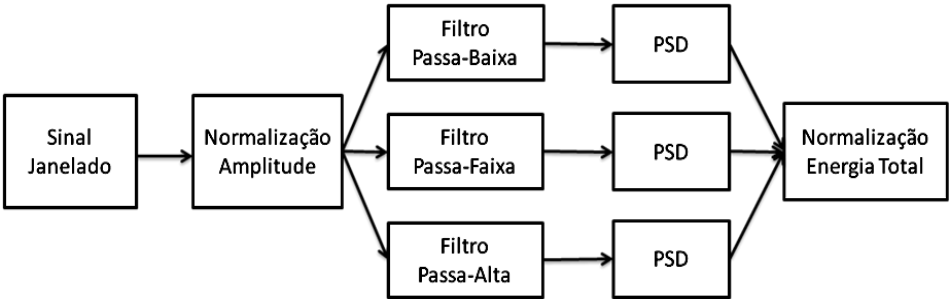


**Figura 2. Energia em faixas de freqüência.**

## 3.2 Coeficientes Mel-cepstral

Para extração destes parâmetros, o sinal é dividido em quadros (janelas) temporais de curta duração (~30 ms) na etapa de pré-processamento.

A Figura 3 mostra as etapas necessárias para a estimação dos MFCC's. Inicialmente o sinal temporal é janelado, em seguida a transformada discreta de Fourier (DFT – *Discrete Fourier transform*) é aplicada em cada janela. Após o mapeamento na escala mel é realizado um processo de filtragem seguido da aplicação da transformada discreta do cosseno (Diniz, et al., 2004).

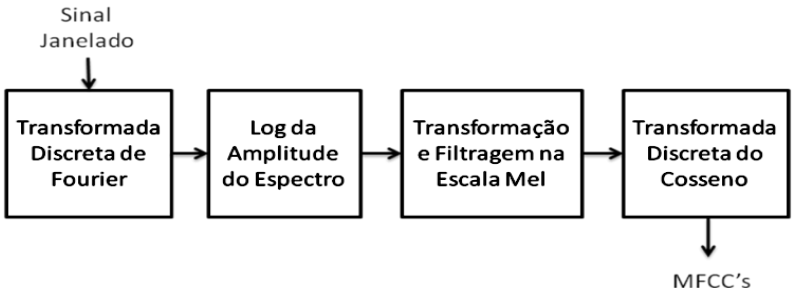


**Figura 3. Fluxograma de extração dos MFCC's.**

## 3.3 Histograma Rítmico

A Figura 4 descreve um sistema para a obtenção do histograma rítmico.

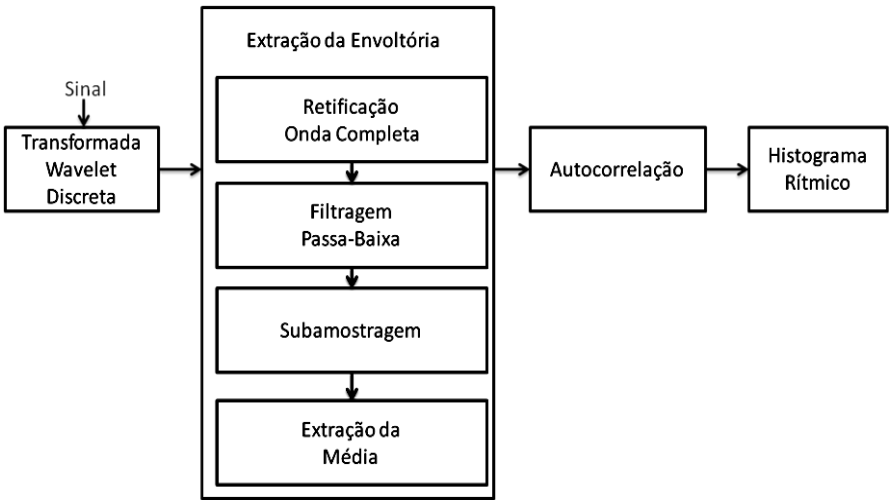


**Figura 4. Fluxograma de criação do Histograma rítmico.**

A descrição das etapas para extração da extração envoltória do sinal encontrada em Tzanetakis e Cook, (2002) é mostrada abaixo.

Retificação de onda completa: aplicado a fim de extrair a envoltória temporal do sinal, conforme Eq.05.

Filtragem Passa – baixa: para garantir que o teorema da amostragem seja preenchido, um filtro passa-baixa é usado como um filtro *anti-aliasing* para reduzir a largura de banda do sinal antes do sinal ser reduzido, o processo global (filtro passa-baixa, então subamostragem) é chamado dizimação. A Eq.06 mostra como se obter a filtragem. Para essa implementação $\alpha = 0.99$.

Redução da resolução (subamostragem), processo de redução da taxa de amostragem de um sinal, descrito na Eq.07. Feito normalmente para reduzir a taxa de dados ou o tamanho dos dados e conseqüentemente o esforço computacional para seu processamento. Para essa implementação k=16. Seguido pela extração da média e autocorrelação.

### 3.4 Vetor de características

A metodologia utilizada para classificação em gêneros musicais adota a extração do vetor de médias e variâncias da taxa de cruzamento por zero, energia em faixas de frequência e dos cinco primeiros coeficientes mel cepstrais além de quatro parâmetros do histograma rítmico (amplitude relativa do primeiro e do segundo picos do histograma, razão da amplitude do segundo pico dividido pela amplitude do primeiro pico, Soma do histograma) totalizando 22 parâmetros adotados para alimentar o classificador neural.

## 3.5 Reconhecimento de padrões

A partir das características extraídas, procedemos ao reconhecimento de padrões. O vetor de características alimentou uma rede neural artificial supervisionada do tipo PMC (Perceptrons de múltiplas camadas). As amostras foram divididas em três grupos: treinamento, teste e validação. Foram treinadas redes neurais com dois tipos de classificadores, a fim de comparar os resultados e definir a arquitetura que provém o melhor desempenho.

Foram utilizados dois tipos de classificadores neurais, um com neurônios de função de ativação linear e outro com neurônio tipo tangente hiperbólica. Ambos classificadores têm uma camada escondida e cinco neurônios de saída (um para cada classe), utilizando treinamento em modo batch com o algoritmo de Lenvenberg-Marquardt. A rede não-linear é chamada R1 e a linear R2. O número de neurônios da camada oculta foi escolhido após testes do desempenho de discriminação. Considera-se para fim de classificação que, um determinado gênero musical é corretamente classificado se o neurônio ativado corresponde a esse gênero, ou seja, o seu valor é o mais alto dentre os outros neurônios. Se não há duvida do resultado a musica é bem classificada (MALHEIRO, 2003). A Figura 5 mostra um diagrama dos classificadores neurais utilizados.

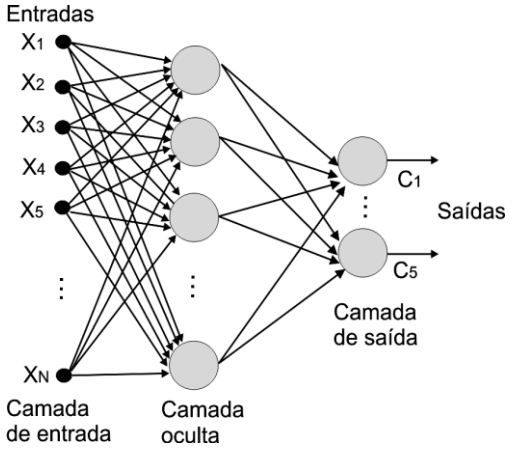


**Figura 5. Diagrama do classificador neural utilizado.**

Para escolha do número de neurônios na camada escondida foram realizadas diversas inicializações da rede neural, variando-se o número de neurônios escondidos. Testou-se várias redes com arquitetura R1 e R2 e descobriu-se que o número total de neurônios na camada intermediária influência fortemente os resultados para redes com arquitetura R1 como pode-se observar na tabela 1. Já nas redes com arquitetura R2 observa-se uma "saturação" a partir da rede com 5 neurônios.

## 4. RESULTADOS

Na Figura 7 são mostrados os histogramas rítmicos estimados para arquivos de diferentes gêneros musicais. O histograma indica os padrões rítmicos mais marcantes da música. Por exemplo, a música de MPB analisada apresenta uma batida mais lenta e marcante. O resultado para o samba também indica batidas lentas, porém há uma batida secundária (auxiliar) de grande importância para o padrão rítmico. Considerando os arquivos de blues e reggae, percebe-se que a batida principal é mais rápida e no blues há ainda uma maior influência de outras batidas auxiliares.

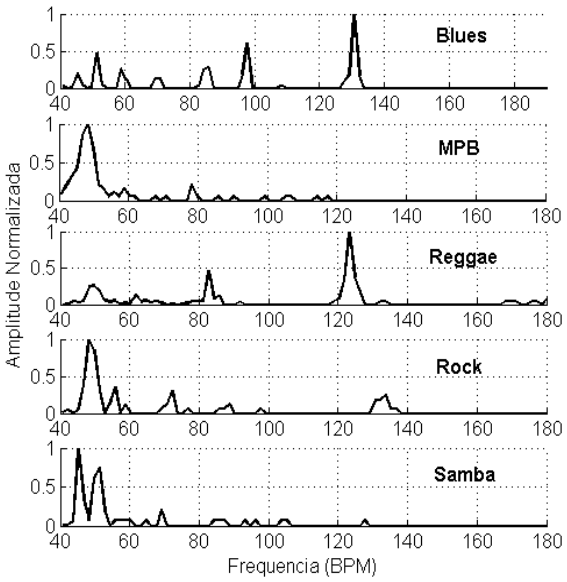


**Figura 7. Histogramas rítmicos obtidos para diferentes gêneros musicais.**

O julgamento do processo de erro e desempenho foi determinante para a escolha uma RN com 13 neurônios na camada escondida e cinco neurônios na camada de saída usando arquitetura R1. Os valores máximos da taxa de acerto médio obtidos variando-se o número de neurônios ocultos são mostrados na Tabela 1.

**Tabela 1. Desempenho (acerto médio em %) obtido variando-se o número de neurônios na camada oculta do classificador neural R1 em (a) e R2 em (b).**

| **Número Neurônios** | **2** | **5** | **8** | **10** | **13** | **14** | **15** |
|---|---|---|---|---|---|---|---|
| **Acerto Médio** | 60.3 | 70.7 | 73.6 | 75.9 | 76.4 | 75.1 | 74.8 |

**(a)**

| **Números Neurônios** | **2** | **5** | **8** | **10** | **13** | **14** | **15** |
|---|---|---|---|---|---|---|---|
| **Acerto Médio** | 47,6 | 56 | 56 | 56 | 56 | 56 | 56 |

**(b)**

A matriz de confusão obtida a partir da utilização de um classificador neural de uma camada oculta com 13 neurônios usando arquitetura R1 e R2 é mostrada na Tabela 2. Pode-se verificar que alguns ritmos como samba e reggae apresentaram maior eficiência de discriminação, o que pode ser justificado por apresentarem características distintas em relação às demais classes analisadas. Uma característica particular do gênero MPB é que as músicas associadas ao mesmo apresentam, muitas vezes, influência forte de outros gêneros como reggae, samba e rock, o que é refletido numa maior confusão entre as referidas classes.

**Tabela 2. Matriz de Confusão para classificador R1em (a) e Matriz de Confusão para classificador R2 em (b).**

| **Gênero** | **Blues** | **MPB** | **Reggae** | **Rock** | **Samba** |
|---|---|---|---|---|---|
| **Blues** | **71.43** | 1.43 | 6.76 | 11.27 | 8.82 |
| **MPB** | 1.41 | **54.93** | 18.92 | 7.04 | 17.65 |
| **Reggae** | 0 | 9.86 | **83.78** | 2.82 | 4.41 |
| **Rock** | 5.71 | 5.63 | 5.41 | **76.06** | 7.35 |
| **Samba** | 0 | 2.99 | 0 | 1.41 | **95.59** |

**(a)**

| **Gênero** | **Blues** | **MPB** | **Reggae** | **Rock** | **Samba** |
|---|---|---|---|---|---|
| **Blues** | **74.28** | 8.57 | 0 | 17.14 | 0 |
| **MPB** | 20 | **54.29** | 5.71 | 20 | 0 |
| **Reggae** | 5.71 | 14.29 | **74.29** | 5.71 | 0 |
| **Rock** | 11.43 | 17.14 | 0 | **71.43** | 0 |
| **Samba** | 2.86 | 71.43 | 8.57 | 0 | **5.71** |

**(b)**

## 5. CONCLUSÃO

Neste trabalho foi proposto um sistema automático para classificação de arquivos de áudio digital em diferentes gêneros musicais. De cada arquivo foram estimadas características como o histograma rítmico, os coeficientes mel-cepstrais e a taxa de cruzamento por zero. Classificadores neurais supervisionados foram utilizados para produzir a discriminação entre os gêneros. Considerando cinco classes musicais diferentes, o sistema (utilizando a rede não-linear) foi capaz de obter boa eficiência de discriminação, produzindo uma taxa de acerto médio da ordem de 76%. A "saturação" da rede com arquitetura R2 (de neurônios lineares) pode ser explicado pelo fato da arquitetura não poder mais diferenciar os padrões, pois estes não eram linearmente separáveis. Em futuros trabalhos pretende-se aumentar o número de gêneros musicas e também avaliar a relevância de cada parâmetro estimado para a discriminação entre as classes. Outro aspecto a ser abordado é o teste de diferentes arquiteturas de classificadores neurais.

## 6. AGRADECIMENTOS

À FAPESB, pelo apoio financeiro, ao IFBA, pela infraestrutura; ao GPEND pelo auxílio técnico.

## 7. REFERÊNCIAS

DINIZ, P. S. R., DA SILVA, E. A. B. e LIMA NETTO, S. **Processamento Digital de Sinais**, Ed. Bookman, Porto Alegre, 2004.

HAYKIN, **S. Neural Networks and Learning Machines**, Prentice Hall, 3rd Ed. New Jersey, 2008.

KIM, H.-G., MOREAU, N. and SIKORA, T. **MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval**, John Wiley & Sons, Ltd., 2005.

LOGAN B. "**Mel frequency cepstral coefficients for music modeling**," in Proc. Int. Symp. Music Inf. Retrival, Olymouth, MA, 2000.

LU, L., ZHANG, H. J. and JIANG, H. **Content Analysis for Audio Classification and Segmentation**, IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 7, 2002.

MALHEIRO, R. M. S. **Sistemas de classificação automática em gênero musicais**, dissertação de mestrado em Engenharia Informática, Universidade de Coimbra, Coimbra, Portugal, 2003.

MALLAT, S. **A Wavelet Tour of Signal Processing,** 2 ed., San Diego, Academic Press, 1998

MANO, F.R. da C. **Classificação e Segmentação de Áudio a partir de Fatores de Escala MPEG**, dissertação de mestrado em Informática, PUC-Rio, Rio de Janeiro-RJ, 2007.

MATITYAHO, B. and FURST, M. **Neural Network Based Model for Classification of Music Type,** Department of Electrical Engineering-Systems Faculty of Engineering, Tel Aviv University, Tel Aviv, Israel, 1995.

MENG A.**Temporal feature integration for music organization**, Ph.D, dissertation, Informatics and Mathematical Modeling Technical Univ. Denmark, Lyngby, Denmark, 2006.

PEREIRA, E. M. **Estudos sobre uma Ferramenta de Classificação Musical**, dissertação de mestrado em Engenharia Elétrica, UNICAMP, Campinas-SP, 2009.

TZANETAKIS, G. and COOK, P. **Musical genre classification of audio signals**, IEEE Transactions on Speech and Audio Processing, 10(5):293-302, 2002.