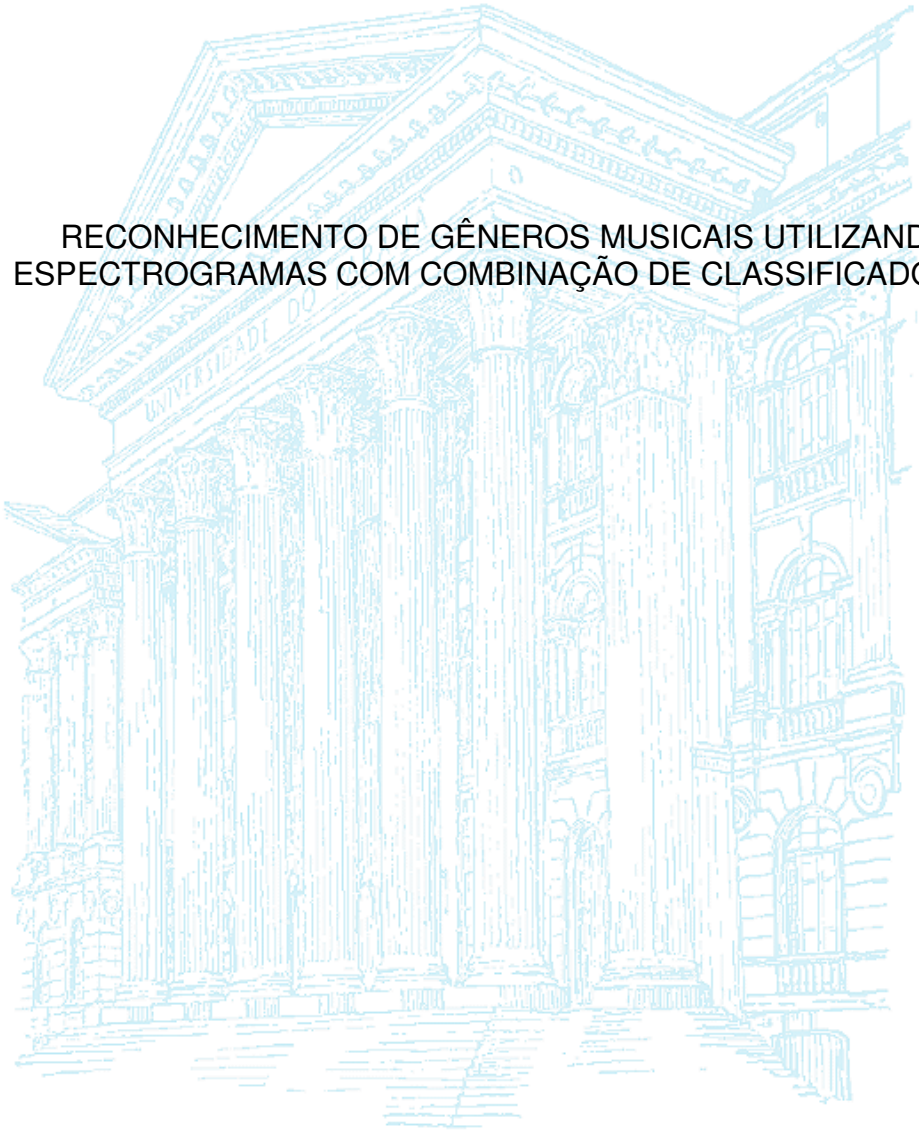
UNIVERSIDADE FEDERAL DO PARANÁ

YANDRE MALDONADO E GOMES DA COSTA

# RECONHECIMENTO DE GÊNEROS MUSICAIS UTILIZANDO ESPECTROGRAMAS COM COMBINAÇÃO DE CLASSIFICADORES



CURITIBA
2013

Y. COSTA RECONHECIMENTO 2013

YANDRE MALDONADO E GOMES DA COSTA

# RECONHECIMENTO DE GÊNEROS MUSICAIS UTILIZANDO ESPECTROGRAMAS COM COMBINAÇÃO DE CLASSIFICADORES

Texto apresentado ao Programa de Pós-Graduação em Informática, Setor de Ciências Exatas da Universidade Federal do Paraná, como requisito parcial para a obtenção do título de Doutor.
Orientador: Prof. Dr. Luiz Eduardo Soares de Oliveira
Co-orientador: Prof. Dr. Alessandro Lameiras Koerich

CURITIBA

2013

YANDRE MALDONADO E GOMES DA COSTA

# RECONHECIMENTO DE GÊNEROS MUSICAIS UTILIZANDO ESPECTROGRAMAS COM COMBINAÇÃO DE CLASSIFICADORES

Texto apresentado ao Programa de Pós-Graduação em Informática, Setor de Ciências Exatas da Universidade Federal do Paraná, como requisito parcial para a obtenção do título de Doutor.
Orientador: Prof. Dr. Luiz Eduardo Soares de Oliveira
Co-orientador: Prof. Dr. Alessandro Lameiras Koerich

CURITIBA

2013

C837r Costa, Yandre Maldonado e Gomes da
Reconhecimento de gêneros musicais utilizando espectogramas com combinação de classificadores / Yandre Maldonado e Gomes da Costa. – Curitiba, 2013.
106f. : il. color. ; 30 cm.

Tese (doutorado) - Universidade Federal do Paraná, Setor de Tecnologia, Programa de Pós-graduação em Informática, 2013.

Orientador: Luiz Eduardo Soares de Oliveira -- Co-orientador: Alessandro Lameiras Koerich.
Bibliografia: p. 99-106.

1. Música - Análise. 2. Classificação - Música. 3. Sistemas de reconhecimento de padrões I. Universidade Federal do Paraná. II. Oliveira, Luiz Eduardo Soares de. III. Koerich, Alessandro Lameiras. IV. Título.

CDD: 006.45

Ministério da Educação
Universidade Federal do Paraná
Programa de Pós-Graduação em Informática

ATA DA DEFESA DE TESE DE DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO DO ALUNO: YANDRE MALDONADO E GOMES DA COSTA

No dia 15 de agosto do ano de dois mil e treze, às 10:00 horas, no Departamento de Informática do Setor de Ciências Exatas da Universidade Federal do Paraná, foi realizada a sessão pública da defesa de Tese de Doutorado em Ciência da Computação do aluno Yandre Maldonado e Gomes da Costa. Estavam presentes, além do candidato, os Membros da Comissão Examinadora composta pelos Professores Luiz Eduardo Soares de Oliveira(Orientador), Alessandro Lameiras Koerich, Aura Conci, Carlos Nascimento Silla Junior e Daniel Weingaertner. Após a apresentação do trabalho do candidato, intitulado *"Reconhecimento de Gêneros Musicais utilizando Espectrogramas e Combinação de Classificadores"*, o mesmo foi arguido pela Comissão. A seguir, a Comissão reuniu-se em local reservado e decidiu, por unanimidade, pela aprovação do candidato condicionado as alterações sugeridas pela mesma. O resultado foi então comunicado ao candidato e aos presentes na sessão pública. A seguir, o Presidente declarou encerrada a sessão da qual eu, Jucélia Miecznikowski, Secretária do Programa de Pós-graduação em Informática, lavrei a presente Ata, que depois de aprovada será assinada por mim, pelo Presidente, e pelos demais membros da Comissão.

Prof. Dr. Luiz Eduardo Soares de Oliveira
**DINF/UFPR** – Orientador

Prof. Dr. Alessandro Lameiras Koerich
**PUC/PR** – Membro Externo

Profa. Dra. Aura Conci
**UFF** – Membro Externo

Prof. Dr. Carlos Nascimento Silla Junior
**UTFPR** – Membro Externo

Prof. Dr. Daniel Weingaertner
**DINF/UFPR** – Membro Interno

Jucélia Miecznikowski
Secretária da PPGInf

Ministério da Educação
Universidade Federal do Paraná
Programa de Pós-Graduação em Informática

**PARECER**

Nós, abaixo assinados, membros da Banca Examinadora da defesa do aluno de Doutorado em Ciência da Computação, Yandre Maldonado e Gomes da Costa, avaliamos a tese de doutorado intitulada *"Reconhecimento de Gêneros Musicais utilizando Espectrogramas e Combinação de Classificadores"*, cuja defesa pública foi realizada no dia 15 de agosto de 2013, às 10:00 horas, no Departamento de Informática do Setor de Ciências Exatas da Universidade Federal do Paraná. Após avaliação, decidimos pela **aprovação** do candidato.

Curitiba, 15 de agosto de 2013.

Prof. Dr. Luiz Eduardo Soares de Oliveira
**DINF/UFPR** – Orientador

Prof. Dr. Alessandro Lameiras Koerich
**PUC/PR** – Membro Externo

Profa. Dra. Aura Conci
**UFF** – Membro Externo

Prof. Dr. Carlos Nascimento Silla Junior
**UTFPR** – Membro Externo

Prof. Dr. Daniel Weingaertner
**DINF/UFPR** – Membro Interno

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
Setor de Ciências Exatas
Coordenação de Pós-Graduação em Informática

# SUMÁRIO

# LISTA DE FIGURAS

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS

| | |
|---|---|
| ACE | *Autonomous Classification Engine* |
| AG | Algoritmo Genético |
| BFS | *Backward Feature Selection* |
| BWWV | *Best-Worst Weighted Vote* |
| DFT | *Discrete Fourier Transform* |
| DWCH | *Daubechies Wavelet Coefficients Histogram* |
| DWPT | *Discrete Wavelet Packet Transform* |
| FFS | *Forward Feature Selection* |
| FFT | *Fast Fourier Transform* |
| FFTC | *Fast Fourier Transform Coefficient* |
| GLCM | *Gray Level Co-occurrence Matrix* |
| GMM | *Gaussian Mixture Models* |
| GTZAN | *Base de músicas criada por George Tzanetakis* |
| GSV | *Gaussian Supper Vector* |
| HMM | *Hidden Markov Model* |
| HOSVD | *High-Order Singular Value Decomposition* |
| IGS | *Inter-Genre Similarity* |
| IIGS | *Iterative Inter-Genre Similarity* |
| IOIHC | *Inter-Onset Interval Histogram Cofficients* |
| ISMIR | *International Society for Music Information Retrieval* |
| k-NN | *k Nearest Neighbor* |
| KNORA | *K Nearest Oracles* |
| LBP | *Local Binary Pattern* |
| LCA | *Local Class Accuracy* |
| LDA | *Linear Discriminant Analysis* |
| LDC | *Linear classifier assuming Densities with equal Covariance matrices* |
| LMD | *Latin Music Database* |
| LPC | *Autoregression coefficients* |
| LPQ | *Local Phase Quantization* |
| LPNTF | *Locality Preserving Non-Negative Tensor Factorization* |
| MARSYAS | *Music Analysis, Retrieval and Synthesis for Audio Signals* |
| MFCC | *Mel Frequency Cepstral Coefficients* |
| MIR | *Multimedia Information Retrieval* |
| MLP | *Multi-Layer Perceptron* |
| MP3 | *MPEG-1/2 Audio Layer 3* |
| MPCA | *Multilinear Principal Component Analysis* |
| MVD | *Modulation Frequency Variance Descriptor* |
| NB | *Naïve Bayes* |
| NMF | *Nonnegative Matrix Factorization* |
| NTF | *Non-Negative Tensor Factorization* |
| OAA | *One Against All* |
| OLA | *Overall Local Accuracy* |

Continua na próxima página

continuação da página anterior

| | |
|---|---|
| OPF | *Optimum Path Forest* |
| OSC | *Octave-based Spectral Contrast* |
| P | *Número de vizinhos utilizados em LBP* |
| PCA | *Principal Component Analysis* |
| PCM | *Pulse Code Modulation* |
| PDC | *Parzen Density Based Classifier* |
| PGM | *Piecewise Gaussian Model* |
| QDC | *Quadratic Classifier assuming normal Densities* |
| R | *Distância entre o pixel e seus vizinhos em LBP* |
| RCEP | *Real Cepstral Coefficient* |
| RH | *Rhythm Histogram* |
| RP | *Rhythm Patterns* |
| RR | *Round Robin* |
| RWC | *Real World Computing* |
| SMPC | *Society for Music Perception and Cognition* |
| SoX | *Sound eXchange* |
| SRC | *Sparse Representation-based Classification* |
| SSD | *Statistical Spectrum Descriptors* |
| STFT | *Short-Time Fourier Transform* |
| SVD | *Singular Value Decomposition* |
| SVM | *Support Vector Machine* |
| UDC | *Quadratic classifier assuming normal Uncorrelated Densities* |
| ZCR | *Zero Crossing Rate* |

# AGRADECIMENTOS

Não é possível expressar em palavras a minha gratidão às pessoas mais importantes, que mais amo ou que estiveram sempre próximas durante o desenvolvimento deste trabalho. De qualquer forma, é justo pelo menos tentar.

Em primeiro lugar, agradeço a Deus, sem o qual não haveria nem sequer fôlego de vida, e Ele tem me dado muito mais do que isso, tem me sustentado e mostrado como é grande e assombroso seu poder.

Aos meus pais Sebastião e Lúcia. Esta, o maior exemplo de amor que eu, e muitos que a conhecem, já pudemos ver. Aquele, presente direta ou indiretamente, Sebastião me ensinou a maior parte das coisas mais importantes que sei sobre a vida.

À minha esposa Daniele, que sempre foi companheira e presente. Foi ela que em muitos momentos dividiu sua vida comigo, tornando minha caminhada menos solitária e dolorosa. Por extensão, à Camila, que é um grande presentinho.

Aos meus irmãos Andrey e Daryne, que não vejo muito por circunstâncias da vida, mas eu amo, me fazem rir e também me amam.

Ao meu orientador Dr. Luiz Eduardo Soares de Oliveira, que foi atencioso, compreensivo e me deu o suporte necessário para desenvolver este trabalho. Luiz sabe e exerce muito bem o sentido da palavra "orientar". Seus conselhos ajudam a resolver os problemas inevitáveis e evitar que outros desnecessários surjam.

Ao meu co-orientador Dr. Alessandro Lameiras Koerich, que sempre me deu dicas preciosas e oportunas. Alessandro foi companheiro e prestativo em alguns momentos difíceis e importantes.

Ao meu supervisor Dr. Fabien Gouyon, que abriu as portas do INESC-Porto para o desenvolvimento de partes decisivas do desenvolvimento deste trabalho. Fabien foi sempre compreensivo e atencioso.

Agradeço também à Fundação Araucária e CAPES que financiaram o desenvolvimento deste projeto. Ao Departamento de Informática da Universidade Estadual de Maringá, que me afastou para a capacitação com dedicação integral, condição *sine qua non* para o desenvolvimento de um trabalho de doutorado.

Enfim, eu sou grato à todos que direta ou indiretamente foram companheiros, estiveram presentes, me fizeram mais feliz e/ou compartilharam algo comigo durante este período. Infelizmente, não é possível citar todas estas pessoas...

# RESUMO

Com a rápida expansão da Internet um imenso volume de dados tem se tornado disponível *on-line*. Entretanto, essa informação não segue um padrão de apresentação e não está disponível de maneira estruturada. Devido a isso, tarefas como busca, recuperação, indexação e sumarização automática dessas informações se tornaram problemas importantes, cujas soluções coadunam no sentido de facilitar o acesso a estes conteúdos. Há algum tempo, a maior parte das informações sobre dados multimídia é organizada e classificada com base em informações textuais. A música digital é um dos mais importantes tipos de dados distribuídos na Internet. Existem muitos estudos a respeito da análise de conteúdo de áudio usando diferentes características e métodos. Um componente fundamental para um sistema de recuperação de informações de áudio baseado em conteúdo é um módulo de classificação automática de gêneros musicais. Os gêneros musicais são rótulos categóricos criados por especialistas humanos e por amadores para determinar ou designar estilos de música. Em alguns trabalhos verificou-se que o gênero musical é um importante atributo para os usuários na organização e recuperação de arquivos de música. Este trabalho propõe o uso de características inovadoras para a representação do conteúdo das músicas, obtidas a partir de imagens de espectrograma geradas a partir do sinal do áudio, para aplicação em tarefas de reconhecimento de gêneros musicais. As imagens de espectrograma apresentam a textura como principal atributo visual. Assim, as características propostas foram obtidas utilizando-se alguns descritores de textura propostos na literatura de processamento de imagens, em particular os descritores *Local Binary Pattern* e *Local Phase Quantization*, pois ambos se destacaram por apresentar um bom desempenho. Também foram investigados os impactos proporcionados pelo uso de uma estratégia de preservação de informações locais, através do zoneamento das imagens. O zoneamento propiciou a criação de múltiplos classificadores, um para cada zona, e os melhores resultados foram obtidos com a fusão das saídas destes classificadores. A maioria dos experimentos foi realizada sobre a base LMD com o uso de *"artist filter"*. O método também foi experimentado sobre a base ISMIR 2004. Os melhores resultados obtidos são comparáveis aos melhores resultados já apresentados na literatura utilizando outras abordagens. Considerando os experimentos com a base LMD e com o uso de *"artist filter"*, os resultados obtidos são superiores ao melhor resultado descrito na literatura até então. Finalmente, seleção dinâmica de classificadores e seleção de características foram avaliadas e mostraram resultados promissores.

# ABSTRACT

With the rapid expansion of the internet, a huge amount of data from different sources has become available online. In most cases, this information is not organized according to some predefined pattern. Thus, tasks related to automatic search, retrieval, indexing and summarization has become important questions, whose solutions could support the access to this content. For some time, textual annotation is used to organize and classify multimedia data. Digital music is among the most common types of data distributed through the internet. There are a number of studies concerning to audio content analysis using different features and methods. Automatic music genre recognition is a crucial task for a content based music information retrieval system. Musical genres are categorical labels created by humans to characterize pieces of music. A musical genre is characterized by the common characteristics shared by its members. These characteristics typically are related to the instrumentation, rhythmic structure, and harmonic content of the music. In some studies it was found that genre is an important attribute which helps users in organizing and retrieving music files. In this work we propose an alternative approach for music genre classification which converts the audio signal into a spectrogram (short-time Fourier representation) and then extract features from this visual representation. Texture is the main visual content in a spectrogram image. Thus, the features to be explored here were taken among some well known texture descriptors presented in the image processing literature, in particular Local Binary Pattern and Local Phase Quantization. Both have shown good performance in works related to different application domains recently presented in the literature. In addition, the effects of local information preserving, by zoning the images, were investigated. The rationale behind the zoning and combining scheme is that music signals may include similar instruments and similar rhythmic patterns which leads to similar areas in the spectrogram images. By zoning the images we can extract local information and try to highlight the specificities of each music genre. A positive side effect obtained with zoning strategy is that one can create a specific classifier to deal with the features extracted from each specific zone. Thus, we can naturally obtain several classifiers. Not by chance, the best obtained results happened by combining these classifiers outputs. Most of the experiments was developed ont the LMD dataset using the artist filter restriction. Some experiments with the ISMIR 2004 dataset were performed as well. With this dataset, the best obtained results are comparable to the best obtained results described in the literature. Regarding to the LMD dataset, the best obtained result is the best ever obtained using artist filter. Finally, dynamic ensemble of classifiers selection (using KNORA) and feature selection (using genetic algorithm) were tested and presented promising results.

# CAPÍTULO 1

# INTRODUÇÃO

A criação de grandes bases de músicas oriundas tanto da restauração de arquivos analógicos existentes, quanto de novos conteúdos tem demandado cada vez mais ferramentas rápidas e confiáveis para análise e descrição deste conteúdo para serem utilizadas em pesquisas, buscas de conteúdo e acesso interativo. Neste contexto, gêneros musicais são descritores cruciais, já que há anos são amplamente utilizados para categorizar música, organizar catálogos musicais, bibliotecas e depósitos de música. A despeito do seu uso, gêneros musicais permanecem como um conceito mal definido, o que torna o problema de classificação automática uma tarefa não trivial.

Há algum tempo, boa parte das informações sobre dados multimídia são organizadas e classificadas com base em meta-informações textuais que são associadas ao seu conteúdo, como é o caso dos rótulos ID3 incorporados aos arquivos de áudio no formato MP3. Recentemente, iniciativas como o desenvolvimento da ferramenta RAMA têm permitido que se estabeleça algum nível de organização entre as informações de conteúdo musical. RAMA é um aplicativo para web que permite a visualização da similaridade entre artistas por meio de uma rede expressa na forma de um grafo conectado [76]. RAMA opera sobre dados tomados da Last.fm (*http://www.lastfm.com.br*), uma rádio web que oferece conteúdo de centenas de milhares de artistas e que possui rótulos associados aos títulos gerados por um universo de três milhões de usuários. 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, o que torna o processo caro, demorado e ainda impreciso devido à subjetividade da percepção humana.

Para exemplificar o problema relacionado a custo e tempo, mencionado no parágrafo anterior, Dannenberg *et al.* [13] reproduzem um relato de Cristopher Weare, da Microsoft, que afirma que uma operação de classificação manual de algumas centenas de milhares de músicas realizada pela empresa exigiu a dedicação em tempo integral de profissionais que, juntos, somaram 30 anos-homem de trabalho. Ainda assim, a rotulação manual é eventualmente empregada em alguns contextos porque uma definição precisa de gênero é muito difícil e muitas músicas se situam no limite entre diferentes gêneros. Diante disto, é oportuno o desenvolvimento de ferramentas para recuperação automática de músicas baseadas em conteúdo.

## 1.1 Motivação

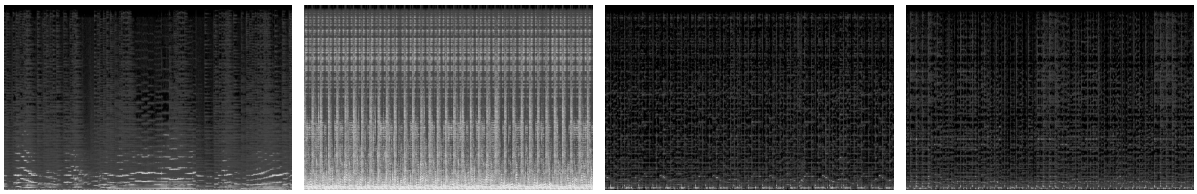
Com a rápida expansão da Internet, um grande volume de dados oriundos de diferentes fontes tem se tornado disponível *on-line*. Estudos apresentados ainda em 2008 apontavam que em 2007 a massa de dados digitais espalhada ao redor do mundo consumia aproximadamente 281 exabytes e que, em 2011, este volume se multiplicaria por dez [25]. 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 acerca dos quais muitas pesquisas têm sido realizadas. Neste contexto, é oportuno o desenvolvimento de pesquisas relacionadas a recuperação automática de informações multimídia, que visa criar ferramentas capazes de organizar e gerenciar essa grande quantidade de informações. A realização destas tarefas de forma automática elimina a necessidade de mão de obra humana para a indexação dos conteúdos, além de evitar problemas relacionados à subjetividade da percepção humana.

## 1.2 Desafios

A definição de gêneros musicais é inerentemente subjetiva e imprecisa. Este trabalho é voltado para a identificação de um novo conjunto de características, obtidas no domínio visual através da exploração de imagens de espectrogramas, que possam ser empregadas em tarefas de classificação automática de gêneros musicais. Espectrograma é uma representação visual do espectro de frequências do som. No seu formato mais comum, ele é representado por um gráfico em que o eixo horizontal representa o tempo e o eixo vertical a frequência. A amplitude do sinal é representada em uma terceira dimensão, descrita pela intensidade da cor de cada ponto da imagem. A figura 1.1 mostra algumas imagens de espectrogramas extraídos de alguns títulos musicais de diferentes gêneros. Por esta amostra, pode-se perceber a complexidade do cenário. As figuras 1.1(a) e 1.1(b) apresentam significativas diferenças, nelas podem ser percebidas claramente a presença de linhas que parecem representar dimensões musicais geralmente associadas aos gêneros nos quais estão classificadas. O espectrograma da figura 1.1(a) foi gerado a partir do sinal de uma música clássica, e nele é possível perceber a presença de linhas predominantemente horizontais, presumidamente relacionadas às estruturas harmônicas, muito presentes nos titulos deste gênero musical. A figura 1.1(b) mostra um espectrograma extraído de uma música eletrônica. Nela percebe-se uma maior presença de linhas quase verticais, relacionadas às batidas, comum nas músicas deste gênero. Em outros casos, como 1.1(c) e 1.1(d) pode-se notar quão difícil é discriminar entre dois gêneros teoricamente não muito distantes em termos de características musicais e que, conseqüentemente, não apresentam

espectrogramas tão diferentes entre si.



(a) Clássica (b) Eletrônica (c) Forró (d) Gaúcha

Figura 1.1: Similaridades e diferenças entre espectrogramas de diferentes gêneros

A maioria absoluta dos experimentos desenvolvidos ao longo deste trabalho foi realizada utilizando-se a base Latin Music Database. Esta base é composta por dez gêneros latino-americanos, sendo cinco oriundos do Brasil e outros cinco oriundos de países da América Central ou América do Sul. Um fato importante a se considerar é que estes países possuem grande similaridade em termos de aspectos culturais, o que inevitavelmente influencia na estrutura harmônica e ritmo das músicas populares nos mesmos. Com isto, entende-se como grande desafio deste trabalho a identificação e extração de características que sejam úteis para representar e discriminar conteúdo musical, especialmente na classificação de gênero, dentro deste cenário e que, ao mesmo tempo, possa ser utilizado com sucesso em outras diferentes bases de dados, também impregnadas da subjetividade inerente a definição de gêneros.

## 1.3 Hipóteses de pesquisa

A hipótese lançada neste trabalho é a de que seja possível representar uma música, para o propósito de classificação de gêneros musicais, através de características extraídas da imagem do espectrograma extraído do sinal do áudio. Investiga-se ainda a hipótese de que, neste cenário em que a natureza musical do sinal é integralmente abstraída, se possa alcançar resultados similares ou superiores aos alcançados por métodos tradicionais.

## 1.4 Objetivos

O principal objetivo deste trabalho é a prospecção de um novo formato de características, obtidas no domínio visual, útil para a aplicação na classificação automática de gêneros musicais. Pretende-se identificar características que possam ser utilizadas isoladamente ou juntamente com outras características já conhecidas para este tipo de tarefa de classificação. Adicionalmente, é importante que este novo formato de características seja versátil e eficiente. Por versátil, entende-se um formato de características que possa ser utilizado com a maior variedade possível de gêneros musicais. Por eficiente, entende-se um

formato que apresente resultados próximos ou melhores do que os obtidos com métodos tradicionalmente empregados neste tipo de tarefa.

Para atingir os objetivos gerais supracitados, serão cumpridos os seguintes objetivos específicos:

- Identificar características de textura que possam ser extraídas de imagens de espectrograma e que tenham potencial para serem utilizadas como descritores em tarefas de classificação;
- Avaliar o desempenho das características identificadas na classificação de gêneros musicais sobre as bases LMD e *ISMIR 2004*;
- Avaliar o possível impacto da preservação de informações espaciais, através do zoneamento das imagens e criação de vários classificadores, no desempenho geral do sistema;
- Investigar a complementaridade das características identificadas com outras tradicionalmente utilizadas neste domínio de aplicação;
- Utilizar técnicas complementares, como seleção de características e seleção dinâmica de classificadores a fim de investigar possibilidades de se obter ganhos nas taxas de reconhecimento do sistema.

## 1.5 Contribuições

O desenvolvimento deste trabalho proporcionou algumas importantes contribuições no contexto da classificação automática de gêneros musicais.

As primeiras contribuições, apresentadas em [7] e [8], mostraram que existem informações no conteúdo de textura presente nas imagens de espectrogramas gerados a partir do sinal do áudio com potencial para serem empregadas com sucesso na tarefa de reconhecimento de gêneros musicais. Nestes trabalhos, as características foram obtidas utilizando-se Gray Level Co-occurrence Matrix (GLCM) e, embora tenha sido empregado apenas o zoneamento linear das imagens para preservação de informações locais, um único classificador foi criado no processo de classificação e a decisão final se dava através do voto majoritário entre os vetores extraídos para as diferentes zonas criadas na imagem. Os resultados obtidos foram suficientes para confirmar a hipótese de que as imagens de espectrograma podem fornecer informações para suportar este tipo de tarefa. A paritir disso, foram realizadas outras investigações variando a forma como o processo de classificação é configurado e o tipo de características utilizadas.

Em [9], foram apresentados resultados obtidos utilizando o descritor de textura Local Binary Pattern (LBP). Estes resultados foram comparados aos obtidos com o uso de

GLCM para a obtenção de características. Adicionalmente, foi introduzida a estratégia que prevê a criação de vários classficadores no processo. Para isto, foi gerado um classificador para cada zona estabelecida na imagem. Diferentes padrões de zoneamento linear foram testados e as saídas dos vários classificadores empregados em cada caso foram fundidas com algumas das mais conhecidas regras de fusão apresentadas na literatura. Os resultados mostraram que o uso de vários classificadores proporciona melhores resultados e, além disso, LBP proporciona resultados superiores aos obtidos com GLCM.

A fim de que não ficassem abertas questões relacionadas à viabilidade de aplicação do método aqui proposto à outras bases de música, foram apresentados em [12] os resultados de experimentos realizados sobre a base de músicas ISMIR 2004, uma base com gêneros bastante diversos daqueles presentes na base LMD. Também foram apresentados resultados obtidos com descritores tradicionalmente utilizados em classificação de gêneros musicais a fim de que se pudesse comparar os desempenhos. Também foram experimentados padrões de zoneamento não lineares, baseados em escalas construídas levando-se em consideração bandas de frequência estabelecidas de acordo com a percepção humana. Os resultados mostraram um desempenho similar ou superior do método aqui proposto aos resultados apresentados em outros trabalhos descritos na literatura tanto para a base LMD quanto para a base ISMIR 2004. Além disso, os resultados obtidos com o método aqui proposto foram superiores aos obtidos com descritores tradicionais e a estratégia de utilizar escalas perceptuais para o zoneamento das imagens pode proporcionar resultados ainda melhores em termos de taxa de reconhecimento.

Adicionalmente, foram apresentados resultados obtidos utilizando um método para seleção dinâmica de classificadores (KNORA) em [10] e resultados obtidos utilizando descritores de textura LPQ e Filtros de Gabor em [11].

## 1.6 Organização

Este trabalho encontra-se organizado da seguinte forma: no capítulo 2 é descrita uma revisão bibliográfica acerca de classificação automática de gêneros musicais; no capítulo 3 são apresentados os principais fundamentos teóricos inerentes às técnicas utilizadas para a realização dos experimentos; o capítulo 4 descreve o método proposto para a construção do sistema de classificação de gêneros musicais; o capítulo 5 apresenta os resultados obtidos utilizando o protocolo para o reconhecimento de gêneros musicais desenvolvido neste trabalho; o capítulo 6 apresenta resultados obtidos em experimentos adicionais, utilizando seleção dinâmica de agrupamento de classificadores e seleção de características com algoritmo genético; por fim, o capítulo 7 mostra as conclusões finais deste trabalho.

# CAPÍTULO 2

# REVISÃO BIBLIOGRÁFICA

A idéia de classificação automática de gêneros musicais como uma tarefa de reconhecimento de padrões ficou conhecida e devidamente caracterizada a partir do trabalho de Tzanetakis e Cook [87]. Neste capítulo serão descritos os trabalhos mais relevantes relatados pela comunidade científica desde então, em sequência predominantemente cronológica. Muito embora já tivesse ocorrido, ainda em 2001, uma tentativa de classificar gêneros musicais automaticamente utilizando as etapas clássicas de reconhecimento de padrões no trabalho de Deshpande [14], não houve repercussão expressiva do mesmo na comunidade acadêmica, talvez pela amostra de músicas bastante limitada utilizada nos experimentos. A referida amostra tinha um total de 157 títulos musicais de apenas três diferentes gêneros.

Lidy *et al.* [50] apresentam várias abordagens já descritas na literatura que caracterizam esforços no sentido de realizar a classificação automática de gêneros musicais. A primeira das abordagens descrita pelos autores é a abordagem baseada em conteúdo, em que o conteúdo dos arquivos de música é analisado e são extraídas do sinal do áudio características que os descrevem. A segunda abordagem é a análise semântica de músicas, que pode ser empregada na classificação de títulos musicais em categorias que não são predominantemente relacionadas a características acústicas. Metadados de comunidade (*community metadata*) caracterizam uma terceira abordagem, na qual utiliza-se filtragem colaborativa e análise de meta-informação fornecida por usuários. Também existem abordagens híbridas, que combinam várias das abordagens descritas anteriormente.

No trabalho de Tzanetakis e Cook [87], foi proposto um conjunto abrangente de características a fim de representar espectro sonoro (*timbral texture*), padrão rítmico (*beat-related*) e a altura da nota (*pitch-related*). Estas características são extraídas diretamente do sinal, portanto a técnica segue a abordagem baseada em conteúdo, assim como a técnica proposta neste trabalho. As características relacionadas à textura de timbre já eram empregadas em tarefas correlacionadas, com propósitos voltados ao reconhecimento de fala, e os outros dois conjuntos propostos foram propostos especificamente com o intuito de representar conteúdo de ritmo e harmonia, presentes em sinais de músicas e supostamente discriminantes na identificação dos gêneros atribuídos às mesmas. Embora a criação musical seja um processo artístico em que muitos compositores de um gênero são influenciados por outros gêneros, percebe-se que músicas de um mesmo gênero compartilham certas características, como a presença de instrumentos similares, padrões de ritmo similares e distribuição de variações de frequência de vibração similares.

O vetor final de características empregado por Tzanetakis e Cook neste trabalho é composto por um total de 30 características, sendo 19 de textura de timbre, seis de conteúdo rítmico e cinco características de conteúdo de vibração. Estas características foram submetidas a três tipos de classificadores: classificador Gaussiano, modelos de mistura Gaussiana (GMM) e k-NN. Os experimentos foram avaliados em uma base de dados contendo mil músicas de dez gêneros distintos, sendo cem músicas de cada gênero, esta base foi denominada GTZAN e disponibiliza um segmento de 30 segundos de cada título musical presente nela. Os gêneros presentes na mesma são: *blues*, clássica, *country*, *disco*, *hip-hop*, *jazz*, *metal*, *pop*, *reggae* e *rock*. O fato de utilizar apenas 30 segundos de cada música produz o interessante efeito colateral de reduzir o tamanho da massa de dados a ser processada. O acerto obtido inicialmente nessa base foi de cerca de 61%, e os autores apontam que apesar da natureza nebulosa das fronteiras entre os diferentes gêneros musicais, a classificação automática dos mesmos pode ser realizada com desempenho comparável à obtida pelos humanos.

O trabalho de Tzanetakis e Cook possui o grande mérito de ter introduzido de forma contundente a classificação de gêneros musicais como uma tarefa de reconhecimento de padrões, e por isso passou a ser referência adotada em praticamente todos os trabalhos relevantes acerca deste domínio que o sucederam. Adicionalmente, os autores ainda apresentaram a primeira base de dados criada especificamente para tarefas de reconhecimento de gêneros musicais.

Outro aspecto interessante do trabalho é que o conjunto de características utilizadas está disponível através do ambiente *MARSYAS*, um software livre para o desenvolvimento e avaliação de aplicações voltadas à computação musical. Tzanetakis e Cook motivaram a pesquisa e desenvolvimento de novas abordagens para a tarefa de classificação automática de gêneros musicais utilizando técnicas de aprendizado de máquina e processamento digital de sinais.

No contexto da classificação automática de gêneros musicais, frequentemente é importante que se tenha uma noção acerca do desempenho do elemento humano neste tipo de tarefa. A fim de preencher a lacuna caracterizada pela falta de trabalhos significativos sobre o reconhecimento de gêneros musicais por parte de humanos, Gjerdingen e Perrott apresentaram o trabalho [26]. O trabalho é um clássico e é tomado como referência para a questão do reconhecimento de gêneros musicais por parte de humanos já desde o trabalho de Tzanetakis e Cook [87]. Os experimentos utilizaram dez gêneros musicais bastante populares nos anos 90: *blues*, *classical*, *country*, *dance*, *jazz*, *latin*, *pop*, *rhythm and blues*, *rap* e *rock*. Foram escolhidos 52 estudantes universitários voluntários que, de forma geral, gostam de ouvir música, mas não são músicos profissionais e nem especialistas no assunto. Os resultados do estudo mostram que, embora haja uma grande variação nas taxas de acerto de cada gênero, houve coincidência entre os rótulos de gênero previamente atribuído às músicas e os gêneros indicados pelos participantes do estudo em 70% dos casos. Este

resultado é particularmente entusiasmante, uma vez que o desempenho de muitos dos sistemas automáticos já criados ou em criação, descritos na literatura, estão próximos dele. O trabalho de Gjerdingen e Perrott cumpriu o importante papel de preencher a lacuna caracterizada pela falta de referência acerca da taxa de acerto de humanos em torno da atribuição de gêneros para títulos musicais.

No lastro de contribuições para a comunidade de pesquisa em Recuperação de Informações Musicais, duas novas bases de dados apropriadas para o uso em tarefas de reconhecimento de gêneros musicais foram disponibilizadas em 2003 e 2004. Trata-se, respectivamente, das bases *RWC* [29] e *ISMIR 2004* [4]. A base *RWC (Real World Computing)* é composta por cem músicas distribuídas entre os seguintes dez gêneros: popular, *rock*, *dance*, *jazz*, *latin*, *classical*, *marches*, *world*, *vocals* e *japanese music*. A base *ISMIR 2004* é composta por 1458 músicas distribuídas entre seis gêneros: *classical*, *electronic*, *jazz/blues*, *metal/punk*, *rock/pop* e *world*.

Li *et al.* [47] realizaram um estudo comparativo entre o conjunto de características propostas por Tzanetakis e Cook e um novo conjunto de características para a classificação automática de gêneros musicais baseada em conteúdo. Estas novas características investigadas representam tanto informações locais quanto informações globais do sinal, extraídas utilizando Histogramas de Coeficientes fornecidos por Daubechies Wavelet (DWCH). Também foi verificado se métodos como análise discriminante linear (LDA) e máquinas de vetores de suporte (*Support Vector Machine* - SVM) teriam um melhor desempenho do que os classificadores utilizados anteriormente. Os experimentos foram realizados em duas bases de dados, a primeira foi a *GTZAN* e a segunda foi uma base contendo 756 músicas de cinco gêneros diferentes: ambiente, clássica, *fusion*, *jazz* e *rock*. Um aspecto importante dessa segunda base de dados é que as características foram extraídas do segmento composto entre o segundo 31 e o segundo 60, ao invés dos primeiros 30 segundos. As conclusões dos experimentos realizados neste trabalho mostram que a melhor taxa de classificação foi obtida com o classificador SVM que melhorou o acerto obtido na primeira base para cerca de 72% com o mesmo conjunto de características, e para cerca de 78% no melhor caso com as características geradas com DWCH. Na segunda base a taxa de acerto obtida foi de 74% utilizando DWCH e 71% utilizando as características do trabalho de Tzanetakis e Cook.

Outro aspecto importante deste trabalho é que foram avaliadas diferentes estratégias de decomposição que são necessárias por classificadores que não lidam naturalmente com problemas multi-classe. Eles avaliaram o classificador SVM utilizando as estratégias um contra todos (*One Against All* - OAA) e *Round Robin* (RR). Os melhores resultados foram alcançados com a estratégia OAA com as características geradas com DWCH. De forma geral, o conjunto de características baseado em DWCH apresentou desempenho superior em relação às características de Tzanetakis e Cook, sendo que a diferença entre as taxas de classificação obtidas foram de 1% (utilizando o k-NN) a 7% utilizando SVM com OAA

para a primeira base e de 1% (utilizando o k-NN) a 5% utilizando SVM com OAA para a segunda base. O mérito destacável deste trabalho foi a introdução de um novo formato de características descritoras de conteúdo de sinal de áudio.

No trabalho de Li e Ogihara [46] foi investigado o uso de uma taxonomia hierárquica para a classificação de gêneros musicais. A principal motivação para o desenvolvimento deste trabalho consistiu no fato de que na maioria dos trabalhos de classificação de gêneros musicais, os gêneros são considerados de forma que não existe nenhuma estrutura que defina relacionamentos entre eles. Este esquema possui limitações devido ao fato de que conforme a indústria da música cresce, o número de gêneros possíveis também cresce e as fronteiras entre eles se tornam nebulosas. O uso de uma hierarquia poderia permitir o emprego da abordagem "dividir para conquistar". Na prática, cada classificador tem que lidar com um problema de classificação mais fácil. Este esquema, torna o erro mais tolerável, pois ele tende a se concentrar em um nível próximo ao do acerto na hierarquia. Para gerar automaticamente uma hierarquia de gêneros, a idéia central utilizada foi a de inferir relacionamentos a partir da matriz de confusão produzida por classificadores eficientes. A matriz de confusão mostra claramente um grau de confusão entre algumas classes. De forma geral, a matriz de confusão oferece uma estratégia independente de domínio para inferir relacionamento entre os gêneros. Essa taxonomia identifica as relações de dependência de diferentes gêneros e fornece valiosas fontes de informação para a classificação de gêneros. Os experimentos foram realizados com as mesmas bases utilizadas no trabalho anteriormente desenvolvido pelos autores e a taxa de classificação aumentou em 0,7 pontos percentuais para a primeira base e 3 pontos percentuais para a segunda base.

Outro trabalho relacionado com a tarefa de classificação automática de gêneros musicais, porém com foco diferente foi apresentado por Hu *et al.* [38]. Neste trabalho são utilizados *reviews* de músicas e técnicas de mineração de textos para realizar a classificação automática dos gêneros. *Reviews* são textos com revisão crítica sobre títulos musicais, elaborados por críticos de música. Foi empregado um classificador Naïve Bayes sobre *reviews* disponíveis *on-line* a fim de identificar não somente o gênero musical, mas também identificar uma avaliação qualitativa da música de acordo com o conteúdo dos *reviews*. Nos experimentos que envolveram o reconhecimento de gêneros musicais foram utilizados 12 diferentes gêneros e a taxa de precisão foi igual a 78,89%. Os autores concluem dizendo que os experimentos foram bem sucedidos e sugerem que a idéia apresentada consiste em uma linha de pesquisa promissora. Um aspecto interessante acerca deste trabalho é o fato de que ele caracteriza uma iniciativa em classificação automática de gêneros musicais diferente das tradicionais baseadas em conteúdo.

A idéia de decomposição e combinação de classificadores foi utilizada para a classificação automática de gêneros musicais no trabalho de Grimaldi *et al.* [30]. Neste trabalho foram realizados experimentos utilizando diferentes estratégias de combinação de classi-

ficadores e seleção de atributos. Os experimentos foram realizados numa base contendo 200 músicas de cinco gêneros (*jazz*, *classical*, *rock*, *heavy metal* e *techno*), para realizar a classificação foi utilizado o método de validação cruzada utilizando 5 folds. Todos os experimentos foram avaliados utilizando apenas o classificador k-NN. Para extrair as características foi utilizada a DWPT (*Discrete Wavelet Packet Transform*) aplicada ao sinal da música inteira. Os melhores resultados alcançados nos experimentos indicaram uma taxa de acerto de 65%. O trabalho caracterizou uma importante iniciativa no sentido de utilizar técnicas mais sofisticadas de reconhecimento de padrões, como seleção de características e combinação de classificadores neste domínio de aplicação, além de introduzir um novo formato de descritor (DWPT).

No trabalho de Costa *et al.* [6] foi proposto um novo método para a classificação automática de gêneros musicais, baseado na extração de características de três segmentos do sinal do 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 redes neurais MLP (*Multi-Layer Perceptron*) e k-NN. Os experimentos foram realizados em uma base contendo 414 músicas de dois gêneros (*rock* e clássica). A conclusão obtida no trabalho foi que o método de combinação proposto não melhorava o desempenho além da classificação individual dos segmentos isolados.

Uma continuação deste trabalho foi apresentada por Koerich e Poitevin [44] em que para realizar a combinação dos classificadores foram utilizadas outras regras de combinação além do voto majoritário, regras estas baseadas nas probabilidades individuais de cada classe fornecida na saída dos classificadores. As regras utilizadas foram máximo, soma, soma ponderada, produto e produto ponderado. A base utilizada foi a mesma do experimento anterior. Uma alteração é que neste trabalho os autores utilizaram apenas redes neurais MLP para fazer a classificação. Os resultados obtidos mostraram uma melhora na taxa de acerto em relação aos segmentos individuais utilizando os dois segmentos melhor classificados e as regras de soma e produto ponderados. 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 das saídas dos classificadores que utilizam características extraídas dos mesmos, o erro cometido em um segmento é diluído e o impacto negativo que ele provocaria na classificação tende a diminuir. Embora tenha potencial para contribuir com a obtenção de resultados positivos, é bastante comum que a estratégia de segmentação não possa ser empregada quando se utiliza uma base de músicas na qual não é disponibilizado o conteúdo dos títulos musicais

por inteiro.

Pampalk *et al.* [68] introduziram o conceito de *"artist filter"*. Com ele, preconiza-se que para uma avaliação adequada do desempenho de sistemas de classificação de gêneros musicais, não pode haver títulos musicais de um mesmo artista nos conjuntos de teste e de treinamento simultaneamente. A recomendação serve para que se evite desenvolver classificadores eficientes em classificar artistas ao invés de gêneros musicais. Os autores mostram que, com o emprego de *"artist filter"* houve caso de redução na taxa de acerto de 72% para 27%, no caso mais extremo.

Dando continuidade a investigação acerca de benefícios potenciais do emprego desta estratégia, Flexer [21] obtém resultados que mostram que *"artist filter"* não somente reduz as taxas de acerto na classificação, pelo fato de reduzir taxas que podem não ser tão realistas, como também reduz as diferenças de acerto entre diferentes técnicas. Mais uma vez o autor descreve resultados de experimentos com e sem o uso de *"artist filter"*. Utilizando características obtidas com MFCC e classificador construído com GMMs, o autor obtem sobre a base ISMIR 2004 taxa de reconhecimento igual a 75,72% sem o uso de *"artist filter"* e 61,22% utilizando o filtro. Adicionalmente, o autor recomenda que os resultados obtidos em classificação musical sem o uso do *"artist filter"* sejam revistos. A partir destes trabalhos, este conceito passou a ser empregado em vários outros desenvolvidos pela comunidade de pesquisa em reconhecimento de gêneros musicais e é um importante instrumento na tentativa de se produzir classificadores mais robustos.

No trabalho de Meng *et al.* [60] são utilizadas características baseadas em três escalas de tempo: as características de tempo curto são computadas utilizando janelas de análise de tamanho 30ms, o significado perceptual deste tipo de característica está relacionado ao timbre (frequência instantânea); as características de tempo médio são computadas utilizando janelas de análise de tamanho 740ms, e estão relacionadas à modulação (instrumentalização); as características de tempo longo são computadas utilizando janelas de análise de tamanho 9,62 segundos e estão relacionados à batida, o humor vocal, etc. Para realizar os experimentos foram considerados dois classificadores: *perceptron* e um classificador Gaussiano. Os experimentos foram realizados em duas bases de dados, mas o propósito destes era verificar o desempenho relativo das características ao invés de verificar o erro no conjunto de dados. A primeira base de dados utilizada contém cem músicas, distribuídas igualmente em cinco gêneros (clássica, *rock*, *jazz*, *pop* e *techno*), já a segunda consiste de 354 músicas de 30 segundos extraídas do "Amazon.com Free-Downloads" e possuem seis gêneros (*classical*, *country*, *jazz*, *rap*, *rock* e *techno*). A integridade da primeira base de dados foi verificada por meio de um teste de audição que envolveu 22 classificadores humanos. Foram realizados diversos experimentos e os melhores resultados computacionais obtidos no conjunto de teste apresentaram erro de apenas 5% sobre a primeira base de dados utilizando a combinação de características de tempo médio e longo enquanto que a classificação feita por humanos apresentou erro de 3%.

Em [91], Yaslan e Cataltepe utilizaram os seguintes classificadores: Fisher, LDC, QDC, UDC, Naïve Bayes, PDC e k-NN. A base utilizada foi a GTZAN e a de extração de características foi feita utilizando o *MARSYAS*. A principal diferença deste trabalho em relação aos anteriores, é que foram avaliadas as características de acordo com o grupo a que elas pertencem, quais sejam: *Beat*, *Mpitch*, STFT e MFCC. Além disso, foram utilizados métodos de FFS (*Forward Feature Selection*) e BFS (*Backward Feature Selection*) para tentar encontrar um melhor subconjunto de características que aumentasse o desempenho dos classificadores. Os resultados obtidos foram positivos e os autores ainda propuseram o uso de um agrupamento combinando a saída dos classificadores que apresentaram os melhores resultados. Essa técnica de combinação também apresentou resultados positivos, e as melhores taxas de acerto encontradas nos experimentos foram de 80%.

Em [22], Flexer *et al.* conseguem 78,19% de reconhecimento sobre a base *ISMIR 2004* utilizando características extraídas com MFCC e fazendo a classificação com GMM. Este resultado foi alcançado em experimentos com validação cruzada utilizando dez folds. Sobre esta mesma base e também utilizando validação cruzada com dez folds, Lidy e Rauber [48] conseguem 80,32% de reconhecimento utilizando características obtidas com *Statistical Spectrum Descriptor (SSD)* e *Rhythm Histogram (RH)* submetidas ao classificador SVM.

Homburg *et al.* apresentam em [36] uma nova base de dados (*HOMBURG set*) para pesquisas em recuperação de informação musical. A base disponibilza um total de 1886 títulos musicais classificados em 9 diferentes gêneros (*Blues, Electronic, Jazz, Pop, Rap/HipHop, Rock, Folk/Country, Alternative, Funk/Soul*). Entretanto, são disponibilizados apenas dez segundos de cada título musical, o que faz com que o potencial de uso da base seja bastante limitado.

Em 2006, foi apresentada a base de músicas CODAICH [59]. Esta base é composta por um total de 20849 títulos distribuídos em 53 gêneros. Embora a base seja composta por uma vasta coleção musical, problemas relacionados a direitos autorais não permitem que sejam disponibilizados os conteúdos das músicas diretamente. Ao invés disso, é disponibilizado um mecanismo a partir do qual se pode extrair características do conteúdo.

Depois do desenvolvimento de alguns trabalhos voltados a classificação automática de gêneros musicais, Aucouturier e Pachet apresentaram em [1], ainda em 2003, o primeiro trabalho de inspeção desta tarefa. Neste trabalho os autores discutem a definição de taxonomias para a tarefa. Os autores concluíram que, em muitos contextos, gêneros musicais são mal definidos (*ill-defined*) e esta limitação seria um fator que potencialmente introduziu interferências negativas nos primeiros trabalhos relacionados à classificação automática de gêneros musicais. Eles classificaram as abordagens para a classificação automática de gêneros em dois tipos (por sinal, os mesmos estabelecidos para sistemas de reconhecimento de padrões em geral): supervisionada e não supervisionada. Neste trabalho eles fizeram uma crítica aos sistemas baseados em janelas de análise por não utilizarem as informações temporais da música. Os autores ainda criticaram o baixo

número de gêneros utilizados em muitos trabalhos apresentados até então.

Os autores ainda sugerem o uso de duas técnicas oriundas da área de mineração de dados conhecidas como filtragem colaborativa e análise de co-ocorrência para determinar a similaridade de músicas. Para a construção de novas bases de dados para o problema, eles sugerem criar bases de dados utilizando compilações de músicas com um mesmo ritmo.

De forma semelhante, Scaringella *et al.* [77] apresentam uma revisão sobre o estado da arte em classificação automática de gêneros musicais. Entre as principais conclusões, os autores apontam que há sérias distorções entre as definições de gêneros musicais apesar de sua grande importância na organização de coleções musicais. Os autores revisam e classificam os principais tipos de características utilizadas em recuperação de informações musicais nas três seguintes categorias: características de timbre, características de melodia/harmonia e características de ritmo. A propósito, a classificação já estabelecida por Tzanetakis e Cook. Descritas estas características, os três principais paradigmas de classificação de gêneros musicais são relacionados, além de suas vantagens e desvantagens: sistemas especialistas, agrupamento não supervisionado e classificação supervisionada. Finalmente, são introduzidas novas técnicas e campos de pesquisa emergentes que investigam a proximidade entre os gêneros musicais, como folksonomia e categorias perceptuais.

Lippens *et al.* [51] desenvolveram um estudo que compara resultados obtidos em classificação automática de gêneros musicais com os resultados obtidos na classificação feita por humanos. Os resultados mostraram que, embora ainda haja bastante espaço para melhorias na classificação automática, a classificação de gênero é inerentemente subjetiva e, então, resultados perfeitos não podem ser esperados, nem na classificação por humanos, nem na classificação automática.

Bergstra *et al.* apresentam em [3] resultados obtidos sobre duas novas bases de músicas. A primeira, a Magnatune, composta por dez gêneros musicais: *Classical*, *New Age*, *Electronic*, *World*, *Ambient*, *Jazz*, *Hip-hop*, *Alt Rock*, *Electro Rock* e *Hark Rock*. A segunda base é a USPOP, composta por seis gêneros: *Country*, *Electronic/Dance*, *New Age*, *Rap/Hip-hop*, *Raggae* e *Rock*. Foram utilizadas várias características: *Fast Fourier Transform Coefficients (FFTCs)*, *Real Cepstral Coefficients (RCEPs)*, *Mel Frequency Cepstral Coeficients (MFCCs)*, *Zero Crossing Rate (ZCR)*, *Spectral Spread*, *Spectral Centroid*, *Spectral Rolloff* e *Autoregression Coefficients (LPC)*. A classificação foi feita com ADABOOST e os melhores resultados obtidos foram 75,1% na base Magnatune e 86,92% na base USPOP.

Ezzaidi e Rouat investigam em [17] o desempenho de MFCCs com classificador GMM sobre a base de músicas RWC. O melhor desempenho obtido apresenta taxa de reconhecimento de 73%.

Fiebrink e Fujinaga [20] discutem a real eficácia do emprego de métodos de seleção de características em recuperação de informações musicais. Os autores argumentam que em trabalhos anteriores de recuperação de informações musicais há certo exagero por parte

dos autores na avaliação da eficácia da seleção de características. Os autores desenvolvem novos experimentos, cujos resultados são explorados para defender uma reavaliação do impacto da seleção de características na taxa de reconhecimento em recuperação de informações musicais.

No trabalho de Mckay e Fujinaga [58] é feita uma análise crítica se a tarefa de classificação automática de gêneros musicais mereceria ou não continuar a ser pesquisada/tratada. Antes de apresentar os argumentos, eles utilizam a definição de Fabbri [18] para definir os gêneros musicais como sendo: "um tipo de música, como ela é aceita por uma comunidade por qualquer razão, propósito ou critério". As principais conclusões apresentadas neste trabalho são:

1. Para aumentar o desempenho dos sistemas de classificação automática de gêneros musicais é necessário utilizar outras características além do timbre, como informações culturais disponíveis na web;

2. Poderia ser permitida a atribuição de mais de um gênero à cada título musical. Neste caso, os rótulos de classe poderiam ter pesos associados;

3. A aquisição de dados para *ground-truth* e sua respectiva classificação têm que ser considerados objetivos prioritários por si só;

4. Deve-se permitir uma estrutura, mesmo que simples, de ontologia mapeando as relações entre os gêneros;

5. Outra questão levantada considera que diferentes partes de uma música podem pertencer a diferentes gêneros, assim como podem ser representações diferentes do mesmo gênero e argumentam que utilizar as médias das características ao longo de longas janelas de análise ou mesmo da música inteira pode ser uma abordagem limitadora. Um alternativa seria permitir a rotulação independente de cada segmento extraído de uma música;

6. De uma perspectiva musicológica, eles desencorajam o uso de técnicas como PCA para a redução de características. Por mais que isso possa promover uma melhora na taxa de acerto. Isso limita a qualidade dos resultados de uma perspectiva teórica, pois são perdidas informações importantes, como quais características são mais úteis em diferentes contextos, e sugerem o uso de FFS e BFS assim como abordagens baseadas em algoritmos genéticos;

7. Por fim, eles apontam para a necessidade de realizarem mais pesquisas no aspecto psicológico da classificação de gêneros musicais realizadas pelas pessoas considerando especialistas, não especialistas, pessoas de diferentes idades, culturas e experiências, pois isso seria benéfico não apenas para melhorar o *ground-truth* da área como

também desenvolver diferentes sistemas para diferentes audiências e suas respectivas necessidades.

Em [49], Lidy *et al.* utilizam SSDs juntamente com descritores simbólicos e submetem estas características à um classificador SVM. O trabalho foi realizado sobre a base de dados ISMIR 2004 e no melhor caso a taxa de acerto foi igual a 81,4%.

Bagci e Erzin [2] investigam o uso de característica de timbre dinâmica e propõem classificadores que utilizam similaridade extra-classe (IGS). A similaridade extra-classe é modelada sobre amostras difíceis de classificar do espaço de características de gêneros musicais. A partir destas amostras é estabelecida uma classe, denominada classe de similaridade extra-classe. Na classificação, amostras situadas nesta classe são eliminadas para reduzir a confusão extra-classe e para melhorar o desempenho da classificação de gêneros. Resultados experimentais obtidos sobre a base GTZAN mostraram que os classificadores propostos alcançam melhor desempenho do que métodos apresentados até então. Na modelagem utilizando IGS atingiu-se, no melhor caso, uma taxa de acerto de 88,60% e com o IGS iterativo (IIGS) obteve-se 92,40%. Estas taxas são significativamente superiores às alcançadas por outros trabalhos apresentados até então que utilizaram a mesma base.

Um aspecto comum a maioria dos trabalhos da literatura é que eles estão normalmente propondo novos métodos de extração de características em conjunto com classificadores bem definidos. Como pode ser visto na proposta do ACE (*Autonomous Classification Engine*) [57], mecanismos de combinação de classificadores foram pouco estudados e utilizados para a tarefa de reconhecimento automático de gêneros musicais. Outro aspecto que só recentemente tem sido investigado neste domínio é o uso de mecanismos de seleção de atributos.

Panagakis *et al.* [69] extraíram características utilizando técnicas de subespaço multilinear: *Non-Negative Tensor Factorization* (NTF), *High-Order Singular Value Decomposition* (HOSVD) e *Multilinear Principal Component Analysis* (MPCA). A classificação foi feita com SVM sobre as bases GTZAN e ISMIR 2004, os melhores resultados obtidos foram 78,20% e 80,95% respectivamente para as duas bases.

Em [35], Holzapfel e Stylianou utilizam *Nonnegative Matrix Factorization (NMF)* para derivar um novo descritor para timbre musical. Estas características são submetidas à um classificador construído com GMM e as melhores taxas de reconhecimento obtidas foram de 83,50% sobre a base ISMIR 2004 e 74% sobre a base GTZAN.

Algumas das bases de música disponíveis publicamente apresentam algumas sérias limitações para o desenvolvimento de trabalhos de recuperação de informações de músicas. A base GTZAN, por exemplo, disponibiliza apenas os primeiros 30 segundos de cada música no formato de áudio PCM. Em casos mais extremos, como na base apresentada em [36], são disponibilizados apenas dez segundos extraídos de segmentos aleatórios de cada título musical. Além disso, algumas bases apresentadas na literatura possuem poucas músicas,

e os gêneros utilizados são em geral os mesmos (*rock* e clássica) e normalmente os gêneros são disjuntos, ou seja, não existem trabalhos com subgêneros realmente próximos como *House* e *Trance*. Dessa forma, tendo em mente o trabalho de Aucouturier e Pachet [1], em que é mostrado que definir uma taxonomia para gêneros é uma tarefa mal formulada, uma possível solução para este problema seria utilizar uma classificação um pouco mais abrangente baseada na percepção humana de como os gêneros são dançados para fazer essa classificação. Apesar de não ser abrangente o suficiente para incluir todos os gêneros musicais possíveis, essa abordagem permitiria a construção de uma base de dados usando características culturais de diversos tipos de música.

Pensando nisso, em 2008 Silla *et al.* [82] apresentam à comunidade de pesquisa em recuperação de informações musicais a *Latin Music Database* (LMD). Uma base constituída originalmente por 3227 títulos musicais de dez diferentes gêneros oriundos de países latino-americanos (Axé, Bachata, Bolero, Forró, Gaúcha, Merengue, Pagode, Salsa, Sertaneja e Tango). A base foi desenvolvida com propósitos também voltados à classificação automática de gêneros musicais e durante o seu desenvolvimento foram observadas várias características desejáveis descritas na literatura [58]. A atribuição de rótulos de gêneros aos títulos foi feita com base na percepção de especialistas humanos, que levaram em consideração inclusive a forma como a música é dançada. Dada a presença de muitos gêneros oriundos de um mesmo país ou de países com grandes semelhanças no que diz respeito aos seus aspectos culturais, esta base se mostrou particularmente desafiadora, e a tentativa de discriminar seus gêneros automaticamente se caracterizou como uma tarefa difícil.

A partir da apresentação da LMD, os mesmos autores desenvolveram outros trabalhos acerca da classificação automática de gêneros musicais utilizando a mesma. No primeiro deles [81], os autores descrevem uma abordagem diferente das convencionalmente utilizadas para classificar gêneros musicais. Múltiplos vetores de características são criados e uma abordagem de reconhecimento de padrões para combiná-los é aplicada. Um conjunto de classificadores bastante conhecidos é utilizado e é adotado um procedimento para combinar seus resultados a fim de se obter a classificação final. Os melhores resultados obtidos alcançaram taxa de acerto de 65,06%.

Em outro trabalho, Silla *et al.* [83] utilizam uma abordagem para a seleção de características no processo de classificação. Nos experimentos realizados, os autores constroem vetores com características tradicionalmente utilizadas em classificação automática de gêneros musicais extraídas de três segmentos dispersos ao longo do sinal da música. Em seguida, são utilizados algoritmos genéticos para a seleção de características. Neste trabalho, além da LMD os experimentos também foram realizados sobre a base ISMIR 2004 e os resultados mostraram que as diferentes características têm importância variada de acordo com o segmento do sinal a partir do qual foram extraídas. Ao final, observou-se que, em conjuntos de vetores com alta dimensionalidade a seleção de características é

muito importante na busca de uma boa relação entre taxa de reconhecimento e esforço computacional.

Silla *et al.* [80] também apresentam um trabalho em que avaliam o desempenho obtido com a combinação de diferentes conjuntos de características extraídas do sinal: histograma de rítmo (RH), descritores estatísticos de espectro (SSD), coeficientes do histograma de intervalos *Inset-Onset* (IOIHC) além de outras características obtidas com o framework MARSYAS. Os autores também utilizam técnicas de seleção de características para verificar um eventual ganho de performance. Os experimentos são desenvolvidos sobre as bases LMD e ISMIR 2004 e as melhores taxas obtidas são 89,53% sobre a base LMD e 82,43% sobre a base ISMIR 2004.

Paradzinets *et al.* [72] apresentam um trabalho no qual utilizam um novo conjunto de características acústicas que chamam *Piecewise Gaussian Model (PGM)* sobre um conjunto com 1873 títulos musicais tomados da *Magnatune database* com os seguintes gêneros: *Classic*, *Dance*, *Jazz*, *Metal*, *Rap* e *Rock*. Os autores criam classificadores especialistas em características criadas com histogramas de batidas, características relacionadas ao timbre e características criadas com PGM. Ao final, o melhor resultado acontece quando se cria um comitê com todos estes classificadores e o melhor resultado alcança taxa de 80,9% de reconhecimento.

Panagakis *et al.* [70] utilizam propriedades da percepção auditiva humana para representar as músicas na classificação de gêneros musicais. Os experimentos são realizados com as bases GTZAN e ISMIR 2004 e os autores apresentam o uso do classificador SRC (*Sparse Representation-based Classification*) para este tipo de problema. Ao final são descritos resultados obtidos com os classificadores SRC, Rede Neural e SVM. Os melhores resultados são obtidos com SRC e as taxas são iguais a 91% para a base GTZAN e 93,56% para a base ISMIR 2004. Em [71], os mesmos autores utilizam características derivadas por *Locality Preserving Non-Negative Tensor Factorization (LPNTF)* e submetem a um classificador SRC. Os experimentos são realizados novamente com as bases GTZAN e ISMIR 2004. As melhores taxas de reconhecimento obtidas são 92.4% e 94.38% respectivamente.

Pohle *et al.* descrevem em [73] o uso de características de rítmo e timbre na classificação de gêneros musicais. Os autores utilizam o classificador k-NN e os melhores resultados obtidos são, 90,4% sobre a base ISMIR 2004 e 57% sobre a *HOMBURG set*.

Seyerlehner e Schedl [78] apresentam características que chamam de *block-level features*. Os autores defendem que estas características apresentam a vantagem de capturar mais informações temporais do que outros tipos de características. Os autores utilizam o classificador SVM e obtêm 82,72% de reconhecimento com a base ISMIR 2004 e 77,96% com a base GTZAN. Em [79], Seyerlehner *et al.* utilizam o mesmo tipo de características também com o classificador SVM e obtêm 88,27% utilizando a base ISMIR 2004 e 85,49% com a base GTZAN. Os autores disputam o concurso MIREX 2010 utilizando o mesmo

método e alcançam 79,86% de taxa de reconhecimento sobre a base LMD utilizando *"artist filter"*, o melhor resultado até então.

Lidy *et al.* [50] discutem a compatibilidade dos métodos tradicionais empregados em recuperação de informações musicais com músicas orientais ou alguns tipos particulares de músicas de etnias peculiares, já que estes estilos musicais possuem características completamente diferentes das músicas ocidentais tradicionalmente utilizadas nestes estudos, geralmente em forma de músicas gravadas em estúdio e masterizadas. Os autores realizam os experimentos sobre três bases: uma com músicas orientais, a LMD e uma coleção de músicas africanas. Considerando as peculiaridades das músicas orientais e de músicas étnicas, tanto em termos de conteúdo musical quanto em termos de características de gravação, os resultados mostraram que as abordagens experimentadas funcionaram, de forma geral, surpreendentemente bem. Este trabalho caracterizou uma importante iniciativa no sentido de verificar a robustez de métodos tradicionais quando aplicados à coleções musicais de estilos diferentes.

Lopes *et al.* [52] apresentam um método que utiliza o que denominam seleção de instâncias de treinamento. Estas instâncias correspondem a vetores com características de tempo curto e de baixo nível extraídas do sinal do áudio e a seleção se dá com base em resultados obtidos com o classificador SVM. Os experimentos foram realizados sobre um conjunto de 900 títulos musicais tomados da LMD e os resultados finais indicam uma pequena melhora nas taxas de acerto no reconhecimento dos gêneros, que foi de 59,6%. Entretanto, os autores indicam que o modelo de classificação foi reduzido significativamente, permitindo uma classificação mais rápida. Embora os resultados deste trabalho apresentem taxas de acerto inferiores às de outros trabalhos realizados sobre a mesma base, deve-se considerar que foi empregado o *"artist filter"* quando da separação dos títulos atribuídos aos conjuntos de treinamento e de teste. Assim, fica caracterizado um grau de dificuldade significativamente superior para a realização da tarefa, de modo que os resultados podem ser considerados dignos de crédito.

Em [55], Marques *et al.* investigam evidências de que as características comuns de baixo nível não são representativas para a classificação de gêneros musicais. Os autores utilizaram 17 características de baixo nível extraídas com o framework MARSYAS e experimentaram diferentes tipos de classificadores nos experimentos. Sobre a base ISMIR 2004 a melhor taxa de reconhecimento foi igual a 79,8% e sobre a base LMD, utilizando *"artist filter"*, a melhor taxa foi de 64,9%. Em [54], os mesmos autores investigam o espaço de características de tempo curto e avaliam a precisão das mesmas em tarefas de classificação de gêneros utilizando novamente diferentes classificadores. Os autores utilizam novamente as mesmas bases de músicas e os melhores resultados obtidos são 83,03% sobre a base ISMIR 2004 fazendo a classificação utilizando HMMs e 71,61% sobre a base LMD com a restrição *"artist filter"* e utilizando mesmo classificador.

Mayer e Rauber [56] investigam a combinação de classificadores criados com caracte-

rísticas tradicionais de áudio, como RH, RP e SSD, com outros criados com características obtidas a partir das letras das músicas. Os autores criam uma base de músicas com 600 títulos de 10 diferentes gêneros, que chamam de base pequena. Sobre esta base, os autores obtêm no melhor caso uma taxa de acerto de 65,83% utilizando o classificador SVM. Para certificar-se da confiabilidade dos resultados, os autores criam uma outra base com 3010 títulos dos mesmos 10 gêneros, que chamam de base grande. Sobre esta base, a melhor taxa de acerto é igual a 75,08% fazendo a fusão das saídas dos classificadores com BWWV.

Marques *et al.* [53] utilizam os classificadores Naïve Bayes, SVM e *Optimum Path Forest (OPF)* sobre um subconjunto de títulos musicais da base GTZAN e sobre a base Magnatagatune. Para os títulos da base GTZAN foram utilizadas 26 diferentes características de MFCCs, para a base Magnatagatune foi utilizado um conjunto de características de timbre disponibilizadas junto com a base. Os resultados obtidos com os três classificadores foram bastante próximos entre si, sendo a taxa mais alta, de 98,72%, obtida com SVM sobre a base GTZAN. Sobre a base Magnatagatune, o melhor desempenho também foi obtido com SVM e a taxa foi de 63,15% de reconhecimento. Ao final, os autores ressaltam que com o uso do classificador OPF, o tempo gasto nas tarefas de treinamento e teste é, em geral, muito inferior ao tempo gasto com os outros classificadores e as taxas de reconhecimento são praticamente as mesmas. Isto pode ser bastante interessante para aplicações de tempo real.

Wu *et al.* experimentam em [90] o uso de características extraídas diretamente do sinal, que chamam características acústicas, juntamente com características extraídas de imagens de espectrogramas gerados a partir do sinal, que chamam de características visuais. Os autores utilizam características acústicas de tempo curto como *Octave-based Spectral Contrast (OSC)* e MFCCs entre outras de tempo longo. O conjunto de características acústicas é utilizada com um método baseado em *Gaussian Supper Vector (GSV)*. Como características visuais, os autores utilizam filtros de Gabor para extrair características da textura presente nos espectrogramas. Utilizando o classificador SVM, os melhores resultados obtidos são, 86,1% sobre a base GTZAN e também 86,1% sobre a base ISMIR 2004.

Também no contexto de representação de conteúdo musical, Orio [67] apresenta uma metodologia baseada em Modelos Escondidos de Markov (*Hidden Markov Models* - HMM) para modelagem estatística de conteúdo de áudio e descreve a aplicação da mesma em duas bases de músicas étnicas, sendo uma com músicas dos balcãs e outra com canções italianas. Em ambos os casos a identificação foi realizada pela modelagem do conteúdo melódico. É importante observar que o trabalho não envolveu classificação de gêneros, e sim a identificação de versões diferentes de uma mesma música. De qualquer forma, a tentativa de representar conteúdo é algo comum aos trabalhos de classificação de gêneros musicais. Embora as coleções utilizadas sejam pouco representativas diante do vasto repertório de músicas étnicas, os resultados indicam que o modelo estatístico e as características

acústicas utilizadas podem ser empregadas em outras coleções, especialmente naquelas em que o conteúdo melódico tenha um papel importante. As taxas de reconhecimento foram boas o suficiente para sugerir a aplicação do método na identificação de gravações alternativas de um mesmo título musical.

Uma grande parte das pesquisas no reconhecimento automático de gêneros musicais foca na distribuição de características de baixo nível do sinal. Algoritmos de aprendizagem de máquina são utilizados para fazer o mapeamento entre estes descritores de baixo nível e conceitos musicais de alto nível como, por exemplo, gêneros musicais. Esta abordagem tem propiciado relativo sucesso, mas está limitada em diferentes aspectos. Por exemplo, a representação de baixo nível oculta os aspectos verdadeiramente relevantes de uma música como, por exemplo, ritmo e harmonia. Foi demonstrado recentemente que estes modelos não correspondem à percepção que os seres humanos têm da música [1].

Além disto, as experiências de reconhecimento automático de gêneros musicais são feitas principalmente com categorias musicais relativamente amplas e gerais, como por exemplo, *pop*, *rock*, clássico, etc. Acredita-se que um entendimento melhor pode ser ganho nos aspectos relevantes do que faz dois títulos musicais similares (ou pertencentes ao mesmo gênero) usando uma categorização muito mais fina dos dados. Adicionalmente, a comunidade de pesquisa tem aceito que outres fontes de informação, além do próprio sinal, como metadados criados por usuários, podem ser bastante uteis e contribuir com a obtenção de melhores taxas no reconhecimento de gêneros musicais.

No trabalho aqui proposto, procura-se um novo conjunto de características de baixo nível que possa ser útil em tarefas de reconhecimento de gêneros musicais baseado em conteúdo. Para isto, investiga-se o uso de características obtidas no domínio de frequências, a partir de imagens de espectrograma. Os trabalhos de Yu e Slotine [92] e de Deshpand *et al.* [14] já haviam tratado previamente da classificação de sinal de áudio com o uso de imagens de espectrogramas extraídos do sinal. O primeiro trabalho é voltado ao reconhecimento de instrumentos musicais, os autores tentam classificar automaticamente o áudio de 8 diferentes instrumentos musicais e conseguem uma taxa geral de acerto próxima de 85%. O segundo trabalho é voltado ao reconhecimento de gêneros musicais, e os autores alcançaram na melhor situação uma taxa média de acerto de 75%, entretanto, foram utilizados apenas três diferentes gêneros musicais e várias questões acerca da exequibilidade da classificação automática de gêneros musicais a partir de características extraídas de imagens de espectrogramas permaneceram em aberto. A busca de respostas para estas questões é objeto deste trabalho.

## 2.1 Conclusões

Este capítulo descreveu o histórico que mostra como evoluíram as tarefas de pesquisa em classificação automática de gêneros musicais como uma tarefa de reconhecimento de

padrões a partir do primeiro trabalho proposto neste sentido, apresentado em 2002. A tabela 2.1 sumariza alguns dados acerca da maior parte dos trabalhos descritos neste capítulo, que realizaram ações de classificação de gêneros musicais.

Tabela 2.1: Síntese dos resultados de trabalhos em classificação automática de gêneros musicais

| **Autores** | **Ano** | **Características** | **Classificador** | **Base/gêneros** | **Melhor acerto** |
|---|---|---|---|---|---|
| Deshpande *et al.* | 2001 | Espectrograma e MFCC | k-NN, SVM e modelo Gaussiano | 157 músicas 3 gêneros | 75,00% |
| Tzanetakis e Cook | 2002 | Textura de timbre, conteúdo rítmico e variações da frequência de vibração | k-NN e GMM | 10 gêneros, 1000 músicas (GTZAN) | 61,00% |
| Grimaldi *et al.* | 2003 | DWPT | k-NN | 200 músicas 5 gêneros | 65,00% |
| Li *et al.* | 2003 | DWCH | SVM | Base 1: 10 gêneros, 1000 músicas (GTZAN)<br>Base 2: 5 gêneros, 756 músicas | 72,00%<br>78,00% |
| Costa *et al.* | 2004 | Textura de timbre e ritmo | MLP e k-NN | 414 músicas, 2 gêneros | 90,30% |
| Lippens *et al.* | 2004 | MFCC e ritmo | classificador Gaussiano | 6 gêneros, 160 títulos | 69,00% |
| Koerich e Poitevin | 2005 | Textura de timbre e ritmo | MLP | 414 músicas, 2 gêneros | 95,97% |
| Li e Ogihara | 2005 | DWCH | SVM | 10 gêneros, 1000 músicas (GTZAN)<br>5 gêneros, 1458 músicas | 72,70%<br>81,00% |
| Hu *et al.* | 2005 | *reviews* | Naïve Bayes | 1800 músicas, 12 gêneros | 78,89% |
| Meng *et al.* | 2005 | Tempo curto, tempo médio e tempo longo. | Perceptron e classificador Gaussiano | Base 1: 100 músicas, 5 gêneros | 95,00% |

continua na próxima página

continuação da página anterior

| **Autores** | **Ano** | **Características** | **Classificador** | **Base/gêneros** | **Melhor acerto** |
|---|---|---|---|---|---|
| | | | | Base 2: 354 músicas, 6 gêneros | 68,00% |
| Flexer *et al.* | 2005 | MFCC | GMM | 6 gêneros, 1458 músicas (ISMIR 2004) | 78,19% |
| Lidy e Rauber | 2005 | SSD e RH | SVM | 10 gêneros, 1000 músicas (GTZAN) | 74,90% |
| | | | | 6 gêneros, 1458 músicas (ISMIR 2004) | 80,32% |
| Yaslan e Cataltepe | 2006 | Beat, Mpitch, STFT e MFCC | Fisher, LDC, QDC, UDC, NB, PDC e k-NN | 10 gêneros, 1000 músicas (GTZAN) | 80,00% |
| Bergstra *et al.* | 2006 | FFTC, RCEP, MFCC, ZCR, *spectral centroid* e LPC | ADABOOST | Base Magnatune | 75,1% |
| | | | | Base USPOP | 86,92% |
| Ezzaidi e Rouat | 2006 | MFCC | GMM | 10 gêneros, 100 músicas (RWC) | 73,00% |
| Lidy *et al.* | 2007 | SSD, *onset* e descritores simbólicos | SVM | 6 gêneros, 1458 músicas (ISMIR 2004) | 81,40% |
| Bagci e Erzin | 2007 | Textura de timbre e conteúdo rítmico | GMM | 10 gêneros, 1000 músicas (GTZAN) | 88,60% (IGS), 92,40% (IIGS) |
| Flexer *et al.* | 2007 | MFCC | GMM | 1458 títulos, 6 gêneros (ISMIR 2004) | 75,72% 61,22%* |
| Panagakis *et al.* | 2008 | NTF, HOSDV e MPCA | SVM | 1458 títulos, 6 gêneros (ISMIR 2004) | 80,95% |
| | | | | 10 gêneros, 1000 músicas (GTZAN) | 78,20% |
| Holzapfel e Stylianou | 2008 | NMF | GMM | 1458 títulos, 6 gêneros (ISMIR 2004) | 83,50% |

continua na próxima página

continuação da página anterior

| **Autores** | **Ano** | **Características** | **Classificador** | **Base/gêneros** | **Melhor acerto** |
|---|---|---|---|---|---|
| | | | | 10 gêneros, 1000 músicas (GTZAN) | 74,00% |
| Silla *et al.* | 2008 | Textura de timbre, conteúdo rítmico e variações da frequência de vibração | J48, 3NN, MLP, NB e SVM | 10 gêneros, 3227 músicas (LMD) | 65,06% |
| Silla *et al.* | 2009 | MARSYAS, IOIHC, RH e SSD | J48, 3NN, MLP, NB e SVM | 10 gêneros, 3227 músicas (LMD) | 84,70% |
| | | | | 6 gêneros, 1458 músicas (ISMIR 2004) | 77,21% |
| Paradzinets *et al.* | 2009 | Historgramas de batidas, timbre e PGM | Rede Neural | 6 gêneros, 1873 músicas (Magnatune) | 80,90% |
| Panagakis *et al.* | 2009 | *Auditory Temporal Modulation* | SRC | 10 gêneros, 1000 músicas (GTZAN) | 91,00% |
| | | | | 6 gêneros, 1458 músicas (ISMIR 2004) | 93,56% |
| Panagakis *et al.* | 2009 | Derivadas por LPNTF | SRC | 10 gêneros, 1000 músicas (GTZAN) | 92,40% |
| | | | | 6 gêneros, 1458 músicas (ISMIR 2004) | 94,38% |
| Pohle *et al.* | 2009 | Rítmo e timbre | k-NN | 6 gêneros, 1458 músicas (ISMIR 2004) | 90,40% |
| | | | | *HOMBURG set* | 57,00% |
| Seyerlehner e Schedl | 2009 | *Block-level* | SVM | 10 gêneros, 1000 músicas (GTZAN) | 77,96% |
| | | | | 6 gêneros, 1458 músicas (ISMIR 2004) | 82,72% |
| Seyerlehner *et al.* | 2010 | *Block-level* | SVM | 10 gêneros, 1000 músicas (GTZAN) | 85,49% |

continua na próxima página

continuação da página anterior

| **Autores** | **Ano** | **Características** | **Classificador** | **Base/gêneros** | **Melhor acerto** |
|---|---|---|---|---|---|
| | | | | 6 gêneros, 1458 músicas (ISMIR 2004) | 88,27% |
| | | | | 10 gêneros, 3227 músicas (LMD) | 79,86%* |
| Lidy *et al.* | 2010 | MARSYAS, IOIHC, RP, SSD, RH, Temporal | SVM, time decomposition | 10 gêneros, 3227 músicas (LMD) | 88,06% |
| | | SSD e MVD | | 6 gêneros, 1458 músicas (ISMIR 2004) | 80,37% |
| | | | | Coleção de mús. africanas: 1024 títulos | 88,57% |
| Lopes *et al.* | 2010 | Tempo curto, baixo nível (MARSYAS) | SVM | 10 gêneros, 900 músicas (LMD) | 59,60%* |
| Marques *et al.* | 2010 | Tempo curto, baixo nível (MARSYAS) | HMM | 10 gêneros, 900 músicas (LMD) | 64,90%* |
| | | | | 6 gêneros, 1458 músicas (ISMIR 2004) | 79,80% |
| Marques *et al.* | 2011 | Tempo curto, baixo nível (MARSYAS) | HMM | 10 gêneros, 900 músicas (LMD) | 71,61%* |
| | | | | 6 gêneros, 1458 músicas (ISMIR 2004) | 83,03% |
| Silla *et al.* | 2011 | IOIHC, RH, SSD e MARSYAS | SVM | 10 gêneros, 3227 músicas (LMD) | 89,53% |
| | | | | 6 gêneros, 1458 músicas (ISMIR 2004) | 82,43% |
| Mayer e Rauber | 2011 | RP, RH, SSD e *Lyrics feature subspace* | SVM | 10 gêneros, 600 músicas | 65,83% |
| | | | | 10 gêneros, 3010 músicas | 74,08% |
| Marques *et al.* | 2011 | MFCC | Naïve Bayes, OPF e SVM | Subconjunto de GTZAN | 98,72% |
| | | *timbral features* | | Magnatagatune | 63,15% |

continua na próxima página

continuação da página anterior

| **Autores** | **Ano** | **Características** | **Classificador** | **Base/gêneros** | **Melhor acerto** |
|---|---|---|---|---|---|
| Wu *et al.* | 2011 | GSV e filtros de Gabor | SVM | 10 gêneros, 1000 músicas (GTZAN) | 86,10% |
| | | | | 6 gêneros, 1458 músicas (ISMIR 2004) | 86,10% |
| Goulart *et al.* | 2012 | Entropia dos frames | SVM | 3 gêneros, 90 músicas | 92,60% |

* *artist filter*

No próximo capítulo será apresentada uma fundamentação teórica com conceitos que sustentam o desenvolvimento desta tese. Serão apresentadas algumas das principais abordagens presentes na literatura para extração de características de textura presentes em imagens digitais. Estas técnicas são potenciais candidatas para suportar a etapa de extração de características do projeto aqui proposto. Adicionalmente, serão descritos os fundamentos de um método proposto recentemente para a seleção dinâmica de agrupamento de classificadores, algumas das abordagens mais conhecidas para combinação de saídas de classificadores e os principais fundamentos de algoritmos genéticos, que serão empregados com o propósito de selecionar características em alguns experimentos adicionais descritos no capítulo 6.

# CAPÍTULO 3

# FUNDAMENTAÇÃO TEÓRICA

Classificação é o um dos problemas abordados pela disciplina de reconhecimento de padrões. Neste contexto, entende-se por classificação o problema de atribuir uma classe $c_i$ à um vetor de características $x$, extraídas de um item a ser classificado, aqui chamado padrão.

Reconhecimento de escrita, reconhecimento de impressões digitais, reconhecimento de fala e reconhecimento de faces são alguns exemplos clássicos de domínios de aplicação de reconhecimento de padrões. Segundo Duda *et al.* [16], a abordagem clássica para o desenvolvimento de sistemas para o reconhecimento de padrões prevê algumas etapas bem definidas, conforme mostra a figura 3.1.

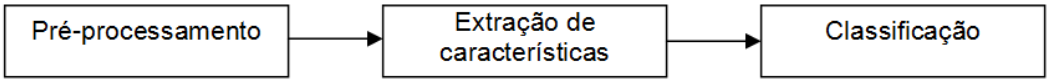



Figura 3.1: Etapas para o reconhecimento de padrões.

Existe uma vasta literatura acerca de cada uma destas etapas e muito pode ser descrito sobre cada uma delas. Entretanto, as etapas de extração de características e de classificação são particularmente desafiadoras e sobre elas, muitos esquemas diferentes vêm sendo propostos.

Trabalhos recentes têm realizado a etapa final, de classificação, empregando diversos classificadores e fazendo a combinação entre os mesmos. Esta abordagem tem apresentado bons resultados em muitos diferentes domínios de aplicação. Muitas vezes, a combinação é feita através de uma regra de fusão, que cumpre o papel de combinar as saídas de vários classificadores. Esta situação é ilustrada na figura 3.2. Nas próximas seções serão descritos alguns aspectos relacionados a extração de características e combinação de classificadores, inclusive no que diz respeito ao domínio para o qual este trabalho é voltado.

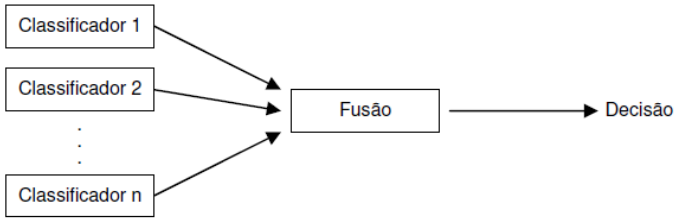



Figura 3.2: Combinação das saídas de classificadores.

## 3.1 Extração de características

A extração de características é uma etapa crucial dentro do desenvolvimento de um sistema de reconhecimento de padrões. Segundo Scaringella *et al.* [77], uma vez que características significativas foram extraídas, qualquer esquema de classificação pode ser utilizado. No caso de aplicações voltadas para a classificação sinais de áudio com conteúdo musical, as características devem ser relacionadas às principais dimensões da música, incluindo melodia, harmonia, ritmo, timbre e localização espacial. A subseção 3.1.1 apresenta os três principais tipos de características de baixo nível definidas e utilizadas nos principais trabalhos de classificação de gêneros musicais baseada em conteúdo presentes na literatura [87], [77] e [1].

Adicionalmente, é importante observar que o principal objetivo desta proposta está relacionado à extração de características de espectrogramas originados a partir do sinal de áudio das músicas. Com isso, observa-se um mapeamento do formato original do sinal para um formato diferente, no domínio visual. Considerando que os espectrogramas gerados caracterizam imagens digitais cujo principal atributo visual é a textura, serão descritas na subseção 3.1.2 algumas das muitas abordagens apresentadas na literatura para extrair características de textura que possam ser utilizadas para descrever o conteúdo das mesmas em sistemas de classificação.

### 3.1.1 Características de baixo nível para representação de conteúdo musical

Ainda em 2006, Scaringella *et al.* [77] apontavam que em aplicações do mundo real, meta-dados descritores de uma nova música raramente estavam disponíveis e era necessário lidar diretamente com uma amostra de áudio. Mais recentemente, iniciativas como a Last.fm passaram a disponibilizar meta-dados atribuídos por usuários a títulos musicais. De qualquer forma, este procedimento não dispensa a necessidade de intervenção humana no processo.

Amostras de áudio, obtidas pela amostragem do som em forma de onda, não podem ser usadas diretamente por sistemas de análise automática. Neste formato, o sinal apresenta uma quantidade de dados muito grande. Assim, o primeiro passo dos sistemas de análise é a extração de algumas características dos dados do áudio para manipular informação mais significativa e reduzir a necessidade de processamento posterior. Os três principais tipos de características de baixo nível utilizadas em trabalhos de classificação automática de gêneros musicais, empregados desde o clássico trabalho de Tzanetakis e Cook [87], são descritos a seguir.

**Relacionadas ao timbre**

Timbre é geralmente definido na literatura como uma característica perceptual que faz com que dois sons com a mesma frequência e intensidade sejam diferentes. Características relacionadas ao timbre analisam a distribuição espectral do sinal, embora algumas delas sejam computadas no domínio do tempo. Estas características são globais no sentido de que elas integram a informação de todas as fontes e instrumentos ao mesmo tempo.

**Relacionadas à variação de frequência (Melodia/Harmonia)**

A harmonia é as vezes referenciada como o elemento vertical de música e a melodia o elemento horizontal. A análise de melodia e harmonia tem sido utilizada há muito tempo por musicólogos para estudar a estrutura de músicas e é bastante sugestiva a idéia de integrar este tipo de análise na modelagem de gêneros.

**Relacionadas ao ritmo**

Não existe uma definição precisa para ritmo. Muitos autores o relacionam a regularidade temporal. De forma geral, a palavra ritmo pode ser usada para fazer referência a todos os aspectos temporais de uma peça musical [77]. Alguns autores sugerem que um classificador automático não deve levar em conta apenas descritores de "timbre global", mas também deve levar em consideração o ritmo. Segundo Tzanetakis e Cook [87], para o reconhecimento do gênero musical, algumas características desejáveis de se representar no vetor de características dizem respeito à regularidade do ritmo, a relação da batida principal com batidas secundárias e a força das batidas secundárias em relação à batida principal. Em seu trabalho eles utilizam um histograma de batidas construído a partir da função de autocorrelação do sinal: verificando o peso de diferentes periodicidades no sinal (e as taxas entre estes pesos). A partir disto tem-se uma idéia da força e complexidade da batida na música.

A tabela 3.1 mostra de forma sucinta as características utilizadas por Tzanetakis e Cook [87] e as categorias a que pertencem de acordo com a taxonomia aqui descrita.

Tabela 3.1: Categorias de características empregadas na classificação de gêneros musicais

| **Timbre** | **Melodia/Harmonia** | **Ritmo** |
|---|---|---|
| Centróide espectral | *Full wave rectification* | Características de tempo real e |
| *Spectral rolloff* | Filtragem passa-baixa | arquivo inteiro |
| Fluxo espectral | *Downsampling* | |
| Cruzamento de zero no | *Mean removal* | |
| domínio de tempo | *Enhenced autocorrelation* | |
| Coeficientes cepstrais de | Detecção de picos e cálculo de | |
| frequência de Mel (MFCC) | histograma | |
| Análise e janela de textura | Características de histograma | |
| Características de baixa energia | de batidas | |

### 3.1.2 Representação de textura

A textura é um importante atributo visual presente nas imagens do mundo real. Assim como cor e forma, a textura é facilmente percebida pelo olho humano e contribui com a identificação de objetos em uma dada cena. Apesar de facilmente percebido pelos humanos, este atributo não possui uma definição formal. Conforme Jain e Farrokhnia [40], a diversidade de texturas naturais e artificiais torna impossível estabelecer uma definição universal para a mesma.

A textura corresponde a um padrão visual, geralmente relacionado à distribuição de pixels em uma região e características do objeto da imagem, como tamanho, brilho e cor. Este atributo geralmente contém informações bastante significativas acerca do conteúdo da imagem e é amplamente explorado em aplicações de visão computacional. Algumas texturas apresentam uma regularidade no que diz respeito a repetição de padrões que aparecem na mesma, enquanto outras não. A figura 3.3(c) mostra um exemplo de textura regular, enquanto os outros exemplos presentes na figura 3.3 são de texturas irregulares.

(a) mármore (b) papel (c) parede (d) madeira (e) areia

Figura 3.3: Amostras de textura.

As texturas são descritas cotidianamente como finas, grossas, granuladas, lisas, etc., implicando na necessidade da definição de algumas características mais precisas para tornar o reconhecimento por máquina possível. De acordo com Tamura *et al.* [86], estas características, correspondentes aos atributos visuais comuns de texturas podem ser estudadas e quantificadas para fins de identificação, diferenciação e classificação de texturas. Estas características são:

- Granularidade: refere-se ao tamanho das células presentes na imagem, eventualmente referida como "espessura". As células podem ser definidas como sendo áreas com aproximadamente o mesmo brilho. Uma textura com células grandes é considerada grossa, enquanto que as texturas finas são aquelas formadas por pequenas células;

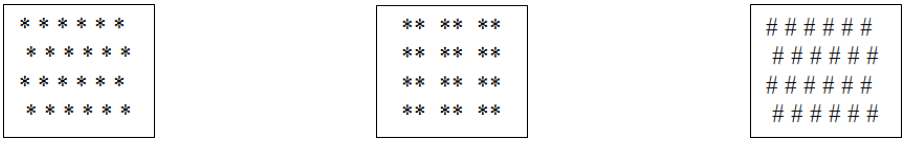
- Contraste: medido pelas variações de tons de cinza presentes na imagem. Uma alta variação destes tons nos limites das células de uma imagem caracteriza uma imagem com alto contraste, uma baixa variação de tons nestes mesmos limites caracteriza baixo contraste;

- Direcionalidade: refere-se ao fato de uma textura ter uma direção principal de ocorrência dos elementos constituintes que pode ser vertical, horizontal, inclinada ou não-direcional;
- Alinhamento: uma imagem pode ou não ter linhas. A presença ou não destas é medida por este atributo;
- Regularidade: diz respeito a regularidade com que os elementos da textura se repetem no espaço;
- Rugosidade: uma textura áspera pode ser identificada visualmente e as imagens destas texturas apresentam contornos que transmitem a sensação de aspereza mesmo sem a possibilidade de tocá-las, eventualmente referida como "aspereza".

Tais características podem ser encontradas no tom e na estrutura de uma textura. O tom é baseado principalmente nas propriedades de intensidade de pixel na primitiva da textura, enquanto a estrutura é baseada no relacionamento espacial entre as primitivas.

Cada pixel é caracterizado pela sua localização e sua propriedade de tom. Uma primitiva de textura é um conjunto de pixels contínuos com alguma propriedade de tom e/ou localização, e pode ser descrita pela sua média de intensidades, intensidade máxima ou mínima, tamanho, forma, etc. O relacionamento espacial entre as primitivas pode ser aleatório, ou pode haver uma dependência mútua entre algumas primitivas. A imagem de textura é então descrita pelo número e tipos de primitivas e pelos seus relacionamentos espaciais [85].

As figuras 3.4(a) e 3.4(b) mostram que o mesmo tipo de primitivas não produz necessariamente a mesma textura. Similarmente, as figuras 3.4(a) e 3.4(c) mostram que o mesmo relacionamento espacial de primitivas não garante uma textura unívoca. Assim, apenas uma destas características não é suficiente para a descrição de textura. O tom e a estrutura da textura não são independentes; texturas sempre exibem tanto tom quanto estrutura mesmo que um deles normalmente predomine sobre o outro, e seja aparentemente mais marcante. O tom pode ser entendido como as propriedades de tom das primitivas, levando em consideração relacionamentos espaciais primitivos. Estrutura refere-se a relacionamentos espaciais de primitivas considerando também suas propriedades de tom.



(a) (b) (c)

Figura 3.4: Diferentes primitivas de textura e relacionamento espacial entre elas. [85]

Se as primitivas de textura em uma imagem são pequenas, tem-se uma textura fina 3.3(e). Se as primitivas de textura são grandes e consistem de vários pixels, tem-se uma

textura grossa 3.3(c). Novamente, esta é uma razão para se utilizar tanto propriedades de tom quanto propriedades de estrutura na descrição de uma textura. Note que a caracterização de uma textura como grossa ou fina depende da escala utilizada.

Adicionalmente, as texturas podem também ser classificadas de acordo com a sua força - a força da textura influencia a escolha do método de descrição de textura. Uma textura é dita fraca quando tem pequenas interações espaciais entre as primitivas, e pode ser descrita adequadamente por freqüências de tipos primitivos que aparecem em alguma vizinhança. Por causa disto, muitas propriedades estatísticas de textura são avaliadas na descrição de texturas fracas. Em texturas fortes, as interações espaciais entre primitivas são um tanto regulares. Para descrever texturas fortes, a freqüência de ocorrência de pares de primitivas em algum relacionamento espacial pode ser suficiente [85].

Este trabalho é voltado para a classificação de gêneros musicais através de características extraídas de imagens de espectrogramas gerados a partir do sinal do áudio. O espectrograma mostra como a densidade do espectro do sinal varia em função do tempo e o conteúdo presente nas imagens de espectrogramas possui como principal atributo visual a textura. A figura 3.5 mostra um exemplo típico de imagem de espectrograma extraída de sinal de áudio e empregada em experimentos aqui realizados, descritos nos capítulos 5 e 6. Considerando este fato, é oportuna a investigação das principais técnicas para a representação de textura propostas na literatura e algumas delas serão detalhadas nesta subseção.

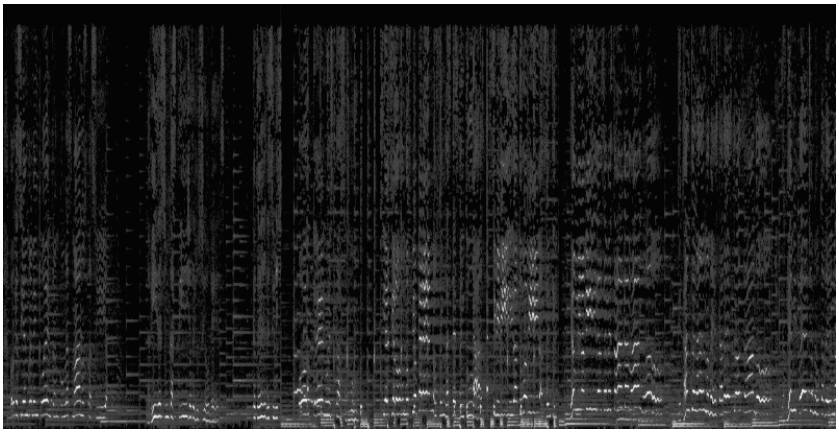


Figura 3.5: Exemplo de imagem digital de espectrograma.

São varias as tentativas de descrever ou caracterizar texturas através de medidas objetivamente extraídas de imagens digitais. Segundo Gonzalez e Woods [28] as técnicas para representação de textura podem ser divididas em três categorias: representação estatística, representação espectral e representação estrutural.

Em princípio, abordagens estruturais seriam mais adequadas para texturas mais dotadas de regularidade, enquanto as abordagens estatísticas caracterizariam modelos em que a textura é vista como uma amostra de um processo estocástico bidimensional que pode ser

descrito por seus parâmetros estatísticos. Esta abordagem tem potencial para melhores resultados quando aplicada em texturas naturais, como grama, água e etc. [19]. Entretanto, Sánchez-Yáñez *et al.* [84] chamam atenção para o fato de que qualquer textura contém tanto características regulares quanto outras de natureza estatística. Na prática, pode-se encontrar texturas entre estes dois extremos, completamente regulares (periódicas) ou completamente aleatórias. Isto explica porque é tão difícil descrever texturas em geral por um único método.

A seguir serão descritas as abordagens para a representação de textura que foram utilizadas nos experimentos desenvolvidos neste trabalho. Além de descritas, as abordagens serão devidamente contextualizadas dentre as três categorias estabelecidas por Gonzalez e Woods.

#### 3.1.2.1 Representação estatística

As técnicas estatísticas para representação de textura se concentram basicamente na extração de medidas estatísticas obtidas a partir da contagem de ocorrências dos níveis de cinza presentes nos pixels da imagem ou obtidas através da forma como pixels de diferentes níveis de cinza se relacionam no espaço bidimensional da imagem. Uma importante observação acerca destas técnicas é que a unidade primitiva a partir da qual as medidas estatísticas são obtidas é o pixel da imagem. A seguir será descrita a Matriz de Co-ocorrência de Niveis de Cinza (GLCM). Criada por Haralick [32], esta é a mais tradicional abordagem estatística para a descrição de texturas e provavelmente a mais utilizada ao longo da história. Por esta razão, esta foi a técnica escolhida dentre as abordagens estatísticas para os experimentos realizados neste trabalho.

**Matriz de co-ocorrência**

A abordagem estatística que utiliza matriz de co-ocorrência permite a caracterização da textura através de medidas estatísticas extraídas das probabilidades de relacionamento espacial entre pixels de diferentes intensidades de cor. Atributos como lisura, rugosidade e granularidade entre outros que podem ser associados à imagem da qual se extrai medidas de textura.

Mesmo tendo sido proposta por Haralick há quase quarenta anos, esta abordagem ainda é empregada para a representação de textura em muitos trabalhos. A seguir será descrita em detalhes a sequência de passos que permite a extração de características utilizando esta abordagem.

O uso de matrizes de co-ocorrência para a extração de características de textura de imagens digitais foi originalmente proposto para a aplicação em imagens em níveis de cinza. Daí seu nome original *Gray Level Co-occurrence Matrix* de onde vem o acrônimo popularmente empregado para designá-la, GLCM.

A idéia fundamental concernente ao método consiste na construção da matriz de co-ocorrência para a posterior extração de medidas estatísticas a partir das mesmas. A matriz construída é uma matriz quadrada, de ordem $N \times N$, na qual $N$ corresponde ao número de tons de cinza utilizados na representação da imagem. Em cada posição da matriz é armazenada a probabilidade de que dois valores de intensidades de cinza estejam envolvidos por uma determinada relação espacial. Parâmetros como a distância $d$ entre os pixels e o ângulo $\theta$ caracterizado pela orientação da reta que passa pelos mesmos definem esta relação espacial. As possíveis orientações para $\theta$ preconizadas por Haralick *et al.* são $0^o, 45^o, 90^o$ e $135^o$, conforme ilustra a Figura 3.6.

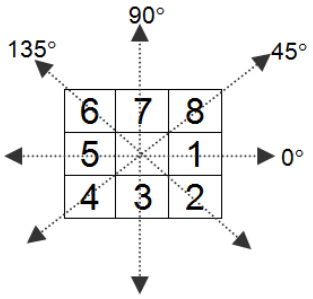



Figura 3.6: Orientações utilizadas para a formação da GLCM.

A fim de ilustrar o processo de construção de uma matriz de co-ocorrência, será descrito um exemplo. Considere que a representação contida na figura 3.7 corresponda à uma matriz de pixels cujos valores das intensidades podem variar entre zero e três.

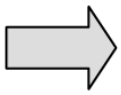
| | | | | |
|---|---|---|---|---|
| 0 | 1 | 2 | 2 | 1 |
| 0 | 1 | 1 | 0 | 3 |
| 3 | 0 | 0 | 3 | 3 |
| 2 | 1 | 3 | 0 | 2 |
| 3 | 1 | 1 | 0 | 2 |

Figura 3.7: Matriz de pixels correspondente à uma imagem.

A partir da representação descrita na figura 3.7, considerando a orientação $\theta$=$0^o$ e distância $d$=1, será composta a matriz de co-ocorrências. De acordo com o método proposto originalmente por Haralick *et al.*, a matriz de co-ocorrência registra na posição $(i, j)$ o número de ocorrências de relação espacial entre um pixel com intensidade $i$ e um pixel com intensidade $j$ considerando a distância $d$ e a orientação $\theta$ independentemente do sentido da relação. Assim, a presença de um pixel de intensidade $j$ imediatamente à direita de um pixel de intensidade $i$ seria contabilizada na matriz com $d$=1 e $\theta$=$90^o$ da mesma forma como a ocorrência da intensidade $j$ imediatamente à esquerda de $i$ seria contabilizada. Com isso, a matriz de co-ocorrência que se forma é simétrica. Depois de contadas as quantidades das relações espaciais, elas são transformadas em probabilidades para a

realização dos processos de extração de características subseqüentes, conforme mostra a figura 3.8.

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | 2 | 4 | 2 | 4 |
| 1 | 4 | 4 | 3 | 2 |
| 2 | 2 | 3 | 2 | 0 |
| 3 | 4 | 2 | 0 | 2 |



| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | 0,05 | 0,1 | 0,05 | 0,1 |
| 1 | 0,1 | 0,1 | 0,075 | 0,05 |
| 2 | 0,05 | 0,075 | 0,05 | 0 |
| 3 | 0,1 | 0,05 | 0 | 0,05 |

Figura 3.8: Matriz de co-ocorrência obtida para $\theta$=0$^o$ e $d$=1.

Haralick *et al.* [31] propuseram originalmente 14 medidas de características de texturas possíveis de se extrair das matrizes de co-ocorrência. Estas características são calculadas a partir de algumas equações que utilizam as probabilidades associadas as posições da matriz de co-ocorrências.

Das 14 características originalmente propostas, sete se consolidaram como características relevantes em processos de descrição de textura. Estas características são: contraste, energia (ou uniformidade), entropia, homogeneidade, momento de terceira ordem, probabilidade máxima e correlação. Sendo $G$ o número de intensidades de cinza utilizado na representação da imagem e $p(i,j)$ a probabilidade de relacinamento entre as intensidades $i$ e $j$, as equações 3.1 a 3.7 descrevem como estas características são encontradas respectivamente.

$$Contraste = \sum_{i=1}^{G}\sum_{j=1}^{G}(i-j)^2\ p(i,j) \tag{3.1}$$

$$Energia = \sum_{i=1}^{G}\sum_{j=1}^{G}(p(i,j))^2 \tag{3.2}$$

$$Entropia = -\sum_{i=1}^{G}\sum_{j=1}^{G}p(i,j)\ log\ p(i,j) \tag{3.3}$$

$$Homogeneidade = \sum_{i=1}^{G}\sum_{j=1}^{G}\frac{p(i,j)}{1+(i-j)^2} \tag{3.4}$$

$$Momento\ de\ terceira\ ordem = \sum_{i=1}^{G}\sum_{j=1}^{G}p(i,j)(i-j)^3 \tag{3.5}$$

$$Probabilidade\ máxima = \sum_{i=1}^{G}\sum_{j=1}^{G}max\ p(i,j) \tag{3.6}$$

$$Correlação = \frac{p(i,j)-\mu_x\mu_y}{\sigma_x^2\sigma_y^2} \tag{3.7}$$

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

#### 3.1.2.2 Representação espectral

As técnicas espectrais são baseadas em propriedades da função de densidade espectral e detectam a periodicidade global além de picos de energia no espectro. Estas técnicas possuem a vantagem de ser invariantes à escala, com isso, independem da resolução espacial segundo a qual a imagem está representada. A seguir serão descritos os Filtros de Gabor, uma das mais importantes abordagens presentes nesta categoria.

**Filtros de Gabor**

Durante muito tempo um sinal podia ser representado em função do tempo ou, alternativamente, em função da frequência através da transformada de Fourier. Entretanto, esta abordagem possuía a limitação de permitir a extração de informações apenas no domínio de frequência e não em função do tempo. Em 1946, Dennis Gabor apresentou os filtros de Gabor, que permitem extrair informações no domínio de frequência e tempo. Em seu trabalho original Gabor buscava a síntese do sinal e preocupou-se em como um sinal poderia ser construído através da combinação linear de funções elementares [42]. Os filtros de Gabor correspondem à um conjunto de funções senoidais complexas, bidimensionais, moduladas por uma função Gaussiana também bidimensional com propriedades muito úteis para a finalidade de classificação de imagens. Na análise de sinais em processamento de imagens, a extração de características exerce um papel importante no qual o principal objetivo é saber "o que está aonde". Com os princípios de Gabor, informações relacionadas a frequência podem informar "o que", enquanto as ligadas ao tempo podem informar "aonde" [42].

A segmentação de textura é uma tarefa difícil e muito importante em muitas aplicações de análise de imagens ou visão computacional e filtros de Gabor têm sido utilizados com êxito para estes propósitos. Existem muitas formas de se implementar filtros de Gabor apresentadas na literatura. Uma possível forma para filtros de Gabor bidimensionais no domínio espacial, portanto apropriados para imagens digitais, é dada pelas equações 3.8 e 3.9 [90].

$$\Psi(x,y) = exp\left( -\left(\frac{x^2+y^2}{2\sigma^2}\right)\right) exp\left(\frac{j2\pi x}{\lambda}\right) \tag{3.8}$$

na qual $j$ é a unidade imaginária, $\sigma$ é o desvio padrão da função Gaussiana e $\lambda$ é o comprimento de onda.

Para uma imagem $I$ de tamanho $M \times N$, e considerando $\Psi(x,y)$ conforme descrito na

equação 3.8, a saída do filtro de Gabor é obtida pela convolução da imagem de entrada com o filtro de Gabor (equação 3.9).

$$\sum_{x}\sum_{y} I(m-x, n-y)\Psi(x, y) \tag{3.9}$$

Filtros de Gabor podem ser utilizados para detectar linhas. Uma vez que a imagem pode conter linhas com diferentes espessuras, é necessário construir filtros de Gabor com diferentes fatores de escala, variando $\lambda$. Adicionalmente, o filtro de Gabor original pode detectar somente linhas verticais, o que não é suficiente em muitos casos, já que é comum a ocorrência de linhas com diferentes orientações nas imagens. Assim, pode-se rotacionar $\Psi(x, y)$ com um ângulo $\theta$ para construir $\Psi(x', y')$ para a detecção de linhas com diferentes orientações. Neste caso, $x'$ e $y'$ podem ser encontrados por 3.10 e 3.11 respectivamente.

$$x' = x\ cos\ \theta\ +\ y\ sen\ \theta \tag{3.10}$$

$$y' = -x\ sen\ \theta\ +\ y\ cos\ \theta \tag{3.11}$$

#### 3.1.2.3 Representação estrutural

As técnicas estruturais descrevem a textura a partir de relacionamentos espaciais entre certas primitivas identificadas na imagem. Com isso, a unidade básica utilizada para tentar caracterizar a textura são estas primitivas, muitas vezes chamadas de *texton*. Uma vez definidos os *textons*, são aplicados processos que procuram avaliar a disposição dos mesmos ao longo da imagem. Em geral, estes métodos funcionam bem para texturas bastante regulares. A seguir será descrita a abordagem LBP (*Local Binary Pattern*) para a representação de textura, uma técnica que vem sendo aplicada com bastante sucesso em diferentes domínios de aplicação.

**LBP**

O acrônimo LBP vem do termo em inglês *Local Binary Pattern* que, em português, seria algo como Padrão Local Binário. Este método foi introduzido inicialmente como uma medida complementar para contraste local da imagem [63]. Posteriormente, o método foi adaptado e se tornou uma abordagem estrutural para descrição de textura, conforme apresentaram Ojala *et al.* [64]. A aplicação de LBP como descritor de textura se baseia no fato de que certos padrões binários locais à região de vizinhança de um pixel são propriedades fundamentais da textura de uma imagem e que o histograma de ocorrência destas características é provavelmente uma poderosa característica de textura.

Neste método, a textura é descrita levando-se em consideração para cada pixel $C$, $P$ vizinhos equidistantes considerando-se uma distância $R$, conforme mostra a figura 3.9.

Um histograma $h$ de padrões LBP é encontrado utilizando-se as diferenças de intensidade entre cada pixel $C$ e seus $P$ vizinhos. Conforme descrito por Ojala *et al.* [64], boa parte da informação sobre características de textura é preservada na distribuição $T$ descrita na equação 3.12.

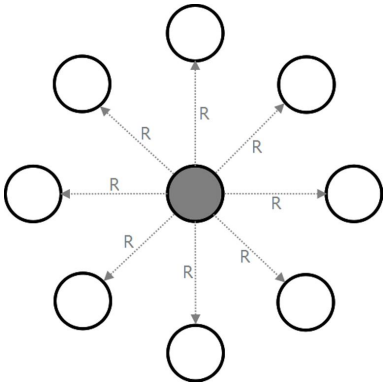



Figura 3.9: Operador LBP. Pixel $C$, círculo escuro ao centro, seus $P$ vizinhos, círculos claros.

$$T \approx (g_0 - g_C, ..., g_{P-1} - g_C) \tag{3.12}$$

na qual $g_C$ é a intensidade nível de cinza do pixel $C$ (pixel central) e $g_0$ a $g_{P-1}$ correspondem as intensidades de nível de cinza dos $P$ vizinhos. Quando um vizinho não corresponde exatamente à posição de um pixel, seu valor é obtido por interpolação.

Considerando o sinal resultante da diferença entre o pixel $C$ e cada vizinho, como descrito na equação 3.13, é definido que: se o sinal é positivo, o resultado é igual a um; caso contrário, o resultado é igual a zero, como descrito na equação 3.14.

$$T \approx (s(g_0 - g_C), ..., s(g_{P-1} - g_C)) \tag{3.13}$$

na qual

$$s(g_i - g_C) = \begin{cases} 1 & \text{se } g_i - g_C \geq 0 \\ 0 & \text{se } g_i - g_C < 0 \end{cases} \tag{3.14}$$

na qual $i = [0, P]$ é o índice dos vizinhos de C.

Com isto, o valor do padrão LBP inerente ao pixel $C$ corrente pode ser obtido através da multiplicação dos elementos binários por um coeficiente binomial. Associando-se um peso binomial $2^P$ a cada $s(g_P - g_C)$, as diferenças presentes na vizinhança são transformadas em um único codigo LBP, um valor $0 \leq C' \leq 2^P$. A equação 3.15 descreve como este código é obtido.

$$LBP_{P,R}(x_C, y_C) = \sum_{P=0}^{P-1} s(g_P - g_C)2^P \tag{3.15}$$

assumindo que $x_C \in \{0, ..., N-1\}$ e $y_C \in \{0, ..., M-1\}$ para uma imagem com dimensões $N \times M$.

Ojala *et al.* [64] introduziram o conceito de uniformidade da sequência obitda no padrão LBP. Este conceito é baseado no número de transições entre zeros e uns presente na sequência associada ao padrão. Um código binário LBP é considerado uniforme se o número de transições é menor ou igual a dois, considerando inclusive que o código é tratado como uma lista circular. Assim, o código representado pela sequência 00100100 não é considerado uniforme, já que contém quatro transições. Por outro lado, o código 00100000 é considerado uniforme, já que apresenta apenas duas transições. A figura 3.10 ilustra este conceito.

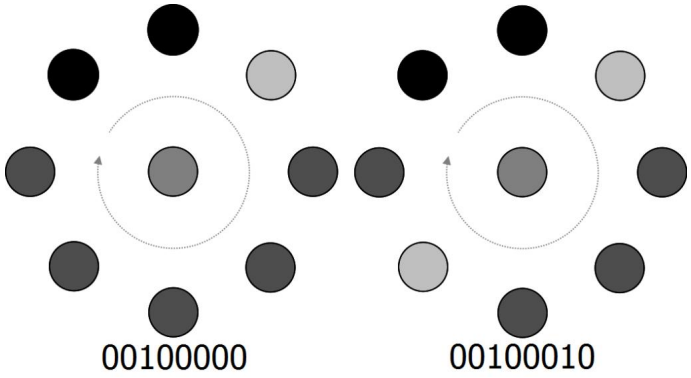



Figura 3.10: Uniformidade do padrão LBP. (a) com apenas duas transições, o padrão é considerado uniforme. (b) com quatro transições, o padrão não é considerado uniforme.

Desta forma, ao invés de utilizar integralmente o historgrama de padrões LBP, cujo tamanho é $2^P$, é possível utilizar apenas os valores associados a padrões uniformes, constituindo um vetor com menor dimensionalidade, com apenas 59 características. Ojala et al. [64] estabeleceram que, além das 58 possíveis combinações uniformes, todos os padrões não uniformes encontrados devem participar de uma coluna adicional do histograma. Por este motivo, o vetor final de padrões LBP construído na versão mais tradicional, em que o número de vizinhos $P$ é igual a 8 e o valor de $R$ é igual a 2, apresenta 59 valores. Esta versão do descritor foi chamada "u2", um rótulo que acompanha os valores do raio $R$ e o tamanho da vizinhança $P$, fazendo a sua descrição final da seguinte forma: $LBP^{u2}_{8,2}$.

Adicionalmente, observou-se nos experimentos realizados neste trabalho que a extração de características com $LBP^{u2}_{8,2}$ é rápida, e precisa o suficiente para a aplicação proposta. O valor de $R$ está relacionado à resolução espacial da imagem. A alteração do valor de $R$ tornaria o processo de extração de características mais lento. Alguns experimentos com valores diferentes para $P$ e $R$ foram realizados, os resultados mostram que a configuração com $R = 2$ e $P = 8$ apresenta a melhor relação custo benefício.

**LPQ**

O borramento é uma forma de degradação de imagens digitais que podem prejudicar consideravelmente a análise das mesmas. Este ruído geralmente tem origem relacionada a problemas de aquisição e, em geral, o uso de algoritmos para removê-los é computacionalmente caro. Pensando nisso, Ojansivu e Heikkila propuseram em [65] um novo método para análise de textura insensível ao borramento. É interessante observar que, embora o método tenha sido criado com este propósito, ele também produz resultados muito bons para imagens não acometidas por este ruído.

O descritor, denominado *Local Phase Quantization* (LPQ) é baseado na propriedade de invariância ao borramento do espectro de fase de Fourier. Ele utiliza a informação de fase local extraída utilizando a 2D DFT computada sobre uma vizinhança retangular, chamada janela local, para cada pixel da imagem. A informação da fase local de uma imagem de tamanho $N \times N$ é dada pela $STFT$ (*Short-time Fourier Transform*) descrita na equação 3.16.

$$\hat{f}_{u_i}(x) = (f \times \Phi_{u_i})x \tag{3.16}$$

sendo o filtro $\Phi_{u_i}$ dado pela equação 3.17

$$\Phi_{u_i} = e^{-j2\pi u_i^T y} | y \in \mathbb{Z}^2 || y || \infty \leq r \tag{3.17}$$

na qual *r=(m - 1)/2*, m é o tamanho da janela local e $u_i$ é um vetor de frequências 2D.

No LPQ são considerados apenas quatro coeficientes complexos que correspondem às frequências 2D: $u_1 = [a, 0]^T$, $u_2 = [0, a]^T$, $u_3 = [a, a]^T$, $u_4 = [a, -a]^T$, em que $a = 1/m$. Por conveniência, a STFT (equação 3.16) é expressa através do vetor de notação conforme a equação 3.18.

$$\hat{f}_{u_i}(x) = w_{u_i}^T f(x) \tag{3.18}$$

sendo $F = [f(x_1), f(x_2), ..., f(x_{n^2})]$ denotado como uma matriz $m^2 \times N^2$ que compreende a vizinhança de todos os pixels na imagem e $w = [w_R, w_I]^T$, em que $w_R = Re[w_{u_1}, w_{u_2}, w_{u_3}, w_{u_4}]$ e $w_I = Im[w_{u_1}, w_{u_2}, w_{u_3}, w_{u_4}]$. O $Re[]$ e $Im[]$, representam, respectivamente, as partes reais e imaginárias de um número complexo e a matriz de transformação $(8 \times N^2)$ é dada por $\hat{F} = wF$.

Ojansivu e Heikkilä assumem que a função $f(x)$ de uma imagem é resultado de um processo de primeira ordem de Markov, em que o coeficiente de correlação entre dois pixels $x_i$ e $x_j$ é relacionado exponencialmente com sua distância $L^2$. Para o vetor $f$ é definida uma matriz de covariância $C$ de tamanho $m^2 \times m^2$, dada pela equação 3.19. A matriz de covariância dos coeficientes de Fourier pode ser obtida por $D = wCw^T$. Considerando

que $D$ não é uma matriz diagonal, os coeficientes são correlatos e podem deixar de ser através de $E = V^T \hat{F}$, sendo $V$ uma matriz ortogonal derivada do valor de decomposição singular (SVD - *Singular Value Decomposition*) da matriz $D$, com $D' = V^T DV$.

$$C_{i,j} = \sigma^{||x_i - x_j||} \tag{3.19}$$

Os coeficientes são quantizados usando-se a equação 3.20, em que $e_{i,j}$ são os componentes de $E$. Estes elementos são transformados de binário para decimal através da equação 3.21 e caracterizam valores inteiros compreendidos entre zero e 255. Então, através de todas as posições da imagem, é composto o vetor de 256 posições que corresponde ao histograma LPQ.

$$q_{i,j} = \begin{cases} 1 & \text{se } e_{i,j} \geq 0 \\ 0 & \text{caso contrário} \end{cases} \tag{3.20}$$

$$b_j = \sum_{i=0}^{7} q_{i,j} 2^i \tag{3.21}$$

## 3.2 Combinação e seleção de classificadores

Dentre os algoritmos mais empregados na etapa de classificação pode-se mencionar árvores de decisão, redes neurais, k-NN (*k Nearest Neighbors*), SVM (*Support Vector Machines*) e LDA (*Linear Discriminant Analisys*) [16]. De forma geral, estes algoritmos foram propostos originalmente com o objetivo de viabilizar a construção de um classificador único capaz de resolver um determinado problema.

Com o passar do tempo, passou a se desenvolver diferentes esquemas de classificação para buscar a solução de um problema de reconhecimento de padrões. Embora algum dos esquemas projetados alcance melhor desempenho do que os outros, os conjuntos de padrões classificados incorretamente pelos diferentes classificadores não necessariamente se sobrepõem. Isto sugere que diferentes projetos de classificadores potencialmente oferecem informação complementar sobre os padrões a serem classificados que poderia ser aproveitada para melhorar o desempenho do classificador selecionado [41].

Estas observações motivaram o interesse relativamente recente em combinação de classificadores. A combinação de classificadores é uma área de pesquisa conhecida na literatura por diferentes nomes: comitê, mistura, agrupamento, *pool*, etc. A idéia é de não contar apenas com um único esquema para tomar a decisão. Ao invés disso, todos os projetos, ou um subconjunto deles, são utilizados para tomar a decisão pela combinação das suas opiniões individuais a fim de produzir uma decisão de consenso. Muitos esquemas de combinação de classificadores têm sido planejados e vem sendo demonstrado experimentalmente que alguns deles superam consistentemente o classificador de melhor

desempenho individualmente [41]. A diversidade entre os classificadores membros de uma combinação é apontada como uma característica muito importante e que deve contribuir muito para a obtenção de bons resultados [45].

Face ao exposto, é válido descrever a forma pela qual Dietterich [15] descreve o problema padrão de aprendizagem supervisionada, já introduzindo notação para a descrição de vários classificadores, dando margem à discussão subsequente acerca de combinação de classificadores, que é a questão central desta seção: à um programa de aprendizagem são dados exemplos de treinamento, da forma $\{(x_1, y_1), ..., (x_m, y_m)\}$ para alguma função *y=f(x)* desconhecida. Os valores $x_i$ são tipicamente vetores da forma $\langle x_{i,1}, x_{i,2}, ..., x_{i,n}\rangle$ cujos componentes são valores discretos ou reais, tais como altura, peso, cor e idade entre outros. Estes valores também são chamados frequentemente de características de $x_i$. Os valores $y$ são tirados de um conjunto discreto de classes $\{1, ..., k\}$, no caso de classificação. Dado um conjunto $S$ de exemplos de treinamento, um algoritmo de aprendizagem produz um classificador. O classificador é uma hipótese acerca da verdadeira função *f*. Dados novos valores de $x$ (novos padrões), ele prediz os valores correspondentes para $y$. Os classificadores serão aqui denotados $h_1, ..., h_L$.

Segundo Dietterich [15], dois classificadores são complementares se cometem erros diferentes para novos padrões. Para ilustrar a importância da diversidade, imagine que se tenha um agrupamento entre três classificadores: $\{h_1, h_2, h_3\}$ e considere um novo padrão $x$. Se os três classificadores forem idênticos (portanto não complementares), então quando $h_1(x)$ estiver errado, $h_2(x)$ e $h_3(x)$ também estarão errados. Entretanto, se os erros cometidos pelos classificadores não são correlacionados, quando $h_1(x)$ estiver incorreto, $h_2(x)$ e $h_3(x)$ podem estar corretos, de forma que o voto majoritário entre os resultados das saídas pode classificar $x$ corretamente.

De forma geral, Dietterich [15] aponta três razões para o fato de que frequentemente é possível construir bons agrupamentos de classificadores. A primeira razão é estatística. Um algoritmo de aprendizagem pode ser visto como algo que busca um espaço $H$ de hipóteses para identificar dentro dele a melhor hipótese. O problema estatístico surge quando a quantidade de dados de treinamento disponível é muito pequena comparada ao tamanho do espaço de hipóteses. Sem dados suficientes, o algoritmo de aprendizagem pode encontrar várias hipóteses diferentes em $H$, todas com a mesma precisão nos dados de treinamento, e escolher a pior hipótese sobre dados desconhecidos. Construindo um agrupamento com vários classificadores, o algoritmo pode fazer uma média entre seus votos e reduzir o risco de escolher o classificador errado. A parte superior esquerda da figura 3.11 ilustra esta situação. A curva externa descreve o espaço de hipóteses *H*. A curva interna descreve o conjunto de hipóteses que apresentam boa precisão nos dados de treinamento. O ponto rotulado com $f$ corresponde à hipótese verdadeira. Os pontos $h_1, h_2, h_3$ e $h_4$ correspondem às saídas dos respectivos classificadores. Pode-se perceber que através da média entre as hipóteses, é possível encontrar uma boa aproximação de *f*.

A segunda razão é computacional. Muitos algoritmos de aprendizagem trabalham realizando uma busca local que pode ficar presa em mínimos locais. Por exemplo, redes neurais empregam algoritmo de descida do gradiente para minimizar uma função de erro sobre os dados de treinamento e algoritmos de árvore de decisão empregam uma regra de divisão gulosa para o crescimento da árvore de decisão. Nos casos em que existem dados de treinamento suficientes (não havendo problema estatístico), ainda pode ser muito difícil em termos computacionais para o algoritmo de aprendizagem encontrar a melhor hipótese. Um agrupamento construído pela execução da busca local a partir de vários pontos de partida diferentes pode proporcionar uma melhor aproximação à verdadeira função desconhecida do que qualquer um dos classificadores individualmente, conforme ilustrado na parte superior direita da figura 3.11. Novamente, o ponto rotulado com $f$ corresponde à hipótese verdadeira. Os pontos $h_1, h_2$ e $h_3$ correspondem às saídas dos respectivos classificadores.

A terceira razão é representacional. Em muitas aplicações de aprendizagem de máquina, a função $f$ não pode ser representada por qualquer uma das hipóteses em $H$. Realizando somas ponderadas das hipóteses tiradas de $H$, pode ser possível expandir o espaço de funções representáveis. A parte inferior da figura 3.11 descreve esta situação. Novamente, o ponto rotulado com $f$ corresponde à hipótese verdadeira. Os pontos $h_1, h_2$ e $h_3$ correspondem às saídas dos respectivos classificadores.

A questão representacional é um tanto sutil, porque existem muitos algoritmos de aprendizagem para os quais $H$ é, em princípio, o espaço de todos os possíveis classificadores. Por exemplo, redes neurais e árvores de decisão são algoritmos bastante flexíveis. Fornecendo dados de treinamento suficientes, eles explorarão o espaço de todos os possíveis classificadores. Entretanto, com uma amostra de treinamento finita, estes algoritmos exploram somente um conjunto finito de hipóteses e param a busca quando encontram uma hipótese que se ajusta aos dados de treinamento. Portanto, na figura 3.11, deve-se considerar o espaço $H$ como um espaço efetivo de hipóteses pesquisadas pelo algoritmo de aprendizagem para um dado conjunto de dados de treinamento.

Estas três questões fundamentais são as três mais importantes formas pelas quais os algoritmos de aprendizagem falham. Portanto, métodos de agrupamento têm o compromisso de reduzir (e talvez eliminar) estas três falhas dos algoritmos de aprendizagem primários.

De forma similar, Jain *et al.* [39] apontam quatro diferentes razões que podem justificar o uso da combinação de classificadores:

- Pode-se ter acesso a diferentes classificadores, cada um desenvolvido em um contexto diferente e com uma representação completamente diferente do mesmo problema.
- As vezes, mais do que um único conjunto de treinamento está disponível, cada um coletado em momentos diferentes ou em ambientes diferentes. Estes conjuntos

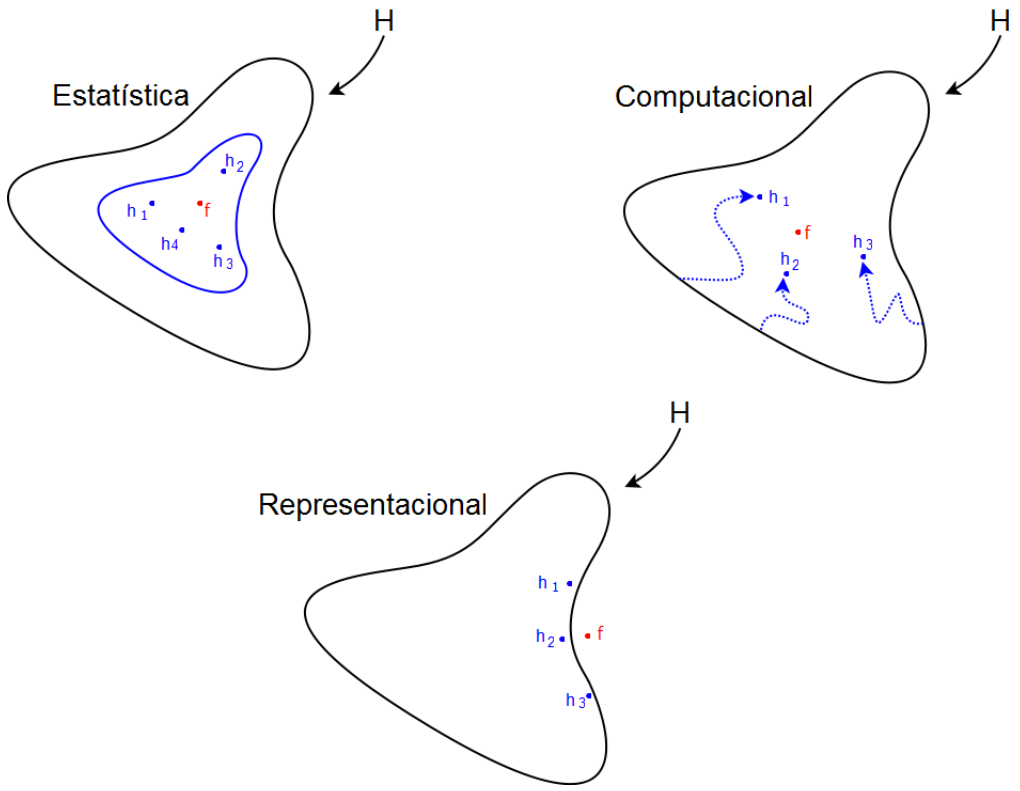



Figura 3.11: As três diferentes razões para combinar classificadores [15].

podem até mesmo utilizar diferentes características.

- Classificadores diferentes, treinados sobre os mesmos dados podem não somente se diferenciar em termos de performance global, mas também podem apresentar grandes diferenças locais. Cada classificador pode ter sua própria região no espaço de características onde obtém melhor desempenho.

- Alguns classificadores, como redes neurais, apresentam resultados diferentes quando inicializados com diferentes parâmetros dada a aleatoriedade inerente aos procedimentos de treinamento. Ao invés de selecionar a melhor rede e descartar as outras, pode-se combinar várias delas tirando proveito de todas as tentativas de aprendizagem a partir dos dados.

Duas abordagens principais para o projeto de agrupamentos de classificadores são claramente definidas na literatura: combinação de classificadores (ou fusão de classificadores) e seleção de classificadores [75]. As próximas seções descrevem os principais aspectos acerca destas abordagens.

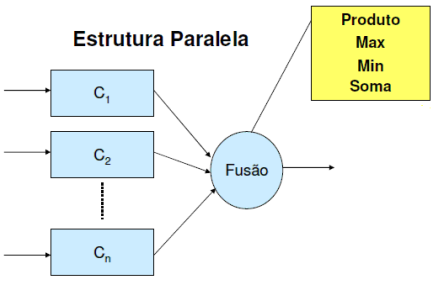
### 3.2.1 Combinação de classificadores

A operação mais comum e mais geral é a combinação das decisões de todos os classificadores membros. Voto majoritário, soma, produto, máximo e mínimo são exemplos de

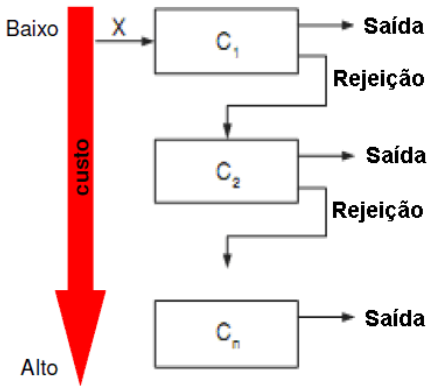
funções utilizadas para combinar decisões de membros de um agrupamento. A fusão de classificadores depende do pressuposto de que todos os membros do agrupamento cometem erros independentes. Quando a condição de independência não é verificada, não se pode garantir que a combinação da decisão de classificadores membros melhorará a performance da classificação final [75].

Jain *et al.* [39] descrevem que os vários esquemas de combinação de múltiplos classificadores podem ser agrupados, de acordo com sua arquitetura, em uma das três seguintes categorias:

- Em paralelo: os classificadores são chamados de forma independente e, posteriormente, seus resultados são combinados. A figura 3.12(a) ilustra a arquitetura em paralelo.
- Em série: classificadores independentes são chamados em uma sequência linear, conforme classificadores vão sendo chamados, o número de possíveis classes para o padrão que está sendo classificado vai diminuindo. A figura 3.12(b) ilustra a arquitetura em série.
- Hierárquico: classificadores independentes são combinados em uma estrutura que é similar à de uma árvore de decisão.




(a) paralelo




(b) série

Figura 3.12: Arquiteturas para combinação de múltiplos classificadores.

As saídas produzidas pelos classificadores podem ser divididas em três níveis: abstrato, ranking e probabilidades. Nas saídas abstratas o classificador gera apenas o rótulo da classe escolhida. Na saída com ranking, o classificador gera uma lista ordenada que indica a sequência de classes possíveis para o padrão corrente, da mais provável para a menos provável. Na saída com probabilidades, são associados valores de probabilidade as saídas. Na sequência do texto serão descritas algumas das regras mais conhecidas para realizar a fusão entre as saídas dos classificadores membros de uma combinação em paralelo. Em cada situação, será indicado em que nível de saída a regra pode ser aplicada.

**Voto majoritário**

Regra mais simples e popular para combinar classificadores. Por esta regra, é feita uma votação entre os resultados produzidos nas saídas dos classificadores envolvidos na combinação. A classe que obtiver o maior número de votos é atribuída ao padrão. Na equação 3.22 é calculada a votação majoritária para uma amostra $x$, na qual $n$ é o número de classificadores, $y_i$ o rótulo de saída do $i$-ésimo classificador em um problema com os possíveis rótulos de classe $\Omega = \omega_1, \omega_2, ..., \omega_c$.

$$mv(x) = \arg\max_{k=1}^{c} \sum_{i=1}^{n} y_{i,k} \tag{3.22}$$

Quando há empate no número de votos, a escolha deve ser aleatória ou deve haver alguma estratégia de rejeição. Além de fácil implementação, esta regra pode ser empregada em saídas abstratas.

**Regra do produto**

Em [41], Kittler *et al.* utilizam teorema de Bayes para demonstrar como chegam à equação 3.23, que permite encontrar o resultado obtido com a fusão das saídas dos classificadores pela regra do produto. Esta regra, assim como as demais que serão descritas na sequência, pode ser utilizada quando as saídas dos classificadores oferecem probabilidades estimadas associadas a cada classe envolvida no problema, uma vez que utiliza as distribuições de probabilidade extraídas pelos classificadores. A regra do produto faz a combinação calculando o produtório entre as probabilidades associadas às saídas dos classificadores $c_i$.

$$pr(x) = \arg\max_{k=1}^{c} \prod_{i=1}^{n} P(\omega_k|y_i(x)) \tag{3.23}$$

Na qual $x$ é o padrão a ser classificado, $n$ é o número de classificadores envolvidos na combinação, $y_i$ o rótulo de saída do $i$-ésimo classificador em um problema com os possíveis rótulos de classe $\Omega = \omega_1, \omega_2, ..., \omega_c$ e $P(\omega_k|y_i(x))$ a probabilidade de que a amostra $x$ pertença à classe $\omega_k$ encontrada pelo $i$-ésimo classificador.

Esta regra é bastante severa, pois a ocorrência de baixa probabilidade para uma classe em um dos classificadores faz com que a probabilidade final associada à ela seja baixa. Assim, ela é indicada em geral para situações críticas, em que o erro não é tolerado. Ainda em Kittler [41], os autores deduzem, a partir da regra do produto, as regras que serão descritas na sequência.

**Regra da soma**

A regra da soma calcula o somatório entre as probabilidades associadas às saídas dos

classificadores $c_i$, dado pela equação 3.24 [41]:

$$sr(x) = \arg\max_{k=1}^{c} \sum_{i=1}^{n} P(\omega_k|y_i(x)) \tag{3.24}$$

Na qual $x$ é o padrão a ser classificado, $n$ é o número de classificadores envolvidos na combinação, $y_i$ o rótulo de saída do $i$-ésimo classificador em um problema com os possíveis rótulos de classe $\Omega = \omega_1, \omega_2, ..., \omega_c$ e $P(\omega_k|y_i(x))$ a probabilidade de que a amostra $x$ pertença à classe $\omega_k$ encontrada pelo $i$-ésimo classificador. Em [41], Kittler *et al.* comparam regras de fusão e, ao final, concluem que a regra da soma apresenta melhores resultados por possuir maior resiliência a erros de estimativa.

**Regra da média**

A regra da média calcula a média entre as probabilidades associadas às saídas dos classificadores, dada pela equação 3.25 [41]:

$$mr(x) = \frac{1}{n} \arg\max_{k=1}^{c} \sum_{i=1}^{n} P(\omega_k|y_i(x)) \tag{3.25}$$

Na qual $x$ é o padrão a ser classificado, $n$ é o número de classificadores envolvidos na combinação, $y_i$ o rótulo de saída do $i$-ésimo classificador em um problema com os possíveis rótulos de classe $\Omega = \omega_1, \omega_2, ..., \omega_c$ e $P(\omega_k|y_i(x))$ a probabilidade de que a amostra $x$ pertença à classe $\omega_k$ encontrada pelo $i$-ésimo classificador. Esta regra produz resultados parecidos aos da regra da soma.

**Regra do máximo**

A regra do máximo utiliza a maior probabilidade dentre as classes, tomando para cada classe a maior probabilidade encontrada dentre todos os classificadores. Dada pela equação 3.26 [41]:

$$max(x) = \arg\max_{k=1}^{c} max_{i=1}^{n} P(\omega_k|y_i(x)) \tag{3.26}$$

Na qual $x$ é o padrão a ser classificado, $n$ é o número de classificadores envolvidos na combinação, $y_i$ o rótulo de saída do $i$-ésimo classificador em um problema com os possíveis rótulos de classe $\Omega = \omega_1, \omega_2, ..., \omega_c$ e $P(\omega_k|y_i(x))$ a probabilidade de que a amostra $x$ pertença à classe $\omega_k$ encontrada pelo $i$-ésimo classificador. Esta regra é de baixa severidade, pois basta que uma classe obtenha bom desempenho em um dos classificadores para que tenha boa chance de ser a escolhida.

**Regra do mínimo**

A regra do mínimo utiliza a probabilidade com maior valor associado às classes, sendo

que às classes é associado o menor valor de probabilidade encontrado entre os diferentes classificadores. Dada pela equação 3.27 [41]:

$$min(x) = \arg\max_{k=1}^{c} min_{i=1}^{n} P(\omega_k|y_i(x)) \tag{3.27}$$

Na qual $x$ é o padrão a ser classificado, $n$ é o número de classificadores envolvidos na combinação, $y_i$ o rótulo de saída do $i$-ésimo classificador em um problema com os possíveis rótulos de classe $\Omega = \omega_1, \omega_2, ..., \omega_c$ e $P(\omega_k|y_i(x))$ a probabilidade de que a amostra $x$ pertença à classe $\omega_k$ encontrada pelo $i$-ésimo classificador. Esta regra é considerada severa.

### 3.2.2 Seleção de classificadores

Seleção de classificadores é uma estratégia que escolhe a partir de um conjunto de classificadores, um classificador (ou um subconjunto de classificadores) para estimar a classe à qual pertença um padrão de teste. Em relação ao momento em que se define o classificador selecionado para realizar a classificação, as técnicas de seleção podem ser divididas em duas categorias: estática e dinâmica. No primeiro caso, regiões de competência são definidas durante a fase de treinamento, enquanto no segundo caso, elas são definidas durante a fase de classificação levando em consideração as características da amostra a ser classificada [75]. Tradicionalmente esta estratégia assume que cada membro do agrupamento é um especialista em alguma região local do espaço de características. O classificador mais preciso localmente é selecionado para estimar a classe à qual pertence cada padrão de teste em particular.

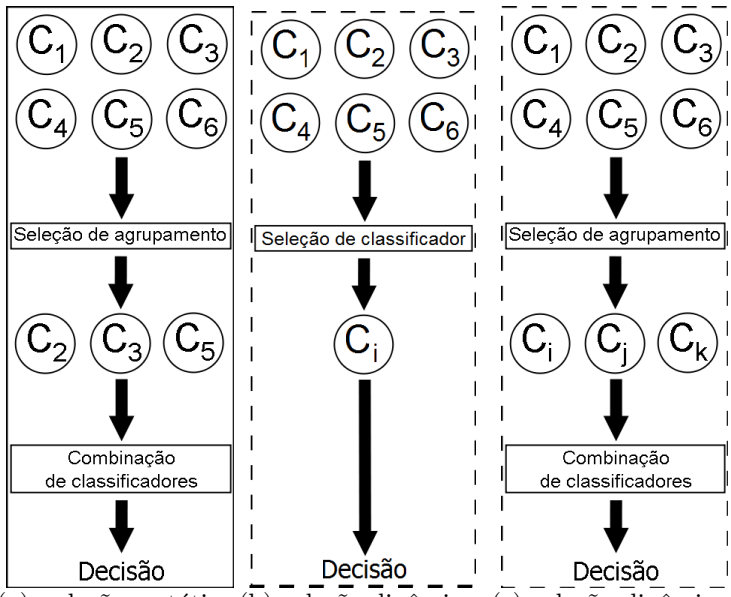
Já em relação a quantidade de classificadores selecionados, é possível encontrar esquemas que selecionem um único classificador para realizar a classificação, ou esquemas que selecionem um agrupamento de classificadores, cujos resultados são combinados posteriormente através de alguma regra de fusão [74], como as descritas na subseção 3.2.1. Ko *et al.* [43] apontam que um ponto crítico da seleção dinâmica de um único classificador é que ela depende da confiabilidade da generalização do mesmo para realizar a classificação pelos outros, já na seleção dinâmica de agrupamento, este risco é diluído entre os vários classificadores selecionados.

A figura 3.13 mostra esquemas tipicamente utilizados em seleção de classificadores. Na figura 3.13(a) é ilustrado o esquema de seleção estática de agrupamento de classificadores. Nele, o conjunto de classificadores a ser utilizado é definido na fase de treinamento e este mesmo conjunto é utilizado para classificar qualquer padrão apresentado ao sistema. As saídas dos classificadores selecionados são combinadas com o uso de algum esquema de fusão.

Na figura 3.13(b) é ilustrado o esquema de seleção dinâmica de classificador. Nele, um

único classificador é selecionado dinamicamente, levando em consideração características do padrão corrente submetido ao sistema.

A figura 3.13(c) mostra o esquema de seleção dinâmica de agrupamento de classificadores. Nele, uma combinação de classificadores é escolhida para cada caso de teste, ou seja, durante a classificação de cada padrão levando em consideração características particulares do mesmo. As saídas dos classificadores selecionados são combinadas com o uso de algum esquema de fusão, como os descritos na subseção 3.2.1, que também pode variar em função dos classificadores escolhidos.




(a) seleção estática de agrupamento (b) seleção dinâmica de classificador (c) seleção dinâmica de agrupamento

Figura 3.13: Esquemas utilizados na seleção de classificadores [43].

A subseção 3.2.2.1 descreve alguns detalhes acerca da seleção dinâmica de classificadores, em especial KNORA, um método para seleção dinâmica de agrupamento de classificadores.

### 3.2.2.1 Seleção dinâmica de classificadores

Existem diferentes métodos propostos na literatura para selecionar dinamicamente classificadores. Alguns dos métodos se propõem a selecionar um único classificador a partir do conjunto de classificadores disponível, enquanto outros selecionam um subconjunto de classificadores. A seguir, será descrito o KNORA, um método recentemente proposto para a seleção dinâmica de um conjunto de classificadores. Este método foi o escolhido para os experimentos de seleção de classificadores realizados neste trabalho por apresentar um bom potencial para explorar as possibilidades de melhoria no desempenho geral do sistema sugeridas pelos oráculos obtidos entre os diferentes conjuntos de classificadores aqui produzidos.

***KNORA***

O método *KNORA*, do inglês *K-Nearest-ORAcles*, foi apresentado por Ko *et al.* [43]. O conceito presente no método é similar aos conceitos apresentados nos métodos *Overall Local Accuracy (OLA)*, *Local Class Accuracy (LCA)*, *A Priori* e *A Posteriori* no que diz respeito ao fato de considerar a vizinhança dos padrões de teste, mas distingue-se destes métodos pelo fato de utilizar propriedades das amostras do conjunto de validação presentes na sua região de vizinhança a fim de identificar o melhor conjunto de classificadores com potencial para classificar corretamente uma dada amostra. Para cada instância de teste, *KNORA* simplesmente encontra seus $K$ vizinhos mais próximos no conjunto de validação, identifica quais classificadores classificam corretamente estes vizinhos no conjunto de validação e utiliza estes classificadores para formar o conjunto empregado na classificação do padrão dado no conjunto de teste. Os autores propõem quatro diferentes esquemas para utilização do *KNORA*: *KNORA-ELIMINATE*, *KNORA-UNION*, *KNORA-ELIMINATE-W* e *KNORA-UNION-W*, maiores detalhes sobre estas variações são apresentados a seguir.

*KNORA-ELIMINATE*

Dados $K$ vizinhos $x_j$ (com $1 \leq j \leq K$) de um padrão $X$ a ser testado, e supondo que um conjunto de classificadores $C(j)$, $1 \leq j \leq K$ classifica corretamente todos os seus $K$ vizinhos mais próximos, então todo classificador $c_i \in C(j)$ pertencente ao conjunto de classificadores $C(j)$ deve submeter um voto para a classificação da amostra $X$. A figura 3.14 ilustra esta estratégia.

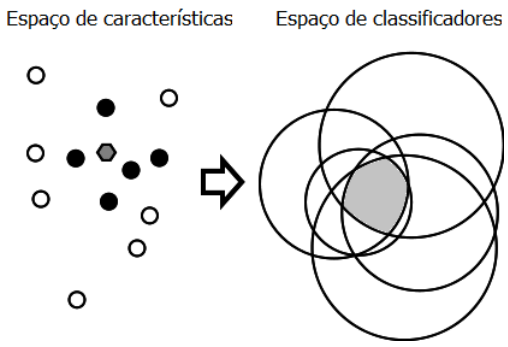



Figura 3.14: KNORA ELIMINATE utiliza apenas os classificadores que classificam corretamente todos os $K$ padrões mais próximos. O hexágono corresponde ao padrão de teste, os padrões do conjunto de validação são os circulares, sendo que os 5 mais próximos estão em preto [43].

*KNORA-UNION*

Dados $K$ vizinhos $x_j$ (com $1 \leq j \leq K$) de um padrão $X$ a ser testado, e supondo que o $j$-ésimo vizinho seja corretamente classificado por um conjunto de classificadores $C(j)$ (com $1 \leq j \leq K$). Então, todo classificador $c_i \in C(j)$ deve submeter um voto para a

classificação do padrão $X$. Observe que, uma vez que todos os $K$ vizinhos mais próximos são considerados, um classificador pode submeter mais do que um voto se ele classifica corretamente mais do que um vizinho. Quanto mais vizinhos um classificador classifica corretamente, mais votos ele submeterá para a classificação do padrão. A figura 3.15 ilustra esta estratégia.

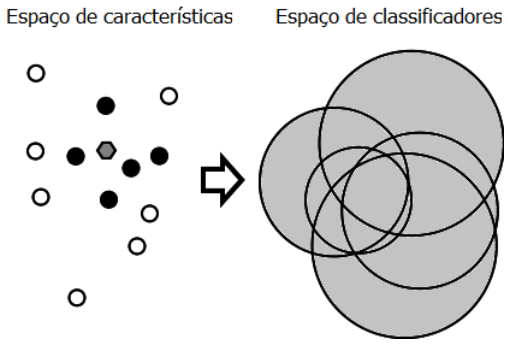



Figura 3.15: KNORA UNION utiliza os classificadores que classificam corretamente algum dos $K$ padrões mais próximos. O hexágono corresponde ao padrão de teste, os padrões do conjunto de validação são os circulares, sendo que os 5 mais próximos estão em preto [43].

*KNORA-ELIMINATE-W*

Este esquema é similar ao KNORA-ELIMINATE, mas cada voto recebe peso inversamente proporcional a distância entre o vizinho $x_j$ e o padrão de teste $X$.

*KNORA-UNION-W*

Este esquema é similar ao KNORA-UNION, mas cada voto recebe peso inversamente proporcional a distância entre o vizinho $x_j$ e o padrão de teste $X$.

## 3.3 Algoritmos Genéticos

Algoritmos Genéticos (AGs) foram criados pelo americano John Henry Holland [33] e são aplicados com bastante sucesso em problemas de busca e otimização. No contexto deste trabalho, AGs podem ser aplicados em tarefas de seleção de características, um problema de otimização.

De acordo com Goldberg [27], no que diz respeito à sua aplicação em problemas de otimização, os aspectos nos quais algoritmos genéticos diferem dos algoritmos tradicionais são:

- Baseiam-se em uma codificação do conjunto das soluções possíveis, e não nos parâmetros da otimização em si;
- Os resultados mostram uma população de soluções e não uma solução única;

- Não necessitam de nenhum conhecimento derivado do problema, apenas de uma forma de avaliação do resultado;
- Não utilizam regras determinísticas e sim transições probabilísticas.

AGs operam sobre soluções potenciais para o problema tratado. Estas soluções potenciais são chamadas indivíduos (ou cromossomos) e um conjunto delas é chamado população. Um elemento do cromossomo (gene) geralmente corresponde a um parâmetro ou dimensão do vetor numérico. Cada elemento pode ser codificado utilizando um ou vários bits. O número total de bits define a dimensão do espaço de busca.

A primeira ação realizada na execução do AG é a inicialização da população, que é feita de maneira aleatória. O tamanho desta população deve ser estabelecido previamente à execução do algoritmo, e geralmente este tamanho está entre algumas dezenas e algumas centenas de indivíduos.

Uma vez inicializada a população, a aptidão ($fitness$) de cada indivíduo pertencente à mesma é calculada. A função $fitness$ mostra quão adequado o indivíduo é a solução do problema, isto é, a adaptabilidade do indivíduo a solução do problema. A partir disto, o próximo passo consiste na reprodução, que nada mais é do que a produção de uma nova população.

A seleção dos indivíduos participantes de uma nova população pode ser feita por diferentes estratégias. De forma geral, as estratégias utilizadas privilegiam os indivíduos com maiores valores encontrados na função de $fitness$. Esta estratégia é bastante oportuna, uma vez que os indivíduos com melhor $fitness$ supostamente estão mais próximos da melhor solução para o problema e, portanto, devem se perpetuar.

Em seguida é realizada a operação de cruzamento, que consiste em trocar porções de sequências de genes dos indivíduos pais para a formação dos filhos. Nem todos os indivíduos são submetidos a operação de cruzamento para que alguns bons indivíduos gerados durante a reprodução sejam preservados. Estes indivíduos são tão somente copiados para a nova população.

Depois do cruzamento é realizada a mutação. Esta operação é realizada a fim de favorecer uma boa cobertura na busca de soluções no espaço dos possíveis estados, que poderia ser dificultada em caso de convergência para mínimos locais. A operação de mutação consiste em alterar um gene de um ou mais indivíduos da população aleatoriamente de acordo com uma probabilidade de mutação ($P_m$). Os valores de $P_m$ utilizados dependem de cada caso. Oliveira *et al.* [66] sugerem que, no caso de seleção de características, uma boa medida para o valor de $P_m$ é 1%.

A partir disto, a população está pronta para uma nova iteração do AG até que algum critério de parada seja satisfeito. Alternativamente, pode-se parar o AG tão logo se alcance um número pré-determinado de iterações.

A sequência de ações descrita anteriormente está expressa no pseudocódigo de um AG clássico, que é descrito a seguir:

$t \leftarrow 0$
Inicializar População (t)
**while** condição de término não for satisfeita **do**
    $t \leftarrow t + 1$
    Seleciona População (t) da População (t-1)
    Cruzamento População (t)
    Mutação População (t)
    Avaliação da População (t)
**end while**

## 3.4 Conclusões

Este capítulo apresentou na seção 3.1, além de uma breve introdução às características tradicionalmente utilizadas para representação de conteúdo musical, uma revisão de algumas das principais técnicas para extração de características de textura presentes em imagens digitais. As abordagens foram divididas de acordo com a classificação proposta pelos autores mais clássicos da literatura acerca de processamento de imagens. Segundo esta classificação, as abordagens se dividem em estatística, estrutural e espectral.

Na seção 3.2 foram descritos alguns dos principais aspectos relacionados ao uso de múltiplos classificadores na tentativa de obter melhores resultados para solucionar um dado problema de classificação. A subseção 3.2.1 descreve as mais conhecidas e utilizadas regras de fusão para a combinação de classificadores em paralelo. A subseção 3.2.2 discorre acerca da seleção de classificadores. São apresentados fundamentos do KNORA, um método para seleção dinâmica de agrupamento de classificadores empregado em alguns experimentos adicionais descritos no capítulo 6.

Diferentes abordagens para a descrição de textura provêem diferentes características para representá-la. Com isto, muitas vezes estas diferentes características, extraídas a partir de uma mesma abordagem ou não, podem ser correlacionadas. Na seção 3.3 foram descritos princípios básicos de AG, uma ferramenta que pode ser utilizada inclusive em tarefas de otimização. AG foi utilizado para seleção de características em experimentos adicionais também descritos no capítulo 6.

# CAPÍTULO 4

# MÉTODO PROPOSTO

Neste capítulo será apresentado o método proposto a fim de se atingir os objetivos descritos na seção 1.4. É importante relembrar que o principal objetivo desta proposta é o de realizar a classificação automática de gêneros musicais utilizando informações extraídas de imagens de espectrograma geradas a partir do sinal de áudio das músicas.

Os sistemas de reconhecimento de padrões são, de forma geral, dotados de três etapas: pré-processamento, extração de características e classificação, conforme ilustrado na figura 3.1.

A etapa de pré-processamento compreende, em geral, tarefas como a segmentação do sinal, a fim de isolar as partes interessantes do mesmo. Adicionalmente, tarefas de eliminação de ruídos também são comumente incluídas no pré-processamento, a fim de que a etapa de extração de características não seja afetada pelos mesmos. A etapa de extração de características depende fundamentalmente do tipo de sinal que se está processando. Em geral, o sinal está descrito em forma de imagem e, por isso, procura-se extrair descritores de atributos visuais como cor, textura e estrutura entre outros. Na última etapa, algoritmos de classificação bastante conhecidos são utilizados sobre os descritores extraídos a fim de se atribuir uma classe para cada padrão submetido ao sistema.

Embora as músicas não estejam descritas originalmente em formato de imagem, o sinal é convertido para este formato, já que a presente proposta trata da classificação de gêneros musicais a partir de espectrogramas. Os espectrogramas representam graficamente dados referentes a um sinal de áudio no domínio de tempo e frequência, e podem ser um instrumento bastante útil para discernir detalhes importantes acerca do mesmo [23].

No que diz respeito especificamente ao método proposto neste trabalho, pode-se identificar as seguintes etapas para realizar a tarefa de classificação: segmentação do sinal, geração das imagens de espectrograma, divisão das imagens em zonas, extração de características, construção de classificadores para cada zona criada e fusão das saídas dos classificadores (classificação). A figura 4.1 ilustra esta sequência de etapas.

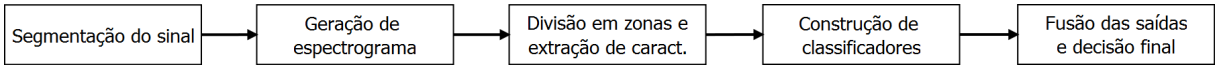



Figura 4.1: Sequência de etapas do método proposto.

A seção 4.1 descreve uma visão geral do método proposto, enquanto as seções 4.2, 4.3, 4.4, 4.5 e 4.6 descrevem detalhes acerca das etapas estabelecidas no método.

## 4.1 Visão Geral

Nesta seção é mostrado o esquema geral da classificação, levando-se em conta desde a etapa de segmentação do sinal, até a fusão das saídas dos diferentes classificadores criados, que leva a produção do resultado final de classificação. A figura 4.2 ilustra as etapas iniciais, que realizam a segmentação do sinal e geração das imagens de espectrograma. Após estas etapas, obtem-se as imagens, que serão matéria prima para o desenvolvimento da tarefa de classificação.

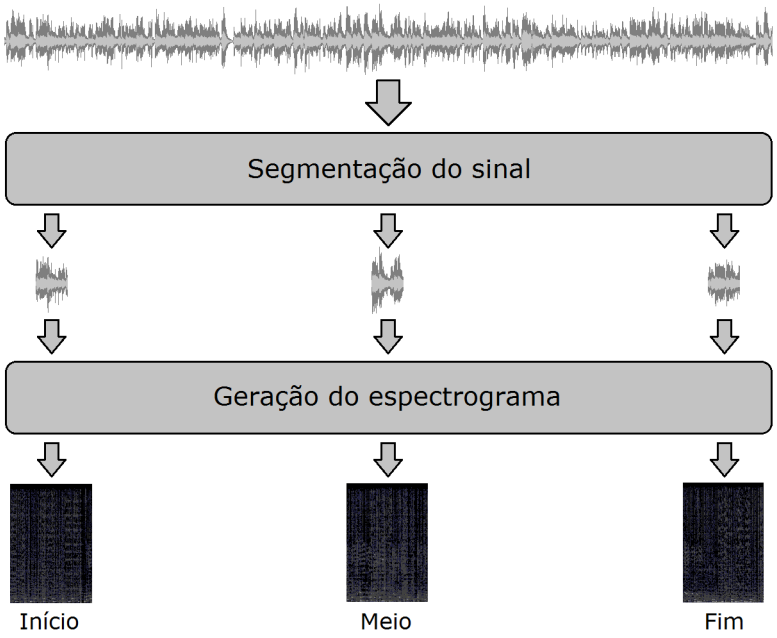



Figura 4.2: Segmentação do sinal e geração dos espectrogramas.

Depois de geradas as imagens de espectrograma, o passo seguinte consiste em criar zonas nas mesmas, a partir das quais serão extraídas características preservando-se alguma informação sobre localização espacial destas. A figura 4.3 ilustra esta etapa.

As características extraídas de cada zona são submetidas à um classificador específico. Em seguida, as predições para as classes colhidas nas saídas dos classificadores são fundidas a fim de se produzir uma decisão final, conforme ilustrado na figura 4.4.

## 4.2 Segmentação do sinal

Inspirados no trabalho de Costa *et al.* [6], os experimentos realizados neste trabalho adotaram uma estratégia de segmentação do sinal. A segmentação do sinal permite reduzir o volume de dados e consequentemente a quantidade de processamento a ser realizado em etapas subsequentes.

Nos experimentos aqui descritos, foi utilizada uma estratégia segundo a qual três segmentos foram extraídos do sinal. O uso de três segmentos é bastante oportuno na

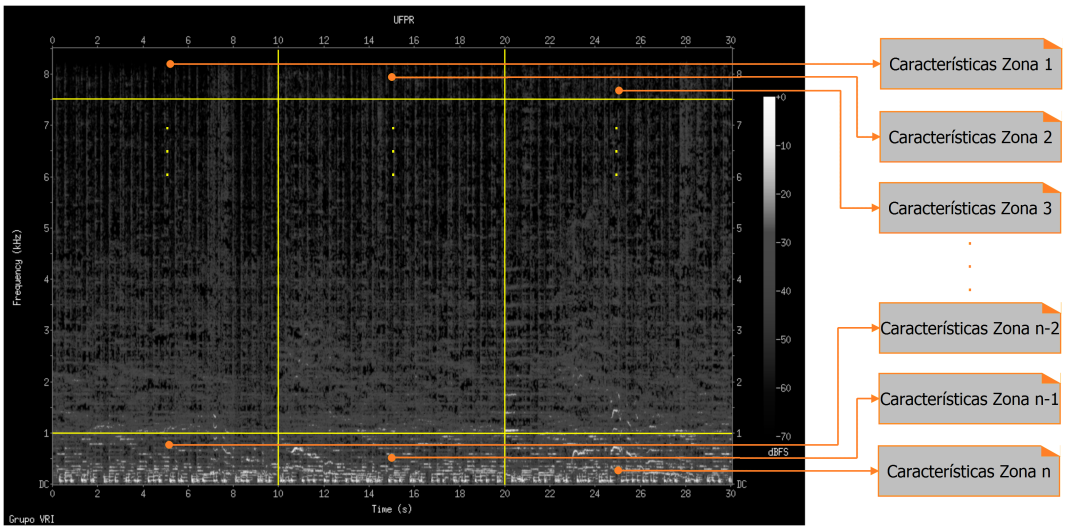



Figura 4.3: Extração de características preservando informações locais.

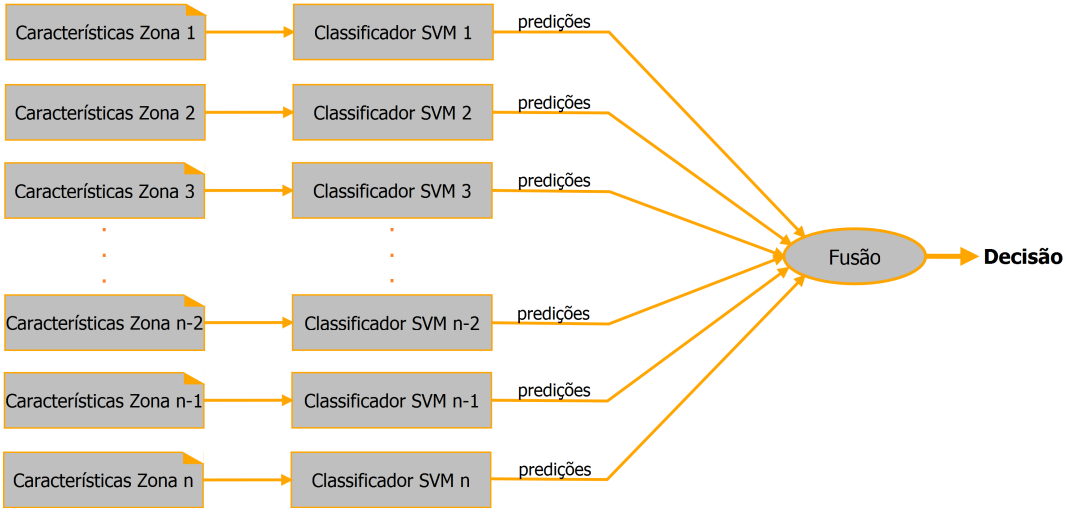



Figura 4.4: Criação de classificadores para as características extraídas de cada zona e fusão das saídas.

medida em que evita o risco de que seja considerado apenas um trecho da música que acidentalmente seja mais parecido com um gênero diferente daquele no qual a música está efetivamente classificada. Adicionalmente, o uso de mais de um segmento favorece a construção de um *pool* de classificadores, o que pode favorecer a obtenção de melhores resultados.

Para a extração dos três segmentos, foram tomadas porções bem distribuídas ao longo do sinal. 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 da 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 4.5, inspirada em [80], ilustra esta estratégia.

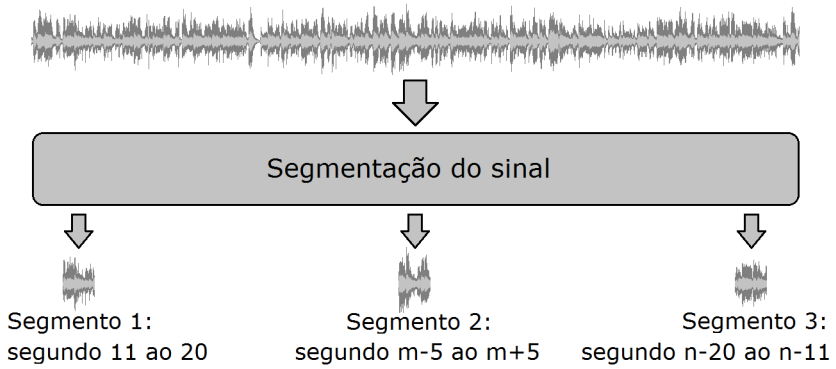



Figura 4.5: Extração de segmentos do sinal.

Na seção 5.4.6 o leitor pode encontrar experimentos específicos em que foram utilizadas apenas informações extraídas do segmento central da música.

## 4.3 Geração do espectrograma

Para a geração dos espectrogramas foi utilizado o software SoX 14.3.0 (*Sound eXchange*), um utilitário disponível em http://sox.sourceforge.net que permite a realização de conversões entre vários diferentes formatos de representação de áudio.

A imagem gerada representa o tempo no eixo horizontal, a frequência no eixo vertical e a intensidade de cor do pixel representa a amplitude do sinal. Através de parâmetros oferecidos pela ferramenta, as resoluções destas três dimensões foram empiricamente ajustadas. No protocolo experimentado, a Transformada Discreta de Fourier foi computada utilizando a janela Hanning de tamanho 1024, que preserva uma boa relação entre as resoluções das duas dimensões da imagem. A figura 4.6 mostra uma imagem típica de um espectrograma, gerado a partir de 30 segundos de música, utilizado nos experimentos realizados.

Depois de extraídas as imagens dos espectrogramas das músicas, elas foram convertidas para níveis de cinza para melhor se adequarem aos processos subseqüentes, em que serão extraídas características de textura das imagens. A maioria das técnicas de processamento de imagens empregadas com este propósito operam sobre imagens em níveis de cinza. Adicionalmente, é importante ressaltar que a principal informação presente nos espectrogramas de interesse para o propósito deste trabalho diz respeito à intensidade de energia do sinal. Esta informação é integralmente preservada com a simples conversão da representação da imagem de uma escala de cores para uma escala de cinza. A figura 4.7 mostra a mesma imagem apresentada na figura 4.6 depois da conversão para a escala

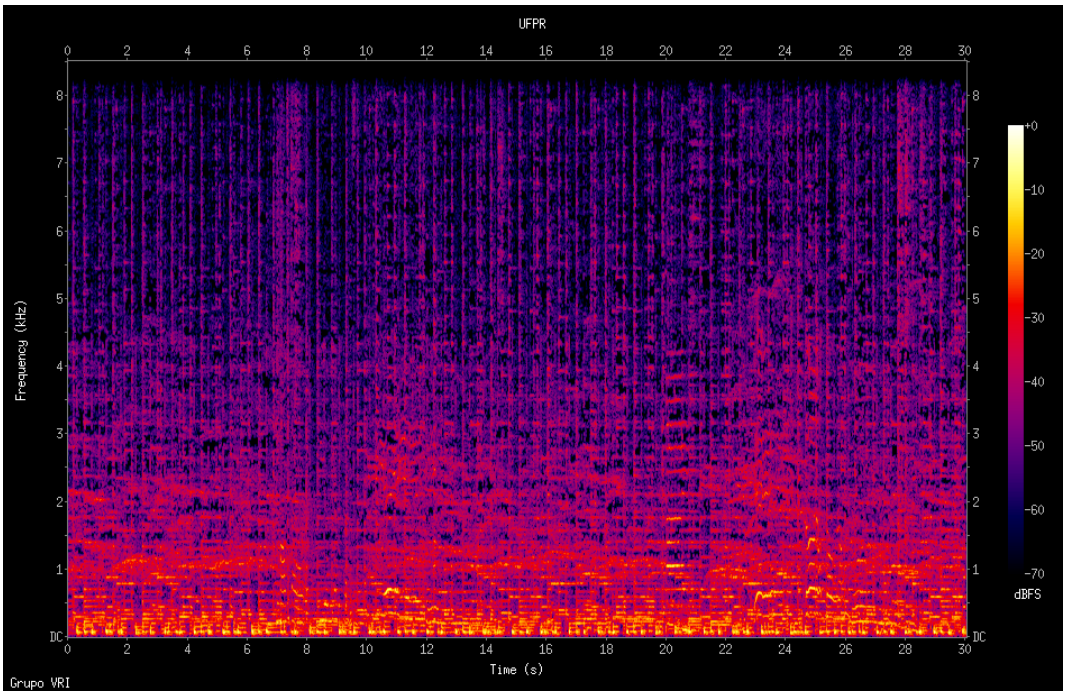



Figura 4.6: Espectrograma colorido gerado a partir do sinal de 30 segundos de música.

de cinza e com a devida segmentação que isola apenas a região de interesse na imagem, excluindo legendas e rótulos dos eixos entre outros.

A conversão da imagem colorida, originalmente representada no espaço $RGB$, para a escala de cinza foi feita conforme descrito na equação 4.1 [28], que leva em consideração o fato de que humanos não percebem as cores igualmente e equipara a luminância da imagem cinza à da imagem colorida original.

$$L = 0,2989 * R + 0,5870 * G + 0,1140 * B \tag{4.1}$$

na qual $L$ corresponde à luminância e será o tom de cinza resultante, $R$ é a intensidade do canal vermelho original, $G$ é a intensidade do canal verde original e $B$ é a intensidade do canal azul original.

## 4.4 Divisão das imagens em zonas

Ao longo do desenvolvimento dos experimentos inerentes a este trabalho, observou-se que a textura presente nas imagens de espectrograma extraídas das músicas não apresentam conteúdo uniforme ao longo dos eixos vertical e horizontal. Com isto, foi proposta uma estratégia que consiste em dividir a imagem em zonas, de forma que seja possível preservar informações locais presentes em regiões específicas da imagem.

Além da preservação de informações locais, a estratégia de divisão em zonas foi bastante oportuna por permitir naturalmente a criação de um *pool* de classificadores, já que

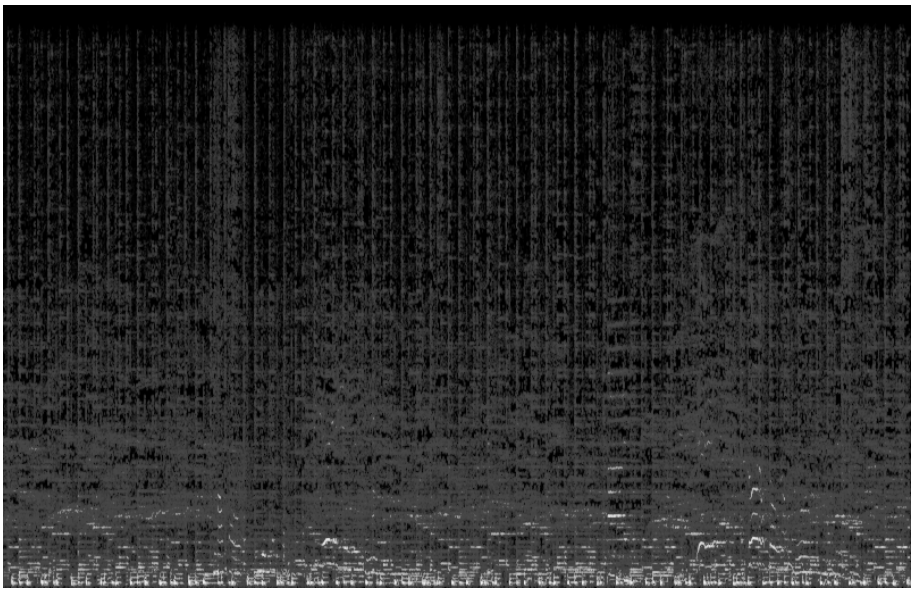


Figura 4.7: Espectrograma em escala de cinza gerado a partir do sinal de 30 segundos de música.

para cada zona criada pode-se estebelecer um classificador específico. O uso do *pool* de classificadores com regras de fusão descritas na literatura (apresentadas na subseção 3.2.1) tem bom potencial para proporcionar soluções eficientes para o problema aqui estudado.

Para implementar esta estratégia, alguns diferentes esquemas de zoneamento foram experimentados. Ao longo do eixo horizontal, o número de zonas criadas foi igual ao número de segmentos extraídos da música. Considerando que na grande maioria dos experimentos realizados foram extraídos três segmentos, foram criadas duas linhas de fronteira entre zonas perpendiculares ao eixo horizontal. Assim, foram caracterizadas três zonas com porções do espectrograma correspondentes a diferentes momentos da música. Além disso, foram testados diferentes padrões de divisão, alguns lineares, outros não, que caracterizam zonas na imagem correspondentes à diferentes bandas de frequência. Para isto, as linhas de fronteira criadas entre estas zonas são perpendiculares ao eixo vertical. O número total de zonas criadas em cada padrão de divisão é igual a $s \times f$, em que $s$ é o número de segmentos extraídos da música e $f$ o número de bandas de frequência (lineares ou não) criadas na estratégia de zoneamento. Também foi experimentada a extração de características sem a criação de zonas correspondentes a zonas de frequência, esta foi chamada extração global de características e com ela foram criados apenas três classificadores, um para cada segmento.

As próximas subseções descrevem em detalhes de todas as alternativas de zoneamento que fazem parte do método aqui proposto.

### 4.4.1 Divisão em zonas lineares

Com a divisão linear, são estabelecidas na imagem do espectrograma zonas de igual tamanho que correspondem a bandas de frequência. Os limites de cada banda criada dependem da quantidade de zonas definidas e do limite de frequência até o qual o sinal das músicas utilizadas apresenta informação relevante.

A figura 4.8 mostra o espectrograma gerado a partir de três segmentos de dez segundos de música, portanto 30 segundos, com divisão das bandas de frequência em dez zonas lineares, criando assim 30 zonas para o espectrograma gerado a partir de uma música.

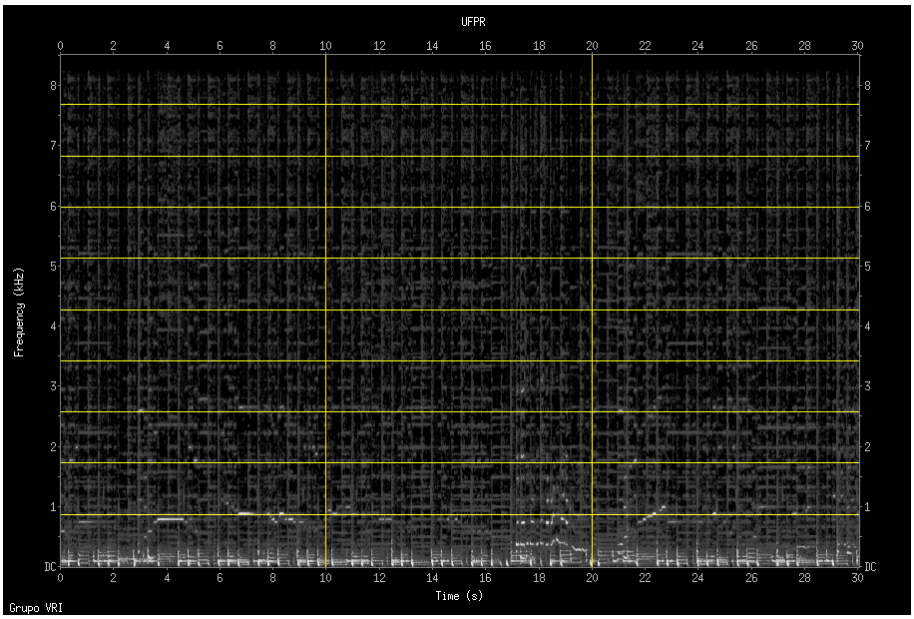



Figura 4.8: Espectrograma dividido em dez zonas lineares por segmento.

Nos resultados apresentados no capítulo 5 foram realizados experimentos com divisão linear das imagens em cinco e dez zonas.

### 4.4.2 Divisão pela escala de Bark

A escala de Bark é uma escala psicoacústica e sua criação se deu em uma tentativa de representar os limites das bandas críticas de audição, segundo as quais a audição humana é capaz de discernir sons e ruídos [94]. Os limites das bandas de frequência em Hz nesta escala são: 0, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500. O número de zonas a serem criadas, e consequentemente o número de classificadores, depende do limite de frequência até o qual a imagem do espectrograma apresenta informações relevantes.

Considerando os limites descritos, pode-se ter a criação de no máximo 24 zonas para a imagem de cada segmento, produzindo um total de 72 classificadores se for considerada a criação de um para a porção da imagem gerada a partir de cada segmento extraído

da música. Para que isto aconteça, é necessário que haja informação relevante acima dos 12000 Hz, o que nem sempre ocorre. A figura 4.9 ilustra a sobreposição das bandas criadas em uma imagem de espectrograma extraída de uma amostra de música tirada de uma base em que o limite de frequência com informação relevante era 8500 Hz. Assim, o número de bandas criadas neste espectrograma foi igual a 22 ao invés de 24.

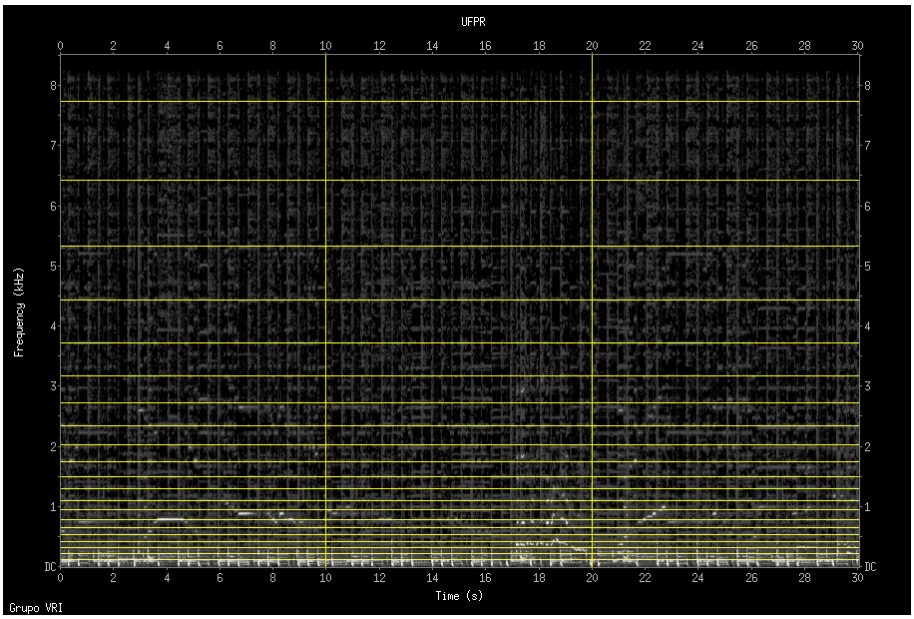



Figura 4.9: Bandas criadas com a divisão da imagem segundo a escala de Bark

### 4.4.3 Divisão pela escala Mel

De acordo com Umesh *et al.* [88], a escala Mel resulta fundamentalmente da psicoacústica, relacionando as frequências reais com as frequências percebidas pelos humanos, semelhantemente a escala de Bark. Nesta escala são estabelecidas 15 bandas de frequência, cujos limites em Hz são: 0, 40, 161, 200, 404, 693, 867, 1000, 2022, 3000, 3393, 4109, 5526, 6500, 7743 e 14000. A figura 4.10 mostra a divisão segundo a escala Mel sobreposta a um espectrograma. Assim como no caso da escala de Bark, o número de zonas criadas deve estar sujeito ao limite até o qual a imagem do espectrograma apresenta conteúdo relevante. Se houver informação relevante acima de 7743 Hz, o que ocorre na maioria dos casos, 15 zonas devem ser criadas para cada segmento e, consequentemente, o número de classificadores criados é igual a 45 considerando a criação de zonas diferentes para a porção da imagem correspondente a cada diferente segmento extraído da música. No caso do exemplo ilustrado, são criadas 15 zonas para cada segmento, já que existe informação relevante até 8500 Hz no sinal.

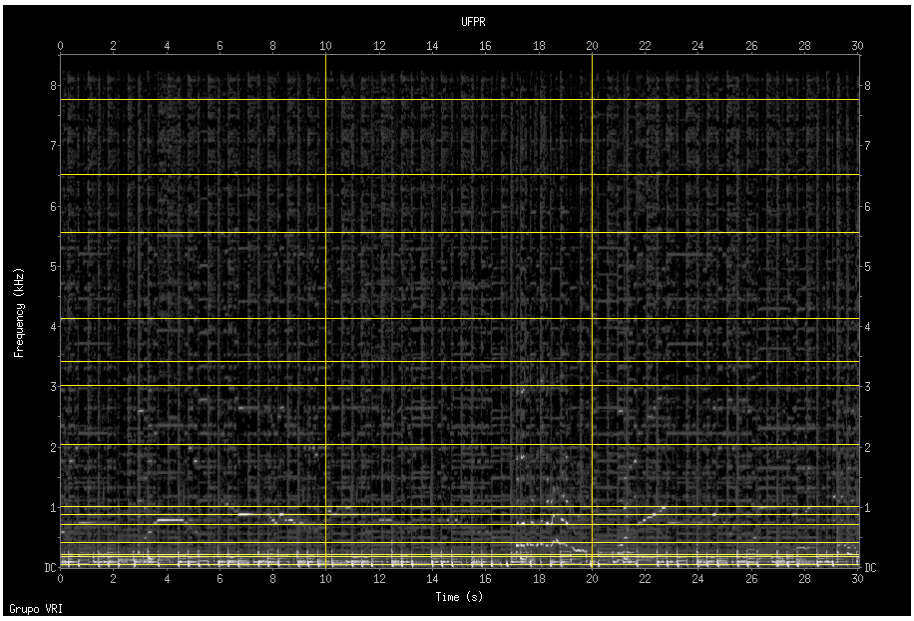



Figura 4.10: Bandas criadas com a divisão da imagem segundo a escala Mel

## 4.5 Extração de características

A textura é destacadamente o principal atributo visual percebido ao se observar uma imagem de espectrograma. Considerando isto, o trabalho aqui proposto utiliza operadores de textura apresentados na literatura para descrever objetivamente o conteúdo destas imagens e aplicá-lo na etapa subsequente, de classificação. A tabela 4.1 mostra dados acerca da abordagem e do número de características (tamanho do vetor) extraídas com cada tipo de operador utilizado nos experimentos.

Tabela 4.1: Dados sobre os descritores utilizados

| **Abordagem** | **Descritor** | **Tamanho do vetor de características** |
|---|---|---|
| Estatística | GLCM | 28 |
| Espectral | Filtros de Gabor | 120 |
| Estrutural | LBP | 59 |
| Estrutural | LPQ | 256 |

Nos experimentos aqui descritos, foi utilizado um mecanismo de normalização segundo o qual os dados são mapeados para o intervalo $[-1,\ 1]$. O valor normalizado para uma característica $x$ é encontrado conforme descrito na equação 4.2.

$$x_{norm} = \frac{2(x - m_i)}{(M_i - m_i) - 1} \tag{4.2}$$

na qual $x$ é o valor da característica antes da normalização, $M_i$ é o valor máximo encontrado para a característica no conjnto de dados e $m_i$ é o valor mínimo encontrado para a característica no conjunto de dados. As próximas subseções descrevem as características exploradas em cada das técnicas utilizadas.

### 4.5.1 GLCM

As características de GLCM utilizadas nos experimentos realizados foram: contraste, energia, entropia, homogeneidade, momento de terceira ordem, probabilidade máxima e correlação (descritas na subseção 3.1.2.1). Estas sete características foram extraídas com distância $d$=1 com quatro diferentes orientações de $\theta$ : $0^o$, $45^o$, $90^o$ *e* $135^o$. Assim, obteve-se um total de 28 características extraídas por zona criada na imagem do espectrograma.

### 4.5.2 Filtros de Gabor

Dentre as técnicas que se enquadram na abordagem espectral, optou-se por utilizar filtros de Gabor, já que os mesmos têm sido empregados com sucesso em diferentes aplicações que envolvem a classificação de textura. Nos experimentos realizados com este descritor, os parâmetros de fator de orientação ($\theta$) e fator de escala ($\lambda$) foram ajustados conforme aplicado em [93]. Com isto, foi utilizada uma máscara de tamanho $64 \times 64$ com oito variações do fator de orientação e cinco do fator de escala, totalizando 40 subimagens. Destas, foram extraídas média, variância e obliquidade, o que proporcionou um vetor com 120 características.

### 4.5.3 LBP

LBP é um poderoso descritor de textura que opera sobre a vizinhança local de cada pixel presente na imagem procurando identificar o padrão binário local presente nesta região. Um histograma que contabiliza as ocorrências de todos os padrões binários previstos é formado, e o vetor final de características corresponde à este histograma normalizado. A variação de LBP originalmente utilizada nos experimentos é $LBP_{8,2}$, a mais difundida e que na maioria dos trabalhos apresentados produz os melhores resultados.

Em $LBP_{8,2}$, os padrões são identificados considerando-se oito vizinhos a uma distância de dois pixel a partir de cada pixel da imagem. Ao final, considerando-se apenas os padrões ditos uniformes (conforme descrito na subseção 3.1.2.3), o histograma final, e consequentemente o vetor de atributos descritores, conta com 59 valores.

Algumas outras variações de LBP foram experimentadas. Conforme ilustra a figura 4.11, o formato da vizinhança utilizada em LBP para capturar os padrões locais podem variar em função de $R$, que corresponde a distância entre o pixel central e os vizinhos a serem tomados, e $P$, que é a quantidade de vizinhos a serem considerados.

Os padrões experimentados foram $LBP_{8,1}$, $LBP_{8,2}$ e $LBP_{16,2}$, sendo o primeiro elemento da dupla o valor de $P$ e o segundo elemento o valor de $R$. As diferentes variações de LBP produzem vetores de atributos com tamanhos diferentes.

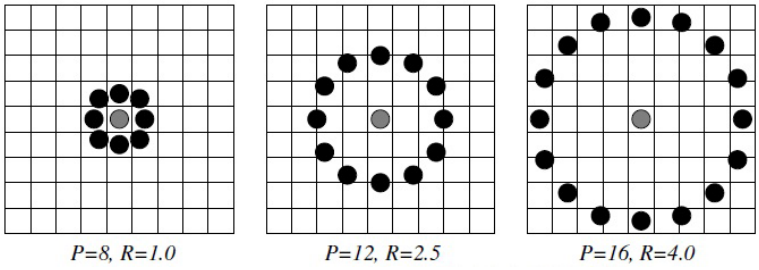



Figura 4.11: Exemplos de possiveis vizinanças utilizadas em LBP [61]

### 4.5.4 LPQ

LPQ é um descritor des textura proposto por Ojansivu e Heikkilä em [65]. Os autores afirmam que, embora o método tenha sido proposto para lidar bem com imagens afetadas por borramento, ele é capaz de produzir bons resultados também em situações em que não há problema com este tipo de ruído. Considerando este fato, optou-se por experimentar este descritor.

Na maioria dos experimentos com LPQ o descritor foi extraído utilizando-se a janela $m$ de tamanho $3 \times 3$. Também foram experimentadas variações do descritor com valores para $m$ igual a cinco, sete, nove e onze. O vetor final de características LPQ corresponde ao histograma construído pelo método, e posui em todas as variações um total de 256 valores.

## 4.6 Classificação

O classificador utilizado nos experimentos descritos neste trabalho é o *Support Vector Machine (SVM)*. Este classificador, apresentado por Vapnik em [89], tem sido utilizado com sucesso em vários trabalhos de classificação nos mais diversos domínios de aplicação. Para empregar este classificador, utilizou-se a biblioteca LIBSVM, desenvolvida por Chang e Lin [5] e disponível em *http://www.csie.ntu.edu.tw/˜cjlin/libsvm/*.

SVM é originalmente um classificador binário. Entretanto, existem algumas diferentes estratégias apresentadas na literatura para utilizá-lo em problemas multiclasse [37], como o problema abordado neste trabalho. Dentre as estratégias apresentadas, a chamda *"one-against-one"*, também conhecida como *pairwise*, é uma das mais utilizadas, e é a empregada para problemas multiclasse em LIBSVM porque permite realizar a classificação de forma mais rápida.

Utilizando *pairwise*, são criados $k(k-1)/2$ classificadores binários, em que $k$ é o número de classes envolvidas no problema. LIBSVM pode, também, fornecer predições para as classes envolvidas no problema. O esquema de votações que produzem as predições em LIBSVM é feito tal como em [37]. As predições são fundamentais no método aqui proposto, pois com o zoneamento das imagens dos espectrogramas, vários classificadores são criados

e, ao final, o que se propõe é que suas saídas sejam fundidas para que se obtenha o resultado final da classificação.

É importante ainda ressaltar que, antes da realizar as tarefas de classificação, os dados foram devidamente normalizados, conforme descrito na seção 4.5. O *kernel* utilizado foi *Radial Basis Function (RBF)* e os parâmetros $C$ (custo) e $\gamma$ foram otimizados utilizando um procedimento *grid-search*.

## 4.7 Avaliação de resultados

Considerando o fato de que o problema abordado neste trabalho é um problema multi-classe, a matriz de confusão é adotada para avaliar os resultados obtidos nos experimentos. A matriz de confusão é uma tabela a partir da qual se pode observar com que intensidade uma classe é confundida com cada uma das outras classes envolvidas no problema.

A taxa de reconhecimento percentual de uma classe $i$ (neste caso um gênero), dada pela equação 4.3, será a medida objetiva utilizada para aferir o desempenho do classificador.

$$taxa\ de\ reconhecimento_i = \frac{100c_i}{t_i} \tag{4.3}$$

na qual $c_i$ é o número de instâncias corretamente classificadas pertencentes à classe $i$, e $t_i$ corresponde ao número total de instâncias pertencentes à classe $i$.

A taxa de reconhecimento percentual geral é dada pela equação 4.4:

$$taxa\ de\ reconhecimento\ geral = 100\frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} t_i} \tag{4.4}$$

na qual $n$ corresponde ao número de classes envolvidas na classificação, $c_i$ é o número de instâncias corretamente classificadas pertencentes à classe $i$, e $t_i$ corresponde ao número total de instâncias pertencentes à classe $i$.

As boas práticas de experimentação em reconhecimento de padrões sugerem que as amostras sejam divididas em conjuntos para treinamento e teste. Usualmente, se utiliza um conjunto ora como teste, ora como treinamento (ou parte do conjunto de treinamento). Assim, é comum que muitas vezes a taxa final de reconhecimento de um experimento seja dada pela taxa média obtida a partir de experimentos realizados com diferentes conjuntos de teste. Por isso, o desvio padrão entre os valores utilizados para o cálculo desta média será utilizado em alguns casos para avaliar a intensidade da dispersão entre eles.

## 4.8 Conclusão

Este capítulo descreveu detalhes do método aqui proposto para a construção de um classificador automático de gêneros musicais baseado em características extraídas de imagens de espectrograma. Inicialmente, uma visão geral do método foi apresentada, na seção 4.1.

Na seção 4.2 foi descrita a estratégia utilizada para a segmentação do sinal. Na seção 4.3 foram descritos alguns detalhes técnicos e parâmetros utilizados para a geração das imagens de espectrograma. Na seção 4.4 são apresentados os detalhes sobre as estratégias de zoneamento utilizadas. Na seção 4.5 foram descritos os parâmetros utilizados para a extração de características de textura com cada uma das diferentes abordagens experimentadas. Na seção 4.6 foram apresentados detalhes sobre a classificação e, em seguida, estratégias para a avaliação dos resultados, na seção 4.7.

O próximo capítulo descreve os resultados obtidos com o esquema de classificação aqui proposto.

# CAPÍTULO 5

# RESULTADOS EXPERIMENTAIS

Este capítulo descreve resultados obtidos utilizando diferentes abordagens para extrair descritores de textura das imagens de espectrograma obtidas a partir do sinal das músicas. Além disso, diferentes padrões de zoneamento foram experimentados a fim de se verificar em que medida a preservação de alguma informação acerca da localização espacial destas informações pode influenciar nos resultados obtidos. A fim de permitir alguma comparação entre os resultados obtidos com o uso de espectrogramas e com o uso de descritores tradicionais, também constam neste capítulo alguns resultados obtidos com descritores extraídos diretamente do sinal. A seção 5.1 descreve as bases de música utilizadas nos experimentos realizados neste trabalho. As seções subsequentes descrevem os resultados obtidos nos experimentos realizados.

## 5.1 Bases de Músicas

Para o desenvolvimento dos experimentos descritos neste trabalho, foram utilizadas duas das principais bases de músicas disponibilizadas à comunidade acadêmica de pesquisa em recuperação de informações musicais, as bases *Latin Music Database (LMD)* e *ISMIR 2004*. Estas bases foram escolhidas pelo fato de terem sido significativamente exploradas em outros trabalhos já apresentados na literatura, o que permite uma melhor comparação dos resultados obtidos. Além disso, é válido ressaltar que estas bases são complementares na medida em que representam dois conjuntos de gêneros disjuntos. Assim, pode-se verificar o desempenho da solução proposta em um universo abrangente de gêneros musicais.

Cabe ainda destacar que outras bases de músicas bastante difundidas na comunidade de pesqusia não puderam ser utilizadas aplicando o protocolo de segmentação do sinal utilizado nos experimentos aqui descritos. Isso ocorreu porque estas bases, em geral, não disponibilizam o conteúdo completo das músicas presentes nelas. As subseções 5.1.1 e 5.1.2 descrevem, respectivamente, alguns detalhes acerca das bases *LMD* e *ISMIR 2004*.

### 5.1.1 *Latin Music Database*

A LMD, apresentada por Silla Jr. *et al.* [82], é uma base de músicas latino-americanas composta por 3227 títulos musicais de 501 artistas diferentes. Os títulos estão disponíveis em formato MP3 e são classificados em dez diferentes gêneros musicais: Axé, Bachata, Bolero, Forró, Gaúcha, Merengue, Pagode, Salsa, Sertaneja e Tango. A tabela 5.1 mostra alguns detalhes acerca do número de títulos musicais disponibilizados e artistas por gênero.

Tabela 5.1: Número de artistas e títulos por gênero na LMD

| Gênero | Número de artistas | Número de títulos |
|---|---|---|
| Axé | 37 | 313 |
| Bachata | 64 | 313 |
| Bolero | 99 | 315 |
| Forró | 27 | 313 |
| Gaúcha | 92 | 311 |
| Merengue | 96 | 315 |
| Pagode | 16 | 307 |
| Salsa | 54 | 311 |
| Sertaneja | 9 | 321 |
| Tango | 7 | 408 |
| **Total** | **501** | **3227** |

A LMD foi construída com base na percepção de especialistas humanos de como as músicas são dançadas. A classificação dos títulos foi feita por dois professores de dança profissionais com mais de dez anos de experiência no ensino de danças de salão brasileiras e latino-americanas. A equipe envolvida no projeto ainda realizou uma segunda rodada de verificação a fim de evitar a ocorrência de eventuais erros de classificação.

Uma importante particularidade desta base e que a torna bastante desafiadora para o desenvolvimento de trabalhos de recuperação de informações musicais, é o fato de que os gêneros que ela apresenta possuem uma significativa similaridade no que diz respeito a instrumentalização, estrutura rítmica e conteúdo harmônico. Isto se deve ao fato de que os gêneros que a compõem são oriundos de um mesmo país ou de países com fortes semelhanças no que diz respeito a aspectos culturais.

Outro fato importante relativo aos experimentos realizados com a base LMD como parte deste trabalho é o uso da restrição conhecida como *"artist filter"* [21]. O *"artist filter"* determina que a divisão dos folds para a ralização das tarefas de treinamento seja feita de forma que não hajam músicas interpretadas por um mesmo artista em folds diferentes. Esta restrição torna o trabalho de classificação mais difícil, e tende a diminuir as taxas de reconhecimento. Por outro lado, ela favorece a construção de classificadores mais robustos, uma vez que evitam que os classificadores construídos aprendem a classificar artistas ao invés de gêneros.

As músicas da base LMD estão originalmente disponíveis em formato MP3, com uma taxa de bits de 352 kbps, amostra de áudio de 16 bits e com taxa de amostragem de áudio de 22,05 kHz. Para a geração dos espectrogramas, apenas um canal do sinal foi utilizado, já que os conteúdos dos dois canais originais são bastante parecidos.

Em função do uso do *"artist filter"*, apenas 900 títulos musicais da LMD puderam ser utilizados nos experimentos. Estes 900 títulos foram divididos em três folds com 300 títulos cada, sendo que em cada fold foram colocados 30 títulos de cada gênero musical presente na base. A divisão em apenas 3 folds também foi imposta em função da opção pelo uso do *"artist filter"*.

A menos quando mencionado algo diferente, a base LMD foi a base utilizada nos experimentos descritos neste trabalho. As músicas da base LMD utilizadas nos experimentos

haviam passado por uma filtragem previamente de forma que o conteúdo do sinal acima dos 8500 Hz foi eliminado, por isso este é o limite empregado nos experimentos feitos com esta base. Esta filtragem foi aplicada para reduzir diferenças entre o gênero Tango e os demais no que diz respeito aos limites de frequência até onde se encontra informação relevante no sinal, já que muitas gravações deste gênero eram muito antigas e este limite era bastante baixo.

Em parte dos experimentos, o padrão de zoneamento adotado é chamado de "extração global". Global no sentido de que nestes experimentos não são criadas zonas correspondentes a bandas de frequência nos espectrogramas extraídos de cada segmento da música. Entretanto, é válido observar que, mesmo neste caso, mais de um classificador é criado, já que para cada segmento extraído da música um classificador é criado. Além dos experimentos com extração global, também são mostrados os resultados obtidos com a divisão linear das imagens de espectrograma em zonas. As imagens foram divididas em cinco zonas e em dez zonas lineares. Com a divisão em cinco zonas, obtem-se um total de 15 classificadores, já que o número de classificadores é igual ao número de segmentos (3 neste caso) multiplicado pelo número de zonas criadas em cada segmento. Para este esquema de divisão sobre as músicas da base LMD, as bandas criadas foram as seguintes:

- Banda 1: de 0 até 1700 Hz;
- Banda 2: de 1700 até 3400 Hz;
- Banda 3: de 3400 até 5100 Hz;
- Banda 4: de 5100 até 6800 Hz;
- Banda 5: de 6800 até 8500 Hz;

Quando são criadas dez zonas lineares nos espectrogramas, produz-se um total de 30 classificadores. Neste caso, as dez bandas de frequência criadas são as seguintes:

- Banda 1: de 0 até 850 Hz;
- Banda 2: de 850 até 1700 Hz;
- Banda 3: de 1700 até 2550 Hz;
- Banda 4: de 2550 até 3400 Hz;
- Banda 5: de 3400 até 4250 Hz;
- Banda 6: de 4250 até 5100 Hz;
- Banda 7: de 5100 até 5950 Hz;

- Banda 8: de 5950 até 6800 Hz;
- Banda 9: de 6800 até 7650 Hz;
- Banda 10: de 7650 até 8500 Hz;

Em alguas situações foram experimentadas divisões não lineares das imagens de espectrograma, utilizando as escalas de Bark e Mel, descritas respectivamente nas subseções 5.4.3 e 5.4.4. Nestes casos, as divisões se deram conforme descrito nestas subseções considerando, entretanto, o limite 8500 Hz, a partir do qual não há informação relevante nas músicas da base LMD utilizadas nos experimentos. Assim, foram criadas 22 zonas quando utilizada a divisão pela escala de Bark e 15 zonas quando utilizada a divisão pela escala Mel.

### 5.1.2 *ISMIR 2004*

A fim de verificar a versatilidade da metodologia proposta, foram realizados alguns experimentos sobre a base *ISMIR 2004* [4]. Esta base foi criada para a realização do concurso *ISMIR 2004* para o desenvolvimento de uma série de tarefas de recuperação de informações musicais e, na ausência de uma variedade de bases à época, acabou tornando-se uma alternativa amplamente utilizada em várias trabalhos apresentados na literatura desde então, conforme descrito no capítulo 2.

A base é composta por um total de 1458 músicas, sendo 729 previamente rotuladas para a formação do conjunto de treinamento e outras 729 destinadas à formação do conjunto de teste. As músicas são classificadas em seis diferentes gêneros musicais, quais sejam: *classical*, *electronic*, *jazz/blues*, *metal/punk*, *rock/pop* e *world*. A tabela 5.2 mostra as quantidades de títulos musicais por gênero presentes nos conjuntos de treino e teste.

Tabela 5.2: Número de títulos por gênero nos conjuntos de treino e teste da base *ISMIR 2004*

| **Gênero** | **Número de títulos no conjunto de treino** | **Número de títulos no conjunto de teste** |
|---|---|---|
| *Classical* | 320 | 320 |
| *Electronic* | 115 | 114 |
| *Jazz/blues* | 26 | 26 |
| *Metal/punk* | 45 | 45 |
| *Rock/pop* | 101 | 102 |
| *World* | 122 | 122 |
| **Total** | **729** | **729** |

As músicas da base *ISMIR 2004* também estão originalmente disponíveis em formato MP3. Entretanto, nem todas as características técnicas são iguais as da base LMD. A taxa de bits é de 706 kbps, a amostra de áudio de 16 bits e com taxa de amostragem de áudio de 44,1 kHz. Para a geração dos espectrogramas, também foi utilizado apenas um canal do sinal.

O limite até o qual as músicas da base *ISMIR 2004* apresenta informações relevantes é 14 kHz. Com isto, as bandas de frequência criadas quando se utilizou o zoneamento linear com cinco zonas para esta base foram as seguintes:

- Banda 1: de 0 até 2800 Hz;
- Banda 2: de 2800 até 5600 Hz;
- Banda 3: de 5600 até 8400 Hz;
- Banda 4: de 8400 até 11200 Hz;
- Banda 5: de 11200 até 14000 Hz;

Quando são criadas dez zonas lineares nos espectrogramas, as bandas de frequência criadas com a base *ISMIR 2004* são as seguintes:

- Banda 1: de 0 até 1400 Hz;
- Banda 2: de 1400 até 2800 Hz;
- Banda 3: de 2800 até 4200 Hz;
- Banda 4: de 4200 até 5600 Hz;
- Banda 5: de 5600 até 7000 Hz;
- Banda 6: de 7000 até 8400 Hz;
- Banda 7: de 8400 até 9800 Hz;
- Banda 8: de 9800 até 11200 Hz;
- Banda 9: de 11200 até 12600 Hz;
- Banda 10: de 12600 até 14000 Hz;

Quando utilizadas divisões não lineares das imagens de espectrograma com a base *ISMIR 2004*, foram criadas 24 zonas com a escala de Bark e 15 com a escala Mel.

A base *ISMIR 2004* foi utilizada apenas com o propósito de verificar a viabilidade de aplicação do método aqui proposto sobre outras bases, além da LMD. Desta forma, optou-se por utilizá-la somente em experimentos realizados com o operador de textura LBP, já que este foi o primeiro dos descritores estudados a apresentar destacado potencial para a obtenção de bons resultados sobre a base LMD.

Os resultados experimentais são apresentados a partir da próxima seção, a iniciar pelos obtidos com o uso de GLCM.

## 5.2 Gray Level Co-occurrence Matrix (GLCM)

Esta seção descreve os resultados obtidos utilizando GLCM como descritor de textura. Inicialmente são apresentados os resultados obtidos de acordo com o protocolo geral descrito no capítulo 4 e com os parâmetros específicos descritos na subseção 4.5.1.

A tabela 5.3 mostra os resultados obtidos com a fusão das saídas dos três classificadores criados utilizando-se as regras de fusão do máximo, do mínimo, do produto e da soma utilizando extração global de características, zoneamento linear com cinco e zoneamento linear com dez zonas. Os resultados referem-se às médias obtidas quando cada um dos três folds figura como conjunto de teste. O desvio padrão entre os três resultados também é apresentado.

Tabela 5.3: Taxas de reconhecimento (%) com GLCM

| **Padrão de zoneamento** | **Regra do máximo** | **Regra do mínimo** | **Regra do produto** | **Regra da soma** |
|---|---|---|---|---|
| Extração global de características | 50,11±6,87 | 54,44±7,75 | 57,44±4,67 | 56,56±4,35 |
| Divisão linear com cinco zonas | 56,33±3,21 | 62,56±2,55 | **70,78±2,22** | 70,00±3,06 |
| Divisão linear com dez zonas | 42,33±7,86 | 50,22±6,00 | 68,11±3,42 | 66,33±4,36 |

Os resultados obtidos com a fusão das saídas dos classificadores mostram a força do *pool* de classificadores. Considerando os resultados produzidos isoladamente por todos os classificadores gerados nos três casos experimentados, o melhor desempenho individual foi de 51,67% de taxa de reconhecimento. Com a fusão das saídas, pode-se observar, no melhor caso, um ganho de quase vinte pontos percentuais na taxa de reconhecimento.

A regra do produto proporcionou os melhores resultados em todos os casos, embora bastante próximos aos produzidos pela regra da soma. Outro aspecto interessante a ser observado é o fato de que com a criação de dez zonas, as taxas de reconhecimento começam a sofrer decréscimo. A matriz de confusão encontrada para o melhor caso, com a fusão pela regra do produto e divisão linear em cinco zonas, é mostrada na tabela 5.4.

Tabela 5.4: Matriz de confusão (%) obtida no melhor caso (regra do produto) com GLCM e divisão linear em cinco zonas

| | (0) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|---|
| (0) Axé | **66,67** | 1,11 | 2,22 | 0,00 | 7,78 | 1,11 | 10,00 | 0,00 | 11,11 | 0,00 |
| (1) Bachata | 1,11 | **86,67** | 2,22 | 0,00 | 3,33 | 2,22 | 0,00 | 2,22 | 2,22 | 0,00 |
| (2) Bolero | 0,00 | 2,22 | **80,00** | 4,44 | 3,33 | 0,00 | 3,33 | 0,00 | 4,44 | 2,22 |
| (3) Forró | 1,11 | 0,00 | 8,89 | **64,44** | 10,00 | 0,00 | 4,44 | 3,33 | 7,78 | 0,00 |
| (4) Gaúcha | 17,78 | 0,00 | 13,33 | 8,89 | **46,67** | 0,00 | 2,22 | 2,22 | 8,89 | 0,00 |
| (5) Merengue | 0,00 | 2,22 | 0,00 | 0,00 | 1,11 | **87,78** | 2,22 | 6,67 | 0,00 | 0,00 |
| (6) Pagode | 10,00 | 0,00 | 13,33 | 4,44 | 5,56 | 0,00 | **53,33** | 4,44 | 8,89 | 0,00 |
| (7) Salsa | 0,00 | 1,11 | 4,44 | 3,33 | 4,44 | 5,56 | 5,56 | **68,89** | 6,67 | 0,00 |
| (8) Sertaneja | 15,56 | 1,11 | 10,00 | 5,56 | 2,22 | 0,00 | 2,22 | 1,11 | **62,22** | 0,00 |
| (9) Tango | 0,00 | 0,00 | 7,78 | 0,00 | 0,00 | 0,00 | 1,11 | 0,00 | 0,00 | **91,11** |

Pode-se constatar alguns importantes focos de confusão na matriz, notadamente dos gêneros gaúcha e sertaneja para o gênero axé. Em contrapartida, gêneros com uma estrutura harmônica melhor definida, como tango e bolero apresentaram taxas de reconhecimento muito boas. Isto sugere que as características utilizadas nestes experimentos

estejam capturando, em alguma medida, dimensões musicais inerentes ao conteúdo do sinal explorado. A subseção 5.2.1 descreve resultados obtidos variando-se o parâmetro $d$ durante a extração de características com GLCM.

### 5.2.1 Variando parâmetro de GLCM

Nesta sequência de experimentos, foi verificado o desempenho de GLCM com diferentes valores para o parâmetro $d$. Conforme descrito na subseção 3.1.2.1, o parâmetro $d$ estebelece a distância entre o pixel central e seus vizinhos que são considerados para a composição da matriz de co-ocorrência de níveis de cinza. A tabela 5.5 mostra as taxas de reconhecimento obtidas em percentual utilizando quatro diferentes regras de fusão e aplicando o zoneamento linear com cinco zonas para os espectrogramas extraídos de cada segmento. Este padrão de zoneamento foi escolhido por ter apresentado o melhor resultado nos experimentos descritos na seção 5.2.

Tabela 5.5: Taxas de reconhecimento (%) com GLCM utilizando cinco zonas lineares e diferentes valores para o parâmetro $d$

| **Valor de $d$** | **Regra do máximo** | **Regra do mínimo** | **Regra do produto** | **Regra da soma** |
|---|---|---|---|---|
| 1* | 56,33±3,21 | 62,56±2,55 | **70,78±2,22** | 70,00±3,06 |
| 2 | 57,11±1,39 | 59,44±0,69 | 70,22±1,35 | 68,78±2,52 |
| 3 | 58,44±0,38 | 60,00±3,48 | 69,67±1,86 | 68,00±2,40 |
| 4 | 55,00±0,33 | 58,44±2,04 | 68,67±2,03 | 67,56±0,84 |
| 5 | 52,44±1,90 | 57,11±1,17 | 66,33±0,67 | 65,67±1,20 |

* Resultados apresentados na tabela 5.3

Os resultados mostram que o uso de outros valores para o parâmetro $d$, diferentes do valor originalmente utilizado, não configuram uma boa solução. Além dos resultados serem próximos entre si, os obtidos com valor de $d$ igual a 1 são superiores aos obtidos com os outros valores experimentados para $d$.

## 5.3 Filtros de Gabor

Esta seção descreve os resultados obtidos utilizando descritores de textura extraídos com Filtros de Gabor. Os resultados descritos foram obtidos de acordo com o protocolo geral descrito no capítulo 4 e com os parâmetros específicos descritos na subseção 4.5.2.

Os resultados obtidos com a fusão das saídas dos classificadores criados com extração global de características e com os padrões de zoneamento linear com a criação de cinco e de dez zonas, utilizando diferentes regras de fusão, estão descritos na tabela 5.6. Novamente, as taxas de reconhecimento apresentadas referem-se a média entre os três diferentes folds criados quando utilizados como conjunto de teste. Os desvios padrão entre os três resultados também são apresntados.

Mais uma vez o conjunto de classificadores produzidos mostrou força para a produção de bons resultados. A melhor taxa de acerto obtida individualmente pelos classificadores

Tabela 5.6: Taxas de reconhecimento (%) com características extraídas com Filtros de Gabor

| **Padrão de zoneamento** | **Regra do máximo** | **Regra do mínimo** | **Regra do produto** | **Regra da soma** |
|---|---|---|---|---|
| Extração global de características | 55,89±9,94 | 56,67±11,60 | 59,78±9,91 | 58,78±9,08 |
| Divisão linear com cinco zonas | 66,22±2,22 | 69,67±2,33 | **74,67±3,79** | 74,11±2,69 |
| Divisão linear com dez zonas | 60,56±1,02 | 65,33±2,85 | 71,78±1,84 | 71,00±0,58 |

criados, considerando todos os padrões de zoneamento, foi igual a 53,78%. Neste caso, a melhoria na taxa de acerto obtida no melhor caso, com fusão pela regra do produto e divisão linear em cinco zonas, foi superior a vinte pontos percentuais.

Assim como nos resultados obtidos com GLCM, os melhores resultados aconteceram quando foi utilizada a regra do produto, acompanhados de perto pelos resultados obtidos com a fusão pela regra da soma. Novamente, a divisão linear com cinco zonas produziu os melhores resultados. A matriz de confusão encontrada para o melhor caso, com a fusão pela regra do produto e divisão linear em cinco zonas, é mostrada na tabela 5.7.

Tabela 5.7: Matriz de confusão (%) obtida no melhor caso (regra do produto) com filtros de Gabor e divisão linear em cinco zonas

| | (0) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|---|
| (0) Axé | **72,22** | 0,00 | 4,44 | 0,00 | 0,00 | 3,33 | 7,78 | 3,33 | 8,89 | 0,00 |
| (1) Bachata | 1,11 | **90,00** | 4,44 | 0,00 | 1,11 | 1,11 | 0,00 | 1,11 | 1,11 | 0,00 |
| (2) Bolero | 0,00 | 1,11 | **80,00** | 0,00 | 8,89 | 0,00 | 1,11 | 2,22 | 3,33 | 3,33 |
| (3) Forró | 0,00 | 0,00 | 3,33 | **65,56** | 12,22 | 1,11 | 0,00 | 10,00 | 7,78 | 0,00 |
| (4) Gaúcha | 17,78 | 1,11 | 12,22 | 6,67 | **48,89** | 1,11 | 2,22 | 6,67 | 3,33 | 0,00 |
| (5) Merengue | 0,00 | 4,44 | 0,00 | 0,00 | 1,11 | **88,89** | 0,00 | 5,56 | 0,00 | 0,00 |
| (6) Pagode | 7,78 | 0,00 | 11,11 | 1,11 | 4,44 | 1,11 | **65,56** | 4,44 | 4,44 | 0,00 |
| (7) Salsa | 0,00 | 0,00 | 4,44 | 1,11 | 4,44 | 2,22 | 1,11 | **84,44** | 2,22 | 0,00 |
| (8) Sertaneja | 20,00 | 1,11 | 4,44 | 0,00 | 3,33 | 0,00 | 3,33 | 5,56 | **62,22** | 0,00 |
| (9) Tango | 0,00 | 0,00 | 11,11 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | **88,89** |

Novamente, os principais focos de confusão se dão dos gêneros gaúcha e sertaneja para o gênero axé. Isto revela que este descritor captura características presentes no conteúdo da textura compatíveis com aquelas capturadas por GLCM. Adicionalmente, os gêneros com melhor desempenho foram bachata, merengue e tango.

## 5.4 Local Binary Pattern (LBP)

Nesta seção são mostrados resultados obtidos utilizando um descritor da abordagem estrutural para a obtenção de descritores de textura, o LBP. A menos quando mencionado o contrário, os descritores foram extraídos utilizando $LBP_{8,2}$, conforme descrito na subseção 4.5.3.

A tabela 5.8 mostra o desempenho do método proposto utilizando $LBP_{8,2}$ para a extração de características utilizando extração global de características, zoneamento linear com cinco zonas e zoneamento linear com dez zonas.

Os resultados obtidos são particularmente animadores. No melhor caso, o desempenho individual do classificador criado para uma zona chegou a 71,22%. Este desempenho indi-

Tabela 5.8: Taxas de reconhecimento (%) com características extraídas com LBP

| **Padrão de zoneamento** | **Regra do máximo** | **Regra do mínimo** | **Regra do produto** | **Regra da soma** |
|---|---|---|---|---|
| Extração global de características | 76,78±0,38 | 77,11±1,71 | 78,67±0,67 | 79,00±1,00 |
| Divisão linear com cinco zonas | 73,56±1,95 | 75,22±2,71 | **79,44±1,35** | 79,22±1,39 |
| Divisão linear com dez zonas | 71,56±1,26 | 72,56±2,67 | 77,78±0,38 | 77,56±1,17 |

vidual já é superior ao melhor desempenho obtido com a fusão de classificadores utilizando GLCM e está próximo do melhor desempenho obtido com a fusão de classificadores construídos utilizando Filtros de Gabor. Este melhor desempenho individual foi obtido com extração global de características. Cabe aqui destacar que os classificadores criados individualmente quando empregada a extração global de características apresentaram muito bom desempenho. Como consequência, o desempenho final alcançado com extração global ficou muito próximo do desempenho obtido quando empregado o zoneamento das imagens. Esta situação destoa sensivelmente do ocorrido com os descritores apresentados anteriormente.

Embora muito próximo do resultado obtido com extração global de características com a fusão pela regra da soma, o melhor desempenho foi obtido novamente utilizando a divisão linear em cinco zonas com a aplicação da regra do produto para a fusão das saídas. A matriz de confusão obtida no melhor caso encontra-se descrita na tabela 5.9.

Tabela 5.9: Matriz de confusão (%) obtida no melhor caso (regra do produto) com LBP e divisão linear em cinco zonas

| | (0) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|---|
| (0) Axé | **77,78** | 0,00 | 2,22 | 0,00 | 0,00 | 1,11 | 11,11 | 2,22 | 5,56 | 0,00 |
| (1) Bachata | 1,11 | **93,33** | 2,22 | 1,11 | 0,00 | 0,00 | 0,00 | 1,11 | 1,11 | 0,00 |
| (2) Bolero | 0,00 | 1,11 | **88,89** | 2,22 | 2,22 | 0,00 | 0,00 | 1,11 | 2,22 | 2,22 |
| (3) Forró | 0,00 | 1,11 | 5,56 | **80,00** | 5,56 | 1,11 | 2,22 | 2,22 | 2,22 | 0,00 |
| (4) Gaúcha | 15,56 | 1,11 | 5,56 | 7,78 | **62,22** | 0,00 | 0,00 | 0,00 | 7,78 | 0,00 |
| (5) Merengue | 1,11 | 3,33 | 0,00 | 0,00 | 0,00 | **93,33** | 0,00 | 2,22 | 0,00 | 0,00 |
| (6) Pagode | 11,11 | 0,00 | 13,33 | 0,00 | 1,11 | 0,00 | **60,00** | 10,00 | 4,44 | 0,00 |
| (7) Salsa | 2,22 | 0,00 | 2,22 | 0,00 | 3,33 | 1,11 | 2,22 | **87,78** | 1,11 | 0,00 |
| (8) Sertaneja | 12,22 | 1,11 | 10,00 | 3,33 | 11,11 | 0,00 | 2,22 | 0,00 | **60,00** | 0,00 |
| (9) Tango | 1,11 | 0,00 | 5,56 | 0,00 | 2,22 | 0,00 | 0,00 | 0,00 | 0,00 | **91,11** |

O principal foco de confusão observado está em classificações equivocadas de gaúcha para axé. De forma geral, as principais confusões aqui apresentadas são similares às ocorridas com o uso dos descritores já apresentados. Entretanto, estas confusões aparecem, em geral, com intensidade reduzida quando utilizado o $LBP_{8,2}$. As melhores taxas de acerto acontecem com bachata, merengue e tango, gêneros que também apresentaram bom desempenho com os descritores descritos anteriormente. Mais uma vez fica reforçado o entendimento de que, em linhas gerais, os diferentes descritores capturam características de textura semelhantes, embora provavelmente em proporções diferentes.

Diante do bom desempenho apresentado por LBP, optou-se por realizar com este descritor uma série de experimentos diferentes que buscam verificar a versatilidade do método ou a existência de alternativas que possibilitem alcançar resultados ainda melhores. Estes experimentos encontram-se descritos nas próximas subseções.

### 5.4.1 Variando parâmetros de LBP

Conforme descrito na seção 4.5.3, podem ser empregados diferentes tipos de vizinhança para a extração de características com LBP. Nesta subseção serão descritos alguns resultados obtidos utilizando LBP com diferentes valores para os parâmetros $R$ e $P$. Neste caso, o único padrão de zoneamento a ser utilizado será o de cinco zonas lineares, já que este produziu os melhores resultados nos experimentos anteriormente descritos nesta seção. Os padrões experimentados foram $LBP_{8,1}$, $LBP_{8,2}$ e $LBP_{16,2}$, sendo o primeiro elemento da dupla o valor de $P$ e o segundo elemento o valor de $R$. Os resultados estão descritos na tabela 5.10.

Tabela 5.10: Taxas de reconhecimento (%) com LBP utilizando cinco zonas lineares e variando os valores de $R$ e $P$

| Padrão LBP | Regra do máximo | Regra do mínimo | Regra do produto | Regra da soma |
|---|---|---|---|---|
| $LBP_{8,1}$ | 71,11±0,96 | 72,56±4,02 | 77,33±1,45 | 76,56±0,69 |
| $LBP^{*}_{8,2}$ | 73,56±1,95 | 75,22±2,71 | 79,44±1,35 | 79,22±1,39 |
| $LBP_{16,2}$ | 74,22±0,69 | 76,44±1,02 | **79,56±1,90** | 79,33±2,03 |

* Resultados apresentados na tabela 5.8

O melhor resultado foi obtido com $LBP_{16,2}$. Entretanto, o resultado é muito próximo do melhor resultado obtido com $LBP_{8,2}$ e é importante ressaltar que o aumento do valor de $P$ faz com que o custo para a extração das características aumente significativamente. Assim, entende-se que $LBP_{8,2}$ seja o padrão mais apropriado para o problema aqui abordado.

### 5.4.2 Escalas não lineares

Nesta seção, são descritos experimentos realizados dividindo-se as imagens dos espectrogramas em zonas não lineares. Para isto, utilizou-se as escalas de Bark e Mel, que criam bandas de frequência de acordo com a percepção humana. O descritor utilizado foi $LBP_{8,2}$.

### 5.4.3 Escala de Bark

A escala de Bark é uma escala psicoacústica e sua criação se deu em uma tentativa de representar os limites das bandas críticas de audição, segundo as quais a audição humana é capaz de discernir sons e ruídos [94]. Os limites das bandas de frequência em Hz são: 0, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500. Uma vez que o sinal das músicas da base LMD só possuem informação relevante até a altura de 8500 Hz, o número de bandas criadas nos espectrogramas será igual a 22 ao invés de 24, conforme descrito na subseção 5.4.3. A figura 4.9 ilustra a sobreposição das bandas criadas utilizando a escala de Bark sobre um espectrograma extraído de uma música da base LMD.

Os resultados obtidos com a fusão das saídas dos 66 classificadores criados utilizando diferentes regras de fusão é mostrado na tabela 5.11.

Tabela 5.11: Taxas de reconhecimento (%) com $LBP_{8,2}$ e divisão da imagem em zonas segundo a escala de Bark

| Padrão de zoneamento | Regra do máximo | Regra do mínimo | Regra do produto | Regra da soma |
|---|---|---|---|---|
| Zoneamento segundo a escala de Bark | 66,44±1,07 | 67,67±1,00 | 55,67±1,73 | **78,00±1,33** |

Os resultados obtidos com a escala de Bark são ligeiramente inferiores aos resultados obtidos nos experimentos em que foram utilizadas escalas lineares. Somando-se a isto o fato de que o uso da escala de Bark produz um número maior de zonas, e consequentemente mais classificadores, observa-se que o uso dela não é recomendável.

### 5.4.4 Escala Mel

De acordo com Umesh *et al.* [88], a escala Mel resulta fundamentalmente da psicoacústica, relacionando as frequências reais com as frequências percebidas pelos humanos, assim como na escala de Bark. Conforme já descrito na subseção 5.4.4, nesta escala são estabelecidas 15 bandas de frequência, cujos limites em Hz são: 0, 40, 161, 200, 404, 693, 867, 1000, 2022, 3000, 3393, 4109, 5526, 6500, 7743 e 14000. A figura 4.10 mostra a divisão segundo a escala Mel sobreposta a um espectrograma extraído de uma música da base LMD. É válido relembrar que o limite superior até o qual as músicas da base LMD apresentam informação é 8500 Hz.

Os resultados obtidos com a fusão das saídas dos classificadores criados utilizando diferentes regras de fusão é mostrado na tabela 5.12.

Tabela 5.12: Taxas de reconhecimento (%) com $LBP_{8,2}$ e divisão da imagem em zonas segundo a escala Mel

| Padrão de zoneamento | Regra do máximo | Regra do mínimo | Regra do produto | Regra da soma |
|---|---|---|---|---|
| Zoneamento segundo a escala Mel | 72,33±3,33 | 71,22±2,34 | **82,33±1,45** | 81,11±1,35 |

Os resultados alcançados com a escala Mel são os melhores obtidos até então, superando ligeiramente os melhores resultados conseguidos nos outros experimentos apresentados até aqui neste trabalho.

### 5.4.5 Base *ISMIR 2004*

A fim de verificar o poder de generalização da metodologia proposta para outras bases de músicas, realizou-se experimentos sobre a base ISMIR 2004. Os experimentos foram realizados utilizando o descritor $LBP_{8,2}$, que apresentou resultados bastante interessantes nos experimentos descritos inicialmente nesta seção. Foram utilizados alguns diferentes

padrões de zoneamento, além da extração global de características. As imagens foram divididas linearmente em cinco e dez zonas de igual tamanho, além dos zoneamentos não lineares com as escalas de Bark e Mel. Os resultados, com quatro diferentes regras de fusão estão descritos na tabela 5.13.

Tabela 5.13: Taxas de reconhecimento (%) sobre a base ISMIR 2004, utilizando $LBP_{8,2}$ e diferentes padrões de zoneamento das imagens

| **Padrão de zoneamento** | **Regra do máximo** | **Regra do mínimo** | **Regra do produto** | **Regra da soma** |
|---|---|---|---|---|
| Extração global | 77,42 | 78,96 | **80,65** | 79,80 |
| Divisão linear em cinco zonas | 72,37 | 75,88 | 77,70 | 77,28 |
| Divisão linear em dez zonas | 71,11 | 76,02 | 78,40 | 77,42 |
| Divisão pela escala de Bark | 64,66 | 71,53 | 68,86 | 70,69 |
| Divisão pela escala Mel | 67,32 | 76,44 | 75,74 | 73,91 |

Os resultados mostram que, de forma geral, obteve-se sobre a base ISMIR 2004 desempenho da metodologia aqui proposta comparável ao desempenho alcançado por outras abordagens apresentadas na literatura. Além disso, os melhor resultado é também próximo ao melhor resultado obtido sobre a base LMD. Isto é importante para eliminar dúvidas quanto à viabilidade de aplicação da metodologia sobre outras bases.

Embora não haja diferença em termos de significância estatística entre os melhores resultados alcançados com e sem zoneamento, o melhor resultado foi alcançado com extração global de características, o que mostra que a preservação de informações locais das características extraídas pode não ser necessáriamente a melhor estratégia quando se aplica a metodologia em casos gerais.

### 5.4.6 Características de um único segmento

Foi verificado o desempenho do esquema de classificação utilizando-se apenas características extraídas do segmento central das músicas. A ideia por trás destes experimentos é de que, em caso de bom desempenho, seria possível reduzir custos tanto no processo de extração de caracteristicas quanto no processo de classificação propriamente dito, uma vez que o número de classificadores ficaria reduzido. A tabela 5.14 mostra os resultados obtidos utilizando-se descritores extraídos com $LBP_{8,2}$ e aplicando diferentes esquemas de zoneamento lineares e não lineares.

Tabela 5.14: Taxas de reconhecimento (%) com $LBP_{8,2}$ e diferentes padrões de zoneamento utilizando apenas o segmento central das músicas

| **Padrão de zoneamento** | **Regra do máximo** | **Regra do mínimo** | **Regra do produto** | **Regra da soma** |
|---|---|---|---|---|
| Extração global* | | 71,22±2,22 | | |
| Cinco zonas lineares | 70,33±2,91 | 71,56±0,69 | 74,78±2,27 | 74,22±1,68 |
| Dez zonas lineares | 69,00±1,86 | 71,44±3,01 | 74,33±0,33 | 73,33±1,20 |
| Escala de Bark | 63,22±3,36 | 67,78±3,37 | 75,78±2,14 | 74,11±2,67 |
| Escala Mel | 69,67±1,20 | 70,11±0,51 | **77,44±0,51** | 76,78±0,84 |

* Neste caso, apenas um classificador é criado

Considerando que apenas um segmento foi utilizado nestes experimentos, pode-se considerar que o melhor desempenho obtido, com divisão pela escala Mel, não é tão inferior

aos melhores desempenhos obtidos em outros experimentos descritos ao longo deste capítulo, com o uso dos três segmentos. Entretanto, as estratégias empregadas nas situações em que se obteve os melhores resultados com o uso dos três segmentos parecem mais recomendáveis, uma vez que ainda guardam alguma margem de superioridade razoável.

## 5.5 Local Phase Quantization (LPQ)

Esta seção descreve os resultados obtidos com LPQ. Os descritores foram extraídos conforme parâmetros descritos na subseção 4.5.4. A tabela 5.15 mostra os resultados obtidos utilizando diferentes regras de fusão com extração global e zoneamento linear com cinco e dez zonas.

Tabela 5.15: Taxas de reconhecimento (%) com LPQ

| **Padrão de zoneamento** | **Regra do máximo** | **Regra do mínimo** | **Regra do produto** | **Regra da soma** |
|---|---|---|---|---|
| Extração global | 73,67±3,53 | 74,22±2,50 | 77,11±3,24 | **77,67±2,85** |
| Divisão linear em cinco zonas | 71,22±0,19 | 73,44±0,19 | 76,33±1,45 | 76,56±0,51 |
| Divisão linear em dez zonas | 67,44±2,01 | 72,78±1,17 | 76,22±1,58 | 75,44±1,68 |

Com o uso de LPQ, o melhor resultado obtido foi produzido com extração global de características e utilizando a regra da soma. O melhor resultado foi superado somente por resultados obtidos com LBP. O classificador de melhor desempenho individual foi o do segmento central quando foi utilizada a extração global de características, e sua taxa de reconhecimento foi igual a 70,11%. A matriz de confusão produzida no melhor caso, extração global com fusão pela regra da soma está descrita na tabela 5.16.

Tabela 5.16: Matriz de confusão (%) obtida no melhor caso (regra da soma) com LPQ e extração global

| | (0) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|---|
| (0) Axé | **67,78** | 0,00 | 1,11 | 0,00 | 4,44 | 1,11 | 14,44 | 4,44 | 6,67 | 0,00 |
| (1) Bachata | 2,22 | **92,22** | 2,22 | 0,00 | 0,00 | 1,11 | 0,00 | 1,11 | 1,11 | 0,00 |
| (2) Bolero | 0,00 | 3,33 | **84,44** | 1,11 | 1,11 | 0,00 | 1,11 | 1,11 | 5,56 | 2,22 |
| (3) Forró | 0,00 | 1,11 | 6,67 | **67,78** | 11,11 | 0,00 | 2,22 | 2,22 | 8,89 | 0,00 |
| (4) Gaúcha | 17,78 | 0,00 | 4,44 | 10,00 | **57,78** | 2,22 | 2,22 | 1,11 | 4,44 | 0,00 |
| (5) Merengue | 1,11 | 2,22 | 0,00 | 0,00 | 2,22 | **91,11** | 1,11 | 2,22 | 0,00 | 0,00 |
| (6) Pagode | 5,56 | 0,00 | 5,56 | 1,11 | 1,11 | 0,00 | **83,33** | 2,22 | 1,11 | 0,00 |
| (7) Salsa | 3,33 | 1,11 | 5,56 | 0,00 | 1,11 | 4,44 | 4,44 | **78,89** | 1,11 | 0,00 |
| (8) Sertaneja | 13,33 | 0,00 | 10,00 | 6,67 | 4,44 | 0,00 | 1,11 | 0,00 | **64,44** | 0,00 |
| (9) Tango | 0,00 | 0,00 | 6,67 | 0,00 | 3,33 | 0,00 | 1,11 | 0,00 | 0,00 | **88,89** |

Em linhas gerais, as confusões guardam semelhança com as apresentadas quando outros descritores de textura foram utilizados. Os gêneros bachata, merengue e tango figuram novamente como os de melhor desempenho. Mais uma vez confirma-se a hipótese de que características de textura, provavelmente associadas a dimensões musicais, estejam sendo capturadas de forma a permitir uma discriminação entre gêneros musicais.

### 5.5.1 Variando parâmetro de LPQ

Nesta subseção são descritos resultados obtidos variando-se o tamanho da janela $m$ na extração de características utilizando o descritor LPQ. Foram realizados experimentos com o valor de $m$ variando entre três e onze e foi utilizada a estratégia de extração global de características (sem divisão das imagens dos espectrogramas em zonas), uma vez que esta foi a que apresentou os melhores resultados nos experimentos descritos na seção 5.5. Os resultados estão descritos na tabela 5.17.

Tabela 5.17: Taxas de reconhecimento (%) com LPQ utilizando extração global de características e variando o tamanho da janela $m$

| **Tamanho da janela $m$** | **Regra do máximo** | **Regra do mínimo** | **Regra do produto** | **Regra da soma** |
|---|---|---|---|---|
| 3* | 73,67±3,53 | 74,22±2,50 | 77,11±3,24 | 77,67±2,85 |
| 5 | 75,33±0,58 | 76,67±0,88 | 80,33±0,67 | 78,67±1,45 |
| 7 | 76,89±2,12 | 77,22±1,68 | **80,78±0,77** | 79,44±1,17 |
| 9 | 74,11±0,69 | 75,56±2,67 | 77,44±2,01 | 77,00±2,31 |
| 11 | 74,11±0,69 | 75,56±2,67 | 77,44±2,01 | 77,00±2,31 |

* Resultados apresentados na tabela 5.15

O melhor resultado foi obtido com o tamanho da janela $m$ igual a sete. A partir deste valor, as taxas de reconhecimento começam a cair. Com isto, entende-se que $m = 7$ seja uma boa medida para o uso de LPQ dentro da metodologia aqui proposta, sobretudo porque quando se utiliza este valor, não há aumento de custo perceptível em termos de tempo de processamento, quando comparado a janelas de tamanhos menores. Adicionalmente, considerando-se o fato de que os melhores resultados obtidos com LPQ utilizam extração global de características, este descritor é fortemente recomendado. Uma vez que a extração global implica em um número reduzido de classificadores, diminuindo a complexidade geral do sistema.

## 5.6 Características visuais e acústicas

Esta seção apresenta resultados obtidos utilizando características acústicas (descritas na subseção 3.1.1), tradicionalmente utilizadas em tarefas de recuperação de informações musicais. Estes experimentos são importantes para comparar os desempenhos obtidos com as características obtidas no domínio visual, propostas neste trabalho, com o desempenho obtido utilizando características tradicionais.

Nestes experimentos foi utilizada a base LMD e foi empregado o framework *MARSYAS* [87] para extrair as características *Spectral Centroid*, *Roll-off*, *Flux*, *Zero Crossing* e 13 diferentes *MFCCs*. Estas 17 características foram colhidas ao longo do sinal de duas maneiras diferentes, conforme feito em [62], o que resultou em 34 valores. Isto feito, médias e desvios padrão destas características foram calculados e utilizados para compor um vetor final cujo tamanho é igual a 68. A taxa de reconhecimento obtida foi de 61,11%, com desvio padrão de 1,85% considerando-se a média entre os três folds utilizados.

No trabalho de Wu *et al.* [90], os autores experimentam o uso de vetores formados com características acústicas concatenadas a características obtidas no domínio visual. Os autores apresentam resultados que sugerem que esta estratégia pode ser positiva na busca por bons resultados na classificação de gêneros musicais. Com isto, decidiu-se realizar experimentos similares utilizando características acústicas e visuais concatenadas em um mesmo vetor. As características utilizadas foram as 59 de $LBP_{8,2}$ e as 68 acústicas mencionadas nesta seção, totalizando 127 características. Para a extração das características visuais, foram empregados três padrões de zoneamento linear das imagnes, global, com cinco zonas e com dez zonas, além de dois padrões de zoneamento não lineares, utilizando as escalas de Bark e Mel. Os resultados obtidos estão descritos na tabela 5.18.

Tabela 5.18: Taxas de reconhecimento (%) utilizando características de $LBP_{8,2}$ com diferentes padrões de zoneamento concatenadas a características acústicas

| **Padrão de zoneamento** | **Regra do máximo** | **Regra do mínimo** | **Regra do produto** | **Regra da soma** |
|---|---|---|---|---|
| Extração global | 76,00±1,70 | 76,11±0,57 | 77,33±1,09 | **77,78±1,50** |
| 5 zonas lineares | 76,00±1,91 | 77,00±1,70 | 77,22±1,10 | 76,22±1,10 |
| 10 zonas lineares | 75,67±1,70 | 75,56±2,20 | 76,11±1,03 | 75,78±1,29 |
| Escala de Bark | 73,44±0,96 | 74,33±1,19 | 72,00±1,36 | 73,00±1,19 |
| Escala Mel | 73,89±0,68 | 74,67±1,52 | 75,56±1,13 | 75,33±0,98 |

O melhor resultado foi obtido com a extração global das características visuais. Esta taxa não é superior aos melhores resultados obtidos utilizando somente características $LBP_{8,2}$. Além de não produzir melhores taxas de reconhecimento, esta estratégia aumentou o tamanho do vetor de características utilizado na classificação. Desta forma, as expectativas iniciais não foram alcançadas com esta estratégia.

## 5.7 Todos os descritores visuais juntos

A fim de investigar a hipótese de que haja complementaridade entre os diferentes descritores de textura utilizados para a extração de características no domínio visual, realizou-se um experimento utilizando um vetor com características visuais de todas as abordagens testadas concatenadas. Para isto, foram construídos vetores com as 28 características de GLCM, 59 características de LBP, 256 características de LPQ e 120 caracterísitcas obtidas com filtros de Gabor. Assim, o vetor final foi composto por 463 características. Foram experimentados dois diferentes padrões de divisão das imagens: extração global e zoneamento linear com cinco zonas. Estes padrões foram os escolhidos porque foram os que apresentaram os melhores resultados quando os quatro diferentes descritores de textura foram experimentados isoladamente. A tabela 5.19 mostra os resultados obtidos.

Ainda na tentativa de identificar eventuais complementaridades, realizou-se um experimento adicionando-se ao vetor de características as 68 características acústicas utilizadas nos experimentos descritos na seção 5.6. Assim, o vetor final foi composto por um total de 531 características. Foi escolhido o padrão de zoneamento com cinco zonas lineares, já

Tabela 5.19: Taxas de reconhecimento (%) as características obtidas com GLCM, LBP, LPQ e filtros de Gabor concatenadas

| Padrão de zoneamento | Regra do máximo | Regra do mínimo | Regra do produto | Regra da soma |
|---|---|---|---|---|
| Extração global | 78,56±0,96 | 77,33±3,18 | 78,67±2,03 | 79,22±1,58 |
| 5 zonas lineares | 77,67±0,88 | 78,11±2,22 | 81,89±3,66 | **82,44±2,27** |

que este apresentou o melhor desempenho nos experimentos descritos na tabela 5.19. A tabela 5.20 mostra os resultados obtidos.

Tabela 5.20: Taxas de reconhecimento (%) as características obtidas com GLCM, LBP, LPQ, filtros de Gabor e as características acústicas concatenadas

| Padrão de zoneamento | Regra do máximo | Regra do mínimo | Regra do produto | Regra da soma |
|---|---|---|---|---|
| 5 zonas lineares | 80,44±3,34 | 80,44±3,86 | 81,89±3,98 | **82,00±3,84** |

O uso das características acústicas concatenadas às obtidas no domínio visual é contraindicado, uma vez que além de aumentar a dimensionalidade do vetor, não contribui com a melhoria das taxas de reconhecimento. Já o uso apenas das características obtidas no domínio visual concatenadas apresenta um desempenho superior aos apresentados individualmente por estas características. Entretanto, deve-se ponderar acerca da dimensionalidade excessiva do vetor utilizado na classificação.

## 5.8 Verificação do tempo de execução

A fim de que se possa ter uma ideia do desempenho do método aqui proposto em termos de tempo, foi calculado o tempo gasto acumulado em todas as etapas previstas para a classificação de uma música com quatro minutos de duração em alguns diferentes cenários. Para isto, foram considerados alguns cenários nos quais se obteve boas taxas de reconhecimento, a saber: extração global de características utilizando o descritor $LBP_{8,2}$, extração de características com zoneamento linear em cinco zonas utilizando o descritor $LBP_{8,2}$, extração de características com zoneamento segundo a escala Mel utilizando o descritor $LBP_{8,2}$ e extração global de características utilizando o descritor LPQ com janela com tamanho igual a sete.

A tabela 5.21 apresenta os resultados obtidos. Os tempos gastos em algumas etapas, tais como: conversão do sinal do formato MP3 para o formato Wave, segmentação do sinal, geração do espectrograma e conversão da imagem do espectrograma para níveis de cinza são comuns em todos os cenários. Já os tempos gastos nas etapas de extração de características, classificação e fusão das saídas dos classificadores variam em função do tipo de característica utilizada e do padrão de zoneamento adotado. Aqui, é válido relembrar que o número de zonas utilizadas determina o número de classificadores que serão criados.

Considerando os resultados obtidos, percebe-se que, em todos os casos, o acúmulo de

Tabela 5.21: Tempo gasto em milisegundos nas diferentes etapas considerando diferentes cenários de classificação

| Descritor/ zoneamento | Conversão MP3-Wav | Segment. do sinal | Geração do espectrog. | Conversão para níveis de cinza | Extração de caracterist. | Classificação | Fusão | **Total** |
|---|---|---|---|---|---|---|---|---|
| $LBP_{8,2}$ global | 350 | 290 | 410 | 70 | 216 | 260 | 20 | **1616** |
| $LBP_{8,2}$ 5 zonas | 350 | 290 | 410 | 70 | 221 | 1140 | 30 | **2511** |
| $LBP_{8,2}$ Mel | 350 | 290 | 410 | 70 | 502 | 3790 | 50 | **5462** |
| LPQ global | 350 | 290 | 410 | 70 | 444 | 580 | 20 | **2164** |

tempo gasto nas diferentes etapas alcança somas aceitáveis para aplicações com expectativa de resposta em tempo real. Mesmo levando-se em consideração o pior caso, em que se utiliza o descritor $LBP_{8,2}$ com a divisão da imagem em zonas segundo a escala Mel, o tempo gasto, de aproximadamente 5,4 segundos é suportável para uma situação em que a classificação instantânea fosse esperada. Outro fato interessante de se observar é que o aumento da quantidade de classificadores, com a criação de um maior número de zonas na imagem, impacta de forma mais importante no tempo de classificação do que o uso de um descritor com maior número de características. Nos experimentos cujos resultados estão descritos nesta seção, o descritor $LBP_{8,2}$ produz um total de 59 características, enquanto LPQ produz 256 características.

## 5.9 Teste estatístico

A fim de verificar se há diferença estatisticamente significantiva entre os diferentes classificadores descritos neste capítulo, foi aplicado o teste de Friedman [24] com o procedimento sequencial de Holm [34]. Foram incluídos no teste os melhores resultados obtidos sobre a base LMD com os descritores de textura GLCM, Filtros de Gabor, LBP e LPQ, além do melhor resultado obtido utilizando vetores com as características obtidas com todos os descritores de textura concatenados, os resultados obtidos com descritores acústicos e os melhores resultados obtidos com a divisão das imagens com escalas não lineares (LBP com escala de Bark e escala Mel).

A tabela 5.22 mostra o $valor-p$ encontrado para cada um dos classificadores utilizando nível de significância $\alpha = 0,05$. Com o procedimento de Holm, os classificadores criados com características acústicas e os classificadores criados com características extraídas com GLCM ficaram abaixo do valor crítico ($p \leq 0,01$) e foram hipóteses rejeitadas. Portanto, há diferença estatisticamente significativa entre os resultados produzidos por estes dois classificadores e os demais.

## 5.10 Conclusão

Este capítulo descreveu resultados obtidos com os quatro diferentes descritores de textura escolhidos para verificação da efetividade do método aqui proposto. Inicialmente, percebe-

Tabela 5.22: $Valor - p$ encontrado para os classificadores comparados

| **Classificador** | $Valor - p$ |
|---|---|
| Características acústicas | **0,00154** |
| GLCM | **0,00766** |
| Filtros de Gabor | 0,04550 |
| Zoneamento com escala de Bark | 0,06675 |
| LBP | 0,31731 |
| LPQ | 0,55966 |
| Zoneamento com escala Mel | 0,93358 |
| Todas as características de textura | 1 |
| Valor crítico: 0,01 | |

se pelos resultados obtidos que independentemente do descritor de textura utilizado, os resultados são comparáveis aos descritos na literatura com o uso de outras abordagens, tradicionalmente baseadas em características acústicas. Além disso, percebe-se ainda que, de forma geral, os diferentes descritores apresentam comportamentos parecidos com a variação dos padrões de zoneamento testados e das regras de fusão aplicadas nas saídas dos classificadores.

A tabela 5.23 mostra os melhores resultados obtidos sobre a base LMD com cada descritor de textura experimentado.

Tabela 5.23: Melhores resultados com cada descritor de textura experimentado

| **Descritor de textura** | **Divisão da imagem** | **Regra de Fusão** | **Taxa de reconhecimento (%)** |
|---|---|---|---|
| GLCM | 5 zonas | Regra do produto | 70,78 |
| Filtros de Gabor | 5 zonas | Regra do produto | 74,67 |
| **LBP** | **Escala Mel** | **Regra do produto** | **82,33** |
| LPQ | Global | Regra do produto | 80,78 |

Na grande maioria dos casos, a regra de fusão do produto produziu os melhores resultados. Em alguns poucos casos a regra da soma proporcionou melhores resultados. Contudo, as diferenças entre os resultados produzidos por estas duas regras foi quase sempre desprezível e, além disso, praticamente não há diferença entre o custos para a aplicação das duas. Isto faz com que ambas as regras sejam boas alternativas no processo de classificação aqui proposto.

No que diz respeito ao padrão de zoneamento, pode-se notar que a preservação de informações locais (estabelecida com o zoneamento) contribui em alguma medida com a produção de melhores resultados. Quando empregada a divisão linear, a criação de cinco zonas proporcionou os melhores resultados. A divisão em um maior número de zonas lineares não parece ser uma boa ideia, já que os resultados com o uso de dez zonas lineares foram, em geral, inferiores aos obtidos com cinco zonas lineares. O melhor resultado geral foi obtido com a divisão da imagem segundo a escala Mel. É importante lembrar o fato de que o número de zonas criadas determina diretamente o número de classificadores a serem criados segundo o método aqui proposto. Conforme descrito na seção 5.8, este fato

traz algum impacto no tempo gasto para realizar a classificação. Entretanto, constatou-se que a diferença não inviabiliza a aplicação do método de classificação em situações nas quais há expectativa por resposta em tempo real.

Em algumas situações, LBP e LPQ produziram resultados que, além de superiores aos obtidos com os demais descritores de textura, são comparáveis aos melhores resultados já descritos na literatura. No caso da base LMD, considerando-se a restrição *"artist filter"* os resultados obtidos são os melhores já apresentados. Sobre a base *ISMIR 2004* também foram obtidos bons resultados, próximos ou superiores a muitos descritos na literatura. Este fato é importante para certificar a versatilidade do método aqui proposto com o uso de $LBP_{8,2}$ e sua capacidade de generalização para outras bases de dados. Em algumas situações, LPQ produz resultados similares e até superiores aos de $LBP_{8,2}$. Entretanto, ao se levar em consideração a menor dimensionalidade do vetor produzido com $LBP_{8,2}$, optou-se por este operador para ser utilizado nos experimentos adicionais com KNORA, que estão descritos no capítulo 6. No mesmo capítulo também estão descritos experimentos envolvendo seleção de características com todos os descritores de textura aqui investigados.

# CAPÍTULO 6

# EXPERIMENTOS ADICIONAIS

Este capítulo descreve experimentos adicionais utilizando algumas técnicas complementares àquelas empregadas nos experimentos descritos no capítulo 5. Foram experimentadas duas abordagens, a primeira consiste no uso de um método para selecionar dinamicamente um agrupamento de classificadores, o KNORA, e a segunda consiste no uso de algoritmos genéticos para a seleção de características. Os experimentos aqui descritos foram realizados sobre a base LMD.

## 6.1 Seleção dinâmica de agrupamento de classificadores com KNORA

O KNORA é um método para seleção dinâmica de agrupamento de classificadores, e foi empregado neste trabalho com o objetivo de explorar a complementaridade identificada entre os diferentes classificadores criados com a estratégia de zoneamento das imagens. Esta complementaridade se revela nas altas taxas de reconhecimento identificadas no limite superior entre os conjuntos de classificadores selecionados para o experimento. Para encontrar o limite superior, deve-se considerar que um padrão é classificado corretamente se qualquer um dos $n$ classificadores envolvidos no sistema for capaz de classificar corretamente este padrão.

Para a realização destes experimentos foram utilizadas características extraídas com LBP com dois diferentes padrões de zoneamento das imagnes que produzem uma grande quantidade de classificadores, a divisão linear com dez zonas (30 classificadores) e a divisão não linear pela escala Mel (45 classificadores). Adicionalmente, foram selecionados 400 títulos musicais da LMD, mantendo-se as restrições impostas pelo *"artist filter"*, para compor o conjunto de validação.

O KNORA foi utilizado com seis variações diferentes. KNORA-ELIMINATE e KNORA-UNION foram experimentados fundindo-se as saídas dos classificadores selecionados pelas regras da soma e do produto. Além disso, KNORA-UNION-W também foi utilizado com fusão das saidas dos classificadores pelas mesmas regras. Entretanto, houve uma adaptação com a qual o peso W foi estabelecido em função da quantidade de vezes que cada classificador foi selecionado. Nos experimentos realizados, o valor de $K$ variou de 1 a 20. As subseções 6.1.1 e 6.1.2 mostram, respectivamente, os resultados obtidos utilizando a divisão linear das imagens em dez zonas e a divisão das imagens em zonas segundo a escala Mel.

### 6.1.1 KNORA com divisão linear em dez zonas

A taxa de reconhecimento identificada no limite superior entre os 30 classificadores criados foi de 98,67%. Em um primeiro momento, este valor bastante elevado gera grandes expectativas de que o uso de KNORA possa trazer significativas melhorias ao desempenho do sistema de classificação. Entretanto, não é o que se verifica nos resultados obtidos, que encontram-se descritos na tabela 6.1. O valor de $K$ indicado na tablea refere-se ao menor valor de $K$ com o qual foi possível obter o melhor desempenho.

Tabela 6.1: Taxas de reconhecimento (%) obtidas com KNORA e divisão linear em dez zonas

| Esquema | Regra do produto | Regra da soma |
|---|---|---|
| Fusão direta* | 77,78 | 77,56 |
| KNORA-ELIMINATE | 77,44 ($K$=5) | 77,44 ($K$=18) |
| KNORA-UNION | 79,11 ($K$=2) | 78,56 ($K$=2) |
| KNORA-UNION W | 79,11 ($K$=2) | **79,33** ($K$=5) |

* sem o uso de KNORA, resultados apresentados na tabela 5.8

As melhorias obtidas no desempenho são bastante discretas quando comparados ao resultado obtido com a fusão direta entre as saídas dos classificadores (sem o uso do KNORA). A frustração das expectativas provavelmente se deva a baixa representatividade da base de validação utilizada nos experimentos.

### 6.1.2 KNORA com divisão segundo a escala Mel

O limite superior entre os 45 classificadores criados neste caso mostram uma excelente taxa de reconhecimento, igual a 99,78%. Contudo, mais uma vez as boas expectativas não se confirmaram, conforme mostram os resultados descritos na tabela 6.2.

Tabela 6.2: Taxas de reconhecimento (%) obtidas com KNORA e zoneamento por escala Mel

| Esquema | Regra do produto | Regra da soma |
|---|---|---|
| Fusão direta* | 82,33 | 81,11 |
| KNORA-ELIMINATE | 81,00 ($K$=19) | 80,22 ($K$=13) |
| KNORA-UNION | 81,00 ($K$=19) | 81,89 ($K$=7) |
| KNORA-UNION W | **83,00** ($K$=7) | 82,11 ($K$=13) |

* sem o uso de KNORA

O melhor resultado geral obtido é o melhor dentre todos os experimentos realizados neste trabalho. Entretanto, o ganho obtido na taxa de reconhecimento parece não justificar o uso desta técnica, que introduz um custo importante ao processo de classificação. A provável explicação para o desempenho abaixo do esperado é a mesma apresentada no experimento descrito anteriormente.

## 6.2 Seleção de características com Algoritmo Genético

Muitos dos vetores de características extraídas no domínio visual utilizadas nos experimentos descritos neste trabalho apresentam um tamanho relativamente grande. Em função disso, optou-se por realizar alguns experimentos utilizando Algoritmos Genéticos a fim de verificar a possibilidade de melhorias no desempenho. Neste contexto, pode-se entender por melhoria no desempenho alcançar uma redução no tamanho dos vetores mantendo os níveis alcançados nas taxas de reconhecimento, ou ainda conseguir alcançar melhores taxas de reconhecimento com um subconjunto das características utilizadas originalmente. Utilizou-se uma função multi-objetivo nas tarefas de seleção de características, de forma que os objetivos eram, minimizar o número de características, maximizando a taxa de acerto.

As próximas subseções descrevem os resultados obtidos com extração global de caracteristicas utilizando os quatro diferentes descritores de textura explorados neste trabalho. Adicionalmente, serão descritos resultados obtidos utilizando vetores com todas as características visuais concatenadas. Serão ainda apresentados resultados obtidos criando-se cinco zonas lineares para a extração de características das imagens. Neste caso, foram utilizadas características obtidas com filtros de Gabor, LBP e vetores com todas as características visuais concatenadas.

Para a seleção de características foram utilizadas, além da base de treinamento, uma base de busca (*search-database*) e uma base de validação (*validation-database*). A fim de evitar que problemas relacionados à ocorrência de *overfitting* pudessem produzir um modelo com baixa capacidade de generalização. Para compor a base de busca foi utilizado o fold com 400 músicas já utilizado em experimentos descritos anteriormente. A base de treinamento foi formada por 600 músicas. A base de validação foi composta por 300 músicas. A *fitness* utilizada foi a minimização do erro na base de busca. O algoritmo foi executado até mil gerações e o tamanho da população utilizada foi igual a 40. A seleção dos indivíduos foi feita pelo método “roleta russa”.

### 6.2.1 Seleção de características com extração global

Nesta seção são descritos os resultados obtidos com a extração global de características. É válido lembrar que, neste caso, são criados três classificadores. Sendo um para cada um dos segmentos extraídos das músicas. Foram realizadas tarefas de seleção de características específicas para cada um dos três classificadores. Com base nas características selecionadas para cada classificador, realizou-se a classificação utilizando-se os mesmos folds utilizados nas tarefas de reconhecimento descritas ao longo deste trabalho. As tabelas descritas nesta seção mostram sempre os resultados obtidos com e sem a seleção de características, a fim de permitir uma melhor comparação entre ambos.

### GLCM

A tabela 6.3 mostra o número de características selecionadas em cada classificador a partir das 28 características extraídas com GLCM originalmente e as taxas médias de reconhecimento obtidas em cada segmento.

Tabela 6.3: Desempenho individual dos classificadores utilizando seleção de características com extração global das características extraídas com GLCM

| | Segmento Inicial | Segmento Central | Segmento Final |
|---|---|---|---|
| **Com seleção** | | | |
| Número de características | 5 | 10 | 4 |
| Taxa de reconhecimento (%) | 48,56 | 54,78 | 37,89 |
| **Sem seleção** | | | |
| Número de características | 28 | 28 | 28 |
| Taxa de reconhecimento (%) | 50,33 | 51,33 | 38,00 |

O resultado final obtido, em termos de taxa de reconhecimento (%), após a fusão por quatro diferentes regras encontra-se descrito na tabela 6.4.

Tabela 6.4: Taxas de reconhecimento (%) com e sem seleção de características extraídas com GLCM

| | Regra do Máximo | Regra do Mínimo | Regra do Produto | Regra da Soma |
|---|---|---|---|---|
| Com seleção | 54,67 | 56,44 | **58,89** | 57,78 |
| Sem seleção* | 50,11 | 54,44 | 57,44 | 56,56 |

* resultados apresentados na tabela 5.3

Com a seleção de características, a melhor taxa de reconhecimento alcançada é ligeiramente superior à melhor taxa alcançada sem a seleção de características. Por outro lado, cabe ressaltar que houve uma significativa redução na quantidade de características utilizadas quando considerada a seleção.

### Filtros de Gabor

Detalhes acerca dos resultados obtidos individualmente por cada classificador, com e sem seleção de características, a partir das 120 características originalmente criadas com filtros de Gabor estão descritos na tabela 6.5.

Tabela 6.5: Desempenho individual dos classificadores utilizando seleção de características com extração global das características extraídas com filtros de Gabor

| | Segmento Inicial | Segmento Central | Segmento Final |
|---|---|---|---|
| **Com seleção** | | | |
| Número de características | 24 | 33 | 22 |
| Taxa de reconhecimento (%) | 47,44 | 48,00 | 45,89 |
| **Sem seleção** | | | |
| Número de características | 120 | 120 | 120 |
| Taxa de reconhecimento (%) | 48,78 | 50,44 | 49,67 |

A tabela 6.6 permite comparar os desempenhos alcançados com e sem a seleção de características extraídas com filtros de Gabor após a fusão das saídas dos classificadores.

Tabela 6.6: Taxas de reconhecimento (%) com e sem seleção de características extraídas com filtros de Gabor

| | Regra do Máximo | Regra do Mínimo | Regra do Produto | Regra da Soma |
|---|---|---|---|---|
| Com seleção | 53,56 | 58,00 | 58,56 | 57,11 |
| Sem seleção* | 55,89 | 56,67 | **59,78** | 58,78 |

* resultados apresentados na tabela 5.6

Embora tenha ocorrido mais uma vez uma significativa redução na quantidade de características utilizadas quando aplicada a seleção, a melhor taxa de reconhecimento obtida ainda é um pouco inferior à melhor taxa obtida sem o uso da seleção de características.

**LBP**

Detalhes acerca dos resultados obtidos individualmente por cada classificador, com e sem seleção de características, a partir das 59 características originalmente criadas com LBP estão descritos na tabela 6.7.

Tabela 6.7: Desempenho individual dos classificadores utilizando seleção de características com extração global das características extraídas com LBP

| | Segmento Inicial | Segmento Central | Segmento Final |
|---|---|---|---|
| **Com seleção** | | | |
| Número de características | 12 | 23 | 10 |
| Taxa de reconhecimento | 61,11 | 70,00 | 59,78 |
| **Sem seleção** | | | |
| Número de características | 59 | 59 | 59 |
| Taxa de reconhecimento | 67,44 | 71,22 | 69,78 |

A tabela 6.8 permite comparar os desempenhos alcançados após a fusão das saídas dos classificadores com e sem a seleção de características extraídas utilizando LBP.

Tabela 6.8: Taxas de reconhecimento (%) com e sem seleção de características extraídas com LBP

| | Regra do Máximo | Regra do Mínimo | Regra do Produto | Regra da Soma |
|---|---|---|---|---|
| Com seleção | 72,22 | 72,67 | 76,00 | 74,11 |
| Sem seleção* | 76,78 | 77,11 | 78,67 | **79,00** |

* resultados apresentados na tabela 5.8

Neste caso, o uso da seleção de características provocou uma queda nas taxas de reconhecimento mais importante, embora o número de características tenha reduzido significativamente mais uma vez.

### LPQ

Os desempenhos individuais de cada classificador, com e sem seleção de características, a partir das 256 características originalmente criadas com LPQ estão descritos na tabela 6.9.

Tabela 6.9: Desempenho individual dos classificadores utilizando seleção de características com extração global das características extraídas com LPQ

| | Segmento Inicial | Segmento Central | Segmento Final |
|---|---|---|---|
| **Com seleção** | | | |
| Número de características | 12 | 17 | 15 |
| Taxa de reconhecimento (%) | 58,22 | 66,00 | 53,89 |
| **Sem seleção** | | | |
| Número de características | 256 | 256 | 256 |
| Taxa de reconhecimento (%) | 68,11 | 72,22 | 70,67 |

A tabela 6.10 permite comparar os desempenhos alcançados com e sem a seleção de características extraídas com LPQ após a fusão das saídas dos classificadores.

Tabela 6.10: Taxas de reconhecimento (%) com e sem seleção de características extraídas com LPQ

| | Regra do Máximo | Regra do Mínimo | Regra do Produto | Regra da Soma |
|---|---|---|---|---|
| Com seleção | 66,56 | 68,22 | 68,22 | 68,22 |
| Sem seleção* | 73,67 | 74,22 | 77,11 | **77,67** |

* resultados apresentados na seção 5.5

Este é o caso em que a queda nas taxas de desempenho foi a mais acentuada quando empregada a seleção de características, de forma que o uso da mesma não se justifica.

### Todas as características

A tabela 6.11 mostra os desempenhos individuais dos classificadores com e sem o uso de seleção de características quando são concatenadas as características extraídas com GLCM, filtros de Gabor, LBP e LPQ, formando um vetor composto originalmente por 463 caracteristicas.

Tabela 6.11: Seleção de características com extração global utilizando características extraídas com GLCM, filtros de Gabor, LBP e LPQ

| | Segmento Inicial | Segmento Central | Segmento Final |
|---|---|---|---|
| **Com seleção** | | | |
| Número de características | 20 | 30 | 43 |
| Taxa de reconhecimento (%) | 66,00 | 69,33 | 66,44 |
| **Sem seleção** | | | |
| Número de características | 463 | 463 | 463 |
| Taxa de reconhecimento (%) | 66,67 | 74,11 | 70,89 |

A tabela 6.12 permite comparar os desempenhos alcançados com e sem a seleção de

características utilizando todos os descritores de textura investigados neste trabalho, após a fusão das saídas dos classificadores.

Tabela 6.12: Taxas de reconhecimento (%) com e sem seleção de características extraídas com GLCM, filtros de Gabor, LBP e LPQ

| | Regra do Máximo | Regra do Mínimo | Regra do Produto | Regra da Soma |
|---|---|---|---|---|
| Com seleção | 75,44 | 74,78 | 77,11 | 77,33 |
| Sem seleção* | 78,56 | 77,33 | 78,67 | **79,22** |

* resultados apresentados na tabela 5.19

Mas uma vez, houve uma discreta queda nas taxas de reconhecimento quando utilizada a seleção de características.

### 6.2.2 Seleção de características com zoneamento linear

Esta seção apresenta os resultados alcançados utilizando zoneamento das imagens quando da extração de características. Foi empregado o zoneamento linear, no qual cinco zonas foram criadas, tal como descrito no capítulo 5 e utilizado em vários experimentos com zoneamento linear descritos ao longo deste trabalho. Neste caso, 15 classificadores são criados, já que são formadas cinco zonas para a imagem gerada a partir de cada um dos três segmentos extraídos da música. Foi executada uma seleção de característcias específica para cada um dos 15 classificadores.

Os experimentos foram realizados com filtros de Gabor, LBP e, adicionalmente, com vetores formados com todas as características de textura exploradas neste trabalho concatenadas. Com base nas características selecionadas para cada classificador, realizou-se a classificação utilizando-se os mesmos folds utilizados nas tarefas de reconhecimento descritas ao longo deste trabalho. As tabelas descritas nesta seção mostram sempre os resultados obtidos com e sem a seleção de características, a fim de permitir uma melhor comparação entre ambos.

**Filtros de Gabor**

A tabela 6.13 mostra o número de características selecionadas em cada um dos 15 classificadores a partir das 120 características extraídas originalmente com filtros de Gabor e as taxas médias de reconhecimento obtidas em cada segmento.

O resultado final obtido, em termos de taxa de reconhecimento (%), após a fusão por quatro diferentes regras encontra-se descrito na tabela 6.14.

Assim como ocorrido na extração global, o uso da seleção de características provocou um decréscimo significativo do número de característica e uma redução bastante discreta na melhor taxa de reconhecimento obtida.

Tabela 6.13: Taxa de reconhecimento(%)/número de características com e sem seleção de características com zoneamento linear utilizando características extraídas com filtros de Gabor

| **Com seleção de características** | | | |
|---|---|---|---|
| **Banda de frequência** | **Segmento inicial** | **Segmento central** | **Segmento final** |
| Banda 5 (6800 a 8500 Hz) | 46,89/7 | 48,55/5 | 46,89/9 |
| Banda 4 (5100 a 6800 Hz) | 48,89/11 | 53,00/15 | 45,67/6 |
| Banda 3 (3400 a 5100 Hz) | 51,78/22 | 52,22/18 | 49,00/12 |
| Banda 2 (1700 a 3400 Hz) | 49,00/29 | 51,67/30 | 48,89/27 |
| Banda 1 (0 a 1700 Hz) | 47,89/27 | 46,44/16 | 47,00/20 |
| **Sem seleção de características** | | | |
| **Banda de frequência** | **Segmento inicial** | **Segmento central** | **Segmento final** |
| Banda 5 (6800 a 8500 Hz) | 48,67/120 | 51,33/120 | 48,22/120 |
| Banda 4 (5100 a 6800 Hz) | 50,33/120 | 51,44/120 | 48,78/120 |
| Banda 3 (3400 a 5100 Hz) | 49,78/120 | 53,00/120 | 51,89/120 |
| Banda 2 (1700 a 3400 Hz) | 49,44/120 | 52,56/120 | 49,00/120 |
| Banda 1 (0 a 1700 Hz) | 50,00/120 | 52,67/120 | 53,78/120 |

Tabela 6.14: Taxas de reconhecimento (%) com e sem seleção de características extraídas com filtros de Gabor

| | Regra do Máximo | Regra do Mínimo | Regra do Produto | Regra da Soma |
|---|---|---|---|---|
| Com seleção | 60,56 | 67,11 | 73,78 | 72,44 |
| Sem seleção* | 66,22 | 69,67 | **74,67** | 74,11 |

* resultados apresentados na tabela 5.6

### LBP

A tabela 6.15 mostra o número de características selecionadas em cada um dos 15 classificadores a partir das 59 características extraídas originalmente com LBP e as taxas médias de reconhecimento obtidas em cada segmento.

Tabela 6.15: Taxa de reconhecimento(%)/número de características com e sem seleção de características com zoneamento linear utilizando características extraídas com LBP

| **Com seleção de características** | | | |
|---|---|---|---|
| **Banda de frequência** | **Segmento inicial** | **Segmento central** | **Segmento final** |
| Banda 5 (6800 a 8500 Hz) | 53,33/9 | 57,11/13 | 49,22/11 |
| Banda 4 (5100 a 6800 Hz) | 56,33/11 | 59,67/22 | 49,67/10 |
| Banda 3 (3400 a 5100 Hz) | 57,78/24 | 58,00/19 | 55,67/13 |
| Banda 2 (1700 a 3400 Hz) | 55,11/17 | 55,56/16 | 54,11/31 |
| Banda 1 (0 a 1700 Hz) | 55,11/23 | 59,22/29 | 60,89/36 |
| **Sem seleção de características** | | | |
| **Banda de frequência** | **Segmento inicial** | **Segmento central** | **Segmento final** |
| Banda 5 (6800 a 8500 Hz) | 55,56/59 | 62,33/59 | 51,11/59 |
| Banda 4 (5100 a 6800 Hz) | 58,22/59 | 62,00/59 | 56,44/59 |
| Banda 3 (3400 a 5100 Hz) | 57,89/59 | 61,11/59 | 59,00/59 |
| Banda 2 (1700 a 3400 Hz) | 58,56/59 | 60,56/59 | 57,56/59 |
| Banda 1 (0 a 1700 Hz) | 57,00/59 | 62,11/59 | 62,00/59 |

O resultado final obtido, em termos de taxa de reconhecimento (%), após a fusão por quatro diferentes regras encontra-se descrito na tabela 6.16.

Tabela 6.16: Taxas de reconhecimento (%) com e sem seleção de características extraídas com LBP

| | Regra do Máximo | Regra do Mínimo | Regra do Produto | Regra da Soma |
|---|---|---|---|---|
| Com seleção | 72,11 | 74,33 | 77,44 | 78,56 |
| Sem seleção | 73,56 | 75,22 | **79,44** | 79,22 |

* resultados apresentados na tabela 5.8

Assim como na extração global, o uso de seleção de caracterísitcas utilizando LBP provocou queda na taxa de reconhecimento. Entretanto, neste caso a queda é menor.

**Todas as características**

A tabela 6.17 mostra o número de características selecionadas em cada um dos 15 classificadores a partir das 463 características originalmente extraídas com GLCM, filtros de Gabor, LBP e LPQ e as taxas médias de reconhecimento obtidas em cada segmento.

Tabela 6.17: Taxa de reconhecimento(%)/número de características com e sem seleção de características com zoneamento linear utilizando características extraídas com GLCM, filtros de Gabor, LBP e LPQ

| **Com seleção de características** | | | |
|---|---|---|---|
| **Banda de frequência** | **Segmento inicial** | **Segmento central** | **Segmento final** |
| Banda 5 (6800 a 8500 Hz) | 56,22/35 | 60,56/36 | 55,78/15 |
| Banda 4 (5100 a 6800 Hz) | 59,67/31 | 59,22/29 | 53,67/9 |
| Banda 3 (3400 a 5100 Hz) | 59,11/86 | 61,33/77 | 57,11/31 |
| Banda 2 (1700 a 3400 Hz) | 58,67/22 | 63,44/78 | 54,67/10 |
| Banda 1 (0 a 1700 Hz) | 57,33/30 | 61,00/34 | 63,33/63 |
| **Sem seleção de características** | | | |
| **Banda de frequência** | **Segmento inicial** | **Segmento central** | **Segmento final** |
| Banda 5 (6800 a 8500 Hz) | 59,56/463 | 65,89/463 | 60,78/463 |
| Banda 4 (5100 a 6800 Hz) | 61,33/463 | 64,11/463 | 62,56/463 |
| Banda 3 (3400 a 5100 Hz) | 63,44/463 | 65,22/463 | 62,67/463 |
| Banda 2 (1700 a 3400 Hz) | 63,11/463 | 63,00/463 | 65,67/463 |
| Banda 1 (0 a 1700 Hz) | 63,11/463 | 67,33/463 | 66,11/463 |

O resultado final obtido, em termos de taxa de reconhecimento (%), após a fusão por quatro diferentes regras encontra-se descrito na tabela 6.18.

Tabela 6.18: Taxas de reconhecimento (%) com e sem seleção de características extraídas com GLCM, filtros de Gabor, LBP e LPQ

| | Regra do Máximo | Regra do Mínimo | Regra do Produto | Regra da Soma |
|---|---|---|---|---|
| Com seleção | 71,67 | 75,33 | 78,11 | 77,67 |
| Sem seleção | 77,67 | 78,11 | 81,89 | **82,44** |

* resultados apresentados na seção 5.7

As taxas de reconhecimento se elevaram em raras situações para alguns classificadores especificamente criados para algumas zonas da imagem. As taxas de reconhecimento obtidas com seleção de características foram no quase sempre inferiores às obtidas sem

o uso da seleção. Entretanto, observou-se uma significativa redução na quantidade de características utilizadas quando aplicada a seleção.

## 6.3 Teste estatístico

O teste de Friedman com o procedimento sequencial de Holm também foi aplicado aos resultados descritos neste capítulo a fim de verificar a ocorrência de diferenças estatisticamente significativas entre os resultados. Neste caso, as comparações foram feitas entre pares de classificadores a fim de verificar diferença estatística entre os resultados obtidos com e sem o uso de KNORA e seleção de características com Algoritmo Genético.

Ao final, constatou-se o que segue:

- Não há diferença estatisticamente significativa entre os resultados obtidos com e sem o uso de KNORA dividindo-se as imagens pela escala Mel;
- Não há diferença estatisticamente significativa entre os resultados obtidos com e sem o uso de KNORA dividindo-se as imagens em 10 zonas lineares;
- Não há diferença estatisticamente significativa entre os resultados obtidos com e sem o uso de AG para a seleção de características dividindo-se as imagens em cinco zonas lineares e utilizando características extraídas com filtros de Gabor;
- Não há diferença estatisticamente significativa entre os resultados obtidos com e sem o uso de AG para a seleção de características dividindo-se as imagens em cinco zonas lineares e utilizando características extraídas com LBP;
- Não há diferença estatisticamente significativa entre os resultados obtidos com e sem o uso de AG para a seleção de características dividindo-se as imagens em cinco zonas lineares e utilizando vetores com as características extraídas com todos os descritores de textura investigados concatenadas;
- Não há diferença estatisticamente significativa entre os resultados obtidos com e sem o uso de AG para a seleção de características, com extração global de características e utilizando GLCM;
- Não há diferença estatisticamente significativa entre os resultados obtidos com e sem o uso de AG para a seleção de características, com extração global de características e utilizando filtros de Gabor;
- Há diferença estatisticamente significativa entre os resultados obtidos com e sem o uso de AG para a seleção de características, com extração global de características e utilizando LBP;

- Há diferença estatisticamente significativa entre os resultados obtidos com e sem o uso de AG para a seleção de características, com extração global de características e utilizando LPQ;

- Há diferença estatisticamente significativa entre os resultados obtidos com e sem o uso de AG para a seleção de características, com extração global de características e utilizando vetores com as características extraídas com todos os descritores de textura investigados concatenadas;

## 6.4 Conclusão

Os resultados obtidos nos experimentos descritos neste capítulo não alcançaram as expectativas que se tinha inicialmente.A melhor taxa de acerto considerando todos os resultados descritos neste trabalho, de 83%, foi obtida em experimentos nos quais foi utilizado o KNORA. Embora o desempenho seja o melhor, é preciso lembrar que, além do custo introduzido naturalmente pelo uso do KNORA, esta taxa foi obtida quando se aplicou o zoneamento segundo a escala de Mel, que traz a desvantagem de produzir um grande número de classificadores. Estes custos adicionais se justificariam se fosse obtida uma taxa de reconhecimento significativamente superior às melhores obtidas nos experimentos, descritos no capítulo 5.

Com relação ao uso de algoritmos genéticos, a situação é de certa forma invertida. Por um lado, as taxas de reconhecimento obtidas foram, no melhor caso, um pouco inferiores as obtidas sem a seleção de características. Por outro lado, percebe-se uma redução considerável no número médio de características utilizadas nos classificadores, o que pode se traduzir em uma redução do custo geral do processo de classificação.

Embora não se tenha alcançado as expectativas iniciais com as abordagens utilizadas nos experimentos descritos neste capítulo, pode-se dizer que as mesmas ainda têm potencial para serem exploradas e produzirem resultados satisfatórios. Em uma situação em que haja maior liberdade para a distribuição dos títulos musicais entre os conjuntos utilizados para validação, treinamento e teste, KNORA poderia obter melhor desempenho com o uso de um outro conjunto de validação. Em relação a Algoritmos Genéticos, a sua natureza não determinística poderia levar a situações em que se pudesse conseguir uma melhor aproximação das expectativas iniciais.

# CAPÍTULO 7

# CONCLUSÃO

Este trabalho se desenvolveu em torno da exploração de imagens de espectrograma com propósito voltado a classificação de gêneros musicais. O atributo visual principal e que pode ser imediatamente percebido ao se observar uma imagem de espectrograma é a textura. Com base neste fato, foram avaliados diferentes descritores de textura, das diferentes abordagens estabelecidas por autores bastante reconhecidos na literatura de processamento de imagens, para representar o conteúdo das imagens nos processos de classificação.

Os descritores LBP e LPQ mostraram os melhores desempenhos individuais, e podem ser vistos como soluções interessantes na medida em que parecem capturar dimensões musicais importantes para a discriminação de gêneros.

Outra importante constatação refere-se ao fato de que a preservação de alguma informação acerca da localização espacial das características extraídas, através de zoneamento das imagens, é uma estratégia que contribui em muitas situações com a melhoria do desempenho geral do sistema.

O uso de múltiplos classificadores também favorece a obtenção de bons resultados. Neste caso, uma solução natural foi estabelecida com a criação de um classificador para cada zona produzida na imagem ou mesmo criando um classificador para cada segmento extraído da música quando se utiliza a extração global de características.

O uso de segmentos tomados de diferentes partes do sinal das músicas também se mostrou adequado para a solução do problema, uma vez que os resultados obtidos pelos segmentos isoladamente foram sempre inferiores aos resultados obtidos pela fusão das saídas dos classificadores criados para diferentes segmentos. É importante observar ainda que o uso de segmentos reduz a quantidade de sinal a ser processada e diminui o potencial danoso de se utilizar na classificação um único trecho da música que seja acidentalmente mais parecido com um gênero diferente daquele ao qual ela realmente pertence.

Outras técnicas complementares, conhecidas e empregadas com sucesso em algumas problemas de classificação também foram tentadas. Neste contexto é válido mencionar o uso do KNORA para a seleção dinâmica de um agrupamento de classificadores, e Algoritmos Genéticos para a seleção de características. Em alguns casos, KNORA até proporcionou uma discreta elevação nas taxas de reconhecimento, mas o método introduz um importante custo adicional ao processo de classificação, que não pode ser desprezado. Julga-se oportuna a realização de novos experimentos para que se chegue a conclusões mais definitivas acerca da viabilidade do uso do método. Já o uso de algoritmos genéticos

para a seleção de características reduz, em muitos casos, consideravelmente o tamanho do vetor de características utilizado na classificação. Entretanto, os resultados finais são, em geral, inferiores aos obtidos quando a seleção de características não é empregada.

As classes estabelecidas em tarefas de classificação de gêneros musicais são muitas e podem variar muito de acordo com o contexto e tipo de usuários envolvidos. Em muitos casos, surgem em pouco espaço de tempo novas classes a partir de classes pré-existentes ou não. O método aqui proposto é baseado em aprendizagem supervisionada, assim como a maioria das propostas apresentadas na literatura. Este fato, impõe uma limitação para o uso destas abordagens em cenários bastante dinâmicos, nos quais o conjunto de classes consideradas sofre mudanças constantemente.

A metodologia aqui proposta foi testada sobre a base LMD com o uso de *"artist filter"*. Esta estratégia diminue o risco de que se crie um classificador especialista em discriminar artistas ao invés de gêneros. Sobre esta base, foram alcançados resultados melhores do que os melhores já apresentados na literatura. Com o operador de textura LBP e utilizando zoneamento segundo a escala Mel, a melhor taxa de reconhecimento obtida foi de 82,33%. Utilizando KNORA, esta taxa ainda subiu, no melhor caso, para 83%. A melhor taxa até então obtida sobre esta base utilizando *"artist filter"* era de 79,86%, apresentada em [79].

Além da LMD, o classificador também foi testado sobre a base *ISMIR 2004* e apresentou resultados comparáveis a outros bons resultados apresentados em outros trabalhos descritos na literatura. O melhor desempenho obtido com esta base, com extração global de características utilizando LBP, apresentou 80,65% de taxa de reconhecimento. Este dado é importante para atestar a capacidade de generalização da metodologia para aplicação em outras bases, com estilos musicais diversos.

A avaliação do tempo consumido para realizar a classificação em algumas situações nas quais se obteve bom desempenho mostrou que, embora haja algum aumento no tempo consumido nas situações em que se utiliza uma maior quantidade de classificadores, a aplicação do método quando se espera resposta em tempo real não fica inviabilizada.

Embora os resultados evidenciem a viabilidade e eficiência da solução aqui investigada, o principal ponto negativo deste trabalho reside na ausência de uma explicação científica para a forma como estas dimensões musicais são capturadas pelas características apresentadas. Adicionalmente, pode-se mencionar como ponto negativo o fato de que a proposta aqui apresentada impõe uma etapa adiconal para a criação das imagens de espectrograma quando comparadas às características tradicionais e, em alguns casos, impõe a necessidade de criação de uma grande quantidade de classificadores para que bons resultados sejam alcançados.

Para finalizar, cabe ainda observar que a hipótese lançada neste trabalho, de que é possível representar uma música para o propósito de classificação de gêneros musicais através de características extraídas de imagem de espectrograma, é verdadeira e os melhores desempenhos obtidos com estas características e relatados neste trabalho ultrapassam as

taxas de acerto médias obtidas por humanos descritas na literatura.

## 7.1 Contribuições

As principais contribuições deste trabalho podem ser sintetizadas da seguinte forma:

- Apresentação de um novo formato de características descritoras de conteúdo de sinal de áudio em que a natureza original do sinal é abstraída e o mesmo é mapeado para o domínio visual (espectrograma);
- Demonstração de que as características propostas podem ser utilizadas com eficiência igual ou superior a de outras características criadas para uso em tarefas de classificação de gêneros musicais;
- Demonstração de que a preservação de informações locais das imagens de espectrograma com a criação de múltiplos classificadores pode contribuir com a obtenção de bons resultados;
- Identificação de taxas de reconhecimento bastante altas no limite superior entre múltiplos classificadores criados, fato que pode viabilizar o alcance de taxas de reconhecimento ainda maiores com o uso de técnicas adequadas para a seleção de classificadores;
- Verificação da possibilidade de redução da dimensionalidade de vetores de características propostos, com o uso de seleção de características, mantendo taxas de reconhecimento semelhantes as obtidas sem a seleção de caracteristicas;

## 7.2 Trabalhos Futuros

Durante o desenvolvimento desta tese, não foi possivel realizar algumas investigações com o objetivo de buscar resposta para algumas questões interessantes. A seguir, são descritas algumas delas:

- Identificar as dimensões musicais capturadas pelas características de textura que apresentaram bom desempenho, notadamente LBP e LPQ;
- Experimentar algum esquema de classificação que permita o uso de *tags* juntamente com as características aqui propostas;
- Experimentar o método KNORA com outros conjuntos de validação ou outros esquemas de seleção dinâmica de classificadores, diferentes do KNORA, a fim de tentar tirar melhor proveito da potencial alta taxa de reconhecimento identificada no limite superior entre os classificadores criados em várias situações.

# BIBLIOGRAFIA

[1] J.J. Aucouturier e F. Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.

[2] U. Bagci e E. Erzin. Automatic classification of musical genres using inter-genre similarity. *IEEE Signal Processing Letters*, 14(8):521–524, 2007.

[3] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, e B. Kégl. Aggregate features and adaboost for music classification. *Machine Learning*, 65(2):473–484, 2006.

[4] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, e N. Wack. ISMIR 2004 audio description contest. Relatório técnico, Music Technology Group, Barcelona, Spain, 2006.

[5] Chih-Chung Chang e Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[6] C.H.L. Costa, J.D. Valle Jr, e A.L. Koerich. Automatic classification of audio data. *IEEE International Conference on Systems, Man, and Cybernetics*, páginas 562–567, The Hague, Netherlands, 2004.

[7] Y. Costa, L. Oliveira, A. Koerich, e F. Gouyon. Classificação de gêneros musicais por texturas no espaço de frequência. *XXXVIII Seminário Integrado de Software e Hardware, Congresso da Sociedade Brasileira de Computação*, Natal, Brazil, 2011.

[8] Y. Costa, L. Oliveira, A. Koerich, e F. Gouyon. Music genre recognition using spectrograms. *International Conference on Systems, Signals and Image Processing*, Sarajevo, Bosnia and Herzegovina, 2011.

[9] Y. Costa, L. Oliveira, A. Koerich, e F. Gouyon. Comparing textural features for music genre classification. *WCCI 2012 IEEE World Congress on Computational Intelligence*, páginas 1867–1872, Brisbane, Australia, 2012.

[10] Y. Costa, L. Oliveira, A. Koerich, e F. Gouyon. Music genre recognition based on visual features with dynamic ensemble of classifiers selection. *International Conference on Systems, Signals and Image Processing*, Bucharest, Romenia, 2013.

[11] Y. Costa, L. Oliveira, A. Koerich, e F. Gouyon. Music genre recognition using gabor filters and lpq texture descriptors. *Iberoamerican Congress on Pattern Recognition*, Havana, Cuba, 2013.

[12] Y. Costa, L. Oliveira, A. Koerich, F. Gouyon, e J. Martins. Music genre classification using lbp textural features. *Signal Processing*, 92(11):2723–2737, 2012.

[13] R. Dannenberg, J. Foote, G. Tzanetakis, e C. Weare. Panel: New directions in music information retrieval. *International Computer Music Conference*, La Habana, Cuba, 2001.

[14] H. Deshpande, R. Singh, e U. Nam. Classification of music signals in the visual domain. *COST-G6 Conference on Digital Audio Effects*, Limerick, Ireland, 2001.

[15] T. Dietterich. Ensemble methods in machine learning. *Multiple classifier systems*, páginas 1–15, 2000.

[16] R.O. Duda, P.E. Hart, e D.G. Stork. *Pattern classification*. John Willey & Sons, 2001.

[17] H. Ezzaidi e J. Rouat. Automatic musical genre classification using divergence and average information measures. *World Academy of Science, Engineering and Technology*, 15, 2006.

[18] F. Fabbri. Browsing music spaces: Categories and the musical mind, 1999.

[19] Olivier D Faugeras e William K Pratt. Decorrelation methods of texture feature extraction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (4):323–332, 1980.

[20] R. Fiebrink e I. Fujinaga. Feature selection pitfalls and music classification. *International Conference on Music Information Retrieval*, páginas 340–341, Victoria, Canada, 2006.

[21] A. Flexer. A closer look on artist filters for musical genre classification. *International Conference on Music Information Retrieval*, 19(122):341–344, 2007.

[22] A. Flexer, E. Pampalk, e G. Widmer. Hidden markov models for spectral similarity of songs. *International Conference on Digital Audio Effects*, Madrid, Espanha, 2005.

[23] M. French e R. Handy. Spectrograms: turning signals into pictures. *Journal of Engineering Technology*, 24(1):32–35, 2007.

[24] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

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

[26] R.O. Gjerdingen e D. Perrott. Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research*, 37(2):93–100, 2008.

[27] D. E. Goldberg. *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley Professional, 1989.

[28] R. C. Gonzalez e R. E. Woods. *Digital Image Processing.* Pearson Prentice Hall, 2008.

[29] M. Goto, H. Hashiguchi, T. Nishimura, e R. Oka. Rwc music database: Music genre database and musical instrument sound database. *International Conference on Music Information Retrieval*, volume 3, páginas 229–230, Washington, USA, 2003.

[30] M. Grimaldi, P. Cunningham, e A. Kokaram. A wavelet packet representation of audio signals for music genre classification using different ensemble and feature selection techniques. *ACM SIGMM international workshop on Multimedia information retrieval*, páginas 102–108, Berkeley, USA, 2003.

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

[32] Robert M Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[33] John H Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and artificial intelligence.* MIT press, 1992.

[34] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, páginas 65–70, 1979.

[35] A. Holzapfel e Y. Stylianou. Musical genre classification using nonnegative matrix factorization-based features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):424–434, 2008.

[36] H. Homburg, I. Mierswa, B. Möller, K. Morik, e M. Wurst. A benchmark dataset for audio classification and clustering. *International Conference on Music Information Retrieval*, páginas 528–31, 2005.

[37] Chih-Wei Hsu e Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[38] X. Hu, J.S. Downie, K. West, e A. Ehmann. Mining music reviews: promising preliminary results. *International Conference on Music Information Retrieval*, páginas 536–539, London, UK, 2005.

[39] A.K. Jain, R.P.W. Duin, e J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):4–37, 2000.

[40] A.K. Jain e F. Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Pattern recognition*, 24(12):1167–1186, 1991.

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

[42] Joni-Kristian Kämäräinen. *Feature Extraction Using Gabor Filters*. Tese de Doutorado, Lappeenranta University of Technology, 2003.

[43] A.H.R. Ko, R. Sabourin, A.S. Britto, et al. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, 41(5):1718–1731, 2008.

[44] A.L. Koerich e C. Poitevin. Combination of Homogeneous Classifiers for Musical Genre Classification. *IEEE International Conference on Systems, Man, and Cybernetics*, páginas 554–559, Waikoloa, Hawaii, 2005.

[45] L.I. Kuncheva e C.J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

[46] T. Li e M. Ogihara. Music genre classification with taxonomy. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, Philadelphia, USA, 2005.

[47] T. Li, M. Ogihara, e Q. Li. A comparative study on content-based music genre classification. *26th annual international ACM SIGIR conference*, páginas 282–289, Toronto, Canada, 2003.

[48] T. Lidy e A. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. *International Conference on Music Information Retrieval*, páginas 34–41, London, UK, 2005.

[49] T. Lidy, A. Rauber, A. Pertusa, e J.M. Inesta. Improving genre classification by combination of audio and symbolic descriptors using a transcription system. *International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[50] T. Lidy, C.N. Silla Jr, O. Cornelis, F. Gouyon, A. Rauber, C.A.A. Kaestner, e A.L. Koerich. On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-western and ethnic music collections. *Signal Processing*, 90:1032–1048, 2010.

[51] S. Lippens, JP Martens, e T. De Mulder. A comparison of human and automatic musical genre classification. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, Montreal, Canada, 2004.

[52] M. Lopes, F. Gouyon, A. L. Koerich, e L. E. S. Oliveira. Selection of training instances for music genre classification. *International Conference on Pattern Recognition*, Istanbul, Turkey, 2010.

[53] C. Marques, I.R. Guiherme, R.Y.M. Nakamura, e J.P. Papa. New trends in musical genre classification using optimum-path forest. *International Conference on Music Information Retrieval*, Miami, USA, 2011.

[54] G. Marques, T. Langlois, F. Gouyon, M. Lopes, e M. Sordo. Short-term feature space and music genre classification. *Journal of New Music Research*, 40(2):127–137, 2011.

[55] G. Marques, M. Lopes, M. Sordo, T. Langlois, e F. Gouyon. Additional evidence that common low-level features of individual audio frames are not representative of music genre. *Sound and Music Computing Conference, Barcelona*, 2010.

[56] R. Mayer e A. Rauber. Musical genre classification by ensembles of audio and lyrics features. *International Conference on Music Information Retrieval*, páginas 675–680, Miami, USA, 2011.

[57] C. McKay, R. Fiebrink, D. Mcennis, B. Li, e I. Fujinaga. ACE: A framework for optimizing music classification. *International Conference on Music Information Retrieval*, páginas 42–9, London, UK, 2005.

[58] C. McKay e I. Fujinaga. Musical genre classification: Is it worth pursuing and how can it be improved. *International Conference on Music Information Retrieval*, páginas 101–6, Victoria, Canada, 2006.

[59] C. McKay, D. McEnnis, e I. Fujinaga. A large publicly accessible prototype audio database for music research. *International Conference on Music Information Retrieval*, páginas 160–3, Victoria, Canada, 2006.

[60] A. Meng, P. Ahrendt, e J. Larsen. Improving music genre classification by short time feature integration. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, Philadelphia, USA, 2005.

[61] Topi Mäenpää. *The Local Binary Pattern Approach to Texture Analysis - Extensions and Applications*. Tese de Doutorado, University of Oulo, 2003.

[62] Steven R Ness, Anthony Theocharis, George Tzanetakis, e Luis Gustavo Martins. Improving automatic music tag annotation using stacked generalization of probabilistic

svm outputs. *Proceedings of the 17th ACM international conference on Multimedia*, páginas 705–708, Beijing, China, 2009.

[63] T. Ojala, M. Pietikäinen, e D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[64] Timo Ojala, Matti Pietikäinen, e Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[65] Ville Ojansivu e Janne Heikkilä. Blur insensitive texture classification using local phase quantization. *Image and Signal Processing*, páginas 236–243, 2008.

[66] Luiz S Oliveira, Robert Sabourin, Flavio Bortolozzi, e Ching Y Suen. A methodology for feature selection using multiobjective genetic algorithms for handwritten digit string recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 17(06):903–929, 2003.

[67] N. Orio. Automatic identification of audio recordings based on statistical modeling. *Signal Processing*, 2009.

[68] E. Pampalk, A. Flexer, e G. Widmer. Improvements of audio-based music similarity and genre classification. *International Conference on Music Information Retrieval*, volume 5, London, UK, 2005.

[69] I. Panagakis, E. Benetos, e C. Kotropoulos. Music genre classification: A multilinear approach. *International Conference on Music Information Retrieval*, páginas 583–588, Philadelphia, USA, 2008.

[70] Y. Panagakis, C. Kotropoulos, e G. R. Arce. Music genre classification via sparse representations of auditory temporal modulations. *Proc. European Signal Processing Conference*, páginas 1–5, Glasgow, Scotland, 2009.

[71] Y. Panagakis, C. Kotropoulos, e G.R. Arce. Music genre classification using locality preserving non-negative tensor factorization and sparse representations. *International Conference on Music Information Retrieval*, páginas 249–254, Kobe, Japan, 2009.

[72] A. Paradzinets, H. Harb, e L. Chen. Multiexpert system for automatic music genre classification. *Teknik Rapor, Ecole Centrale de Lyon, Departement MathInfo*, 2009.

[73] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, e G. Widmer. On rhythm and general music similarity. *International Conference on Music Information Retrieval*, Kobe, Japan, 2009.

[74] D. Ruta e B. Gabrys. Classifier selection for majority voting. *Information fusion*, 6(1):63–81, 2005.

[75] E. M. Santos. *Static and Dynamic Overproduction and Selection of Classifier Ensembles with Genetic Algorithms*. Tese de Doutorado, Université du Québec, 2008.

[76] L. Sarmento, E.C. Oliveira, F. Gouyon, e B. G. Costa. Visualizing Networks of Music Artists with RAMA. 2009.

[77] N. Scaringella, G. Zoia, e D. Mlynek. Automatic genre classification of music content: a survey. *Signal Processing Magazine, IEEE*, 23(2):133–141, 2006.

[78] K. Seyerlehner e M. Schedl. Block-level audio feature for music genre classification. *Proc. of the 5th Annual Music Information Retrieval Evaluation eXchange (MIREX-09)*, 2009.

[79] K. Seyerlehner, M. Schedl, T. Pohle, e P. Knees. Using block-level features for genre classification, tag classification and music similarity estimation. *6th Annual Music Information Retrieval Evaluation eXchange (MIREX 2010)*, 2010.

[80] C.N. Silla Jr, C.A.A. Kaestner, e A.L. Koerich. Classificação de gêneros musicais utilizando vetores de característica híbridos. *13o Simpósio Brasileiro de Computação Musical (SBCM2011)*, páginas 32–44, Vitória, Brazil, 2011.

[81] C.N. Silla Jr, A.L. Koerich, e C.A.A. Kaestner. A machine learning approach to automatic music genre classification. *Journal of the Brazilian Computer Society*, 14:7–18, 2008.

[82] C.N. Silla Jr, A.L. Koerich, e C.A.A. Kaestner. The latin music database. *International Conference on Music Information Retrieval*, páginas 451–456, Philadelphia, USA, 2008.

[83] C.N. Silla Jr, A.L. Koerich, e C.A.A. Kaestner. A Feature Selection Approach for Automatic Music Genre Classification. *International Journal of Semantic Computing*, 3:1–26, 2009.

[84] Raúl E Sánchez-Yáñez, Evguenii V Kurmyshev, e Francisco J Cuevas. A framework for texture classification using the coordinated clusters representation. *Pattern Recognition Letters*, 24(1):21–31, 2003.

[85] M. Sonka, V. Hlavac, e R. Boyle. *Image Processing, Analysis, and Machine Vision*. PWS Publishing, 1999.

[86] H. Tamura, S. Mori, e T. Yamawaki. Textural Features Corresponding to Visual Perception. *IEEE Trans. Systems, Man, and Cybernetics*, 8(6):460–473, 1978.

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

[88] S. Umesh, L. Cohen, e D. Nelson. Fitting the mel scale. *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, páginas 217–220, Phoenix, USA, 1999.

[89] VN Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, USA, 1982.

[90] M.J. Wu, Z.S. Chen, J.S.R. Jang, J.M. Ren, Y.H. Li, e C.H. Lu. Combining visual and acoustic features for music genre classification. *International Conference on Machine Learning and Applications*, volume 2, páginas 124–129, Honolulu, Hawaii, 2011.

[91] Y. Yaslan e Z. Cataltepe. Audio music genre classification using different classifiers and feature selection methods. *International Conferenceon Pattern Recognition*, volume 2, páginas 573–576, Hong Kong, China, 2006.

[92] G. Yu e J.J. Slotine. Audio classification from time-frequency texture. páginas 1677–1680, Taipei, Taiwan, 2009.

[93] Jianke Zhu, Steven CH Hoi, Michael R Lyu, e Shuicheng Yan. Near-duplicate keyframe retrieval by nonrigid image matching. *ACM international conference on Multimedia*, páginas 41–50, Vancouver, Canada, 2008.

[94] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *Acoustical Society of America Journal*, 33:248, 1961.