Classificação de Gêneros Musicais por Texturas no Espaço de Frequência

Yandre M. G. Costa[1,2], Luiz S. Oliveira[2], Alessandro L. Koerich[3], Fabien Gouyon[4]

[1]Departamento de Informática
Universidade Estadual de Maringá (UEM)*
Av. Colombo, 5790 – 87020-900 – Maringá – PR – Brazil

[2]Departamento de Informática
Universidade Federal do Paraná (UFPR)†
Curitiba – PR – Brazil

[3]Programa de Pós-graduação em Informática
Pontifícia Universidade Católica do Paraná (PUC-PR)
Curitiba – PR – Brazil

[4]Instituto de Engenharia de Sistemas e Computadores do Porto – (INESC-Porto)
Porto – Portugal

yandre@din.uem.br, lesoliveira@inf.ufpr.br, alekoe@ppgia.pucpr.br,

fgouyon@inescporto.pt

Abstract. *This paper describes an attempt to perform automatic musical genre classification based on spectrograms extracted from segments of digital music pieces, taken from the "Latin Music Database". Feature vectors with textural characteristics were extracted from digital images of spectrograms by using gray level co-occurrence matrix. The recognition rate obtained with a Support Vector Machine classifier was 60,11%. This rate is slightly higher than other obtained with different approaches recently performed over the same database. In addition, the classifier proposed here was combined with another classifier. The results obtained show an recognition rate about 66,11% and the upper limit found combining this two classifiers is about 75%.*

Resumo. *Este artigo descreve a classificação de gêneros musicais utilizando espectrogramas extraídos de músicas. Foram gerados espectrogramas a partir do sinal de áudio das músicas, tomadas da "Latin Music Database". A partir destes espectrogramas foram extraídas características de textura com o uso de matriz de co-ocorrência. Foi utilizado Support Vector Machine para a classificação e o desempenho final do esquema de classificação proposto atingiu uma taxa média de acerto de 60,11%. Este desempenho é superior ao de outras abordagens recentemente aplicadas sobre a mesma base. O classificador aqui proposto foi ainda combinado com um outro classificador descrito na literatura. A combinação atingiu taxa de acerto de 66,11%, com limite superior de 75%.*

*Este trabalho é financiado pela Fundação Araucária (Chamada de projetos 01/2009 - protocolo 16.767), †CNPq e CAPES.

1. Introdução

Com a rápida expansão da Internet uma imensa massa de dados oriundos de diferentes fontes tem se tornado disponível *on-line*. A gestão da informação em grandes volumes de dados multimídia está arrolada entre os grandes desafios da pesquisa em computação no Brasil [Lucena et al. 2006], estabelecidos no seminário realizado em 2006 e com isso o desenvolvimento des pesquisas que busquem soluções neste contexto se mostra bastante oportuno. Estudos apontam que, em 2007, a massa de dados digitais espalhada ao redor do mundo consumia aproximadamente 281 exabytes e que, em 2011, esse volume deve se multiplicar por 10 [Gantz et al. 2008]. Porém, boa parte destas informações não segue um padrão de apresentação e não está disponível de maneira estruturada, o que torna muito difícil fazer uso adequado das mesmas.

Devido a isso, tarefas como busca, recuperação, indexação, extração e sumarização automática dessas informações se tornaram problemas importantes sobre os quais muitas pesquisas têm sido realizadas. Neste contexto uma área de pesquisa que tem crescido nos últimos anos é a de recuperação de informações multimídia, que visa criar ferramentas capazes de organizar e gerenciar essa grande quantidade de informações. No momento, a maior parte das informações sobre dados multimídia são organizadas e classificadas baseadas em meta-informações textuais que são associadas ao seu conteúdo, como é o caso dos rótulos ID3 incorporados nos arquivos de áudio no formato MP3. Apesar destas informações serem relevantes para as tarefas de indexação, busca e recuperação, elas dependem da intervenção humana para gerá-las e, posteriormente, associá-las aos arquivos multimídia.

A classificação automática de gêneros musicais pode auxiliar ou substituir o usuário humano nesse processo, assim como prover um componente importante para um sistema de recuperação de informações para músicas, podendo também tornar menos subjetivo este processo de atribuição. Estudos sobre comportamento de usuários de Sistemas de Recuperação de Informação Multimídia, também conhecidos como sistemas MIR (*Multimidia Information Retrieval*), indicam que o gênero musical é o segundo item mais fornecido nas *queries* de busca.

A idéia de classificação automática de gêneros musicais como uma tarefa de reconhecimento de padrões de sinais de músicas foi apresentada em [Tzanetakis and Cook 2002]. Neste trabalho, foi proposto um conjunto abrangente de características para representar o sinal de áudio. Os experimentos foram avaliados em uma base de dados contendo 1.000 músicas de 10 gêneros distintos, sendo 100 músicas de cada gênero. O acerto obtido inicialmente nessa base foi de cerca de 60%. É importante observar que os experimentos foram realizados utilizando apenas os primeiros 30 segundos de cada música.

Em [Li et al. 2003], foi realizado um estudo comparativo para a classificação automática de gêneros musicais baseada em conteúdo entre o conjunto de características propostas por Tzanetakis e Cook e um novo conjunto de características. Os experimentos foram realizados em duas bases de dados, a primeira foi a mesma utilizada nos experimentos de Tzanetakis e Cook e uma segunda base contendo 756 músicas de cinco gêneros. Um aspecto importante dessa segunda base de dados é que as características foram extraídas do segmento composto pelo segundo 31 ao segundo 60, em vez dos primeiros trinta segundos. As conclusões dos experimentos realizados neste trabalho mostram a

melhor taxa de classificação obtida foi de 78% com a primeira base de dados e de 74% com a segunda base.

No trabalho de [Costa et al. 2004] foi proposto um novo método para a classificação automática de gêneros musicais, baseado na extração de características de três segmentos do sinal de áudio. As características foram extraídas do início, meio e fim da música. Para cada segmento foi treinado um classificador. As saídas fornecidas por cada classificador individualmente foram combinadas utilizando a regra de votação majoritária. Os classificadores utilizados foram MLP (Redes Neurais do tipo *Multi-Layer Perceptron*) e k-NN (k *Nearest Neighbor*). Uma continuação deste trabalho foi apresentada por [Koerich and Poitevin 2005] e os resultados obtidos mostraram uma melhora na taxa de acerto em relação aos segmentos individuais. O uso de mais de um segmento passou a ser utilizado em outros trabalhos subsequentes. Uma vantagem bastante clara que se obtém com o uso desta abordagem consiste no fato de que ela permite colher uma amostragem melhor do sinal, podendo captar variações presentes ao longo do mesmo que eventualmente um único segmento não conseguiria captar. Um único segmento de uma música pode ter características mais próximas às de um gênero diferente daquele identificado em seu rótulo, utilizando mais de um segmento e fazendo a fusão dos descritores extraídos de cada um deles, o erro cometido em um segmento é diluído e o impacto que ele provocaria na classificação tende a diminuir.

Um problema correlacionado foi apresentado por [Yu and Slotine 2009]. Neste trabalho os autores apresentam uma tentativa de reconhecer instrumentos musicais através de imagens de texturas caracterizadas nos espectrogramas obtidos a partir do sinal de áudio colhido de 8 tipos diferentes de instrumentos musicais. Neste trabalho, em que não ocorre a mistura polifônica naturalmente presente no sinal de músicas, os resultados mostram um acerto em torno de 85% para o método. Este trabalho inspirou o desenvolvimento dos experimentos que serão relatados no presente artigo, na medida em que suscitou a hipótese de que hajam informações presentes nas imagens caracterizadas pelos espectrogramas gerados a partir do sinal de áudio de musicas que permitam classificá-las de acordo com seus respectivos gêneros musicais.

Em [Lopes et al. 2010] é apresentado um método para o reconhecimento de gêneros musicais, aplicado sobre gêneros latinos, baseado na seleção de instâncias de vetores com características de tempo curto e características de baixo nível do sinal de áudio de músicas. O método empregou o *"artist filter"* na formação dos *datasets*. Proposto em [Pampalk et al. 2005], o *"artist filter"* determina que os títulos musicais de um mesmo interprete sejam todos atribuídos à um mesmo conjunto quando da divisão da base de músicas. Com isto, títulos de um mesmo interprete não são comparados em momento algum, pois nunca estarão simultaneamente no conjunto de treino e de teste. Esta técnica aumenta consideravelmente a dificuldade de se alcançar boas taxas de reconhecimento, entretanto, pode conferir maior robustez ao classificador, induzindo a classificação a ser orientada efetivamente a gênero, em vez de ser orientada a interprete. A base de dados empregada foi a *Latin Music Database (LMD)*, uma base particularmente desafiadora. Ao final, obteve-se taxa de acerto de 59,6%.

Este trabalho descreve uma tentativa de automatizar o reconhecimento de gêneros musicais latinos através da análise de texturas presentes nas imagens de espectrogramas gerados a partir do sinal de áudio de músicas. Os espectrogramas podem repre-

sentar dados referentes a um sinal de áudio no domínio de tempo e frequência, e pode ser um instrumento bastante útil para discernir detalhes importantes acerca do mesmo [French and Handy 2007]. A extração de atributos descritores das imagens dos espectrogramas caracteriza um novo formato de atributo descritor. A fim de verificar a complementaridade deste formato com outros empregados em trabalhos de classificação de gêneros musicais, foi realizada a combinação do classificador aqui proposto com o classificador descrito em [Lopes et al. 2010].

A principal contribuição deste trabalho é introduzir uma investigação acerca da complementaridade de um formato de descritor, extraído de espectrogramas gerados a partir do sinal de áudio, em relação a outros formatos de descritores já utilizados para os propósitos de classificação automática de gêneros musicais. Neste sentido, serão apresentadas estratégias estabelecidas dentro deste processo, como a divisão da imagem do espectrograma em zonas, fato que permitiu discriminar informações do sinal representantes de diferentes faixas de frequência, proporcionando melhores resultados em relação a taxas de acerto. Os experimentos baseados no classificador aqui proposto atingiram taxas médias de acerto de 60,11%, com desvio padrão de 9,06. Esta taxa é comparável a de outros trabalhos recentes aplicados sobre esta mesma base, utilizando o mesmo conjunto de dados. Resultados experimentais aqui descritos ainda mostram que esse tipo de estratégia é capaz de gerar resultados complementares aos sistemas tradicionais de extração de características de áudio, possibilitando o aumento das taxas de reconhecimento através da combinação de classificadores. A combinação do classificador aqui proposto com o classificador apresentado em [Lopes et al. 2010] gerou taxa média de acerto de 66,11% e o limite superior encontrado para a combinação foi de 75%.

A organização deste trabalho encontra-se da seguinte forma: na seção 2 são descritos detalhes acerca da extração de características a partir das imagens de espectrogramas empregada neste trabalho; a seção 3 aborda o esquema de classificação aqui proposto; a seção 4 mostra os resultados experimentais; e a seção 5 apresenta as conclusões e possíveis trabalhos futuros.

2. Extração de Características

Esta seção descreve detalhes acerca da extração de características das musicas com o uso de espectrogramas. Inicialmente, será abordada a geração dos espectrogramas a partir do sinal de áudio. Em seguida, serão detalhadas as técnicas de processamento de imagens empregadas para se obter atributos descritores das características.

2.1. Fonte de Dados para Geração dos Espectrogramas

Para realizar os experimentos aqui descritos, tomou-se como fonte de dados a *Latin Music Database*, uma base particularmente desafiadora, composta por músicas latinas de 10 diferentes gêneros musicais (Axé, Bachata, Bolero, Forró, Gaúcha, Merengue, Pagode, Salsa, Sertaneja, Tango) apresentada em [Silla Jr et al. 2008]. A LMD é baseada na percepção de especialistas humanos e foi desenvolvida com propósitos também voltados à classificação automática de gêneros musicais. No total, foram tomadas 900 musicas desta base, divididas em 3 conjuntos, cada um com 300 músicas, doravante denominados *dataset1*, *dataset2* e *dataset3*. Em cada um destes conjuntos foram incluídas 30 músicas de cada gênero e foi empregado o *"artist filter"* na distribuição dos títulos entre os conjuntos. Devido ao uso do *"artist filter"* não foi possível utilizar todas as músicas da LMD

no experimento, já que existe uma diferença significativa entre as quantidades de músicas de cada intérprete presentes na base.

Adicionalmente, adotou-se a estratégia descrita em [Costa et al. 2004] na qual os autores utilizam três diferentes segmentos dispersos ao longo do sinal do áudio. Para isto, foram tomados segmentos do início, meio e final de cada música. A fim de evitar que efeitos como *"fade in"*, *"fade out"* e vibração de platéia em músicas gravadas ao vivo tornassem trechos da amostra pouco discriminantes, utilizou-se como amostra do início da música o segmento compreendido entre o segundo 11 e o segundo 20 da música, e como amostra do final da música o segmento compreendido entre o segundo n-20 e o segundo n-11, sendo n a duração da música em segundos. O segmento central foi extraído do intervalo compreendido entre o segundo m-5 e o segundo m+5, sendo m o segundo que se encontra exatamente no meio do sinal da música. A figura 1 ilustra esta estratégia.

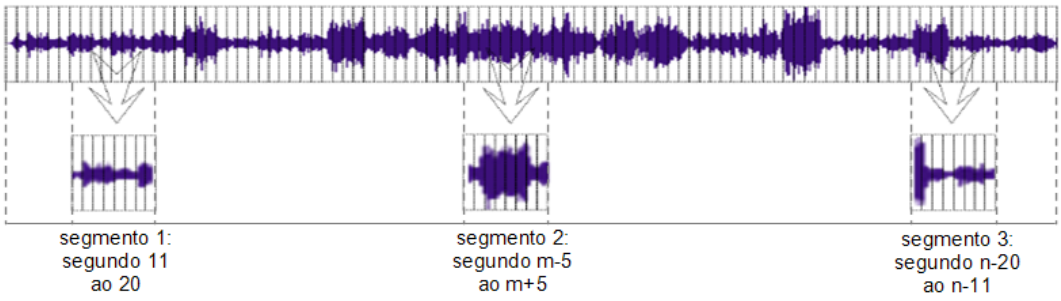

Figura 1. Extração de segmentos do sinal.

Foram gerados espectrogramas a partir de arquivos de áudio com a concatenação dos três segmentos para cada música utilizando-se o software SoX 14.3.0 (*Sound eXchange*), disponível em *http://sox.sourceforge.net*. As imagens dos espectrogramas foram geradas com valores *default* de largura (800 pixels) e altura (550 pixels).

Depois de extraídas as imagens dos espectrogramas das músicas, elas foram convertidas para níveis de cinza para melhor se adequarem aos processos subsequentes. A figura 2 mostra três exemplos de espectrograma utilizados nos experimentos. Estes espectrogramas referem-se a músicas de diferentes gêneros musicais e foram tomados aleatoriamente da base.

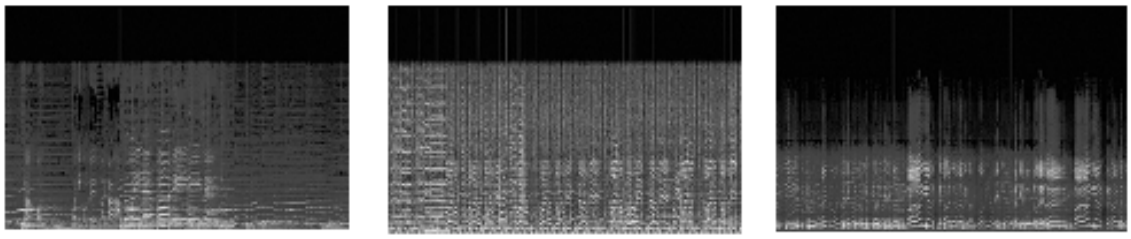

Figura 2. Exemplos de espectrograma.

2.2. Divisão das imagens dos espectrogramas em zonas

Depois de segmentadas as imagens dos espectrogramas, a fim de se remover legendas e outras áreas que não a de interesse, foi estabelecida uma divisão das imagens em zonas

para que se pudesse preservar parcialmente informações espaciais relacionadas às características a serem extraídas.

As imagens dos espectrogramas dos 3 segmentos foram divididas em 10 zonas. Esta divisão foi estabelecida após a realização de diversos testes com quantidades de zonas variando de 1 a 40. Os resultados indicaram a melhor taxa de acerto na situação em que os espectrogramas foram divididos em 10 zonas. O gráfico mostrado na figura 3 permite verificar a importância do zoneamento das imagens empregadas neste processo de classificação. O gráfico apresenta a evolução das taxas de acerto obtidas com os diferentes divisões em zonas testadas: 1 zona, 3 zonas, 5 zonas, 8 zonas, 10 zonas, 15 zonas, 20 zonas, 30 zonas e 40 zonas.

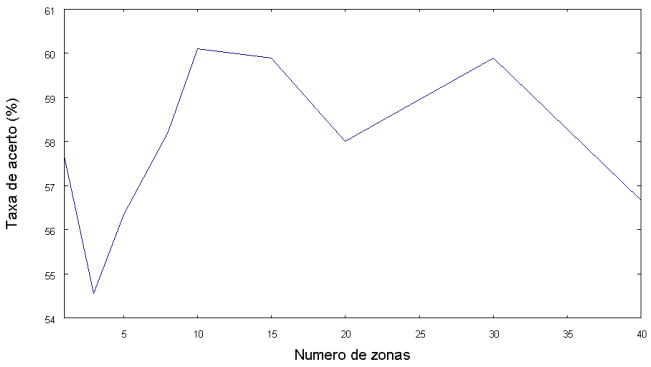

Figura 3. Evolução das taxas de acerto em função de diferentes quantidades de zonas.

Além de melhor taxa de acerto, a divisão em 10 zonas produz descritores com custo de armazenamento e tempo de processamento consideravelmente reduzidos em relação as outras divisões que apresentaram taxas de acerto comparáveis às dela. Com a divisão em zonas foi possível discriminar as características presentes em diferentes faixas de frequência representadas no espectrograma. Esta divisão está ilustrada na figura 4. De cada zona foram extraídos valores correspondentes às características descritas na próxima subseção.

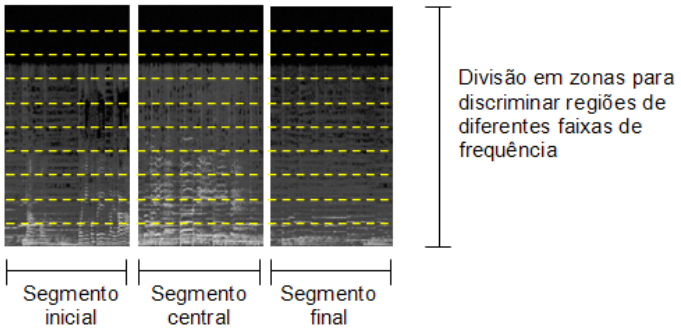

Figura 4. Esquema de divisão das imagens dos espectrogramas em zonas.

2.3. Matriz de Co-ocorrência

Partindo-se do princípio de que podem existir características texturais presentes nas imagens dos espectrogramas que sejam úteis na classificação de gêneros musicais, optou-se

por extraí-las empregando-se matrizes de co-ocorrência. Matriz de co-ocorrência (MC) é um método estatístico para descrição de texturas baseado na ocorrência repetida de níveis de cinza em uma textura, foi proposto em [Haralick et al. 1973] e ainda é empregado com sucesso em diferentes domínios de aplicação.

A partir da matriz de co-ocorrência, pode-se extrair várias medidas relacionadas a características de uma textura. Neste trabalho, para descrever as texturas encontradas nos espectrogramas, foram extraídas as medidas de contraste, energia, entropia, correlação, homogeneidade, momento de terceira ordem e probabilidade máxima. Este conjunto de características foi escolhido depois de alguns testes envolvendo algumas diferentes combinações de características, ao final, esta combinação apresentou o melhor desempenho.

A MC para imagens em tons de cinza armazena a probabilidade de que dois valores de intensidades de cinza estejam envolvidos por uma determinada relação espacial. Nos experimentos aqui descritos, os tons de cinza foram quantizados para 64 valores. Assim, cada MC gerada foi uma matriz de ordem 64×64. Parâmetros como a distância *d* entre os pixels e o ângulo θ caracterizado pela orientação da reta que passa pelos pixels são importantes na caracterização da relação espacial. Nos experimentos aqui descritos, os melhores resultados em termos de taxa de acerto foram obtidos com *d*=1, e foram utilizadas as orientações $0^o, 45^o, 90^o$ e 135^o para o ângulo θ. Sendo *p(i,j)* a probabilidade de ocorrência das intensidades de cinza *i* e *j* observando a distância *d*=1 e com um dado ângulo θ, as características utilizadas neste trabalho são encontradas pelas seguintes equações:

$$Contraste = \sum_{i=1}^{64}\sum_{j=1}^{64}(i-j)^2\, p(i,j) \quad (1)$$

$$Energia = \sum_{i=1}^{64}\sum_{j=1}^{64}(p(i,j))^2 \quad (2)$$

$$Entropia = -\sum_{i=1}^{64}\sum_{j=1}^{64} p(i,j)\; log\; p(i,j) \quad (3)$$

$$Homogeneidade = \sum_{i=1}^{64}\sum_{j=1}^{64}\frac{p(i,j)}{1+(i-j)^2} \quad (4)$$

$$Momento\ de\ terceira\ ordem = \sum_{i=1}^{64}\sum_{j=1}^{64} p(i,j)(i-j)^3 \quad (5)$$

$$Probabilidade\ máxima = \sum_{i=1}^{64}\sum_{j=1}^{64} max\; p(i,j) \quad (6)$$

$$\textit{Correlação} = \frac{p(i,j)-\mu_x\mu_y}{\sigma_x^2\sigma_y^2} \quad (7)$$

onde $\mu_x = \sum_{i=1}^{64} i \times p_x(i),\ p_x(i) = \sum_{j=1}^{64} p(i,j),\ \sigma_x^2 = \sum_{i=1}^{64}(i-\mu_x)^2 p_x(i)$, $\mu_y = \sum_{j=1}^{64} j \times p_y(j),\ p_y(j) = \sum_{i=1}^{64} p(i,j)$ *e* $\sigma_y^2 = \sum_{j=1}^{64}(j-\mu_y)^2 p_y(j)$.

Depois de encontrados os valores de cada uma destas características para cada uma das zonas descritas na subseção 2.2, foram finalmente formados os vetores com atributos descritores das músicas para a realização subsequente do processo de classificação. Para cada zona da imagem do espectrograma foram extraídas as 7 características em cada uma das 4 orientações previstas nesta subseção. Assim, para cada zona foi formado um vetor com um total de 28 características. A próxima seção descreve o emprego destes vetores no processo de classificação.

3. Esquema Geral da classificação Baseada em Espectrogramas

O classificador empregado neste trabalho foi o *Support Vector Machine* (SVM)[Vapnik 2000]. Para isto, foi utilizado o software LIBSVM [Chang and Lin 2001]. Para a execução das tarefas de classificação, foram tomados cuidados apropriados visando a normalização dos dados a fim de evitar que eventuais discrepâncias entre os valores de mesmas características em vetores diferentes prejudicassem o processo de classificação. A normalização foi feita de forma que os valores ficassem compreendidos em uma escala de variação entre -1 e 1. Além disso, os parâmetros C e γ do kernel Gaussiano foram otimizados utilizando uma busca gulosa com validação cruzada a fim de se alcançar melhores resultados.

Foi realizada a validação cruzada entre os três *datasets*, com 300 músicas cada, previamente descritos. Os conjuntos de teste e treinamento utilizados nos experimentos foram formados pelos vetores de características extraídos de cada zona da imagem do espectrograma e a divisão de conjuntos utilizada foi a mesma empregada em [Lopes et al. 2010]. Assim, considerando que de cada música foram extraídos 3 segmentos, de cada segmento de música foi extraído um espectrograma e que, a partir de cada espectrograma foram criadas 10 zonas, o conjunto de teste foi formado por 9.000 vetores enquanto o conjunto de treinamento foi formado por 18.000 vetores em cada execução do classificador. A figura 5 ilustra o esquema geral da classificação.

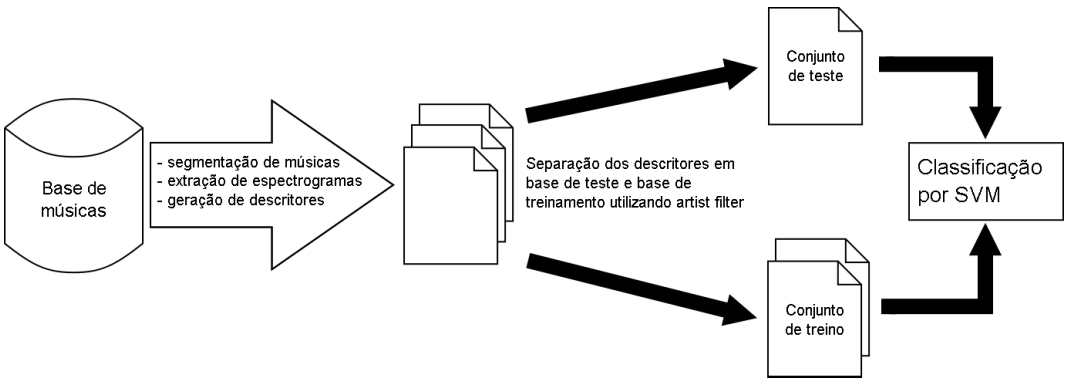

Figura 5. Esquema geral da classificação automática de gêneros musicais.

Depois de executado o classificador SVM para cada um dos cruzamentos previstos na validação cruzada, quais sejam: dataset1 $\times$ dataset2+dataset3, dataset2 $\times$ dataset1+dataset3 e dataset3 $\times$ dataset1+dataset2, obteve-se como resultado uma classe, neste caso gênero musical, atribuída ao vetor extraído de cada zona. A classificação final de cada música foi encontrada através da votação entre as classes atribuídas para as 30 zonas da mesma. Assim, o gênero para o qual foi atribuído o maior número de zonas

foi escolhido como gênero ao qual a música pertence. A figura 6 mostra o esquema de votação.

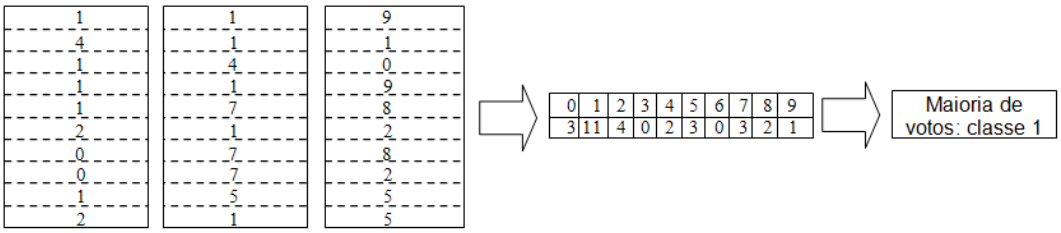

Figura 6. Votação para a decisão final.

Este esquema de votação é frequentemente empregado em trabalhos similares, que envolvem o reconhecimento de gêneros musicais. Em geral, ele oferece bons resultados, já que é bastante comum que partes de uma música sejam parecidas com um gênero ao qual ela não pertence. O esquema de votação ajuda a evitar algumas confusões que poderiam ser causadas por este motivo, já que toma a decisão com base na maioria das zonas criadas.

A próxima seção descreve detalhes acerca da combinação deste classificador com outro baseado na seleção de instâncias de vetores com catacterísticas de tempo curto e características de baixo nível do sinal de áudio de músicas.

4. Resultados Experimentais

A LMD tem como característica o fato de reunir muitos gêneros com significativa similaridade entre si no que diz respeito a instrumentalização, estrutura rítmica e conteúdo harmônico. Isto acontece porque muitos gêneros presentes na base são originários de um mesmo país ou de países com grandes semelhanças no que diz respeito a aspectos culturais. Este fato faz com que a tentativa de discriminar tais gêneros automaticamente seja particularmente desafiadora. As próximas subseções descrevem os resultados obtidos.

4.1. Resultados obtidos com o classificador proposto

A tabela 1 mostra o percentual médio de acerto obtido por gênero com a estratégia de classificação aqui apresentada, seus respectivos desvios padrão e a média geral. Considerando que este trabalho tem como objetivo verificar a complementaridade de um método de classificação baseado em características particulares em relação a outras já conhecidas, não houve preocupação voltada à realização de testes estatísticos mais apurados que permitissem uma efetiva comparação com resultados já apresentados na literatura.

A tabela 2 mostra a matriz de confusão acumulada após a execução da validação cruzada entre os 3 *datasets*. A matriz de confusão mostra o número de vezes que os títulos musicais pertencentes aos gêneros identificados nas linhas foram classificados em cada um dos gêneros identificados pelas colunas. A diagonal principal da matriz corresponde aos casos em que houve acerto. Nesta matriz pode-se perceber a presença de uma taxa de confusão especialmente acentuada entre os gêneros de origem brasileira. É importante lembrar que o número total de títulos pertencentes a cada gênero, considerando os três *datasets* envolvidos nos experimentos, é igual a 90.

Gênero	**Taxa de acerto**	σ
axé	73,33%	8,82
bachata	82,22%	15,03
bolero	64,44%	8,39
forró	65,56%	8,39
gaúcha	35,56%	5,09
merengue	80,00%	6,67
pagode	46,67%	17,64
salsa	42,22%	6,94
sertaneja	17,78%	6,94
tango	93,33%	6,67
geral	**60,11%**	**9,06**

Tabela 1. Desempenhos médios por gênero e seus respectivos desvios padrão

	axé	bac	bol	for	gaú	mer	pag	sal	ser	tan
axé	66	2	3	0	0	3	8	0	3	5
bachata	1	74	4	1	0	2	4	2	0	2
bolero	1	4	58	7	2	0	5	0	0	13
forró	0	0	8	59	5	0	7	4	4	3
gaúcha	20	2	10	17	32	1	1	3	1	3
merengue	2	2	3	1	2	72	3	3	0	2
pagode	12	0	15	6	1	0	42	5	3	6
salsa	2	6	15	8	1	11	5	38	2	2
sertaneja	31	2	21	10	3	0	3	2	16	2
tango	1	0	4	1	0	0	0	0	0	84

Tabela 2. Matriz de confusão acumulada depois da validação cruzada

De forma geral, os resultados obtidos nos experimentos mostram taxas de acerto comparáveis às do primeiro trabalho realizado com a LMD [Silla Jr et al. 2008]. É importante observar que naquele trabalho não foi empregado o *"artist filter"* durante a divisão dos conjuntos para posterior classificação. Além disso, os experimentos aqui descritos apresentaram uma taxa de acerto média ligeiramente superior quando comparados com outros trabalhos recentemente realizados sobre a mesma base de dados, como [Lopes et al. 2010]. Entretanto, houve uma redução significativa do desvio padrão relacionado a taxa média de acerto entre os diferentes gêneros envolvidos no estudo.

4.2. Resultados obtidos com a combinação de classificadores

A fim de verificar em que medida os atributos descritores extraídos com o método aqui proposto podem ser complementares com os de abordagens mais tradicionais, realizou-se a combinação das saídas deste classificador com as do proposto em [Lopes et al. 2010]. Antes de descrever maiores detalhes acerca da combinação realizada, é válido ressaltar que os *datasets* empregados nos experimentos descritos neste trabalho foram compostos exatamente pelas mesmas músicas presentes nos *datasets* utilizados nos experimentos descritos em [Lopes et al. 2010]. Assim, entende-se que os resultados podem ser comparados e combinados de forma consistente e correta.

Para realizar a combinação foi necessário associar à saída dos classificadores probabilidades que indicam o grau de confiabilidade da mesma. Para o classificador aqui proposto (classificador 1), baseado em características extraídas dos espectrogramas, a probabilidade associada à cada gênero foi encontrada através da divisão do número de zonas atribuídas à este gênero pelo número total de zonas associadas a cada música.

Em [Lopes et al. 2010], o número total de instâncias extraídas de cada música no processo de classificação proposto é igual a 646. Para o cálculo das probabilidades associadas à cada gênero neste classificador (classificador 2), tomou-se o número de instâncias atribuídas ao gênero e dividiu-se por 646.

Foram testadas duas diferentes regras de combinação para atribuir a classificação final a cada título musical, as regras MAX e SUM [Kittler et al. 2002]. Na primeira, quando houve divergência entre as classes atribuídas pelos dois classificadores, foi estabelecida como classe aquela cujo classificador apresentou maior probabilidade associada à saída. Na segunda, em caso de divergência foi estabelecida como classe aquela cuja soma das probabilidades indicadas pelos dois classificadores apresentou maior valor. A figura 7 mostra o esquema empregado na combinação dos classificadores.

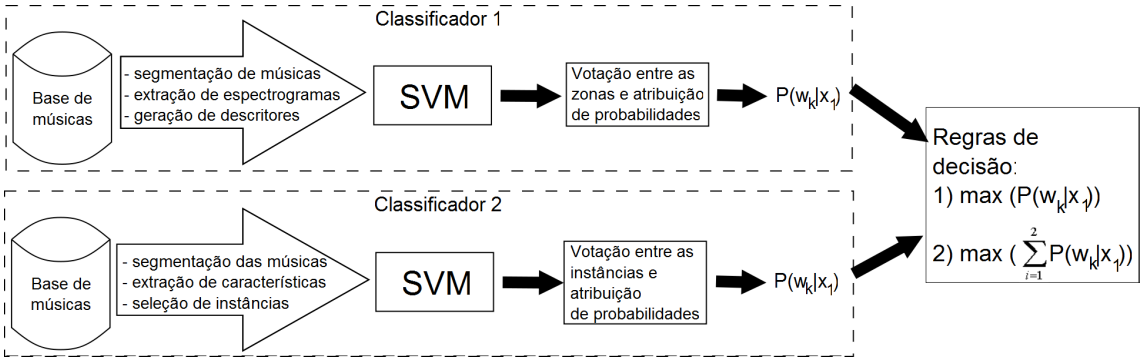

Figura 7. Esquema empregado na combinação dos classificadores.

A tabela 3 mostra as taxas médias de acerto obtidas por gênero e também a taxa geral quando foram empregadas as regras de decisão MAX e SUM.

Os resultados obtidos mostram que os atributos descritores propostos neste trabalho, baseados em características texturais extraídas de espectrogramas, não somente possuem algum poder de discriminação de gêneros musicais, como também possuem alguma complementaridade em relação a outros formatos de atributos descritores já propostos. Levando-se em consideração o fato de que a combinação dos classificadores apresentou as piores taxas médias de acerto nos gêneros brasileiros, o que já era esperado devido ao fato de que estes gêneros não possuem diferenças significativas em termos de instrumentalização, estrutura rítmica e conteúdo harmônico, acredita-se que este esquema pode proporcionar resultados ainda melhores em outras bases de músicas.

O desempenho médio geral de 66,11% de acerto, obtido com a regra de combinação SUM, é particularmente animador e com ele suscita naturalmente a hipótese de que outros esquemas de combinação ou a combinação com outros classificadores, diferentes dos aqui utilizados, pode trazer resultados ainda melhores. A tabela 4 mostra os limites superiores de acerto médios obtidos com os dois classificadores aqui empregados. Este resultado correponde à média entre as três diferetes situações em que cada

Gênero	Regra MAX		Regra SUM	
	Taxa de acerto	σ	Taxa de acerto	σ
axé	64,44%	19,53	74,44%	15,03
bachata	92,22%	6,94	90,00%	6,67
bolero	75,56%	15,40	77,78%	16,44
forró	44,44%	24,57	46,67%	26,03
gaúcha	47,78%	1,92	52,22%	3,85
merengue	85,56%	10,72	88,89%	8,39
pagode	56,67%	8,82	57,78%	1,92
salsa	47,78%	9,62	46,67%	12,02
sertaneja	36,67%	34,80	35,56%	24,11
tango	92,22%	6,94	91,11%	8,39
geral	**64,33%**	**13,93**	**66,11%**	**12,29**

Tabela 3. Desempenhos médios por gênero e geral na combinação dos classificadores com as regras MAX e SUM

um dos diferentes *datases* foi utilizado como conjunto de teste. O limite superior é encontrado considerando como corretamente classificados os itens em que pelo menos um dos classificadores envolvidos na combinação foi capaz de realizar a classificação corretamente. Com base nestes dados, que mostram que a média geral de acerto pode chegar a 75%, percebe-se que vale a pena investir em experimentos que envolvam a combinação dinâmica de classificadores.

Gênero	Taxa de acerto	σ
axé	81,11%	13,47
bachata	93,33%	6,67
bolero	81,11%	5,09
forró	66,67%	6,67
gaúcha	60,00%	12,02
merengue	88,89%	6,94
pagode	75,56%	3,85
salsa	62,22%	11,71
sertaneja	46,67%	17,64
tango	94,44%	5,09
geral	**75,00%**	**8,91**

Tabela 4. Limite superior combinando os dois classificadores

5. Conclusões e Trabalhos Futuros

O uso da LMD com o emprego de *"artist filter"* introduziu uma dificuldade adicional ao processo de classificação proposto neste trabalho. Mesmo assim, os resultados obtidos são comparáveis aos obtidos em outros trabalhos e, sob certos aspectos, ligeiramente superiores. A taxa de acerto média foi de 60,11% com desvio padrão de 9,06.

A combinação do classificador aqui proposto com um outro classificador, baseado num descritor de formato diferente, produziu resultados bastante animadores, 66,11% no

melhor caso (regra de combinação SUM). Este resultado mostra a presença de complementaridade do formato de descritor aqui proposto com outros já existentes. Existem boas perspectivas de se alcançar, em trabalhos futuros, taxas de acerto ainda melhores realizando a combinação com outros formatos de descritores ou mesmo utilizando outros esquemas de combinação com o mesmo classificador, já que o limite superior encontrado para esta combinação foi de 75%.

Trabalhos futuros envolvendo metaclasses também podem ser realizados, já que a matriz de confusão resultante dos experimentos mostra um alto índice de confusão especialmente entre músicas de gêneros brasileiros. A criação de metaclasses permitiria investir em técnicas especificamente voltadas a discriminação entre estes gêneros, buscando uma melhoria no desempenho final do classificador.

Adicionalmente, pretende-se aplicar os métodos aqui descritos em outras bases de músicas, a fim de identificar o comportamento destes métodos em outros cenários.

Referências

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Costa, C., Valle Jr, J., and Koerich, A. (2004). Automatic classification of audio data. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 562–567.

French, M. and Handy, R. (2007). Spectrograms: turning signals into pictures. *Journal of engineering technology*, 24(1):p. 32–35.

Gantz, J., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., and Toncheva, A. (2008). The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011. *IDC white paper, sponsored by EMC*.

Haralick, R., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on systems, man and cybernetics*, 3(6):p. 610–621.

Kittler, J., Hatef, M., Duin, R., and Matas, J. (2002). On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(3):p. 226–239.

Koerich, A. and Poitevin, C. (2005). Combination of Homogeneous Classifiers for Musical Genre Classification. *IEEE International Conference on Systems, Man, and Cybernetics*, pages 554–559.

Li, T., Ogihara, M., and Li, Q. (2003). A comparative study on content-based music genre classification. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 282–289. ACM.

Lopes, M., Gouyon, F., Koerich, A. L., and Oliveira, L. E. S. (2010). Selection of training instances for music genre classification. *ICPR 2010 - 20th International Conference on Pattern Recognition*.

Lucena, C., Medeiros, C., Lucchesi, C., Maldonado, J., Almeida, V., and outros (2006). Grandes Desafios da Pesquisa em Computação no Brasil - 2006-2016. *Relatório sobre o seminário Grandes Desafios da Pesquisa em Computação. São Paulo. SBC/CAPES/FAPESP.*

Pampalk, E., Flexer, A., and Widmer, G. (2005). Improvements of audio-based music similarity and genre classification. In *proc. ISMIR*, volume 5.

Silla Jr, C., Koerich, A., and Kaestner, C. (2008). The latin music database. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 451–456.

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):p. 293–302.

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer Verlag.

Yu, G. and Slotine, J. (2009). Audio classification from time-frequency texture. *ICASSP 2009 - International Conference on Acoustics, Speech and Signal Processing*, pages 1677–1680.