UNIVERSIDADE DE CAXIAS DO SUL

LUIZ EDUARDO ROMANO

RECONHECIMENTO AUTOMÁTICO DE GÊNEROS MUSICAIS UTILIZANDO CLASSIFICADORES BASEADOS EM MÚLTIPLAS CARACTERÍSTICAS

CAXIAS DO SUL

2014

UNIVERSIDADE DE CAXIAS DO SUL

CENTRO DE COMPUTAÇÃO E TECNOLOGIA DA INFORMAÇÃO

LUIZ EDUARDO ROMANO

RECONHECIMENTO AUTOMÁTICO DE GÊNEROS MUSICAIS UTILIZANDO CLASSIFICADORES BASEADOS EM MÚLTIPLAS CARACTERÍSTICAS

Trabalho de Conclusão do Curso de Bacharelado em Ciência da Computação pela Universidade de Caxias do Sul.
Área de concentração: Computação.
Orientador: Prof. Dr. André Gustavo Adami.

CAXIAS DO SUL

2014

LUIZ EDUARDO ROMANO

RECONHECIMENTO AUTOMÁTICO DE GÊNEROS MUSICAIS UTILIZANDO CLASSIFICADORES BASEADOS EM MÚLTIPLAS CARACTERÍSTICAS

Trabalho de Conclusão do Curso de Bacharelado em Ciência da Computação pela Universidade de Caxias do Sul.
Área de concentração: Computação.
Orientador: Prof. Dr. André Gustavo Adami.

Aprovado em __/__/____

Banca Examinadora

_____
Prof. Dr. André Gustavo Adami
Universidade de Caxias do Sul – UCS

_____
Prof.ª Dr.ª Adriana Miorelli Adami
Universidade de Caxias do Sul – UCS

_____
Prof.ª Dr.ª Elisa Boff
Universidade de Caxias do Sul – UCS

AGRADECIMENTOS

Com a finalização deste trabalho de conclusão de curso, gostaria de prestar alguns agradecimentos:

- Ao meu orientador, André Gustavo Adami, pelas várias horas dedicadas no auxílio da elaboração deste trabalho. Agradeço pelos conselhos, dicas e correções que acresceram na qualidade deste trabalho.

- Aos meus pais, amigos e familiares, pela compreensão das longas horas tomadas para a realização deste trabalho.

- Aos que fazem música, uma das maiores artes da humanidade.

RESUMO

O reconhecimento automático de gêneros musicais é um importante problema de pesquisa que tem recebido muita atenção de pesquisadores e profissionais da música digital. Os benefícios deste reconhecimento podem ser aplicados em diversas situações, como na organização de bases de áudios digitais, na construção de novos mecanismos de buscas e recomendações de músicas, entre outros. Porém, este campo de estudo ainda se encontra em fase de evolução, e os modelos atuais propostos para a classificação de gêneros musicais estão longe do ideal em termos de desempenho e reconhecimento.

Este trabalho de conclusão propõe um método para reconhecimento automático de gêneros musicais baseado no conteúdo do áudio. Para isto, uma visão geral da área de reconhecimento de gêneros musicais é apresentada, incluindo processamento de sinal e reconhecimento de padrões. Esta revisão bibliográfica descreve principalmente as características de áudio mais utilizadas, seus significados e formas de cálculo. Ainda, são apresentados alguns classificadores e a forma de funcionamento de cada um deles.

O método proposto neste trabalho está focado na variação da arquitetura básica dos sistemas de reconhecimento de gêneros musicais, envolvendo técnicas como extração por segmentos, agrupamento de características, sistemas multiexpert e combinação de classificadores. Ainda, é analisado o desempenho de um classificador paramétrico e não paramétrico, e é implementada uma técnica de alteração de probabilidades baseado no desempenho dos classificadores em cada gênero.

Palavras-chaves: Reconhecimento de Gêneros. Reconhecimento de Padrões. Processamento de Sinal.

ABSTRACT

Automatic music genre recognition is an important research topic that has received much attention from researchers and digital musical professionals. The benefits of this recognition can be applied in various situations, such as in the organization of digital audio databases, in building new search engines and music recommendations, among others. However, this field of study is still in phase of evolution, and the current models proposed for music genre classification are far from ideal in terms of performance and recognition.

This work proposes a method to automatically recognize musical genres based in the audio content. For this, an overview of the music genre recognition's area is realized, including signal processing and pattern recognition. This literature review describes mainly the most used audio features, their meanings and ways to calculate them. Even, are presented some classifiers and the functional way of each one.

The method proposed in this work is focused on variations of the basic architecture of the music genre recognize systems, involving techniques like extraction per segment, features assembly, multiexpert systems and ensemble of classifiers. Even, it is analyzed the performance of a parametric and non-parametric classifier, and it is implemented a technique to alter probabilities based on the performance of the classifiers in each genre.

Keywords: Genre Recognition. Pattern Recognition. Signal Processing.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BPM	Batidas por minuto
DWT	Discrete Wavelet Transform
DFT	Discrete Fourier Transform
FT	Fourier Transform
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
Hz	Hertz
ISMIR	International Conference of Music Information Retrieval
KNN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
LOG	Logistic Regression
MFCC	Mel-Frequency Cepstral Coefficients
MIR	Music Information Retrieval
MLP	Multilayer Perceptron
MP3	Moving Picture Experts Group 1 (MPEG) Audio Layer 3
RBF	Radial Basis Function
SOM	Self-Organizing Map
SVM	Support Vector Machines
STFT	Short-Time Fourier Transform

SUMÁRIO

1. INTRODUÇÃO

Nos últimos anos, através da expansão da internet e do desenvolvimento da informação e tecnologias multimídia, a música digital cresceu consideravelmente. Este tipo de informação está se tornando cada vez mais disponível às pessoas a partir de diversas fontes de mídias (FU *et al.*, 2011). Esta expansão da música digital trouxe à tona a necessidade de desenvolvimento de ferramentas capazes de extrair informações relevantes do áudio para gerenciar estas coleções (PANAGAKIS *et al.*, 2008). Barreira (2010) argumenta que uma correta classificação de cada música pode ser a chave para manter uma base de dados bem organizada e estruturada. Muitas informações podem ser extraídas de uma música, inclusive o seu gênero, que é provavelmente o mais popular descritor de conteúdo de músicas (AUCOUTURIER e PACHET, 2005). Isto se deve principalmente ao fato de que o gênero permite agrupar músicas parecidas, e de que as pessoas só se interessam por certos tipos de músicas. Logo, se uma pessoa gosta de uma música classificada como rock, provavelmente gostará de outra música que também seja classificada como rock. Esta similaridade musical permite facilmente que as pessoas descubram quais músicas são do seu interesse (e quais não são). Logo, um sistema de classificação por gênero musical poderia permitir que as pessoas procurassem somente pelas músicas em que estão interessadas, de forma mais rápida e intuitiva.

Este trabalho procura identificar uma forma para que uma máquina possa fazer automaticamente a classificação de músicas em gêneros, baseado no conteúdo do áudio das canções. São buscadas maneiras de combinar características do áudio e classificadores, além de formas de estruturar esta classificação. O objetivo é identificar maneiras para que estas partes possam se complementar para atingir uma taxa de reconhecimento melhor.

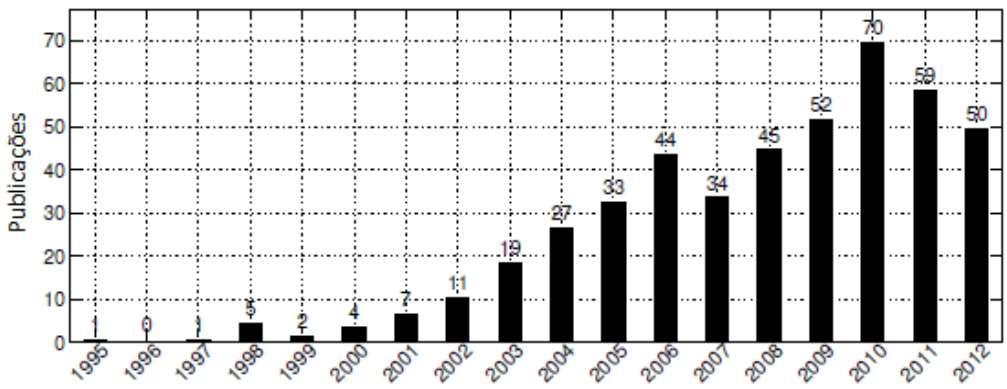
1.1 MOTIVAÇÃO

Atualmente, a organização e classificação de arquivos multimídia são realizadas através de informações textuais que são atribuídas a estes arquivos. Geralmente esta atribuição é feita por algum especialista na área ou até mesmo por

usuários. McKay e Fujinaga (2004) argumentam que esta metodologia, realizada de forma manual, é lenta e difícil de gerenciar. Especificamente para a classificação de músicas em gêneros musicais, esta forma de organização possui certos problemas. Benetos e Kotropoulos (2010) explicam que esta classificação é determinada principalmente pelo gosto de quem atribui o gênero, o que pode variar de pessoa a pessoa. Ainda, há a possibilidade de mais de um gênero musical ser associado a uma mesma música, tornando esta classificação totalmente imprecisa. Este método também apresenta outro problema, pois esta classificação em gêneros é muitas vezes realizada para o artista, e não para as músicas, isto é, se um artista é classificado em determinado gênero, todas as suas músicas são automaticamente classificadas como pertencentes àquele gênero (BARREIRA, 2010). Esta não é a melhor forma, pois os artistas podem não seguir a mesma referência musical em toda a sua carreira, e seus álbuns podem possuir uma mistura de vários gêneros. Desta maneira, fica difícil perceber quais músicas destes artistas pertencem a cada gênero.

Por estas condições, classificar bases de dados se tornou um problema real. Termens (2009) aponta a necessidade de criação de novas metodologias para organizar, estruturar, descobrir, recomendar e classificar músicas. Logo, ferramentas deste tipo tem aumentado o interesse de produtores e usuários de música digital (HARTMANN, 2011). Panagakis *et al.* (2008) confirmam que a classificação de músicas em gêneros distintos têm se tornado um atrativo tópico em pesquisas na área de recuperação de informações musicais. A Figura 1.1 ressalta este recente crescimento das pesquisas nesta área, mostrando a necessidade de novas soluções.

Figura 1.1 - Quantidade anual de publicações na área de reconhecimento de gêneros musicais (adaptado de Sturm, 2012).



Fonte: Autor.

Várias são as vantagens do uso de um classificador automático de gêneros musicais. Na prática, esta classificação é vantajosa na organização de grandes coleções de músicas, como em bibliotecas, na internet, em rádios, entre outros (HARTMANN, 2011). Este tipo de ferramenta pode trazer outros benefícios, como o auxílio na criação de bases de dados de músicas, criação de listas de reprodução por gênero, além de fornecer novos mecanismos de buscas e recomendações de músicas (KOSINA, 2002). Ainda, a classificação de gêneros musicais possui importantes aplicações nas áreas de produção de mídia, estações de rádios, gerenciamento de arquivos e entretenimento (BAGCI e ERZIN, 2007). Outra aplicação interessante seria a equalização automática de músicas. Atualmente, muitos tocadores de áudio possuem parâmetros de equalização pré-definidos, agrupados por gêneros musicais. Um reconhecimento automático de gêneros poderia permitir uma equalização automática em tempo real baseada no gênero da música, atribuindo maior qualidade na reprodução do áudio. Kosina (2002) complementa que, além das vantagens práticas, o reconhecimento de gêneros musicais pode auxiliar também no entendimento da percepção humana na música, entendimento este que ainda é muito limitado. Os resultados de uma pesquisa em processamento de músicas podem aumentar o nosso conhecimento sobre os princípios de nossas próprias percepções.

1.2 ESTRUTURA DO TRABALHO

Este trabalho está dividido em seis capítulos. O Capítulo 2 aborda a história e os princípios destes estudos, além dos conceitos principais de um sistema de reconhecimento de gêneros musicais. O Capítulo 3 trata das características de áudio mais utilizadas e seus objetivos, e o Capítulo 4 apresenta os métodos de classificação mais utilizados. O Capítulo 5 apresenta a proposta de um sistema de classificação de gêneros musicais, detalhando informações como a base de dados utilizada e a arquitetura do sistema. Por fim, o Capítulo 6 apresenta as considerações finais deste trabalho.

2. RECONHECIMENTO DE GÊNEROS MUSICAIS

Este capítulo aborda os princípios básicos para a construção de um sistema de reconhecimento de gêneros musicais e áreas relacionadas com a classificação de músicas em gêneros. A Seção 2.1 trata da definição de gênero musical e a diferença entre gênero e estilo. A Seção 2.2 apresenta outras áreas de estudo relacionadas com o foco deste trabalho, como pesquisas envolvendo recuperação de informações musicais. A Seção 2.3 mostra as definições e principais conceitos da classificação automática de gêneros musicais, e a Seção 2.4 apresenta trabalhos correlatos.

2.1 GÊNERO MUSICAL

Antes de aprofundar o estudo no reconhecimento de gêneros musicais, é interessante entender a distinção entre "gênero musical" e "estilo musical", pois são termos geralmente confundidos e que muitas vezes se misturam. Segundo Moore (2001), estilo se refere às maneiras de articulação de gestos musicais, enquanto que gênero se refere à identidade e o contexto destes gestos. Isto é, o gênero identifica a intenção de criar um tipo particular de experiência musical (um "o quê"). Já o estilo identifica os meios para esta experiência ser alcançada (um "como"). Por isto podemos perceber que existem vários artistas pertencentes a um mesmo gênero musical, mas geralmente cada um possui um estilo musical próprio para mostrar este gênero.

Fabbri (1981) define gênero musical como um conjunto de eventos musicais (possíveis ou reais) que tem seu curso governado por um conjunto de regras aceitas pela sociedade. Este conjunto de eventos musicais pode ser qualquer atividade que envolva som, e o conjunto de regras pode agrupar regras formais, técnicas, semióticas, sociais e ideológicas, entre outras. Simplificando um pouco a ideia e restringindo este significado apenas para a produção de um som, podemos dizer que um gênero musical é o agrupamento de sons que possuem elementos musicais em comum, como estrutura, timbre, ritmo, instrumentalização, entre outros. Um dos objetivos principais da classificação de áudio por gêneros é justamente descobrir os elementos musicais que agrupam músicas similares e as distinguem de outros gêneros.

2.2 RECUPERAÇÃO DE INFORMAÇÕES MUSICAIS

O reconhecimento de gêneros musicais faz parte da tarefa da Recuperação de Informações Musicais (*Music Information Retrieval* - MIR). A MIR é uma emergente área de estudo direcionada para atender as necessidades dos usuários de músicas, cobrindo diferentes aspectos relacionados ao gerenciamento, fácil acesso e utilização de músicas (ORIO, 2006). Segundo Termens (2009), os pioneiros nesta área foram Kassler (1966) e Lincoln (1967). De acordo com eles, MIR pode ser definida como a tarefa de extrair, de uma grande quantidade de dados, porções que mostram que alguma afirmação musical em particular é verdadeira. Kassler (1966) e Lincoln (1967) explicam que as três ideias principais que mapeiam os objetivos da MIR são: (1) a eliminação da transcrição manual, (2) a criação de uma linguagem efetiva para música, e (3) formas econômicas para visualização de músicas.

MIR é uma tarefa que envolve várias disciplinas, e trabalha com várias possíveis aplicações envolvendo áudio e música. Kosina (2002) lista e explica algumas áreas que estão dentro do âmbito de estudo da MIR:

a) Reconhecimento da fala: tem como objetivo atribuir a um computador a habilidade de analisar e entender a fala humana. A funcionalidade básica deste tipo de sistema é decidir que fonema está sendo falado no momento. Esta tarefa é geralmente complicada devido ao grande número de ambiguidades na linguagem falada;
b) Reconhecimento de locutor: o propósito é identificar quem falou em um determinado áudio dado um conjunto de locutores conhecidos. Problemas encontrados nesta área são a distinção de voz imitada e original e a separação da voz de sons externos;
c) Análise de áudio de vídeos: a busca nesta área é analisar o áudio de vídeos para encontrar diversos tipos de sons como choros, tiros, conversas, entre outros. Assim é possível, por exemplo, classificar vídeos em programas de TV, comerciais, noticiários, esportes, entre outros.

Orio (2006) também explica outras aplicações que são objetos de pesquisa da MIR:

a) Busca de músicas: permitir que sons possam ser encontrados através de pesquisas por melodias e/ou harmonias;
b) Classificação de áudio: nesta área o propósito é encontrar características do áudio e usá-las para agrupar sons parecidos. Este agrupamento pode auxiliar em buscas de músicas, criação automática de listas de reprodução, gerenciamento e organização de bases de áudio, entre outras aplicações.

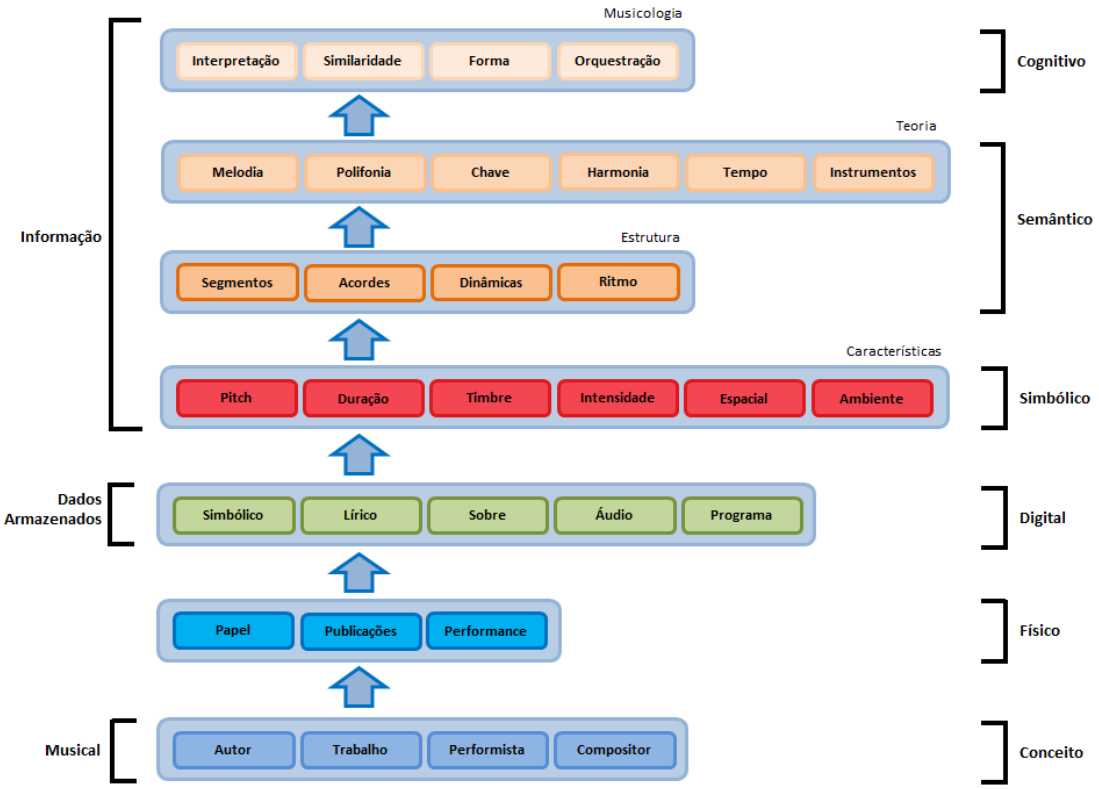
Dentro da área da classificação de áudio, vários são os critérios que podem ser adotados para o agrupamento de músicas. Fu *et al.* (2011) listam e detalham as principais formas de categorização:

a) Classificação por gênero: é a forma de classificação musical mais estudada na MIR, onde o objetivo é classificar músicas através de gêneros, agrupando sons que apresentem elementos musicais parecidos;
b) Classificação pela sensação (emoção): agrupar músicas em diferentes categorias que representem a emoção passada pela música, como felicidade, tristeza, raiva, melancolia, entre outros. É uma forma de classificação que apresenta certos obstáculos, como a falta de bases de dados para comparações entre estudos. Além disto, não há uma certeza sobre qual a emoção passada por uma determinada música, pois esta sensação depende muito de cada pessoa;
c) Identificação de artista: envolve reconhecimento de um artista, cantor ou mesmo compositor. Visto que os artistas possuem estilos distintos para tocar, cantar ou compor, é possível então distinguir quais músicas são tocadas, cantadas ou compostas por estas pessoas;
d) Reconhecimento de instrumentos: possui como objetivo identificar quais instrumentos estão sendo tocados em um determinado trecho de uma música. Atualmente as pesquisas nesta área estão focadas em identificar instrumentos em músicas solo, com um único instrumento. Um dos problemas nesta classificação é justamente a análise de músicas com vários instrumentos (polifônicas), visto a grande quantidade de combinações instrumentais encontradas;
e) Anotação musical: o propósito desta área não é exatamente classificar músicas, mas sim facilitar esta classificação. O objetivo é encontrar uma forma de mapear o conteúdo do áudio em texto. Assim, as demais formas de

classificar músicas poderiam ser desenvolvidas analisando textos, e não mais o áudio.

O núcleo principal da MIR é a informação musical, especificamente sua extração e representação. Isto ocorre através do desenvolvimento de algoritmos capazes de converter dados em informações e, algumas vezes, até em conhecimento (HERRERA-BOYER e GOUYOUN, 2013). A Figura 2.1 mostra um mapa proposto por Fingerhut e Donin (2006), simplificado por Termens (2009), que apresenta todas as disciplinas e informações relacionadas com a MIR.

Figura 2.1 - Mapa simplificado da MIR (adaptado de Termens, 2009).



Fonte: Autor.

As informações da esquerda (musical, dados armazenados e informação) são os conhecimentos que possuímos na nossa mente, arquivadas digitalmente ou apenas guardadas de outra forma. À direita (conceito, físico, digital, simbólico, semântico e cognitivo), são as disciplinas que estão relacionadas com os dados em cada nível de abstração. Os dados são abstraídos em várias etapas. Primeiramente, um músico (autor, compositor, entre outros) desenvolve seu trabalho. Este trabalho pode estar dividido em

diversas partes, como a letra da música, sua notação musical, sua gravação em áudio digital, entre outros. A partir destas partes, é possível extrair diferentes características. Quando agrupadas, estas características formam estruturas mais complexas, como um acorde, um ritmo ou um segmento da música, por exemplo. Estas estruturas, por sua vez, permitem identificar características teóricas da música, como harmonia, melodia, tempo, instrumentos utilizados, entre outros. Por fim, através destas características teóricas, obtêm-se informações musicológicas, como interpretação, orquestração, similaridade e forma. Para o foco deste trabalho, o reconhecimento automático de gêneros musicais, os dados que são necessários são os dos níveis digital, simbólico e semântico.

2.3 RECONHECIMENTO DE GÊNEROS MUSICAIS

O objetivo dos sistemas de reconhecimento de gêneros musicais é categorizar corretamente um sinal de áudio de gênero desconhecido em um gênero musical previamente conhecido a partir das características deste áudio. Para isto, é necessária uma extração de características relevantes do som, e fazer uso destas informações para que, através de uma análise computacional, seja possível identificar a que gênero musical pertence o sinal analisado (KOSINA, 2002; BARREIRA, 2010).

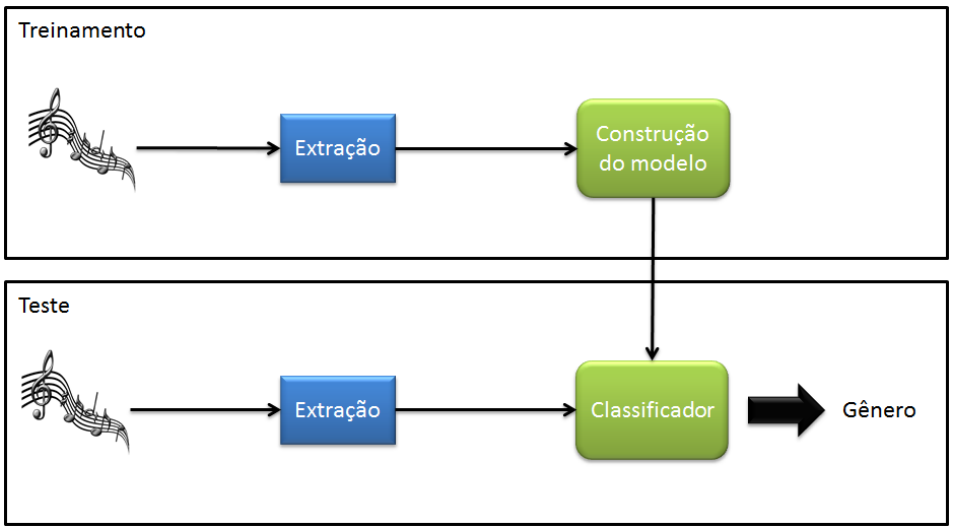
Um sistema de reconhecimento de gêneros musicais pode trabalhar de duas diferentes formas quando analisa músicas de gêneros musicais diferentes daqueles conhecidos pelo sistema. Estas formas de análise definem como o sistema irá categorizar as músicas analisadas. Nesta questão, as possíveis abordagens que determinam o tipo do sistema são:

- Aberto: determina-se um critério mínimo de probabilidade para os gêneros. Se nenhum dos gêneros conhecidos pelo sistema alcançar esta probabilidade mínima, então a música é classificada como pertencente a outro gênero. Por exemplo, se um sistema trabalha com os gêneros Metal, Eletrônica e Jazz, e uma música do gênero Pop é analisada, esta música será provavelmente classificada como pertencente a outro gênero;
- Fechado: toda música é classificada entre algum dos gêneros conhecidos pelo sistema, sendo escolhido o gênero mais parecido com a música

analisada. Utilizando o exemplo anterior, se um sistema trabalha somente com os gêneros Metal, Eletrônica e Jazz, e uma música do gênero Pop é analisada, esta música será classificada entre algum dos três gêneros envolvidos (provavelmente Eletrônica).

Existem propriedades e métodos comuns para se construir um sistema de reconhecimento de gêneros musicais. Este processo de construção se divide basicamente em quatro partes: (1) coleta da base de dados, (2) extração de características, (3) aprendizado de máquina e (4) avaliação dos resultados obtidos (TERMENS, 2009). A extração de características se preocupa em obter dados da entrada, enquanto que o aprendizado de máquina procura encontrar combinações e padrões através dos dados obtidos na extração (KOSINA, 2002). A Figura 2.2 mostra uma representação de um sistema de reconhecimento de gêneros musicais.

Figura 2.2 - Visão geral de um sistema de reconhecimento de gêneros musicais.



Fonte: Autor.

Este modelo de sistema divide-se em duas fases: treinamento e teste. Na fase de treinamento, o sinal do áudio passa pela fase de extração de características, onde recebe diversas modificações e mudanças na sua forma de representação. Após, as características do áudio são selecionadas e extraídas, sendo utilizadas em um algoritmo de aprendizado de máquina (gerando um modelo ou padrão), que durante a fase de treinamento analisa os dados e procura distinguir as informações conforme o gênero dos dados recebidos. Na fase de teste o sinal do áudio também passa pela fase de extração, e as características também são selecionadas e extraídas. O algoritmo de aprendizado,

com o treinamento da primeira fase, passa a ser um classificador. Este classificador analisa os dados recebidos e gera uma saída (a partir do modelo ou padrão estimado no treinamento), que é neste caso um gênero musical. Termens (2009) explica que grande parte dos métodos e trabalhos já realizados são baseados neste modelo de sistema. Segundo ele, alguns autores focam em partes específicas do sistema, aumentando o desempenho do reconhecedor em aplicações específicas. Outros autores muitas vezes utilizam o mesmo modelo de sistema para comparar o desempenho de diferentes bases de dados, características e classificadores.

O desempenho humano em classificar gêneros musicais já foi analisado no trabalho de Perrot e Gjerdigen (1999). Neste trabalho, um grupo de 52 estudantes passou algumas semanas escutando diversas músicas, previamente selecionadas pelos pesquisadores, envolvendo dez gêneros musicais diferentes: Blues, Country, Clássica, Dance, Jazz, Latinas, Pop, R&B, Rap e Rock. Após esse período, os pesquisadores extraíram pequenos trechos de 3 s das músicas, e pediram para que os participantes escutassem e classificassem aquele trecho em algum dos gêneros musicais envolvidos no âmbito da pesquisa. A taxa de acerto foi de aproximadamente 70%. No teste realizado com trechos de até 250 ms a taxa de reconhecimento ficou em aproximadamente 44%. Neste trabalho também foi observado que a utilização de trechos maiores que 3 segundos não trouxeram aumento significativo na taxa de acerto dos participantes.

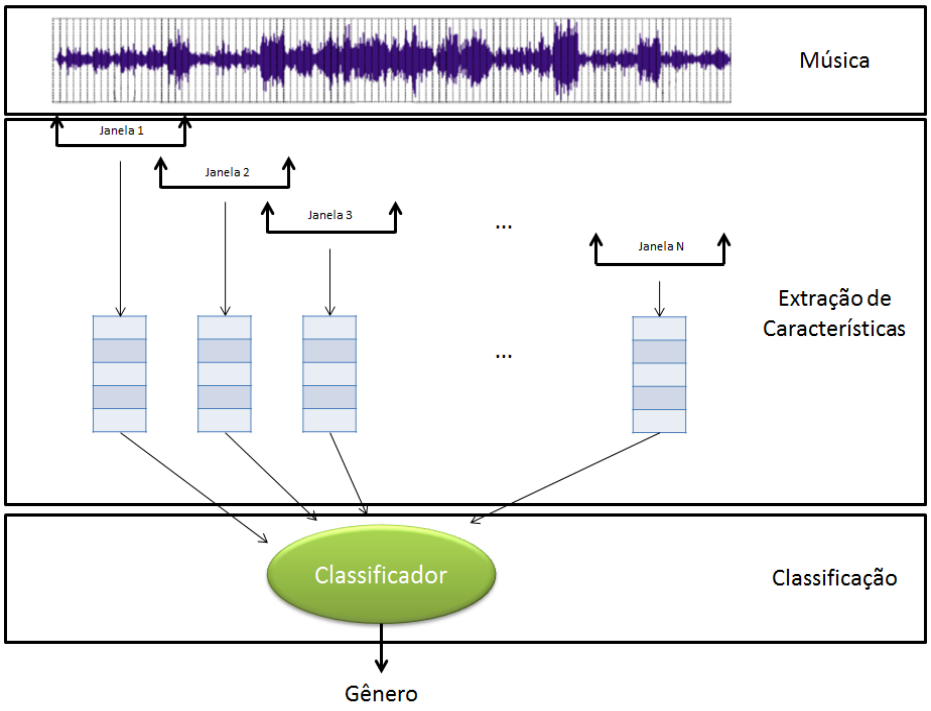
2.4 TRABALHOS CORRELATOS

Na análise de classificação automática de gêneros musicais usando computadores, um dos primeiros trabalhos mais importantes publicados na área foi o de Tzanetakis e Cook (2002). Eles utilizaram características do sinal do áudio como timbre, ritmo e pitch, conseguindo uma taxa de acerto de 61% para dez gêneros. Estes autores ainda disponibilizaram a base de dados utilizada em seus trabalhos (base esta posteriormente apelidada de GTZAN Genre Collection), permitindo que outros pesquisadores pudessem utilizar esta base para realizar seus trabalhos e comparar resultados. No mesmo ano, Kosina (2002) desenvolveu um sistema de classificação de gêneros musicais, chamado MUGRAT, conseguindo 88% de reconhecimento para três

gêneros. Orio (2006) também menciona outro evento que teve grande destaque, a Conferência Internacional de Recuperação de Informações Musicais (ISMIR) de 2004, que reuniu vários estudos e métodos de reconhecimento de gêneros musicais, popularizando esta área de pesquisa.

Desde então, diversos trabalhos foram realizados, aplicando diferentes ideias e variações ao sistema básico de reconhecimento de gêneros musicais. Uma das variações, por exemplo, se refere ao tamanho do trecho da música a ser processada. Alguns trabalhos como os de Tzanetakis e Cook (2002) e Bergstra *et al.* (2006) utilizaram o método mostrado na Figura 2.3, onde todo o conteúdo da música é analisada. Neste modelo, o sinal da música é dividido em pequenos trechos de tamanhos iguais (geralmente de 10 ms a 100 ms), chamados de janelas. Para cada janela, características são extraídas e passadas para a entrada de um classificador, que analisa os dados e retorna como saída um gênero musical. A desvantagem deste modelo é o tempo de execução, pois a análise de toda uma música é um processo muito demorado e que envolve uma grande quantidade de dados. Por outro lado, garante que todos os elementos da música sejam analisados, além de ser uma abordagem simples, por não envolver algoritmos de segmentação no arquivo.

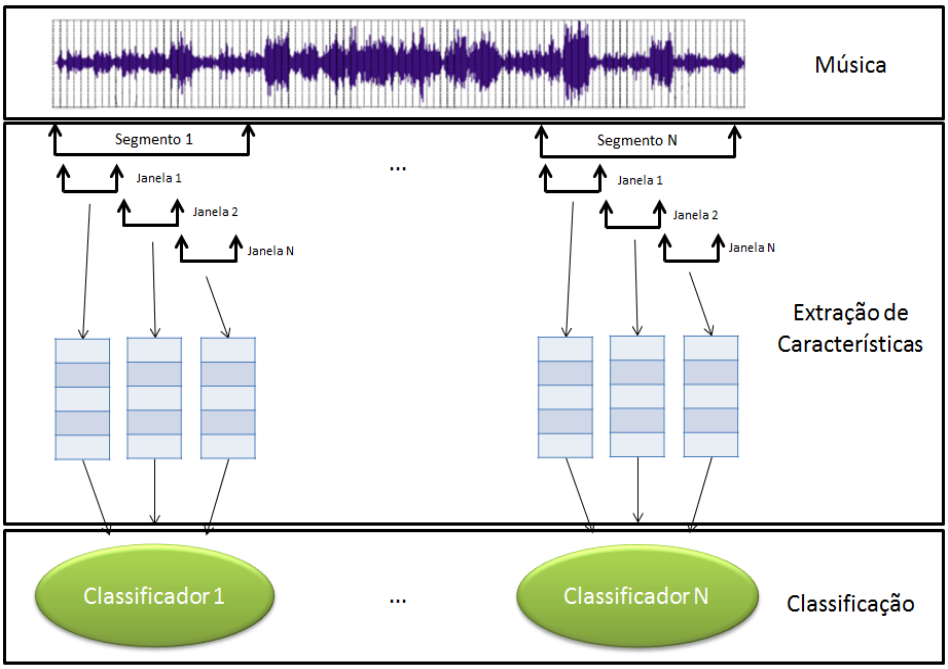
Figura 2.3 - Modelo de análise do sinal inteiro.



Fonte: Autor.

Outra forma de explorar a música sem ter a demora de processar todo o sinal é analisar apenas alguns segmentos do áudio, como exemplificado na Figura 2.4. Scaringella *et al.* (2006) explicam que um pequeno segmento do áudio geralmente contém informações suficientes para caracterizar o conteúdo completo de uma canção, pois em muitos gêneros a estrutura da música se repete com frequência. Neste modelo, o sinal é dividido em *n* segmentos, geralmente de tamanhos iguais. Para cada segmento, é realizada a divisão em trechos menores (janelas), onde as características são extraídas de cada janela e depois passadas para a entrada de um classificador específico para aquele segmento. Scaringella *et al.* (2006) afirmam que a maioria dos métodos utilizam apenas um segmento da música, geralmente de 30 segundos de duração, retirado após os 30 segundos iniciais da música para evitar introduções que não reflitam o verdadeiro gênero da canção. Silla *et al.* (2007) realizaram uma análise diferente, extraindo dados de três partes do áudio (começo, meio e fim da música, com trechos de 30 segundos cada), utilizando classificadores diferentes para cada parte e combinando os resultados. Neste trabalho, foi alcançada uma taxa média de acerto 3% maior do que o melhor resultado obtido individualmente (55,15% sem a combinação contra 58,07% utilizando três segmentos, para 10 gêneros musicais).

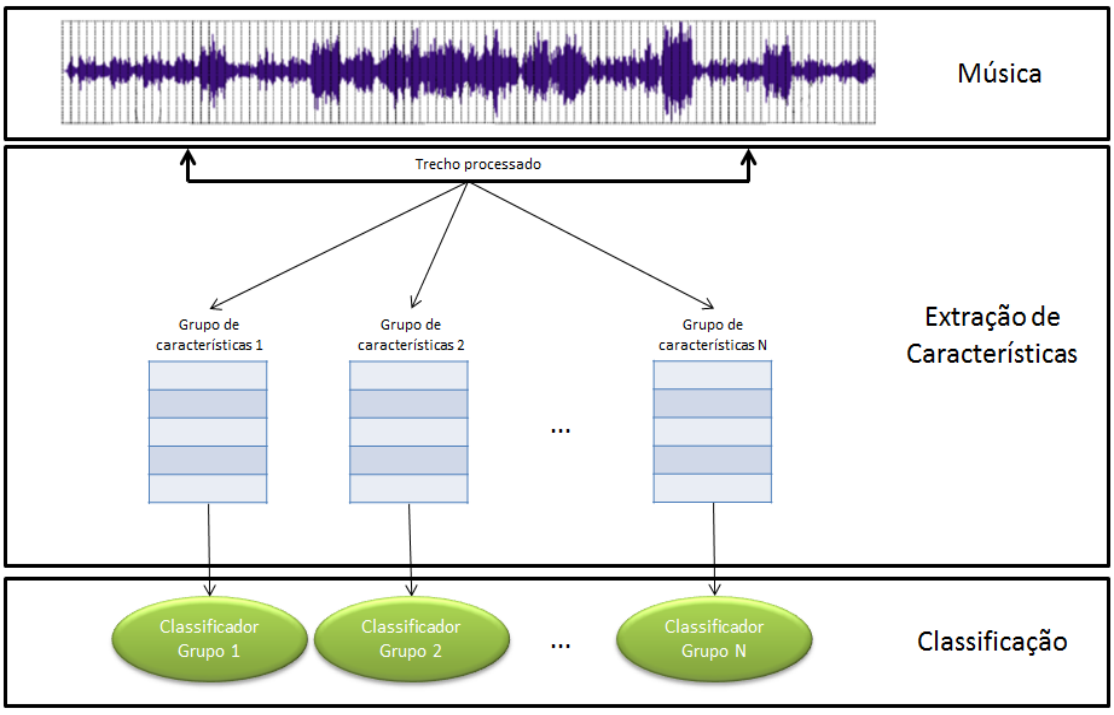
Figura 2.4 - Modelo de análise por segmentos.



Fonte: Autor.

Também é possível modelar um sistema de reconhecimento de gêneros musicais utilizando a classificação de agrupamento de características, conforme exemplificado na Figura 2.5. A ideia é, a partir do trecho da música analisada, extrair características e agrupá-las em *n* grupos. Após, para cada grupo é utilizado um classificador diferente, e o resultado da análise pode ser obtido de cada classificador individualmente ou da combinação deles. Seguindo este modelo, Paradzinets *et al.* (2009) construíram um sistema onde as características do áudio foram extraídas e agrupadas em três grupos (acústicas, rítmicas e timbre), onde cada grupo foi analisado por um classificador diferente, que retornava as probabilidades da música pertencer a cada gênero. Neste trabalho também foi utilizado o conceito de sistemas *multiexpert*, onde a saída de cada classificador foi utilizada como um novo dado de entrada para um novo classificador, que combinava os resultados e definia o gênero da música. Através desta combinação, os autores constataram um aumento de 12% no reconhecimento em comparação ao desempenho do melhor classificador individual (54,60% sem a combinação e 66,70% utilizando grupos de características, para 6 gêneros musicais).

Figura 2.5 - Modelo de análise por grupos de características.



Fonte: Autor.

Outra questão nos sistemas de reconhecimento de gêneros musicais se refere à forma como as características de uma música são passadas para o classificador. Conforme mostrado nos exemplos de arquitetura anteriores, a música é dividida em pequenos trechos (janelas), e para cada janela, características são extraídas. Logo, se uma música é dividida em *n* janelas, *n* conjuntos de características são obtidos para representar a música analisada. Segundo Fu *et al.* (2011), as principais formas de passar estas características para o classificador são:

- Agrupar todos os *n* conjuntos de características em um único conjunto, que representará toda a música analisada, utilizando os valores médios de cada característica nos conjuntos;
- Passar cada um dos *n* conjuntos para o classificador;
- Utilizar medidas de similaridade para cada par de músicas de mesmo gênero. Nesta forma, para cada música, características são extraídas e após um modelo probabilístico é usado para estimar a distribuição destas características. Em seguida, a similaridade entre cada par de músicas é calculada comparando os modelos, utilizando algum critério de divergência. Esta similaridade é então passada para o classificador.

Alternativamente, diversas formas de construção de um sistema de reconhecimento de gêneros musicais também foram apresentadas. Lampropoulos *et al.* (2005), por exemplo, realizaram um método de separação do sinal do áudio em vários pedaços (representando instrumentos de corda, sopro e percussão) e extraíram dados de cada uma destas partes. Costa *et al.* (2011) propuseram um método que ao invés de extrair os dados diretamente do sinal do áudio, os dados foram extraídos de uma imagem do sinal, utilizando processamento de imagem. Já Ariyaratne e Zhang (2012) desenvolveram um modelo de sistema hierárquico, na qual os gêneros foram divididos em grupos menores e organizados de forma hierárquica. Os dados do áudio então foram processados e o classificador decidia para qual grupo o dado era mais similar, e este processo se repetia para todos os subgrupos dentro do grupo escolhido, até o classificador decidir o gênero final.

Com o objetivo de complementar as informações modeladas por diferentes classificadores, diversos trabalhos utilizaram fusão de classificadores no reconhecimento de gêneros. Diversas são as formas de combinar estes resultados, e algumas destas regras, explicadas por Silla *et al.* (2007), são: (1) voto majoritário, que

simplesmente conta o gênero que mais apareceu nos resultados; (2) regra da soma, que soma as probabilidades de cada classe em cada classificador e; (3) regra do produto, que multiplica as probabilidades de cada classe em cada classificador. Chathuranga e Jayaratne (2013) alegam que outra forma de combinar resultados é atribuir pesos para cada classificador individual. Este peso é calculado conforme o desempenho de cada classificador durante a fase de treinamento, associando pesos maiores para classificadores com melhor desempenho. A partir disto, pode ser aplicada a regra da soma ou do produto, multiplicando cada probabilidade com o peso do classificador que originou o resultado. Costa *et al.* (2004) utilizaram a combinação de classificadores com a regra do voto majoritário, onde foi alcançado uma taxa de acerto 0,8% maior que a taxa de acerto do melhor classificador individual (89,5% sem e 90,3% com a combinação, para 2 gêneros musicais). A fusão de classificadores também aparece nos trabalhos de Chathuranga e Jayaratne (2013), onde foi utilizado pesos para cada classificador e constatado um aumento médio de 3% no desempenho (para 10 gêneros musicais); e de Yaslan e Cataltepe (2006), que obtiveram um aumento de 4% no desempenho do sistema após a fusão dos classificadores, para 10 gêneros musicais.

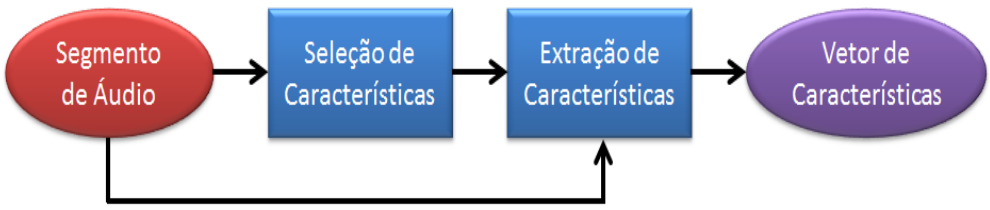
3. EXTRAÇÃO DE CARACTERÍSTICAS

Este capítulo apresenta a etapa de extração de características a partir do áudio, mostrando as características mais utilizadas, e suas formas de extração e representação. A Seção 3.1 apresenta o processo de extração e mostra uma forma de categorizar as principais características. A Seção 3.2 explica as características de baixo nível, abordando conceitos, formas de extração e significados musicais de cada uma delas. Além disto, também é feito um levantamento bibliográfico mostrando o uso e desempenho destas características em trabalhos relacionados. A Seção 3.3 aborda as características de médio nível, seguindo os mesmos princípios da Seção 3.2.

3.1 CATEGORIZAÇÃO DAS CARACTERÍSTICAS

A extração de características (*feature extraction*) no reconhecimento de gêneros musicais é o processo onde, a partir de um segmento de um áudio, dados são extraídos e convertidos em uma forma numérica compactada, referenciada por vetor de características, que representa o trecho analisado. É a parte que engloba conhecimentos de diversos campos de estudo, mas principalmente das áreas da música, psicoacústica e processamento de sinal (NOROWI *et al.*, 2005). A Figura 3.1 mostra o processo de construção de um vetor representando as características de um trecho de uma música analisada.

Figura 3.1 - Processo de construção de um vetor de características.



Fonte: Autor.

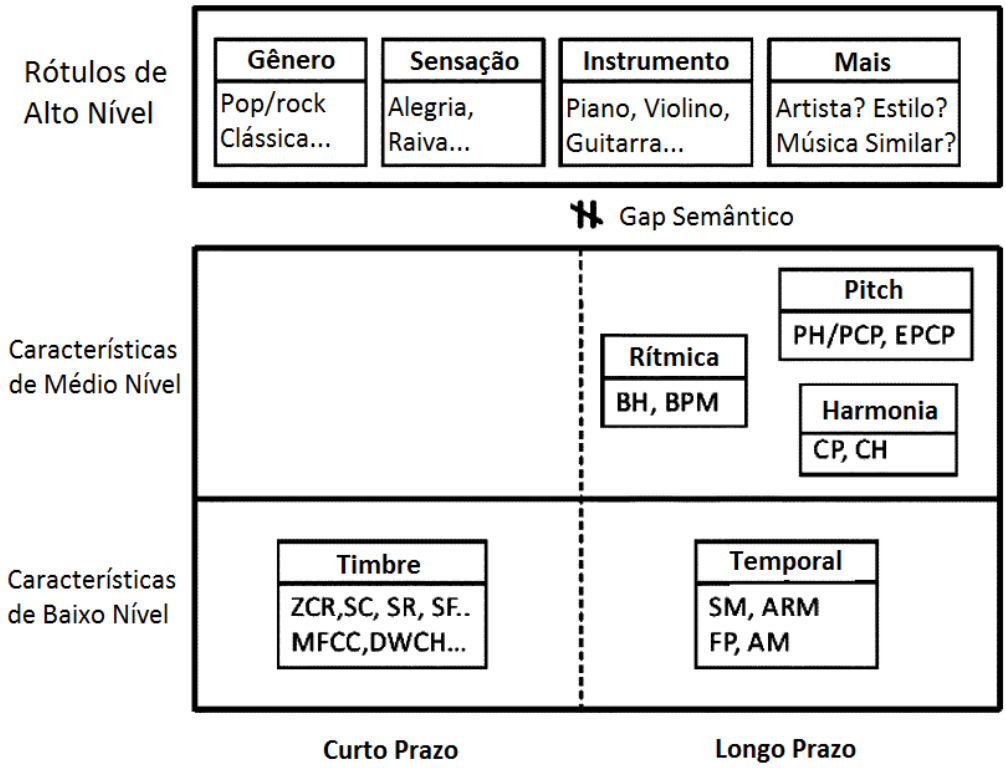
Duas abordagens podem ser tomadas em relação ao processo de extração de características. Na primeira, pode-se realizar uma seleção de características que farão

parte do sistema. A seleção de características (*feature selection*) define quais são as *d* variáveis em um conjunto de *p* características que melhor contribuem para a distinção das classes, assim descartando as outras (*p* – *d*) características (WEBB, 2002). A seleção de características pode ajudar na classificação escolhendo as informações mais promissoras, mas não é necessariamente obrigatória. Na segunda abordagem as *d* variáveis são previamente selecionadas através da teoria ou de estudos já realizados, e por isso são utilizadas diretamente para a construção de sistemas de reconhecimento. Assim, independente da abordagem selecionada, a extração de características produzirá um vetor de características com *d* dimensões representando um segmento da música analisada.

Diversas características do áudio são identificadas e utilizadas em vários trabalhos, e existem diferentes taxonomias para categorizar estas características (NOROWI *et al.*, 2005; FU *et al.*, 2011). Scaringella *et al.* (2006), por exemplo, dividiram as características em três grupos: timbre, melodia e harmonia, e rítmicas. Já Meng *et al.* (2007) utilizaram outros grupos: *short-time*, *medium-time* e *long-time*, que se referem ao tempo de áudio a ser processado para obter tais características.

Fu *et al.* (2011) propuseram outra forma de agrupar as características do áudio, conforme mostrado na Figura 3.2. Neste modelo a divisão é feita em três grandes grupos: características de baixo, médio e alto nível. As de alto nível são informações abstratas que não podem ser obtidas diretamente de uma análise do áudio. Este grupo representa a forma como as pessoas interpretam e entendem música (através de gêneros, instrumentos, vocais, entre outros). As características de baixo nível são aquelas obtidas através de técnicas de processamento de sinal, enquanto que as de médio nível requerem uma análise mais complexa do sinal, e representam propriedades baseadas no sistema de audição humana. Estes dois grupos possuem várias classes que consistem de diferentes características. Ainda, estas classes são divididas em dois grupos: de curto e longo prazo, se referindo ao tamanho dos segmentos de áudio em que estas características são analisadas. Características de curto prazo frequentemente são capturadas em trechos de 10 ms a 100 ms, enquanto que características de longo prazo são normalmente extraídas de trechos maiores.

Figura 3.2 - Categorização das características do áudio (adaptado de Fu *et al.*, 2011).



Fonte: Autor.

3.2 CARACTERÍSTICAS DE BAIXO NÍVEL

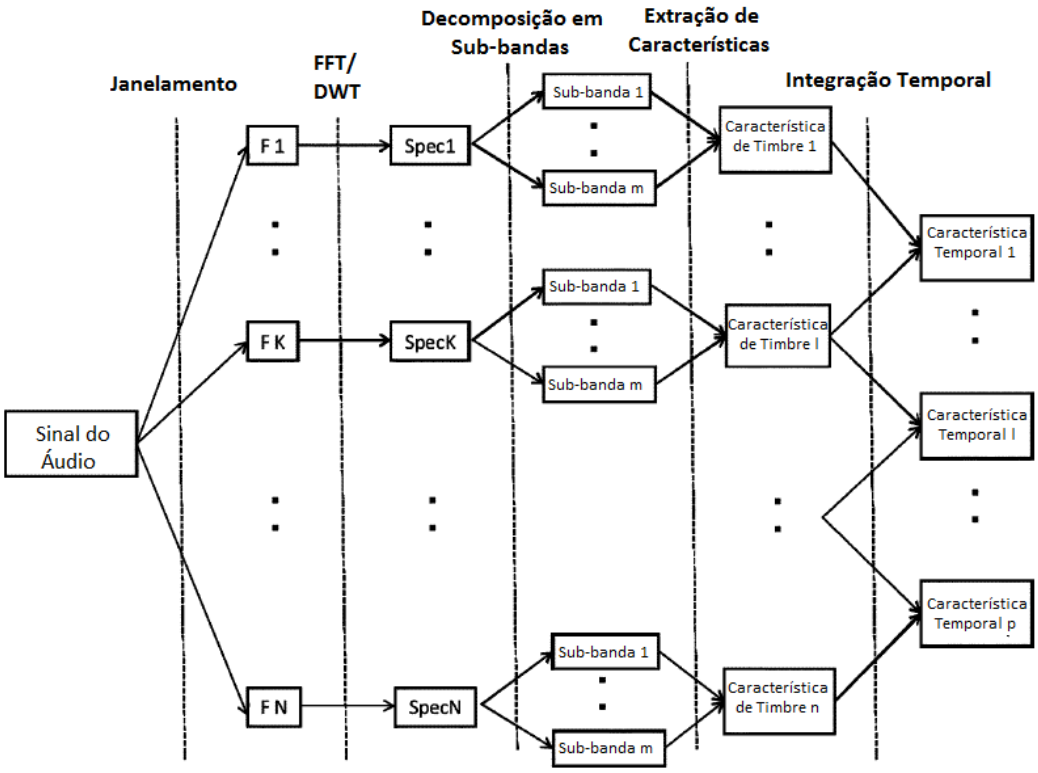
Fu *et al.* (2011) dividiram as características de baixo nível em duas classes: relacionadas ao timbre e temporais. Estas características são obtidas diretamente através de várias técnicas de processamento de sinal como a transformada de Fourier, análise espectral, modelamento auto regressivo, entre outros (FU *et al.*, 2011). A transformada de Fourier, por exemplo, uma das técnicas mais utilizadas, é definida como a conversão de um sinal de sua representação no domínio tempo $x(t)$ para uma representação correspondente no domínio frequência $X(f)$ (BOSI e GOLDBERG, 2003). Para sequências de duração finita, a representação de Fourier em tempo discreto pode ser obtida através da transformada de Fourier discreta (em Inglês, *Discrete Fourier Transform* - DFT) (OPPENHEIM e SCHAFER, 2012), definida por

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi nk/N}, 0 \leq k < N$$

onde *N* é o número de amostras do sinal. Além disso, existe uma classe de algoritmos eficientes para a implementação computacional da DFT, chamados de algoritmos FFT (do Inglês *Fast Fourier Transform*), que também são utilizados para obter a transformada (GOLD *et al.*, 2011; OPPENHEIM e SCHAFER, 2012).

A transformada de Fourier é aplicada sobre curtos segmentos de áudio produzindo amostras em frequência, das quais características são extraídas. A Figura 3.3 mostra os passos para a extração destes tipos de características. Li *et al.* (2003) explicam que para extrair estas características, o sinal do áudio é dividido em segmentos de curta duração (F1, F2, ...FK, ..., FN) chamados de janelas. As características (*Timbre Features* 1, 2, ...l, ..., n) são estimadas das amostras de frequência produzidas pela aplicação da transformada de Fourier (*Subband* 1, 2, ..., m) para cada janela e estatísticas como média e variância destas características são então calculadas (*Temporal Features* 1, 2, ...i, ..., p).

Figura 3.3 - Extração de características de baixo nível (adaptado de Fu *et al.*, 2011).



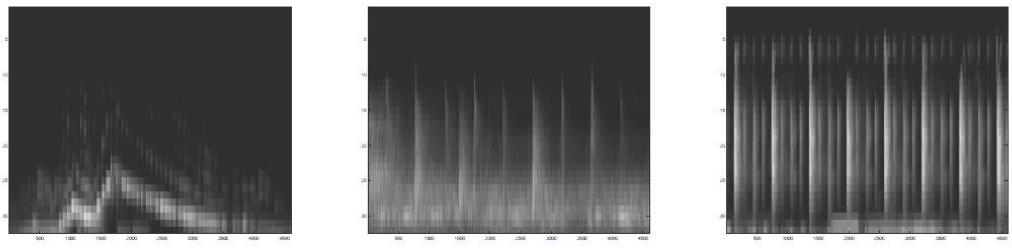
Fonte: Autor.

3.2.1 TIMBRE

Timbre é o termo que descreve a qualidade de um som, onde diferentes timbres são produzidos por diferentes tipos de fontes sonoras, como diferentes vozes e instrumentos musicais (FU *et al.*, 2011). Para melhor entendimento, timbre na música pode ser comparado à cor em imagens. As características de timbre foram originadas do reconhecimento de voz, e são usadas para diferenciar misturas de sons que possivelmente possuem os mesmos conteúdos rítmicos, duração, intensidade e pitch (LI *et al.*, 2003). Estas características são determinadas por fatores como: (1) instrumento, (2) nível dinâmico, (3) articulação / técnica de tocar o instrumento e (4) acústica do ambiente / pós-processamento (PLUMBEY e DIXON, 2012).

Kosina (2002) argumenta que uma boa forma de trabalhar com informações de timbre é transformar a saída da transformada de Fourier para um domínio visual utilizando espectrogramas. O espectrograma é um gráfico de relação tempo-frequência, onde o eixo *x* mostra o tempo *t* da música, enquanto que o eixo *y* mostra a frequência *f* (KOSINA, 2002). A Figura 3.4 mostra exemplos de espectrogramas.

Figura 3.4 - Espectrograma de trechos musicais de três gêneros diferentes.



(a) Classical (b) Heavy Metal (c) Dance

Fonte: Kosina (2002).

As principais características relacionadas ao timbre, como são calculadas e a informação musical obtida através delas são:

- ***Spectral Centroid*:** é o ponto balanceado do espectro, uma medida associada frequentemente com a noção do brilho espectral (SILLA *et al.*, 2004). Segundo Chathuranga e Jayaratne (2013), *spectral centroid* é a média ponderada das frequências sobre o tempo, sendo calculado pela equação

$$C_t = \frac{\sum_{n=1}^{N} X_t[n] * n}{\sum_{n=1}^{N} X_t[n]}$$

onde $X_t[n]$ é o valor da transformada de Fourier no quadro *t* e faixa de frequência *n*. Malheiro *et al.* (2005) explicam que o *spectral centroid* é utilizado para discriminar instrumentos musicais. É uma característica bastante utilizada, como por exemplo, nos trabalhos de Banitabeli-Dehkordi *et al.* (2012), Lamya e Houacine (2007), Aryaratne e Zhang (2012), Tzanetakis e Cook (2002), entre outros.

- ***Spectral Rolloff*:** é o ponto R_t onde a frequência que está abaixo de alguma porcentagem P_o (geralmente 85%) da energia do espectro reside (NOROWI *et al.*, 2005). O ponto R_t é estimado de acordo com a expressão

$$\sum_{n=1}^{R_t} X_t[n] = P_o * \sum_{n=1}^{N} X_t[n], \quad 0 \leq P_o \leq 1$$

onde $X_t[n]$ é a magnitude da transformada de Fourier no quadro *t* e faixa de frequência *n*. Tzanetakis e Cook (2002) explicam que o *spectral rolloff* é outra forma de medir a frequência média do espectro e tem objetivos similares ao *spectral centroid*. Utilizada, por exemplo, nos trabalhos de Banitabeli-Dehkordi *et al.* (2012), Lamya e Houacine (2007), Aryaratne e Zhang (2012), Tzanetakis e Cook (2002).

- ***Spectral Flux*:** é a diferença entre duas amplitudes normalizadas de sucessivas distribuições espectrais, utilizada para calcular a quantidade de mudanças no espectro no decorrer do tempo (LI *et al.*, 2006). É definida pela equação

$$F_t = \sum_{n=1}^{N} (X_t[n] - X_{t-1}[n])^2$$

onde $X_t[n]$ e $X_{t-1}[n]$ representam magnitudes do espectrograma de frequência *n* nos tempos *t* e *t*-1. Chathuranga e Jayaratne (2013) explicam que o valor do *spectral flux* indica as variações na música (se é uma música mais constante ou se apresenta mudanças bruscas). Utilizada por exemplo, nos trabalhos de Banitabeli-Dehkordi *et al.* (2012), Lamya e Houacine (2007), Tzanetakis e Cook (2002).

- ***Zero-Crossing Rate*:** esta taxa é calculada contando o número de vezes que a onda do áudio cruza o eixo x para cada unidade de tempo (CHATHURANGA e JAYARATNE, 2013). É estimada através da equação

$$Z_t = \frac{1}{2}\sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])|$$

onde a função $sign()$ retorna o valor 0 (zero) para valores negativos e o valor 1 (um) para valores positivos, e $x[n]$ denota o domínio do sinal no tempo *t*. Xu *et al.* (2005) argumentam que esta característica é sensível aos vocais e instrumentos de percussão, onde valores altos indicam maior presença destes sons. Utilizada por exemplo, nos trabalhos de Banitabeli-Dehkordi *et al.* (2012), Lamya e Houacine (2007), Aryaratne e Zhang (2012), Tzanetakis e Cook (2002).

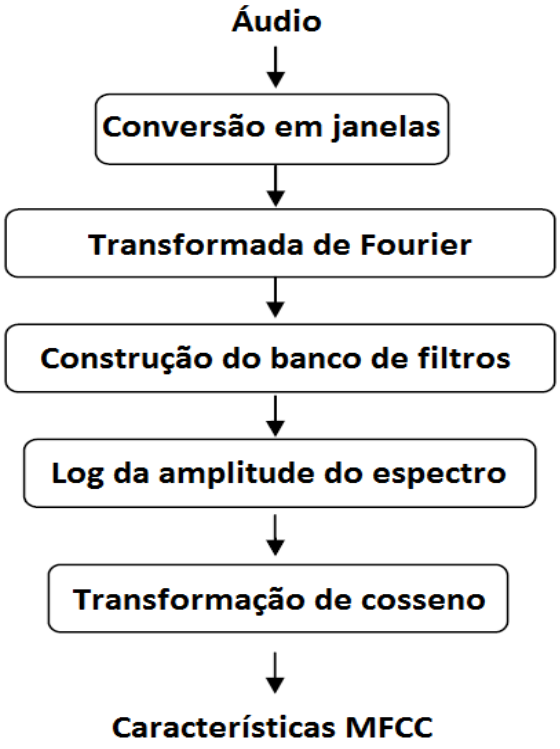
- ***Low Energy*:** é a porcentagem de janelas que têm menos energia do que a energia média de todas as janelas (SILLA *et al.*, 2004). Tzanetakis e Cook (2002) explicam que, por exemplo, músicas com partes silenciosas ou muito calmas terão mais *low energy*. Utilizada nos trabalhos de Banitabeli-Dehkordi *et al.* (2012), Lamya e Houacine (2007), Tzanetakis e Cook (2002).

É interessante observar que as características *spectral centroid*, *spectral rolloff*, *spectral flux*, *zero-crossing rate* e *low energy* são frequentemente utilizadas juntas nos trabalhos de reconhecimento de gêneros musicais. Isto acontece, por exemplo, nos trabalhos de Koerich e Poitevin (2005), Tzanetakis *et al.* (2001), Pohle *et al.* (2004) e Lamya e Houacine (2007).

COEFICIENTES CEPSTRAIS DA FREQUÊNCIA MEL

Os Coeficientes Cepstrais da Frequência Mel (do Inglês *Mel Frequency Cepstral Coefficients* – MFCC), proposto por Davis e Mermelstein (1980), são uma representação compacta do espectro do sinal do áudio, com o objetivo de aproximar a distribuição não linear da largura de banda da audição humana com frequência (BURRED e LERCH, 2003; LEON e MARTINEZ, 2012). É uma das características mais utilizadas em reconhecimento de gêneros musicais (TZANETAKIS E COOK, 2002; MENG *et al.*, 2007; SILLA *et al.*, 2007; LEE *et al.*, 2009; BENETOS E KOTROPOULOS, 2010). Seu fluxo de obtenção é mostrado na Figura 3.5.

Figura 3.5 - Fluxo para cálculo do MFCC (adaptado de Chu, 2009).



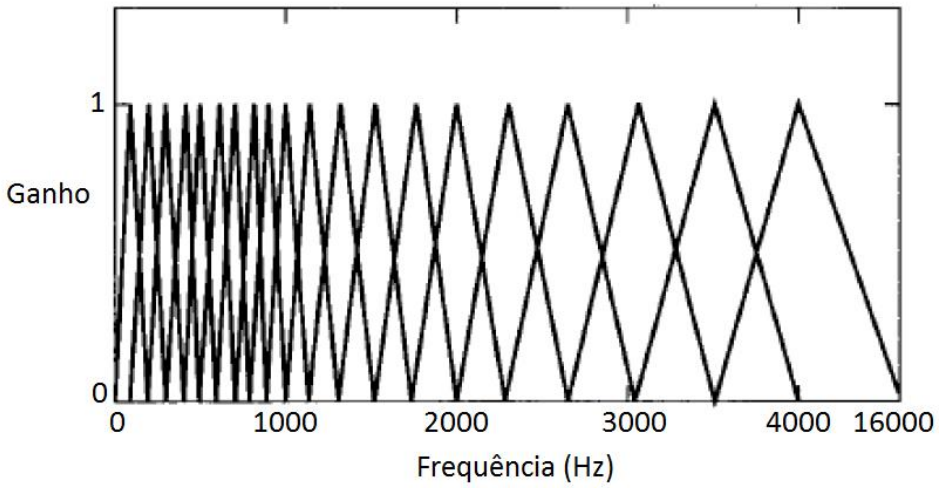
Fonte: Autor.

Segundo Huang *et al.* (2001), primeiramente o áudio é dividido em janelas, e em cada janela é aplicada a transformada de Fourier discreta no sinal. Em seguida, é definido um conjunto de M filtros $(m = 1, 2, \dots, M)$, onde para um conjunto de frequências *f* um filtro *m* é um filtro triangular dado por:

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \dfrac{2(k - f[m-1])}{(f[m+1] - f[m-1])(f[m] - f[m-1])} & f[m-1] \leq k \leq f[m] \\ \dfrac{2(f[m+1] - k)}{(f[m+1] - f[m-1])(f[m+1] - f[m])} & f[m] \leq k \leq f[m+1] \\ 0 & k > f[m+1] \end{cases}.$$

(HUANG *et al.*, 2001; SCHAFER, 2008). Geralmente a construção dos filtros é feita aplicando filtros lineares nas frequências baixas, e filtros logarítmicos nas frequências altas, como exemplificado na Figura 3.6. A filtragem é feita desta forma devido a maior percepção humana das frequências baixas do que das frequências altas. Geralmente, a filtragem é linear abaixo de 1 Khz, e logarítmica acima deste ponto (GOLD *et al.*, 2011; LEON E MARTINEZ, 2012).

Figura 3.6 - Filtragem do MFCC (adaptado de Gold *et al.*, 2011).



Fonte: Autor.

Após, segundo Huang *et al.* (2001) e Schafer (2008), é calculado o logaritmo da magnitude do espectro, e os valores das faixas pré-determinadas são convertidos em novos valores utilizando o conjunto de filtros através de:

$$S[m] = \ln\left[\sum_{k=0}^{N-1} |X[k]|^2 \, H_m[k]\right], \quad 0 \leq m \leq M.$$

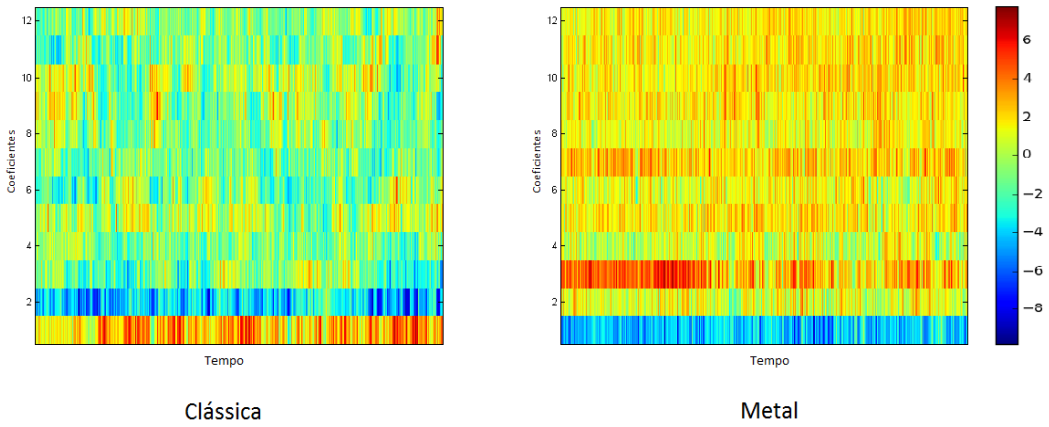
Em seguida, é aplicada uma transformação de cosseno discreta nas saídas dos *M* filtros:

$$c[n] = \sum_{m=0}^{M-1} S[m]\cos(\pi\, n(m + 1/2)/M), \quad 0 \leq n < M$$

onde *M* varia para diferentes implementações de 24 a 40 (HUANG *et al.*, 2001; SCHAFER, 2008). Para o reconhecimento de gêneros musicais, geralmente os 13 primeiros coeficientes são utilizados (TZANETAKIS e COOK, 2002).

A Figura 3.7 mostra as diferenças do MFCC para dois gêneros musicais. O eixo *x* é o tempo da música, enquanto que o eixo *y* são os coeficientes do MFCC (neste exemplo são os coeficientes 1 ao 12). As cores mostram os valores do MFCC em cada tempo e coeficiente; cores com tonalidade azul mais forte representam valores negativos menores, enquanto que cores mais próximas do vermelho representam valores positivos maiores.

Figura 3.7 - Características MFCC para os gêneros clássica e metal (adaptado de Barreira, 2010).



Fonte: Autor.

3.2.2 TEMPORAL

Características temporais são aquelas que capturam a evolução temporal do sinal. Conforme mostrado na Figura 3.3, características temporais são usualmente estimadas a partir das características de timbre, que são extraídas de várias janelas, podendo ser integradas para criar uma característica temporal (FU *et al.*, 2011). A diferença entre características de timbre e temporais é que as de timbre são extraídas diretamente do áudio, enquanto que as temporais são extraídas das próprias

características de timbre, utilizando janelas maiores (MENG *et al.*, 2007; FU *et al.*, 2011). As principais características temporais são:

- ***Statistical Moments*:** utiliza medidas estatísticas, podendo calcular, por exemplo, a média, variância, covariância e distorção das características de timbre (TERMENS, 2009). Utilizada nos trabalhos de Meng *et al.* (2007) e Lim *et al.* (2012).

- ***Amplitude Modulation*:** são estimadas diretamente da amplitude do sinal original do áudio e representam a evolução temporal do áudio sobre o tempo (CHATHURANGA e JAYARATNE, 2013). Alguns possíveis cálculos são computar os valores mínimos e máximos, média, variância, entre outros. Usada nos trabalhos de Meng *et al.* (2007), Chaturanga e Jayaratne (2013), Lim *et al.* (2012).

- ***Autoregressive Modeling*:** o modelo auto regressivo lida com a correlação entre as características sobre o tempo (MENG e SHAWE-TAYLOR, 2005). Matematicamente, é definido como:

$$x_n = \sum_{p=1}^{K} \boldsymbol{A}_p \boldsymbol{x}_{n-p} + \boldsymbol{v} + \boldsymbol{u}_n$$

onde *K* é a ordem do modelo, $\boldsymbol{A}_p$ é uma matriz de coeficientes auto regressivos, $\boldsymbol{v}$ é um vetor de termos de intersecção e $\boldsymbol{u}_n$ é o processo do curso do ruído. Utilizada nos trabalhos de Meng *et al.* (2007), Almeida *et al.* (2012), Jothilakshmi e Kathiresan (2012).

3.2.3 DESEMPENHO DAS CARACTERÍSTICAS DE BAIXO NÍVEL

A Tabela 3.1 apresenta a utilização das principais características de baixo nível em alguns trabalhos, e a melhor taxa de acerto obtida utilizando estas características. A partir da tabela, podemos concluir que:

- O uso das características de timbre, como *spectral centroid* e *spectral rolloff*, quando utilizadas individualmente, apresentam resultados inferiores

do que quando unidas. Por exemplo, comparando as pesquisas de Jothilakshmi e Kathiresan (2012) e Li *et al.* (2003), onde ambos trabalham com 5 gêneros musicais, se percebe uma diferença significativa na taxa de acerto quando utilizada as características de timbre individualmente e em conjunto;

- O MFCC, de um modo geral, apresenta uma alta taxa de acerto, conforme nota-se nos trabalhos de West e Cox (2004), Reed e Lee (2006) e Leon e Martinez (2012);
- A combinação de características apresenta melhores resultados do que o uso individual delas. No trabalho de Li *et al.* (2003) por exemplo, a combinação elevou a taxa de acerto para mais de 3% quando comparado ao melhor desempenho individual.

Tabela 3.1 - Desempenho de características de baixo nível.

TRABALHO	CARACTERÍSTICAS	NÚMERO DE GÊNEROS	TAXA DE ACERTO
Li *et al.* (2003)	Spectral Centroid + Spectral Rolloff + Spectral Flux + Zero-Crossing Rate + Low Energy	5	62,19%
	MFCC	5	67,46%
	Spectral Centroid + Spectral Rolloff + Spectral Flux + Zero-Crossing Rate + Low Energy + MFCC	5	70,63%
Reed e Lee (2006)	MFCC	5	69,32%
Jothilakshmi e Kathiresan (2012)	Spectral Centroid	5	25,00%
	Spectral Rolloff	5	28,00%
	MFCC	5	37,50%
	MFCC + Spectral Centroid	5	41,25%
McKinney e Breebaart (2003)	Spectral Centroid + Spectral Rolloff + Zero-Crossing Rate	6	58,00%
	MFCC	6	61,00%
West e Cox (2004)	MFCC	6	65,00%
Leon e Martinez (2012)	MFCC	10	79,36%
Tzanetakis *et al.* (2001)	Spectral Centroid + Spectral Rolloff + Spectral Flux + Zero-Crossing Rate + Low Energy	15	51,00%

3.3 CARACTERÍSTICAS DE MÉDIO NÍVEL

Características de médio nível são aquelas que procuram capturar as propriedades intrínsecas da música que humanos percebem e apreciam. São características identificadas facilmente por qualquer pessoa que escute música, porém são mais difíceis de serem extraídas e analisadas do sinal do áudio. Todas estas características requerem uma análise de janelas maiores da música (FU *et al.*, 2011).

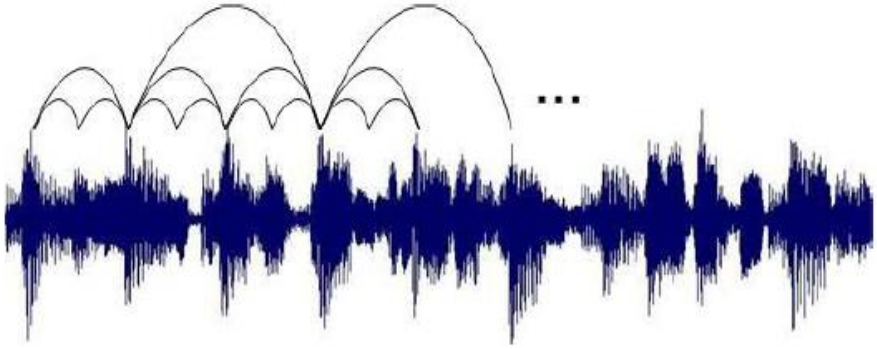
Conforme Fu *et al.* (2011), as características de médio nível podem ser divididas em três classes: rítmicas, pitch e harmônicas. Estas classes são analisadas nas subseções a seguir.

3.3.1 RÍTMICAS

As características relacionadas à batida e à estrutura rítmica de uma música são frequentemente uma boa forma de identificar o seu gênero (SILLA *et al.*, 2004; NOROWI *et al.*, 2005). Porém, são difíceis de serem analisadas devido à complexidade da representação de um ritmo. Apesar disto, as características rítmicas aparecem em vários trabalhos (TZANETAKIS e COOK, 2002; BAGCI e ERZIN, 2007; MENG *et al.*, 2007; CHATHURANGA e JAYARATNE, 2013).

O ritmo é uma estrutura métrica composta por um conjunto hierárquico de pulsos, que são uma sequência espaçada regular de acentos (batidas), onde cada pulso define um nível métrico (PLUMBEY e DIXON, 2012). A Figura 3.8 mostra uma estrutura métrica de uma música. É interessante notar os vários níveis de pulsações existentes, montando uma estrutura hierárquica de pulsações. Portanto, as características rítmicas procuram capturar informações desta métrica.

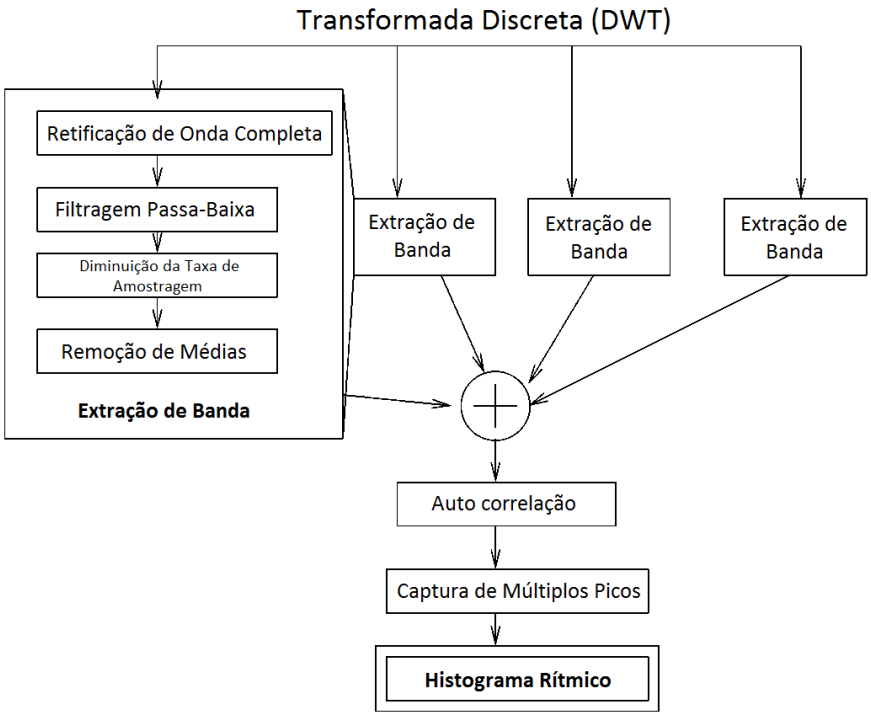
Figura 3.8 - Estrutura métrica de uma música.



Fonte: Plumbey e Dixon (2012).

O histograma da batida é uma forma de representar a estrutura rítmica da música sobre o tempo. Tzanetakis e Cook (2002) propuseram um método baseado na transformada Wavelet para calcular o histograma, conforme mostrado na Figura 3.9.

Figura 3.9 - Diagrama de montagem do histograma rítmico (adaptado de Tzanetakis e Cook, 2002).



Fonte: Autor.

Primeiramente, o sinal é analisado em janelas que geralmente são grandes, de aproximadamente 4 segundos. A necessidade de janelas maiores é para capturar uma estrutura rítmica completa (TZANETAKIS e COOK, 2002; CHATHURANGA e

JAYARATNE, 2013). Segundo Tzanetakis e Cook (2002), para cada janela o sinal é decomposto em um número de bandas de frequência utilizando uma transformada discreta, geralmente a DWT (*Discrete Wavelet Transform*). A DWT é uma transformação que provê uma compacta representação do sinal em tempo e frequência, com o objetivo de calcular de forma eficiente uma decomposição do sinal em frequência. Após, para cada banda, a amplitude no domínio do tempo é extraída em *y* realizando os seguintes passos no áudio da banda *x*:

- Retificação de onda completa, de equação:

  $$y[n] = |x[n]|$$

  para extrair a informação temporal do sinal ao invés do próprio domínio do tempo;
- A fim de suavizar o sinal, um filtro passa-baixa é aplicado no sinal, conforme:

  $$y[n] = (1 - \alpha)x[n] + \alpha y[n-1]$$

  com um filtro de valor α;
- Com o objetivo de diminuir o tempo computacional da auto-correlação sem afetar o algoritmo, a taxa de amostragem do sinal é diminuída através de:

  $$y[n] = x[kn]$$

  onde *k* é o fator de redução do sinal;
- Com o objetivo de centralizar o sinal próximo à zero para a fase da auto-correlação, a média do sinal é removida:

  $$y[n] = x[n] - E[x[n]]$$

Após, os resultados de cada banda são somados, reconstruindo um novo sinal *x* com as bandas modificadas. Em seguida, é aplicada uma auto-correlação neste novo sinal:

$$y[k] = \frac{1}{N}\sum_{n} x[n]x[n-k]$$

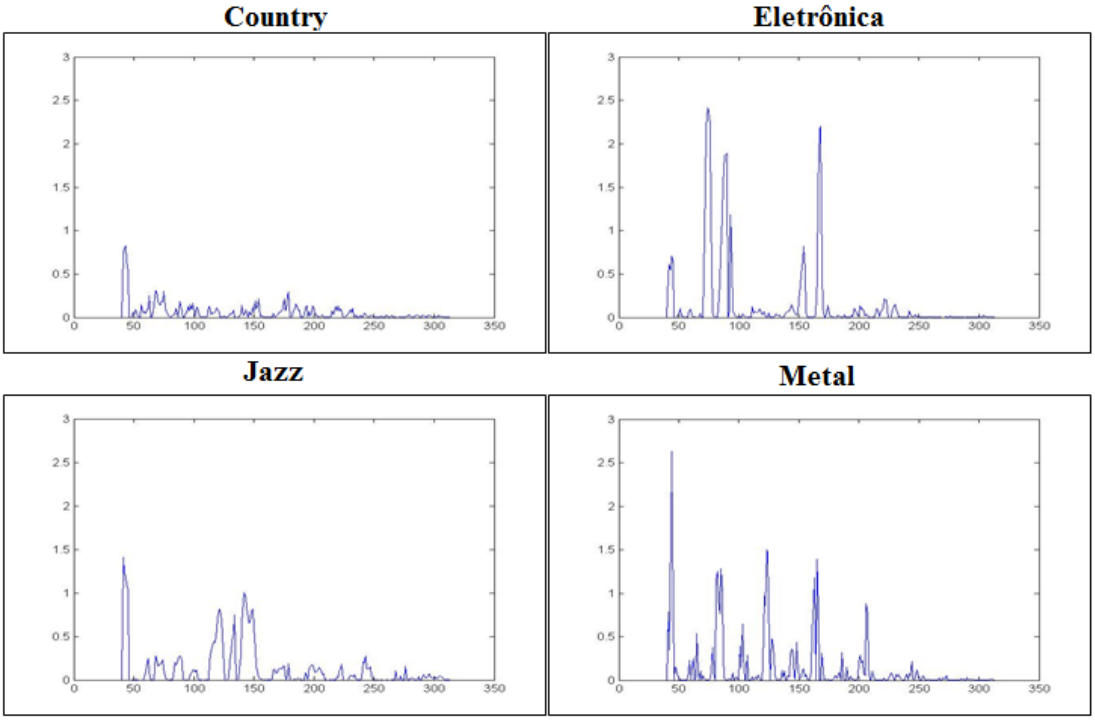
que resulta nas periodicidades da batida. Depois o resultado da auto-correlação é reforçado para reduzir o efeito de inteiros múltiplos das periodicidades. Para isto, o resultado da auto-correlação é transformado em valores positivos. Em seguida, seu domínio de tempo é multiplicado por um fator *w* e em seguida subtraído do resultado original da auto-correlação:

$$y[n] = x[n] - x[wn].$$

O mesmo processo é repetido com outros fatores inteiros, removendo assim picos repetidos em intervalos regulares. Ao final dos cálculos, têm-se as forças das batidas em várias velocidades na janela de áudio analisada. O processo é repetido para as outras janelas.

Em seguida, é feita a junção das forças das batidas de cada janela para formar o histograma da batida. Neste ponto, várias podem ser as formas de construção. Tzanetakis e Cook (2002), por exemplo, escolheram os três maiores picos de cada janela e adicionaram estes valores no histograma, capturando assim somente as batidas mais fortes. Já Chathuranga e Jayaratne (2013) somaram todas as bandas e fizeram desta soma o histograma, capturando assim todas as forças de batida. A Figura 3.10 mostra o histograma rítmico para quatro músicas de diferentes gêneros musicais de uma das bases de dados utilizadas neste trabalho. O eixo *x* mostra o número de batidas por minuto (BPM) da música, enquanto que o eixo *y* mede a força da batida.

Figura 3.10 - Exemplos de histogramas de batida.



Fonte: Autor.

A partir do histograma, várias informações podem ser calculadas. Silla *et al.* (2004) listam algumas características que podem ser extraídas:

a) Amplitude relativa (amplitude dividida pela soma das amplitudes) do primeiro e segundo picos do histograma: indica quão distintas são as batidas comparadas com o resto do sinal;
b) Razão da amplitude do segundo pico dividida pela amplitude do primeiro: expressa a relação entre a batida principal e a primeira batida auxiliar;
c) Período entre o primeiro e segundo picos em BPM: indica quão rápida é a música;
d) Soma do histograma: representa a força da batida.

Grimaldi *et al.* (2006) também apresentam outras informações que podem ser extraídas, como posição, intensidade e largura dos principais picos, além da quantidade total de picos do histograma.

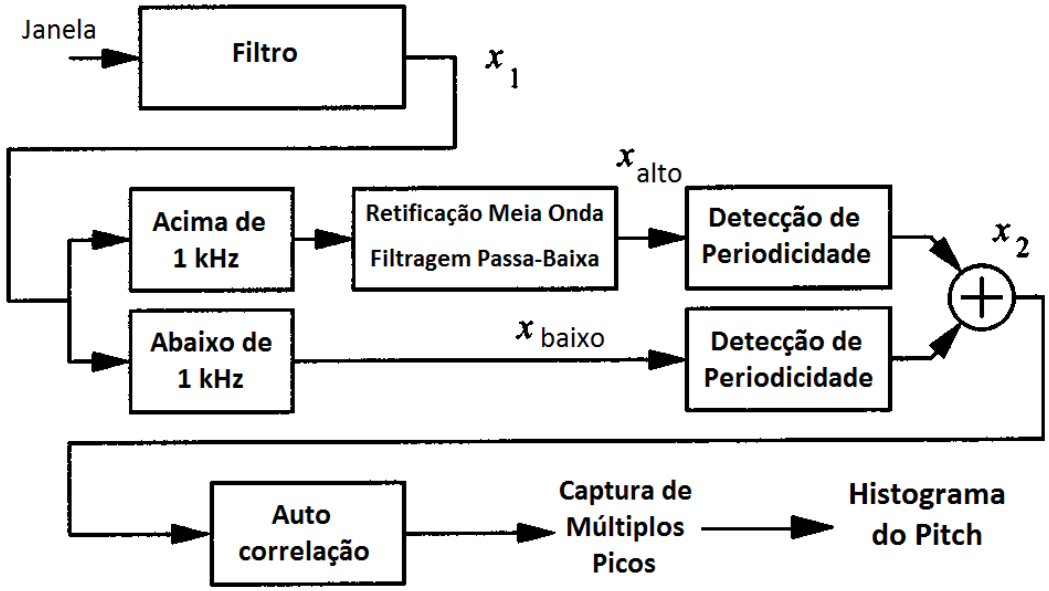
3.3.2 PITCH

Pitch é a percepção da frequência fundamental de uma nota musical. É utilizada também para a obtenção da melodia ou harmonia de uma música (PLUMBEY e DIXON, 2012). Porém, conforme Chathuranga e Jayaratne (2013), apesar de pitch estar relacionado a frequência fundamental, eles não são equivalentes, pois frequência é um conceito objetivo e científico, enquanto que pitch é subjetivo. Dado um determinado trecho de música polifônica (múltiplas notas tocadas simultaneamente), o pitch é utilizado para determinar qual nota é a que recebe mais destaque e é percebida por uma pessoa.

As características relacionadas ao pitch são frequentemente analisadas nos trabalhos de reconhecimento de gêneros musicais (BAGCI e ERZIN, 2007; TZANETAKIS e COOK, 2002). Para a extração destas características, é necessário o uso de um algoritmo de detecção de pitch, o qual pode ser posteriormente representado de várias maneiras, como através de um histograma de múltiplos pitch ou por um mapeamento do pitch em classes principais.

Uma das formas de uso de informações do pitch no reconhecimento de gêneros é através de um histograma, que faz uso de um algoritmo de detecção de múltiplos pitch. Este modelo, proposto por Tolonen e Karjalainen (2000) e adaptado por Tzanetakis e Cook (2002), é representado pela Figura 3.11.

Figura 3.11 - Diagrama de montagem do histograma do pitch (adaptado de Tolonen e Karjalainen, 2000).



Fonte: Autor.

O processo de obtenção do histograma do pitch proposto por Tolonen e Karjalainen (2000) utiliza a auto-correlação da periodicidade do sinal filtrado. Primeiro, o sinal é dividido em pequenas janelas, geralmente de 100 ms ou menos. Para cada janela, o sinal é filtrado para remover a correlação do sinal. Em seguida, o sinal é decomposto em duas bandas de frequência (abaixo e acima de 1 kHz) e envelopes de amplitude são extraídos para a banda alta aplicando retificação de meia onda e filtragem passa-baixa. Após, é calculada a detecção de periodicidade de cada banda através de uma transformada de Fourier discreta (DFT), compressão da magnitude da representação espectral, e uma transformada inversa (IDFT). O sinal x_2 é obtido pela equação:

$$x_2 = IDFT(|DFT(x_{baixo})|^k) + IDFT(|DFT(x_{alto})|^k)$$
$$= IDFT(|DFT(x_{baixo})|^k + |DFT(x_{alto})|^k)$$

onde x_{baixo} e x_{alto} são as bandas abaixo e acima de 1 kHz e k é o parâmetro que determina a compressão da magnitude. As bandas então são unidas. Na adaptação de Tzanetakis e Cook (2002), em seguida a auto-correlação das bandas somadas é estimada. Finalmente, o sinal é reforçado para reduzir o efeito de múltiplos inteiros no pico das frequências (método similar ao utilizado na construção do histograma rítmico). Assim, para cada janela, têm-se a força do pitch nas mais diversas frequências. Após as

forças do pitch de cada janela são unidas para formar o histograma (esta junção é semelhante à realizada no histograma rítmico, onde se podem capturar os picos dominantes de cada janela ou somar a força de todas elas).

Outro conceito importante é a divisão do histograma do pitch em classes, onde cada classe representa uma faixa de frequências, indicando uma nota musical. Dependendo do algoritmo utilizado para detecção do pitch, notas iguais em oitavas diferentes (por exemplo, uma nota "dó" mais grave e outra mais aguda) podem ficar ou não em classes iguais (FU *et al.*, 2011).

A partir do histograma, várias informações podem ser obtidas. Tzanetakis e Cook (2002) listam algumas destas informações:

a) Amplitude do pico máximo: indica a classe dominante na música, correspondendo à tônica ou acorde dominante;
b) Período do pico máximo: indica a quantidade de oitavas da classe dominante;
c) Intervalo entre os dois picos mais predominantes: corresponde à relação de intervalo entre os acordes;
d) Soma de todo o histograma: representa a força da detecção do pitch.

3.3.3 HARMÔNICAS

Características harmônicas são obtidas através da identificação de notas nas músicas. Scaringella *et al.* (2006) afirmam que as características harmônicas são difíceis de serem obtidas, devido justamente à complexidade de identificar notas musicais. Durante a música, vários são os instrumentos e fontes de áudio que emitem ao mesmo tempo diversas notas (muitas vezes diferentes), tornando complicada a separação destas notas. Devido à dificuldade de obtenção destas características, elas não são muito utilizadas, embora apareçam nos trabalhos de Ariyaratne e Zhang (2012) e Benetos e Kotropoulos (2010).

Esta classe é representada principalmente por duas características:

- **Melodia:** é uma sequência de notas, usualmente representando a tonalidade de um pedaço de música. Nesta característica, além da sequência de notas,

pode ser identificado também o intervalo entre elas (PLUMBEY e DIXON, 2012).

- **Harmonia:** obtida através da progressão de acordes, a harmonia se refere às relações entre sequências de acordes, que são a combinação de duas ou mais notas tocadas simultaneamente (PLUMBEY e DIXON, 2012). Fu *et al.* (2011) complementam que, em contraste da melodia, que captura a informação horizontal da música, a harmonia se preocupa em explorar a dimensão vertical da canção.

3.3.4 DESEMPENHO DAS CARACTERÍSTICAS DE MÉDIO NÍVEL

A Tabela 3.2 mostra a utilização e a importância das características de médio nível para a melhora do desempenho dos sistemas de reconhecimento de gêneros musicais. Com base nos resultados apresentados, é possível concluir que:

- O desempenho individual das características de médio nível fica aquém se comparado às características de baixo nível, conforme Li e Ogihara (2006);
- Características rítmicas, quando utilizadas individualmente, apresentam, de modo geral, baixo desempenho, enquanto que características relacionadas ao pitch apresentam melhor desempenho, como mostram os resultados de Yaslan e Cataltepe (2006);
- O uso combinado de características de médio nível eleva a taxa de acerto. No trabalho de Li e Ogihara (2006), o uso combinado de características rítmicas e de pitch elevou a taxa de acerto para pouco mais de 6% quando comparado ao desempenho individual destas;
- O uso de características de médio nível associada com as de baixo nível melhora o desempenho, quando comparado ao uso individual de cada uma delas. No trabalho de Li *et al.* (2003), por exemplo, a combinação de todas as características aumentou o desempenho em mais de 20% quando comparado ao melhor desempenho individual das de médio nível.

Tabela 3.2 - Desempenho de características de médio nível.

TRABALHO	CARACTERÍSTICAS	NÚMERO DE GÊNEROS	TAXA DE ACERTO
Li *et al.* (2003)	Rítmicas	5	44,52%
	Pitch	5	39,37%
	MFCC + Rítmicas + Pitch + Spectral Centroid + Spectral Rolloff + Spectral Flux + Zero-Crossing Rate + Low Energy	5	69,19%
Paradzinets *et al.* (2009)	Rítmicas	6	66,00%
Tzanetakis *et al.* (2001)	Rítmicas	10	42,00%
Tzanetakis e Cook (2002)	Pitch	10	23,00%
	Rítmicas	10	28,00%
	Spectral Centroid + Spectral Rolloff + Spectral Flux + Zero-Crossing Rate + Low Energy	10	45,00%
	MFCC	10	47,00%
	MFCC + Rítmicas + Pitch + Spectral Centroid + Spectral Rolloff + Spectral Flux + Zero-Crossing Rate + Low Energy	10	59,00%
Basili *et al.* (2004)	Melodia e harmonia	10	60,00%
Li e Ogihara (2006)	Rítmicas	10	26,50%
	Pitch	10	36,60%
	MFCC	10	58,40%
	Spectral Centroid + Spectral Rolloff + Spectral Flux + Zero-Crossing Rate + Low Energy	10	61,80%
	MFCC + Pitch	10	64,40%
	Rítmicas + Pitch	10	42,70%
	MFCC + Rítmicas	10	60,40%
	MFCC + Rítmicas + Pitch + Spectral Centroid + Spectral Rolloff + Spectral Flux + Zero-Crossing Rate + Low Energy	10	71,90%
Yaslan e Cataltepe (2006)	Rítmicas	10	29,00%
	Pitch	10	70,00%

4. CLASSIFICAÇÃO

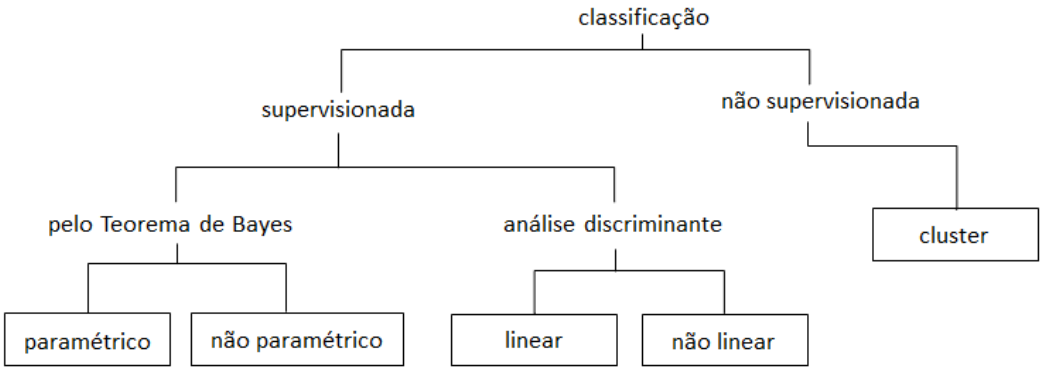
Este capítulo apresenta a etapa de classificação, mostrando os conceitos principais de um sistema de classificação e algoritmos mais utilizados. A Seção 4.1 mostra o conceito básico de um classificador e apresenta uma forma de categorizar os classificadores. A Seção 4.2 apresenta a classificação não supervisionada e os algoritmos de clusterização. A Seção 4.3 apresenta a classificação supervisionada, mostrando os tipos de classificadores, a forma de funcionamento de cada um destes tipos, e alguns dos algoritmos mais utilizados no reconhecimento de gêneros musicais. Por fim, a Seção 4.4 mostra uma revisão bibliográfica com a utilização e desempenho de alguns classificadores em estudos relacionados com este trabalho.

4.1 CLASSIFICADORES

Classificadores são operadores de mapeamento vários-para-um que projetam diversas características em uma dimensão de decisão. A forma exata de mapeamento e de como é feita a decisão depende do tipo do classificador (KIL e SHIN, 1996). A função dos classificadores é de classificar dados em determinadas categorias (ou classes), que são uma coleção de objetos que são similares, mas não necessariamente idênticos, e que podem ser distinguíveis de outras classes (DOUGHERTY, 2013).

Os classificadores podem ser agrupados usando a metodologia mostrada na Figura 4.1 (WEBB, 2002). O primeiro nível de agrupamento divide os classificadores em dois grandes grupos: supervisionados e não supervisionados. Esta divisão trata da forma de como os classificadores são construídos.

Figura 4.1 - Organização dos classificadores (adaptado de Webb, 2002).



Fonte: Autor.

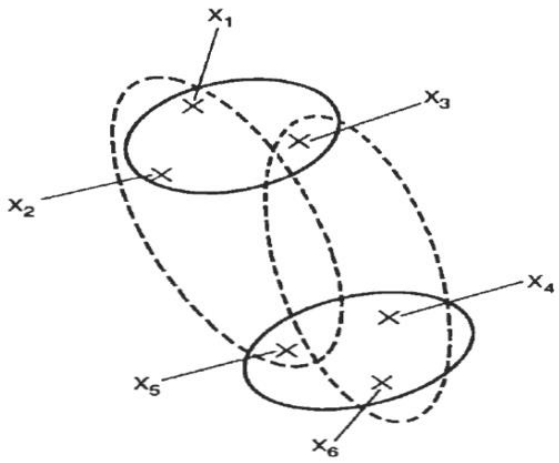
4.2 CLASSIFICAÇÃO NÃO SUPERVISIONADA

A classificação não supervisionada é aquela em que a classe dos dados de entrada não é informada ou é desconhecida. Com isso, cabe ao classificador utilizar as próprias características dos dados para distinguir um grupo dos outros (WEBB, 2002). Para Scaringella *et al.* (2006), uma vantagem é não precisar ter uma taxonomia fixa para a classificação, taxonomia esta que pode conter ambiguidades ou inconsistências. Por outro lado, Theodoridis e Koutroumbas (2003) explicam que um dos problemas desta classificação é definir qual é a similaridade a ser usada para diferenciar os dados, e escolher uma medida apropriada para ela.

4.2.1 CLUSTERIZAÇÃO

Os algoritmos de classificação não supervisionados são conhecidos como algoritmos de clusterização. São algoritmos iterativos que leem dados (geralmente vetores) e, baseados em critérios que definem medidas para separar as classes, agrupam estes dados (FUKUNAGA, 1990). A Figura 4.2 mostra um exemplo do funcionamento destes algoritmos. A partir dos dados recebidos (x_1 a x_6), o algoritmo procura agrupar estes dados, e vários podem ser os resultados da montagem destes grupos.

Figura 4.2 - Exemplo de clusterização.



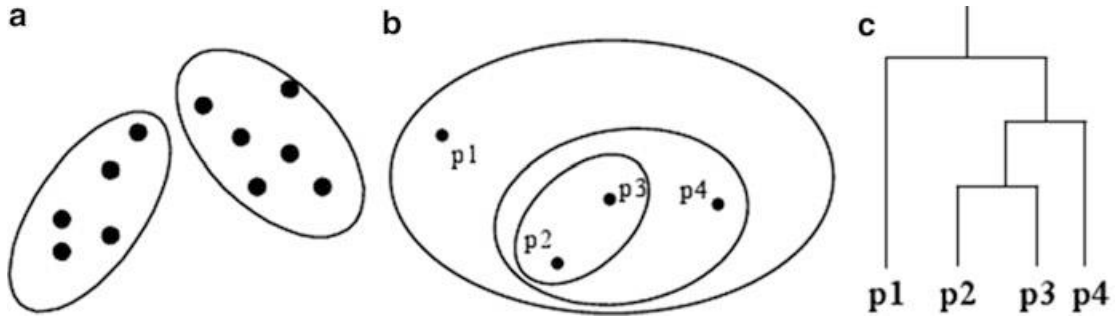
Fonte: Fukunaga (1990).

Para a construção de um cluster é necessário primeiramente, após a seleção dos dados a serem utilizados, definir a medida de proximidade, que representa a distância entre os pontos no espaço. Existem várias formas de calcular estas medidas de proximidade. Dados dois vetores de características representados pelos pontos (x_1, y_1) e (x_2, y_2) em um espaço bidimensional, Dougherty (2013) lista algumas distâncias:

a) Norma L_1 (Manhattan ou distância cidade-bloco), dada por $|x_1 - y_1| + |x_2 - y_2|$;
b) Norma L_2 (distância Euclidiana), de fórmula $\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$;
c) Norma L_∞ (Chebyshev ou distância do tabuleiro de xadrez), dado por $max\{x_1 - y_1, x_2 - y_2\}$.

O passo seguinte é definir o critério de clusterização, que indica como os dados serão agrupados (THEODORIDIS e KOUTROUMBAS, 2003). Deste critério, Dougherty (2013) mostra que duas são as principais formas de agrupamento: (1) partição, onde os dados são divididos em grupos diferentes e um dado pertence a um único grupo; e (2) hierárquico, onde grupos podem estar dentro de outros grupos. A Figura 4.3 mostra as duas formas de agrupamento, onde a Figura 4.3a apresenta a forma de partição, a Figura 4.3b apresenta o agrupamento hierárquico e a Figura 4.3c mostra também o agrupamento hierárquico representado na forma de um dendrograma. O dendrograma é uma representação de uma árvore hierárquica que guarda a sequência de partições realizadas nos dados (DOUGHERTY, 2013).

Figura 4.3 - Formas de agrupamento de um cluster: (a) partição e (b) hierárquico, representado por um (c) dendrograma.



Fonte: Dougherty (2013).

Outra questão a ser decidida é o critério de parada do cluster. O objetivo é que para cada partição, as amostras sejam de alguma forma, mais similares entre elas do que às amostras das outras partições. Assim, é necessário um critério que defina a qualidade aceitável de similaridade em cada partição dos dados, e este critério deve ser maximizado durante a construção do cluster. Vários são os critérios utilizados, como por exemplo, a soma do erro, variância mínima e dispersão (DUDA *et al.*, 2001).

As seguintes etapas para a construção de um cluster são a escolha do algoritmo de clusterização, validação e interpretação dos resultados. Em muitos casos, outra etapa necessária antes da construção do cluster é a tendência à clusterização, que inclui vários testes para indicar se os dados disponíveis são aptos para serem classificados em um cluster (THEODORIDIS e KOUTROUMBAS, 2003).

Conforme Scaringella *et al.* (2006), o algoritmo de clusterização mais conhecido é o K-means. No reconhecimento de gêneros musicais, o K-means já foi utilizado no trabalho de Haggblade *et al.* (2012), onde foi obtido uma taxa de 83% de acerto para 4 gêneros musicais.

4.3 CLASSIFICAÇÃO SUPERVISIONADA

Diferentemente da classificação não supervisionada, onde não é informada previamente a classe dos dados de entrada, na classificação supervisionada os dados possuem rótulos que indicam qual é a classe de cada dado de entrada (WEBB, 2002). Este tipo de classificação é baseado nestes rótulos, e procura mapear os dados de entrada em um algoritmo de aprendizado de máquina. O sistema é treinado com base

nestes rótulos, e depois é utilizado para classificar itens que não foram utilizados no treinamento (SCARINGELLA *et al.*, 2006).

A classificação supervisionada pode ser dividida em dois grandes grupos: que utilizam o Teorema de Bayes e os que utilizam uma função discriminante (WEBB, 2002).

4.3.1 TEOREMA DE BAYES

Conforme Kil e Shin (1996) e Webb (2002), para um problema de decisão envolvendo duas ou mais classes, o objetivo de qualquer decisão estatística é comparar a probabilidade de ocorrer a classe *c*, dado um vetor de características *y*. A classe com maior probabilidade é a escolhida, seguindo, para duas classes, a ideia de

$$y \in c_i \quad se\ p(c_i|y) > p\left(c_j\middle|y\right), i \neq j.$$

Uma das possíveis formas de calcular essa probabilidade é através do Teorema de Bayes, dado por

$$p(c_i|y) = \frac{p(y|c_i)P(c_i)}{p(y)}$$

onde $p(y|c_i)$ é a probabilidade do dado *y* pertencer a classe c_i, $P(c_i)$ é a probabilidade da classe c_i ocorrer, e $p(y)$ é a probabilidade do dado *y* (geralmente a probabilidade do dado *y* é conhecida e é uma constante).

Segundo Kil e Shin (1996), para escolher a classe com a maior probabilidade, muitas vezes é necessário atribuir custos para cada decisão tomada. Para o Teorema de Bayes, quatro são os riscos quando são consideradas duas classes:

1. R_{11} = risco em atribuir *y* para c_1 quando $y \in c_1$;
2. R_{12} = risco em atribuir *y* para c_2 quando $y \in c_1$;
3. R_{21} = risco em atribuir *y* para c_1 quando $y \in c_2$;
4. R_{22} = risco em atribuir *y* para c_2 quando $y \in c_2$;

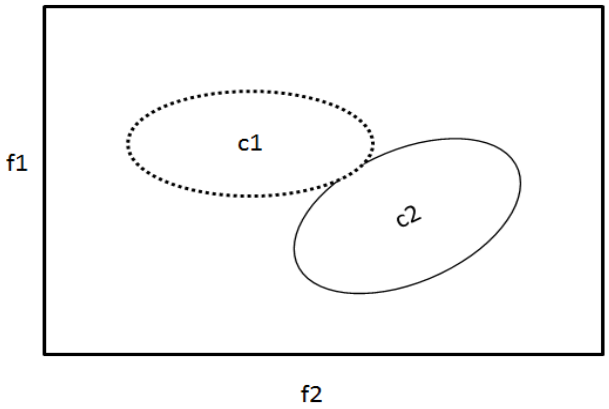
Desta forma, é possível penalizar o classificador quando decisões incorretas são tomadas, minimizando o custo ou a probabilidade de erros. O risco total é calculado somando os quatro riscos.

Seguindo a categorização de Webb (2002), os classificadores que seguem o Teorema de Bayes podem ser divididos quanto à definição da função de densidade probabilística: paramétricos e não paramétricos.

CLASSIFICADORES PARAMÉTRICOS

Os classificadores paramétricos são aqueles que requerem funções de densidade de probabilidade para modelar cada classe, e estimam parâmetros destas funções, como média e desvio padrão, para separar as classes (DOUGHERTY, 2013). Webb (2002) complementa que os classificadores paramétricos assumem que é possível determinar a função de densidade da classe, mas seus parâmetros são desconhecidos, e por isso devem ser estimados. A Figura 4.4 mostra o objetivo dos classificadores paramétricos que, tendo as classes c_1 e c_2, representadas em um plano bidimensional de dimensões f_1 e f_2, determinam o domínio de cada classe, representado neste exemplo pelos polígonos tracejado e contínuo, respectivamente.

Figura 4.4 - Exemplo de separação de classes em um classificador paramétrico.

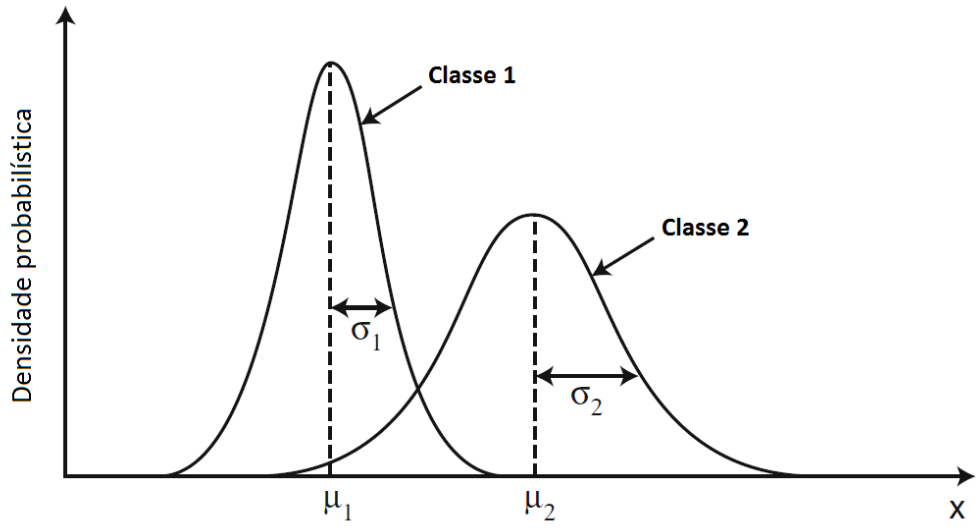


Fonte: Kil e Shin (1996).

Segundo Dougherty (2013), os parâmetros das funções de probabilidade das classes são determinados através do conjunto de características do treinamento. Na Figura 4.5, por exemplo, para duas classes c_1 e c_2, através de uma característica x, é possível estimar as funções de probabilidade de cada classe para aquela característica,

isto é, $p(x|c_1)$ e $p(x|c_2)$. Os parâmetros μ e σ determinam a média e a variância das funções de probabilidade de cada classe.

Figura 4.5 – Exemplo de funções de probabilidade (adaptado de Dougherty, 2013).



Fonte: Autor.

Os métodos paramétricos tendem a ser lentos no treinamento, mas muito rápidos na classificação dos dados de teste (DOUGHERTY, 2013). Kil e Shin (1996) listam outras características dos classificadores paramétricos:

- Geralmente são simples;
- Utilizam o cálculo de erro do Teorema de Bayes;
- Precisam estimar a função de probabilidade da classe.

O classificador paramétrico mais conhecido é o Modelo de Misturas Gaussianas (em Inglês, *Gaussian Mixture Model* - GMM). O GMM assume para cada classe a existência de uma função de densidade de probabilidade expressa como uma mistura de um número de distribuições multidimensionais Gaussianas (LI *et al.*, 2003). Um modelo de misturas gaussianas é a soma dos pesos das densidades de *M* componentes gaussianos, representado pela equação

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} w_i g(\boldsymbol{x}|\boldsymbol{m}_i, \Sigma_i)$$

onde $\boldsymbol{x}$ é o vetor de dados ou características, $w_i, i = 1, \dots, M$, são os pesos das misturas e $g(\boldsymbol{x}|\boldsymbol{m}_i, \Sigma_i), i = 1, \dots, M$, são as densidades dos componentes gaussianos, onde cada densidade é uma função gaussiana de forma

$$g(\boldsymbol{x}|\boldsymbol{m}_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_i)' \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{m}_i)\right\}$$

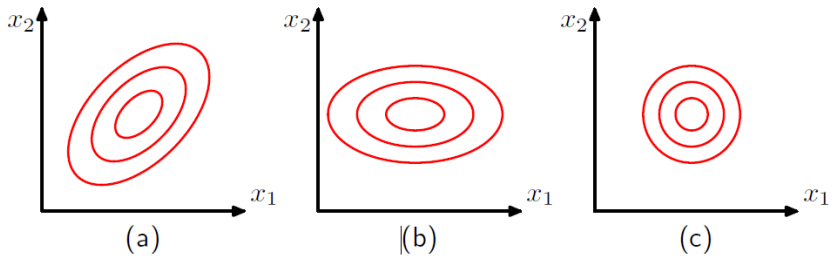
com vetor de médias $\boldsymbol{m}_i$ e matriz de covariância Σ_i, com pesos das misturas que satisfaçam $\sum_{i=1}^{M} w_i = 1$ (WEBB, 2002; REYNOLDS, 2001).

A covariância é uma medida de similaridade entre duas variáveis e de como elas variam uma em relação à outra. Quando não há similaridade, elas são independentes. Para um par de variáveis X_1 e X_2, a covariância entre elas é definida como

$$\sigma_{12}^2 = Cov(X_1, X_2) = E[(X_1 - \boldsymbol{m}_1) . (X_2 - \boldsymbol{m}_2)]$$

onde $\boldsymbol{m}_1, \boldsymbol{m}_2$ são as respectivas médias (ou valores esperados, $E[X_1], E[X_2]$). Logo, no GMM, a matriz de covariância provê uma forma sucinta de sumarizar as covariâncias entre todos os pares de variáveis do modelo (DOUGHERTY, 2013). Bishop (2006) explica que para um GMM com D dimensões, é necessário calcular $D(D + 1) / 2$ parâmetros para a matriz de covariância, o que é uma tarefa computacionalmente complexa à medida que o número de dimensões aumenta. Para resolver este problema, podem ser utilizadas formas restritas da matriz de covariância. Uma forma é considerar que a matriz de covariância é diagonal, tendo assim a matriz com apenas D parâmetros. Outra maneira é restringir a matriz de covariância a ser proporcional à matriz de identidade (covariância esférica ou isotrópica), resultando no cálculo de apenas 1 parâmetro para a matriz. Apesar de tornar o treinamento de um GMM mais rápido, a restrição da matriz de covariância limita a forma da função de densidade de probabilidade e diminui a capacidade do modelo de correlacionar os dados. A Figura 4.6 representa os contornos de probabilidades de densidade para três distribuições gaussianas de duas dimensões com cada tipo de matriz de covariância: a) completa; b) diagonal, com contornos elípticos alinhados com as coordenadas dos eixos; e c) esférica, proporcional a matriz de identidade.

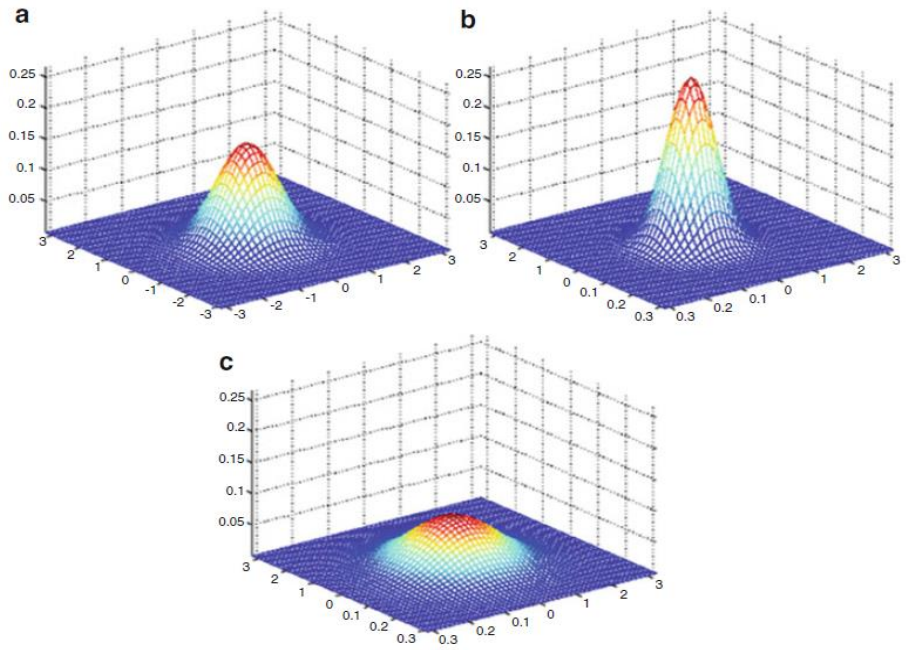
Figura 4.6 - Exemplos de funções de densidade de três GMM: completa, diagonal e esférica (isotrópica).



Fonte: Bishop (2006).

A Figura 4.7 mostra exemplos de GMM de duas dimensões com diferentes matrizes de covariância Σ. No exemplo a), $\Sigma = I$, onde I é a matriz de identidade; já em b) $\Sigma = 0{,}6\ I$; e em c) $\Sigma = 2\ I$.

Figura 4.7 - Exemplos de GMM.



Fonte: Dougherty (2013).

O algoritmo iterativo de maximização de expectativa (EM) é geralmente utilizado para estimar estes parâmetros, de forma a melhor modelar um número fixo de componentes gaussianas aos dados (BISHOP, 2006). A ideia básica do EM é, começando com um modelo inicial λ e para uma sequência de T vetores de treinamento

$X = \{x_1, \ldots, x_T\}$, estimar um novo modelo $\bar{\lambda}$ onde $p(X|\bar{\lambda}) \geq p(X|\lambda)$ (REYNOLDS, 2001).

O EM começa através da inicialização dos *M* componentes gaussianos, e em cada um deles a matriz de covariância é a matriz de identidade, os pesos das misturas são iguais a $1/M$, e as *M* médias são escolhidas aleatoriamente a partir dos dados (ALDER, 2001; HASTIE e TIBSHIRANI, 2008). Conforme Reynolds (2001), em seguida o EM é iterado, e a cada iteração é feita uma reestimativa destes parâmetros:

- Pesos das misturas:

$$\overline{w_i} = \frac{1}{T}\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)$$

- Médias:

$$\overline{\boldsymbol{m}_i} = \frac{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)\boldsymbol{x}_t}{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_{t,}\lambda)}$$

- Covariância:

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)x_t^2}{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_{t,}\lambda)} - \overline{m}_i^2$$

onde σ_i^2, x_t, e m_i referem-se a elementos arbitrários dos vetores $\boldsymbol{\sigma}_i^2$, $\boldsymbol{x}_t$, e $\boldsymbol{m}_i$, respectivamente. A probabilidade para o componente *i* é dada por

$$\Pr(i|x_t, \lambda) = \frac{w_i\ g(\boldsymbol{x}_t|\boldsymbol{m}_i, \Sigma_i)}{\sum_{k=1}^{M} w_k\ g(\boldsymbol{x}_t|\boldsymbol{m}_k, \Sigma_k)}.$$

A reestimativa dos parâmetros é realizada até que o algoritmo convirja, isto é, até que não haja mudança significativa nos números sendo estimados.

Apesar do algoritmo EM ser bastante utilizado, ele apresenta certos problemas, pois é muito vulnerável aos valores de inicialização, feitos de forma aleatória. Por isso muitas vezes é interessante rodar o algoritmo várias vezes para que se possa verificar o desempenho do algoritmo com valores diferentes, e achar a melhor configuração (ALDER, 2001).

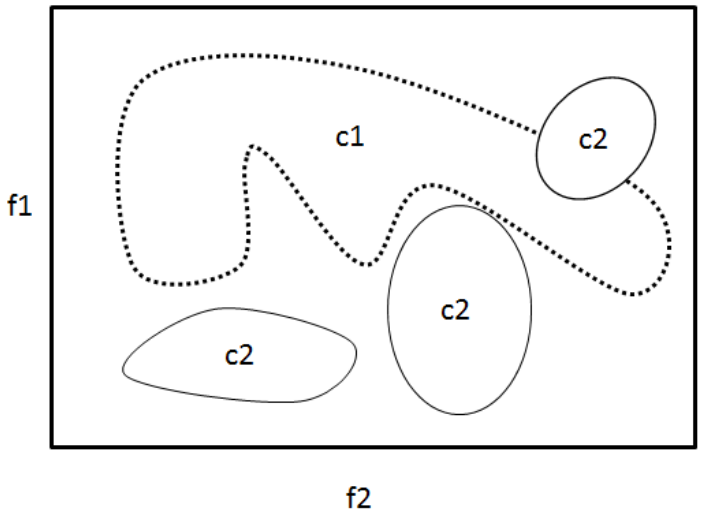
No reconhecimento de gêneros musicais, o GMM já foi utilizado, por exemplo, por Lee *et al.* (2009), Tzanetakis e Cook (2002) e Li e Ogihara (2006), utilizando 10 gêneros musicais, onde os resultados encontrados variaram entre 61% e 83%.

CLASSIFICADORES NÃO PARAMÉTRICOS

Diferentemente dos classificadores paramétricos que utilizam uma distribuição probabilística, os classificadores não paramétricos são utilizados quando estas probabilidades não são conhecidas (DOUGHERTY, 2013). Os classificadores não paramétricos utilizam os dados de treinamento para realizar uma aproximação da função de densidade de probabilidade desconhecida, através de métodos como o histograma de aproximação, *k-nearest neighbors*, expansão de funções base, entre outros (WEBB, 2002).

O uso de classificadores não paramétricos é interessante para situações onde há sobreposição de classes ou casos onde elas não são linearmente separáveis. Também são úteis quando as classes são altamente multimodais, isto é, apresentam amostras em várias regiões da dimensão de decisão, mescladas com amostras de outras classes (KIL e SHIN, 1996). A Figura 4.8 mostra um exemplo das aproximações das funções de densidade das classes. A partir de duas classes c_1 e c_2, representadas em um plano bidimensional de dimensões f_1 e f_2, os classificadores não paramétricos estimam um domínio (representado pelo polígono tracejado e pelos polígonos contínuos, respectivamente) para as classes, podendo inclusive a formar pequenos subdomínios para a mesma classe.

Figura 4.8 - Exemplo de separação de classes em um classificador não paramétrico.



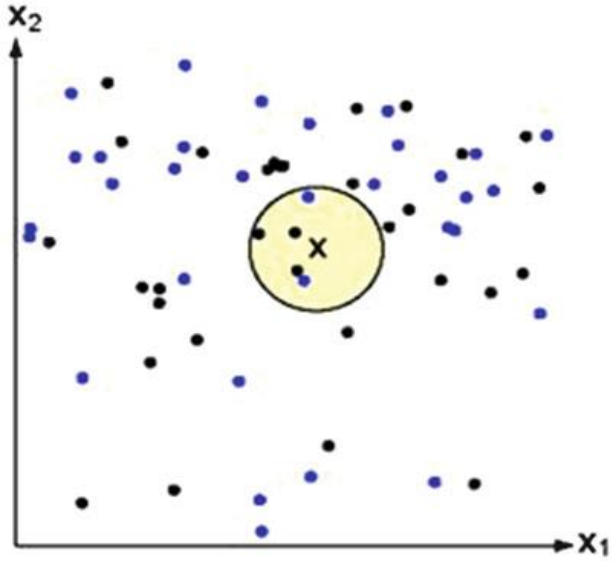
Fonte: Kil e Shin (1996).

As principais características dos classificadores não paramétricos são (KIL e SHIN, 1996):

- Realizam uma aproximação da função de densidade através dos dados;
- Geralmente realizam o treinamento de forma rápida, mas a classificação é mais demorada;
- Podem precisar de uma grande quantidade de dados de treinamento.

Um dos classificadores não paramétricos mais utilizados no reconhecimento de gêneros musicais é o *k-Nearest Neighbors* (kNN). Conforme Scaringella *et al.* (2006), o kNN se baseia na ideia de que um pequeno número de vizinhos influencia a decisão em um ponto. O processo do kNN começa no ponto de teste e a partir dele é verificado o raio de alcance até encontrar *k* amostras de treinamento. A partir disto, o ponto de teste é categorizado pela classe que mais aparece nas amostras de treinamento vizinhas (DOUGHERTY, 2013). A Figura 4.9 exemplifica o funcionamento do kNN onde, neste caso, com $k = 5$, o ponto *x* é categorizado na classe preta. Em casos de empate na quantidade de classes dos vizinhos, é atribuída aleatoriamente uma classe ao ponto. Este classificador foi utilizado nos trabalhos de Silla *et al.* (2007), Lee *et al.* (2009), Tzanetakis e Cook (2002), Li e Ogihara (2006) produzindo uma taxa de acerto entre 60% e 80% para 10 gêneros musicais.

Figura 4.9 – Exemplo de funcionamento do kNN.



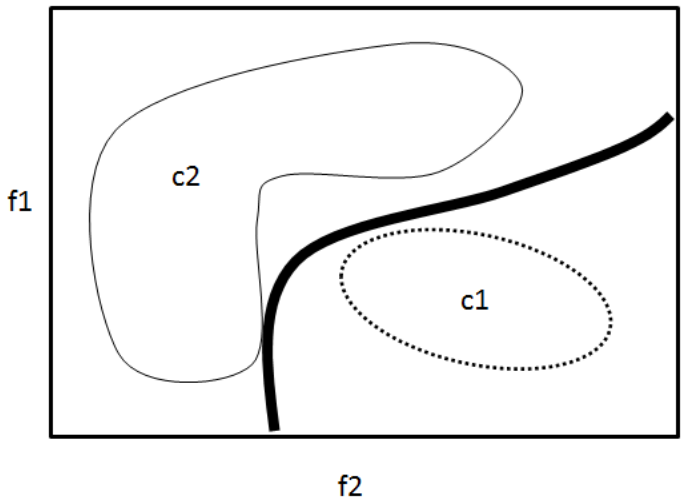
Fonte: Dougherty (2013).

4.3.2 ANÁLISE DISCRIMINANTE

O objetivo dos classificadores que utilizam a análise discriminante é modelar funções que consigam separar as classes (WEBB, 2002). Estes classificadores, também denominados de classificadores de decisões de fronteiras, procuram particionar a dimensão de características em várias regiões (KIL e SHIN, 1996). A Figura 4.10 mostra a ideia básica do funcionamento da análise discriminante, que apresenta duas classes c_1 e c_2, representadas em um plano bidimensional de dimensões f_1 e f_2, sendo particionadas por uma função separadora, representada pela linha mais espessa. Kil e Shin (1996) complementam que as características principais dos classificadores que seguem a análise discriminante são:

- Procuram estimar a função separadora que minimize ao máximo o erro;
- Requerem um longo período de treinamento;
- Possivelmente convergem para o mínimo local (convergem para uma boa solução, mas não necessariamente a melhor).

Figura 4.10 - Exemplo de funcionamento de classificadores baseados na análise discriminante.



Fonte: Kil e Shin (1996).

Estes classificadores podem ser divididos em dois tipos, conforme a função utilizada para separar as classes: linear e não linear.

CLASSIFICADORES LINEARES

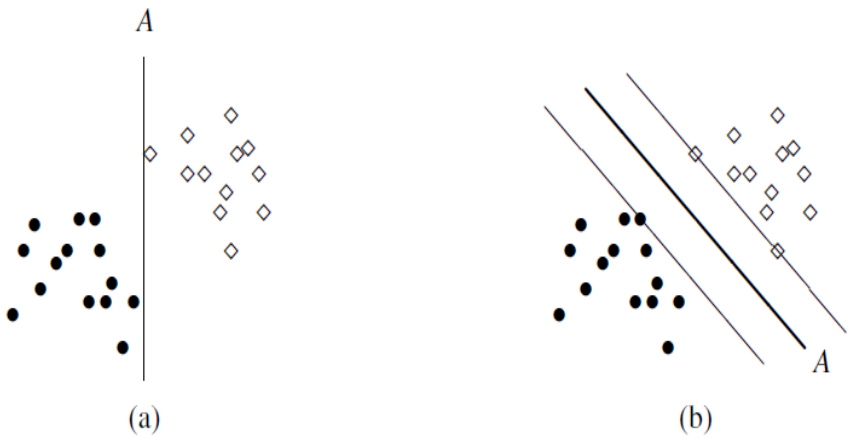
Os classificadores lineares são aqueles que, a partir das amostras do treinamento, procuram construir funções lineares para dividir as classes. A classificação de um objeto é obtida através de uma transformação linear onde os parâmetros são determinados por rotinas de otimização (WEBB, 2002). Theodoridis e Koutroumbas (2003) argumentam que uma boa vantagem destes classificadores é a simplicidade e o baixo custo computacional. Por outro lado, podem não obter bons desempenhos em situações onde as classes não estão linearmente separadas (como mostrado no exemplo anterior da Figura 4.10). Dougherty (2013) explica que em alguns casos não linearmente separáveis, é possível transformar as entradas para espaços com dimensões maiores, onde as classes podem ser linearmente separáveis.

Alguns exemplos de classificadores lineares mais utilizados são:

- ***Support Vector Machines*** **(SVM):** este classificador procura encontrar um hiperplano que classifique corretamente todos os vetores de treinamento (THEODORIDIS e KOUTROUMBAS, 2003). Webb (2002) complementa que o SVM mapeia os vetores de características para outro espaço dimensional maior onde o hiperplano possa ser construído, dando maior margem e aumentando a distância entre as classes.

  A Figura 4.11 mostra a ideia básica do SVM linear, onde a Figura 4.11b mostra a nova separação do SVM, onde a linha maior é o hiperplano de separação e as linhas mais finas identificam as margens.

Figura 4.11 - Exemplo de funcionamento do SVM linear.

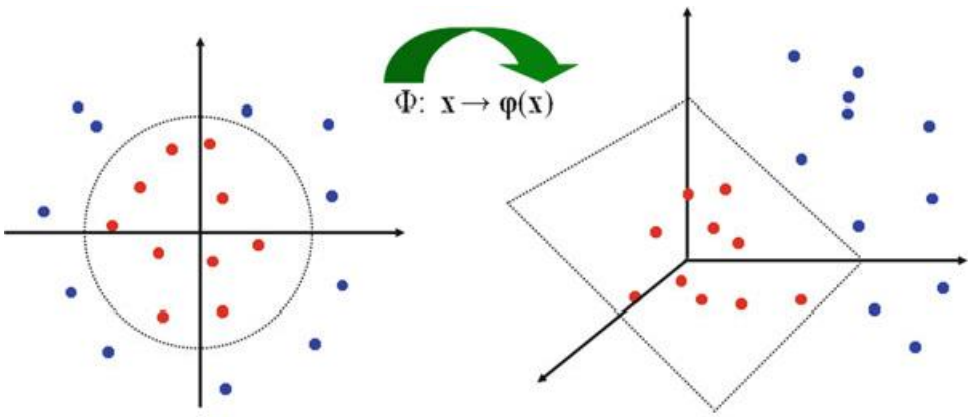


Fonte: Webb (2002).

Apesar de o SVM estar aqui citado como exemplo de classificador linear, ele também pode ser utilizado para hiperplanos não lineares. Uma possível solução é a introdução de variáveis de folgas (*slack variables*), permitindo margens mais leves (*soft margins*) (DOUGHERTY, 2013). Outra solução é mapear as características para um espaço dimensional maior, como mostrado na Figura 4.12. Isto é feito através de multiplicações internas entre vetores $\boldsymbol{x}_i^T \boldsymbol{x}_j$ utilizando uma função não linear $\boldsymbol{\varphi}$. Webb (2002) explica que as funções típicas utilizadas nestas multiplicações são:

- Polinomiais, de forma matemática igual a $(1 + \boldsymbol{x}^T \boldsymbol{y})^d$;
- Gaussianas, de equação $\exp(-|\boldsymbol{x} - \boldsymbol{y}|^2 / \sigma^2)$;
- Sigmoides, de fórmula $\tanh(k\boldsymbol{x}^T \boldsymbol{y} - \delta)$.

Figura 4.12 - Transformação de um SVM não linear para um espaço dimensional maior.



Fonte: Dougherty (2013).

Conforme Dougherty (2013) e Webb (2002), a equação de um hiperplano separador é dada por

$$\boldsymbol{w}^T\,\boldsymbol{\varphi}(\boldsymbol{x}) + b = 0,$$

onde $\boldsymbol{w}$ e b são parâmetros do modelo, com regra de decisão de que se esta equação resultar em valor positivo, então o dado $\boldsymbol{x}$ pertence a classe 1 com o valor correspondente de $y_i = 1$. Caso contrário, então o dado pertence a classe 2 com $y_i = -1$. A partir disto, segundo Dougherty (2013), também é possível calcular uma distância r de uma amostra $\boldsymbol{x}_i$ para o hiperplano através da equação

$$r = \frac{(\boldsymbol{w}^T\,\boldsymbol{\varphi}(\boldsymbol{x}) + b)}{\|\boldsymbol{w}\|}.$$

As amostras mais próximas do hiperplano são chamadas de vetores de suporte, e a margem de classificação dos separadores é a distância entre os vetores de suporte de diferentes classes. Segundo Webb (2002), o SVM determina a margem máxima através da maximização de um Lagrange. De acordo com Duda *et al.* (2001), a otimização de Lagrange procura a posição x_0 da extremidade de uma função escalar. Isto é feito através da equação

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \boldsymbol{\varphi}^T(\boldsymbol{x}_i)\boldsymbol{\varphi}(\boldsymbol{x}_j)$$

onde $y_i = \pm 1,\ \ i = 1, \dots, n$ são valores de indicador das classes e $\alpha_i, i = 1, \dots, n$, são multiplicadores Lagrange satisfazendo

$$0 \le \alpha_i \le C$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

para um parâmetro de regularização C que penaliza os erros de treinamento. A partir disto, a classificação de um novo dado $\boldsymbol{x}$ é feita de acordo com o sinal resultante de

$$g(\boldsymbol{x}) = \sum_{i \in SV} \alpha_i y_i \boldsymbol{\varphi}^T(\boldsymbol{x}_i)\boldsymbol{\varphi}(\boldsymbol{x}) + \boldsymbol{b}$$

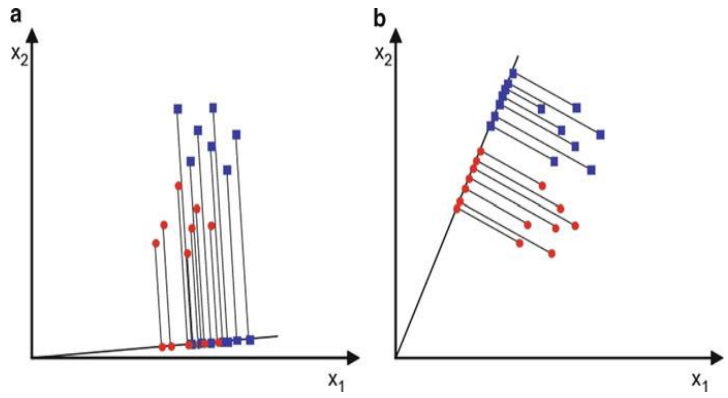
em que SV é o conjunto de vetores de suporte com valores associados de α_i satisfazendo $0 \le \alpha_i \le C$.

Portanto, a construção de um modelo SVM envolve a escolha da função kernel (para casos não lineares), os parâmetros do kernel e a escolha do parâmetro de regularização *C* (WEBB, 2002).

O SVM é bastante utilizado em trabalhos de reconhecimento de gêneros musicais. Aparece nos trabalhos de Silla *et al.* (2007), Benetos e Kotropoulos (2010) e Ariyaratne e Zhang (2012) onde, utilizando bases com 10 gêneros musicais, taxas de reconhecimento entre 64% e 77% foram obtidas.

- ***Linear Discriminant Analysis* (LDA):** a ideia básica do LDA é encontrar uma transformação linear que melhor separa as classes e fazer a classificação neste espaço transformado usando métricas como a distância Euclidiana (SCARINGELLA *et al.*, 2006). Dougherty (2013) complementa que o objetivo é encontrar uma direção em que, quando os dados são projetados nela, as amostras das classes estejam o mais separado possível, como mostrado na Figura 4.13.

Figura 4.13 - Exemplo de funcionamento do LDA.



Fonte: Dougherty (2013).

O LDA já foi utilizado em alguns trabalhos de reconhecimento de gêneros musicais, como nos trabalhos de Lee *et al.* (2009) e Li e Ogihara (2006), obtendo, respectivamente, uma taxa de reconhecimento de 82% e 71% para 10 gêneros.

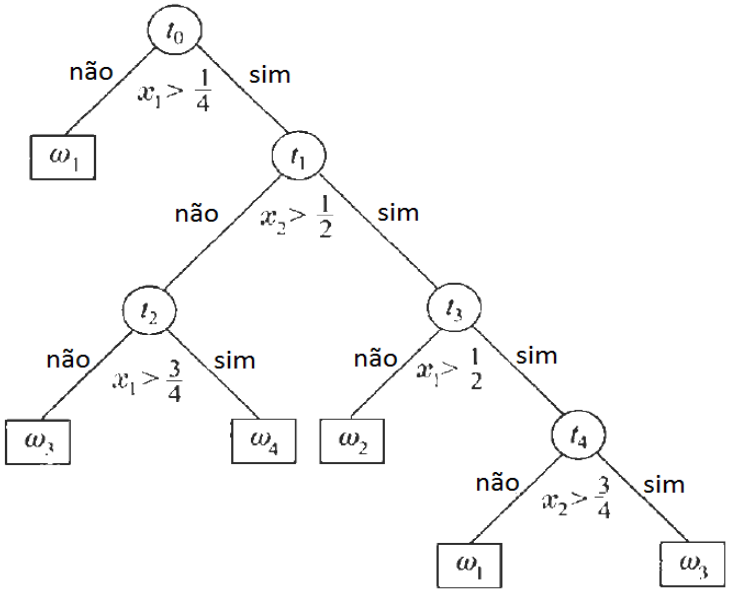
CLASSIFICADORES NÃO LINEARES

Os classificadores não lineares seguem a mesma lógica dos classificadores lineares, isto é, procuram delimitar fronteiras para separar as classes. Os classificadores não lineares, por sua vez, procuram separar estas classes utilizando funções não lineares. Os principais métodos de classificadores não lineares são: (1) kernel, utilizados onde o modelo é linear, mas a separação é não linear; (2) de projeção, onde as funções não lineares são construídas através da projeção linear dos dados; e (3) baseados em árvore, onde o objetivo é recursivamente particionar o espaço dos dados (WEBB, 2002).

Alguns classificadores não lineares mais utilizados são:

- ***Árvores de Decisão*:** as árvores de decisão são sistemas de decisão de vários estágios em que as classes vão sendo sequencialmente rejeitadas até o classificador alcançar a classe aceitável. Possuem nodos que direcionam a decisão do classificador, e folhas que determinam uma classe (THEODORIDIS e KOUTROUMBAS, 2003). Ainda, conforme Theodoridis e Koutroumbas (2003), alguns fatores devem ser adotados para a construção de uma árvore de decisão: (1) as questões a serem usadas para decidir se a classe deve ser aceita ou rejeitada; (2) um critério para dividir as classes; (3) uma regra para parar a decisão e um nodo para ser o último; e (4) uma regra que categoriza uma folha para uma determinada classe. A Figura 4.14 mostra um exemplo do funcionamento de uma árvore de decisão.

Figura 4.14 - Exemplo de funcionamento de uma árvore de decisão (adaptado de Theodoridis e Koutroumbas, 2003).

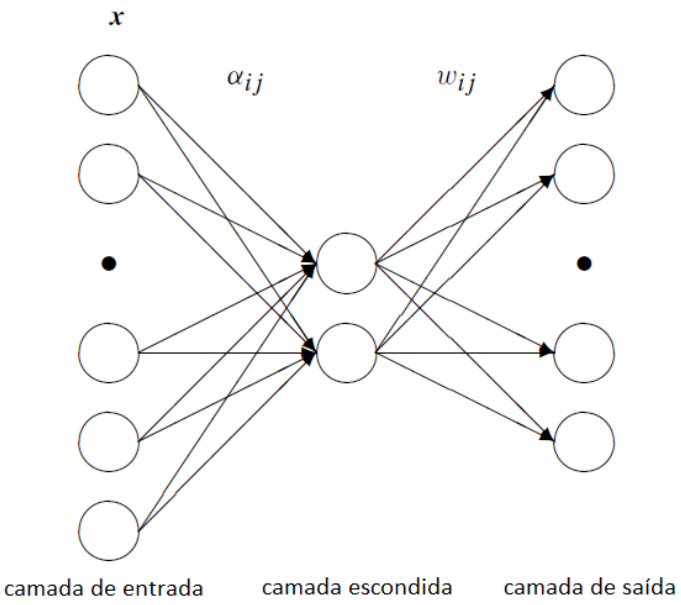


Fonte: Autor.

As árvores de decisão foram utilizadas nos trabalhos de Silla *et al.* (2004), Silla *et al.* (2007), Ariyaratne e Zhang (2012), entre outros. Nestes estudos, as taxas de reconhecimento variaram entre 47% e 66% para bases com 10 gêneros.

- ***Multilayer Perceptron* (MLP):** o MLP produz uma transformação de um vetor de características *x* para um espaço de *n* dimensões, sendo *n* o número de ligações entre as camadas da rede (WEBB, 2002). Sua estrutura é exemplificada na Figura 4.15, que mostra as três estruturas principais do MLP: a camada de entrada que recebe o vetor com os dados, e possui pesos entre as ligações dessa camada com a camada escondida, responsável por fazer a combinação dos pesos e realizar a transformação não linear. Por fim, a camada de saída forma uma combinação linear das saídas da camada escondida e gera a saída final do classificador. Em um MLP pode haver várias camadas escondidas, cada uma realizando uma transformação de um produto escalar da saída da camada anterior com os pesos entre as ligações. Ainda, é possível indicar para que a camada de saída possa fazer uma combinação não linear das saídas da camada escondida.

Figura 4.15 - Exemplo de funcionamento do MLP (adaptado de Webb, 2002).



Fonte: Autor.

O uso do MLP já foi reportado nos trabalhos de Silla *et al.* (2007), Lee *et al.* (2009), Ariyaratne e Zhang (2012), entre outros. Nestes trabalhos, as taxas de reconhecimento obtidas foram de 59%, 77% e 74%, respectivamente, para bases com 10 gêneros musicais.

4.4 DESEMPENHO DOS CLASSIFICADORES

A Tabela 4.1 mostra o desempenho obtido em alguns trabalhos utilizando os classificadores citados. Também é mostrado o uso de outros classificadores que, apesar de não terem sido explicados neste trabalho, são apresentados para fins comparativos. Algumas observações a serem consideradas são:

- Árvores de decisão geralmente possuem um baixo desempenho;
- Os classificadores kNN e GMM, apesar de obterem desempenhos melhores que as árvores de decisão, ainda possuem baixos desempenhos se comparado a outros classificadores;
- Os melhores desempenhos foram obtidos com MLP e SVM.

Tabela 4.1 - Comparativo de desempenho entre classificadores.

TRABALHO	CLASSIFICADOR	NÚMERO DE GÊNEROS	TAXA DE ACERTO
Reed e Lee (2006)	Hidden Markov	5	69,32%
Basili *et al.* (2004)	J48 (Árvore de Decisão)	6	60,00%
West e Cox (2004)	LDA	6	60,00%
	GMM	6	64,00%
Tzanetakis e Cook (2002)	kNN	10	60,00%
	GMM	10	61,00%
Li *et al.* (2003)	kNN	10	62,10%
	LDA	10	71,30%
	SVM	10	78,50%
Silla *et al.* (2007)	J48 (Árvore de Decisão)	10	44,44%
	Naive Bayes	10	47,76%
	kNN	10	56,26%
	MLP	10	56,40%
	SVM	10	63,50%
Benetos e Kotropoulos (2010)	MLP	10	77,00%
	SVM	10	77,20%
Ariyaratne e Zhang (2012)	J48 (Árvore de Decisão)	10	53,30%
	LOG	10	58,60%
	SVM	10	62,10%
	MLP	10	71,20%
Madjarov *et al.* (2012)	kNN	10	55,91%
	MLP	10	67,05%
	SVM	10	69,09%
Wülfing e Riedmiller (2012)	SVM	10	83,37%
Burred e Lerch (2003)	GMM	17	58,71%

5. RECONHECIMENTO UTILIZANDO CLASSIFICADORES BASEADOS EM MÚLTIPLAS CARACTERÍSTICAS

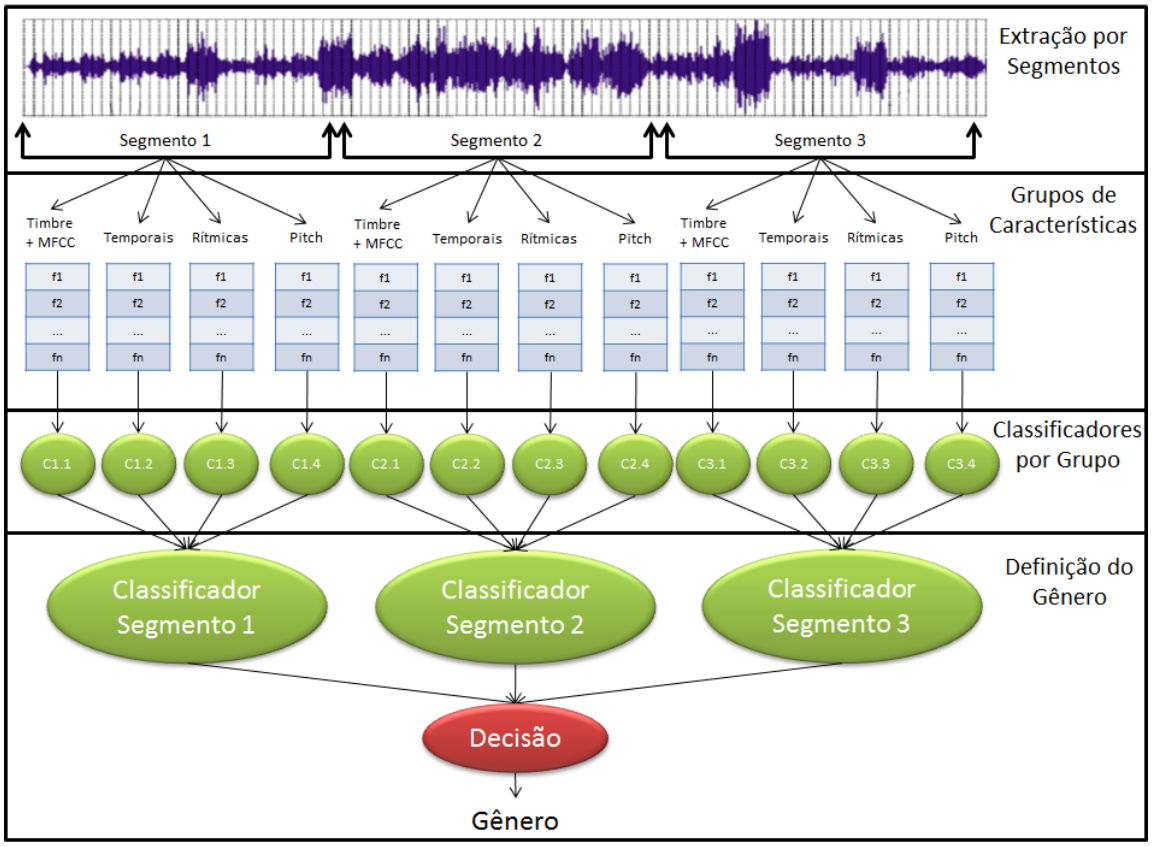
Neste capítulo são detalhadas todas as características do sistema proposto. A Seção 5.1 mostra a arquitetura do sistema, explicando como funciona cada etapa, as características extraídas, classificadores utilizados, e a forma de processamento destas informações para fazer a classificação dos gêneros musicais. A Seção 5.2 explana como foi realizado o desenvolvimento, especificando as bases de dados utilizadas no treinamento e teste do sistema, além de explicar como foram configurados os parâmetros de cada uma das etapas da arquitetura proposta. Por fim, as Seções 5.3 e 5.4 mostram os resultados obtidos nos mais diversos testes aplicados no sistema, além de considerações sobre estes resultados.

5.1 ARQUITETURA DO SISTEMA

A arquitetura do sistema proposto procura fazer uma mescla dos modelos de extração por segmentos com o agrupamento de características. Esta arquitetura se divide em quatro etapas, conforme mostrado na Figura 5.1:

a) **Extração por Segmentos:** divisão da música analisada em três segmentos;
b) **Grupos de Características:** para cada segmento, características são extraídas e divididas em grupos;
c) **Classificadores por Grupo:** uso de um classificador para cada grupo de características;
d) **Definição do Gênero:** combinação dos classificadores de cada segmento para definição do gênero da música analisada.

Figura 5.1 - Arquitetura do sistema proposto.



Fonte: Autor.

Para o sistema proposto, foi utilizada a abordagem de sistema fechado. Pouquíssimos são os trabalhos publicados que utilizaram sistemas abertos, e muito disso se deve à dificuldade de definir um critério mínimo de aceitação. Além disso, como a classificação de gêneros ainda não alcançou taxas de reconhecimento aceitáveis para uma aplicação real trabalhando somente com gêneros específicos, a utilização de um sistema aberto só dificultaria ainda mais o reconhecimento.

5.1.1 EXTRAÇÃO POR SEGMENTOS

O sistema proposto divide a música em três segmentos: no início, no meio e no fim da música. Este modelo de extração analisando três segmentos já foi utilizada em alguns trabalhos de reconhecimento de gêneros musicais, elevando a taxa de acerto. No trabalho de Costa *et al.* (2004), foi observado um aumento de 0,8% no acerto quando comparado ao desempenho do melhor segmento individual. Já para Silla *et al.* (2007), o

aumento médio foi de 3%. No trabalho de Costa *et al.* (2012), o uso dos segmentos teve uma importância muito significativa, elevando, no melhor caso, a taxa de acerto em 10% quando comparado ao desempenho do melhor segmento individual.

O modelo de extração por segmentos resolve dois dos principais problemas da análise completa de uma música, que são a demora no tempo de execução e a enorme quantidade de dados envolvidos. Porém, a extração de um único segmento representa certos riscos. Sabe-se que, apesar de uma música pertencer a um determinado gênero musical, ela pode conter elementos que pareçam ser de outro gênero. Por exemplo, uma música de Metal pode conter violinos ou instrumentos clássicos; músicas Pop podem ter um trecho mais dançante em que a batida muda, e é feito uso de efeitos eletrônicos; uma música de Jazz pode conter durante sua execução um solo de piano que pareça pertencer a uma música Clássica. Estes casos confundem o reconhecedor e podem induzir o sistema a classificar um gênero erroneamente. Além destas situações, no exemplo de uma classificação de músicas que foram gravadas ao vivo, o trecho analisado pode conter ruídos como sons da plateia, aplausos, entre outros, dificultando a análise da música. A forma de extração utilizando três segmentos no início, meio e fim da música evita estes problemas (se um segmento não for uma boa amostra do gênero da música, os outros dois compensam aquele segmento), e pode ajudar a aumentar a taxa de reconhecimento. Além disto, o uso deste modelo serve como uma estatística para definir qual o melhor segmento (início, meio e fim) que representa o gênero de uma música.

5.1.2 GRUPOS DE CARACTERÍSTICAS

Nesta etapa, para cada segmento, características são extraídas e organizadas em grupos, onde cada um possui um classificador próprio, que realiza o processamento somente daquelas características. Alguns trabalhos utilizaram o agrupamento de características, melhorando a taxa de acerto. Lee *et al.* (2009) dividiram as características em 3 grupos: MFCC, *Octave-Based Spectral Contrast* e *Normalized Audio Spectral Envelope*. Em algumas configurações do sistema testadas, o agrupamento de características apresentou resultados até 5% piores do que o melhor grupo individual, mas na maioria dos testes realizados a taxa de acerto melhorou, encontrando no melhor caso um acréscimo de mais de 9% quando comparado ao melhor

classificador individual. Já Paradzinets *et al.* (2009) utilizaram grupos para características de timbre, rítmicas e acústicas. A combinação dos grupos mostrou uma taxa de acerto 12% maior do que o melhor classificador individual. Balti e Frigui (2012) utilizaram três grupos de características (temporais, espectrais e rítmicas), onde cada grupo foi avaliado por um cluster próprio. Para cinco testes realizados, em um dos testes a combinação dos grupos apresentou resultado inferior de 2,6% em relação ao melhor desempenho individual. Nos demais, a taxa de reconhecimento aumentou para mais de 20% comparado com o reconhecimento do melhor classificador individual.

A ideia seguida pelo agrupamento de características é de que algumas características são mais discriminatórias para certos gêneros do que outros. Por exemplo, o ritmo de uma música é bastante distinto quando comparamos músicas Latinas e Hip Hop, mas não é tão simples diferenciar quando comparamos Metal e Rock. E esta ideia está evidenciada nas pesquisas de reconhecimento já publicadas. No trabalho de Meng *et al.* (2007) por exemplo, onde foi utilizado características temporais, o reconhecimento para Country foi de 72,70%, enquanto que Rock obteve apenas 29,10%. Paradzinets *et al.* (2009) utilizaram um classificador apenas para características rítmicas e testaram em duas bases de dados diferentes. Em ambos os testes, constatou-se uma alta taxa de acerto para música Clássica (82,80% e 89,70%), enquanto que o acerto para Rock foi bem inferior (54,40% e 52,70%). Já no trabalho de Lee *et al.* (2009), utilizando apenas MFCC, foi alcançado 92,81% de reconhecimento para músicas Clássicas, mas Rock obteve somente 63,63%. O mesmo aconteceu no trabalho de Reed e Lee (2006), onde utilizando somente MFCC o reconhecimento de músicas Clássicas foi de 92,90% e de Rock foi de 58,50%. Resultado similar foi encontrado no trabalho de Meng e Shawe-Taylor (2005) que, também utilizando apenas MFCC, obteve a pior taxa de acerto no gênero Rock (20% contra 54% alcançado no melhor gênero reconhecido, Country). Estas diferenças de reconhecimento entre um gênero e outro sugerem que as características possuem certas facilidades para reconhecer alguns gêneros e dificuldades para outros. O objetivo de separar em vários grupos é aproveitar os benefícios de cada característica.

Os grupos de características utilizados para o sistema proposto são:

1. **Relacionadas ao timbre e MFCC:** são extraídas características de baixo nível da classe *short-term*, isto é, que são analisadas em janelas menores. As

características analisadas para este grupo são: *spectral centroid*, *spectral rolloff*, *spectral flux*, *zero-crossing rate* e MFCC.

2. **Temporais:** é formado por características temporais, obtidas pela junção das características de timbre e MFCC. Neste grupo é formado um único vetor de características por segmento, utilizando principalmente momentos estatísticos. As características que constituem este grupo são: l*ow energy* e média e variância do *spectral centroid*, *spectral rolloff*, *spectral flux*, *zero-crossing rate* e MFCC.

3. **Rítmicas:** é constituído pelas características relacionadas com a batida da música. Para este grupo, foi construído o histograma rítmico utilizando a técnica explicada na Seção 3.3.1, a mesma utilizada no trabalho de Tzanetakis e Cook (2002). Em seguida, o histograma é analisado, e é montado um vetor por segmento, contendo diversas informações extraídas do histograma, que serão descritas na Seção 5.2.2.

4. **Pitch:** é constituído com características relacionadas ao pitch. Para isto, foi construído um histograma do pitch utilizando a técnica de Tolonen e Karjalainen (2000), explicada na Seção 3.3.2, também utilizada no trabalho de Tzanetakis e Cook (2002). Após a construção do histograma, este é analisado e é montado um vetor por segmento, com diversas informações de análise deste histograma, que serão descritas na Seção 5.2.2.

5.1.3 CLASSIFICADORES POR GRUPO

Seguindo a ideia do agrupamento de característica, cada um dos quatro grupos de cada segmento possui um classificador próprio (representado na Figura 5.1 pelos classificadores C1.1 a C1.4 para o primeiro segmento, C2.1 a C2.4 para o segundo, e C3.1 a C3.4 para o terceiro segmento), que analisa somente as características daquele grupo. Nesta fase, foi testado o uso do classificador paramétrico GMM e do não paramétrico SVM. Estes classificadores foram escolhidos por apresentarem melhores

taxas de acerto quando comparados com os demais classificadores em trabalhos já publicados. O SVM é restrito para a classificação de duas classes, portanto, para torná-lo apto a resolver problemas com mais de duas classes, foi utilizado à abordagem do "um contra todos". Nesta abordagem, para *x* classes, são feitas *x* classificações, e em cada classificação uma classe diferente é comparada com todas as outras. Já no GMM, para cada classificador, há um GMM para cada classe, modelando os dados pertinentes ao gênero.

SUPERVETOR

Um dos desafios encontrados durante o desenvolvimento da solução foi com relação ao uso do SVM no Grupo 1 de características (relacionadas ao timbre e MFCC), devido ao excesso de dados. Como as características deste grupo são formadas a partir da informação extraída diretamente de cada janela (que é de curta duração), cada segmento de 30 segundos contém uma quantidade muito grande de dados. Multiplicando esta quantidade para cada música utilizada no treinamento, isto equivale a um grande número de amostras, quantidade que um SVM não consegue suportar. Para contornar esse problema, foi utilizado o conceito de supervetores. A ideia do supervetor é combinar o uso do GMM com o SVM, onde os dados primeiramente são processados pelo GMM (que consegue lidar bem com grandes quantidades de amostras) e após as médias das misturas são concatenadas formando supervetores. Por fim, estes supervetores são passados para o SVM (GOLD *et al.*, 2011).

Na implementação da proposta de solução, exemplificada na Figura 5.2, supervetores foram utilizados da seguinte forma: para cada música, os dados relacionados ao timbre e MFCC foram extraídos normalmente. Após, estes dados foram passados para o GMM, que modela aqueles dados daquela música. Em seguida, as médias dos componentes das misturas foram concatenadas, tornando-se as novas características que foram passadas para o SVM, conforme representado na equação:

$$\boldsymbol{x} = [\boldsymbol{m}_1, \boldsymbol{m}_2, \dots, \boldsymbol{m}_N]$$

onde $\boldsymbol{x}$ são os novos dados utilizados no SVM, $\boldsymbol{m}_1, \boldsymbol{m}_2, \dots, \boldsymbol{m}_N$, são as médias dos componentes do GMM e $N = MC$, para *M* igual ao número de componentes gaussianas e *C* a quantidade de características. Assim, para cada música analisada, tem-se uma

amostra de características para ela. Supervetores apresentam-se como uma boa solução para a utilização de SVM com muitas amostras. Embora não haja relatos do uso desta técnica em reconhecimento de gêneros musicais, supervetores aparecem em trabalhos de reconhecimento de imagens e de locutor, como nos trabalhos de Zhou *et al.* (2009) e Campbell *et al.* (2005).

Figura 5.2 - Exemplo do uso de supervetores.

Fonte: Autor.

SAÍDA DOS CLASSIFICADORES

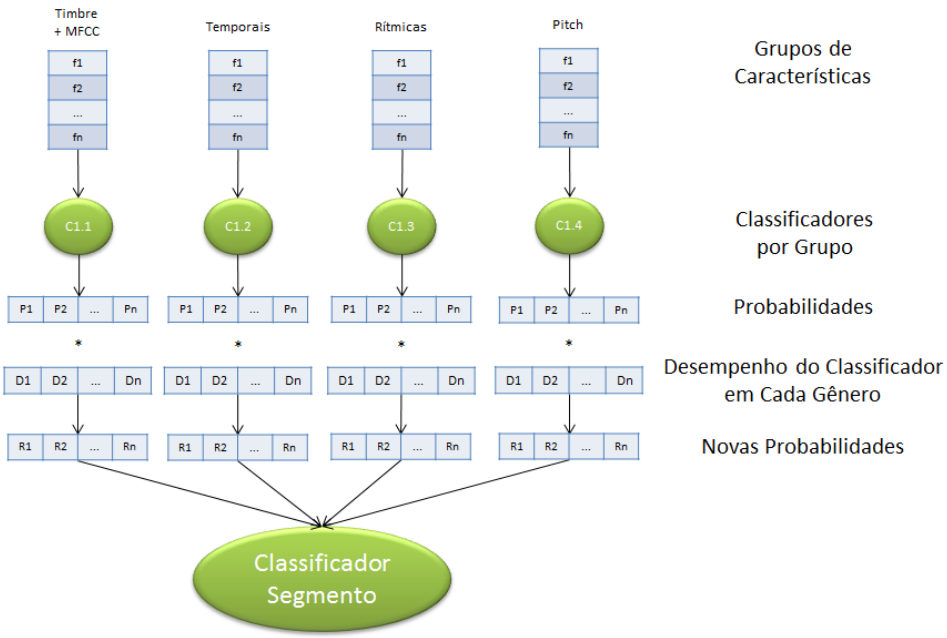
A saída dos classificadores desta fase são probabilidades do vetor de entrada pertencer a cada um dos gêneros envolvidos. Logo, para *x* gêneros analisados, a saída de cada classificador é um vetor de *x* dimensões. Os vetores de cada grupo são normalizados entre zero e um (para padronizar a faixa de valores das probabilidades, independentemente do tipo de classificador utilizado) e, após, unidos, formando um vetor de *4x* dimensões, que é passado para o classificador principal do segmento.

Para cada classificador foi utilizada uma abordagem diferente para capturar a probabilidade das amostras pertencerem a cada um dos gêneros. Para o GMM, foi calculada a função de densidade de probabilidade de cada classe. Já para o SVM, a abordagem foi diferente. Como o SVM não é um classificador que trabalha diretamente com probabilidades, as chances de uma amostra pertencer a uma determinada classe foram simuladas através da distância da amostra para o hiperplano separador. Como para o SVM foi utilizada a abordagem do "um contra todos", em um grupo de classificadores há um classificador SVM para cada gênero (Clássica contra o resto, Country contra o resto, etc.). Cada classificador utiliza o mesmo conjunto de dados, com a diferença em que em cada um deles um gênero é marcado como uma classe e os outros pertinentes a outra classe. A partir disto, dada uma amostra, sua distância é

calculada em cada um dos classificadores. Quanto mais distante a amostra do hiperplano separador, maior a chance daquela amostra pertencer ao gênero analisado naquele classificador. O uso da distância da amostra para o hiperplano é uma das maneiras de estimar probabilidades para o SVM, e foi originalmente proposto no trabalho de Madevska-Bogdanova *et al.* (2004).

Para a passagem das probabilidades para o classificador principal do segmento, são aplicadas duas abordagens. Uma das abordagens é a forma mais comum, que é sem alterá-las. A outra forma é feita utilizando pesos nos classificadores, conforme mostrada na Figura 5.3. Nesta técnica, o desempenho de cada classificador de grupo de características é calculado para cada gênero musical durante a fase de treinamento. Tanto no treinamento como no teste, as probabilidades de cada gênero são multiplicadas pelo respectivo peso do gênero naquele classificador. Esta técnica segue a ideia de que as características são melhores para distinguir certos gêneros e piores para outros. Com isso, pretende-se que os classificadores que determinam melhor certos gêneros tenham um peso maior na decisão daquele gênero, e vice-versa. Por exemplo, se um determinado classificador tem um desempenho muito ruim em reconhecer determinado gênero, suas probabilidades para aquele gênero provavelmente estarão incorretas e atrapalharão na decisão. Desta forma, procura-se amenizar este problema.

Figura 5.3 - Alteração das probabilidades baseada no desempenho de cada gênero em cada classificador.



Fonte: Autor.

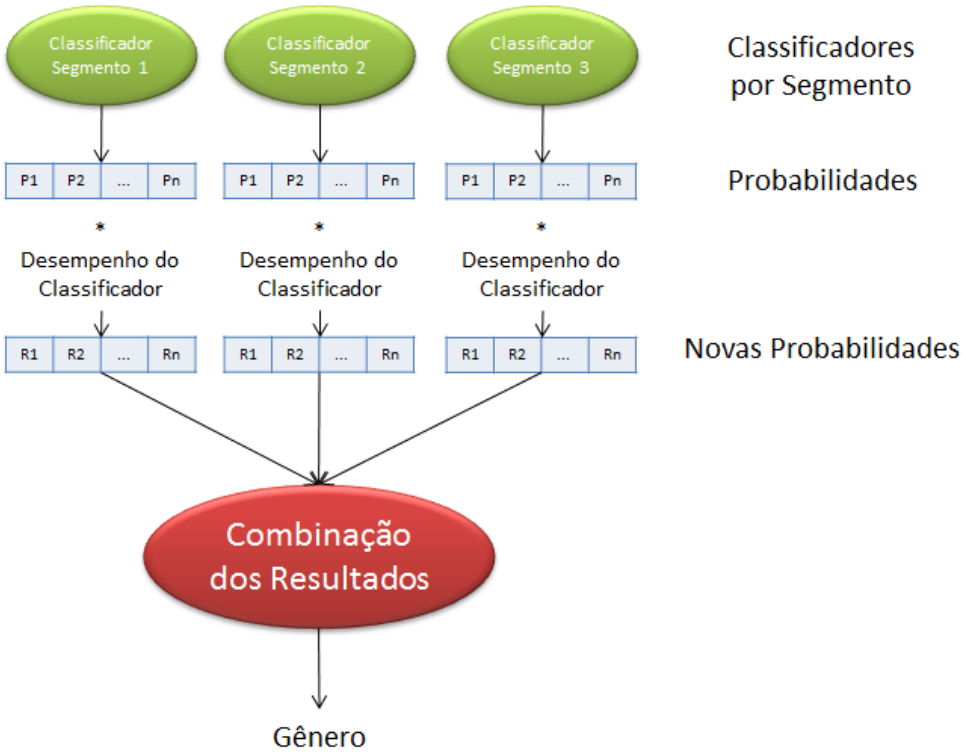
5.1.4 DEFINIÇÃO DO GÊNERO

Na última etapa do sistema proposto, os classificadores de cada um dos três segmentos recebem a junção das saídas dos quatro classificadores de cada grupo de características do segmento, e retornam novas probabilidades. Assim como nos classificadores por grupo, foram testados como classificadores principais de cada segmento o GMM (paramétrico) e o SVM (não paramétrico), devido ao bom desempenho destes. As probabilidades de cada um dos três classificadores são combinadas e é definido o gênero da música. Foram testadas quatro técnicas de combinação de probabilidades, que são comumente utilizadas nos trabalhos de reconhecimento de gêneros musicais para combinar resultados de vários classificadores (SILLA *et al.*, 2007; COSTA *et al.*, 2012; CHATHURANGA e JAYARATNE, 2013):

1. Maior valor: são analisadas todas as probabilidades, e a maior, independente do segmento de origem, é a escolhida. Em caso de empate, é escolhido um gênero aleatoriamente;
2. Regra da soma: as probabilidades de cada gênero oriundas de cada segmento são somadas, e o gênero com o maior valor é escolhido. Em caso de empate, é escolhido um gênero aleatoriamente;
3. Regra do produto: as probabilidades de cada gênero são multiplicadas, e o gênero com o maior valor é escolhido. Em caso de empate, é escolhido um gênero aleatoriamente;
4. Peso do classificador: neste método, exemplificado na Figura 5.4, é atribuído um peso para cada classificador. Este peso é calculado durante a fase de treinamento, e reflete a taxa de reconhecimento daquele classificador. Na fase de testes, as probabilidades dos classificadores são multiplicadas pelo respectivo peso do classificador, resultando em novas probabilidades (WEBB, 2002). Após, podem ser aplicadas as três regras anteriores nas novas probabilidades. Em caso de empate, é escolhido, dentre os gêneros empatados, aquele que obteve maior valor no classificador com melhor taxa de acerto. Esta técnica é muito interessante, pois permite ajustar a influência de cada classificador individual na escolha final do gênero,

permitindo que um classificador com melhor taxa de acerto tenha maior influência do que um classificador que obteve desempenho inferior.

Figura 5.4 - Alteração das probabilidades baseada no desempenho geral dos classificadores.



Fonte: Autor.

A combinação de classificadores desempenha um papel importante no reconhecimento e influencia consideravelmente a taxa de acerto. Por isto várias regras foram testadas. Em trabalhos já publicados, em que foi utilizada a combinação de vários classificadores, a técnica que obteve os melhores resultados foi a regra do produto utilizando os pesos de cada classificador, e a pior regra foi a do maior valor.

5.2 DESENVOLVIMENTO

Esta seção apresenta as bases de músicas utilizadas nos treinamentos e testes, além de apresentar como foram configurados os parâmetros do sistema.

5.2.1 BASE DE DADOS

Duas bases de dados foram utilizadas no desenvolvimento deste trabalho: uma delas, chamada de “Experimental”, foi construída pelo autor deste trabalho. A outra base, chamada de “GTZAN”, foi construída por Tzanetakis e Cook (2002), e é comumente utilizada nos trabalhos de reconhecimento de gêneros musicais.

BASE EXPERIMENTAL

A base de dados Experimental foi construída a partir de músicas adquiridas no site Last.fm[1], uma rádio online que possui uma seção de downloads gratuitos de músicas que são autorizadas por seus respectivos artistas. No próprio site as músicas são agrupadas por gêneros musicais. Esta classificação é feita pelos usuários e membros do site, que atribuem gêneros aos artistas. Os gêneros mais atribuídos recebem destaque e classificam o artista. Algumas músicas dos gêneros Metal, Pop e Rock também foram adquiridas do acervo musical do autor deste trabalho.

A base é composta por 5000 músicas, divididas igualmente em 10 gêneros musicais diferentes (Clássica, Country, Eletrônica, Hip Hop, Jazz, Latinas, Metal, Pop, Reggae, Rock) com 500 músicas cada. Estes gêneros foram escolhidos por serem muito conhecidos e por apresentarem um vasto acervo musical, facilitando a construção da base. A base possui músicas completas, isto é, não tendo apenas trechos das músicas, e apresenta tanto canções instrumentais como com vocais.

Para tentar diversificar e tornar a base o mais abrangente possível, os seguintes critérios foram adotados para a escolha das músicas:

a) Músicas com menos de 60 segundos de duração não foram incluídas, por não apresentarem conteúdo suficiente para treinamento de um sistema;
b) Foi evitado sempre que possível à escolha de músicas que foram gravadas ao vivo, pois estas músicas apresentam outros sons externos (som da plateia, aplausos, entre outros) que podem atrapalhar o processamento da música;

[1] www.last.fm

c) Procurou-se diversificar ao máximo a quantidade de artistas por gênero. Muitas vezes artistas pertencentes ao mesmo gênero musical trabalham com instrumentos e afinações diferentes, e apresentam características peculiares. Quanto mais artistas diferentes, mais se está abrangendo o gênero musical.

d) Também foi buscado escolher músicas de vários subgêneros dentro de um gênero musical. Neste quesito o site Last.fm facilitou bastante, já que no site artistas de Classic Rock ou Modern Rock, por exemplo, também são classificados apenas como Rock, aparecendo em uma busca apenas pelo gênero principal.

A base possui um total de 336 h 24 m 51 s de duração, com aproximadamente 19,26 GB de tamanho para armazenamento. A Tabela 5.1 mostra detalhadamente a base de dados em cada gênero musical. A frequência de amostragem do áudio das músicas é de 44.100 Hz, e todas elas se encontram no formato MP3. Apesar das músicas estarem compactadas (o que pode atrapalhar na extração das características devido à perda de dados desta compactação), este formato de arquivo foi escolhido por ser o mais usado no dia-a-dia, e o mais provável de ser utilizado em uma aplicação real de reconhecimento de gêneros musicais.

Tabela 5.1 - Informações detalhadas da base de dados por gênero.

GÊNERO	MÚSICAS	ARTISTAS	DURAÇÃO TOTAL
Clássica	500	300	32h 34m 51s
Country	500	273	30h 15m 52s
Eletrônica	500	391	34h 48m 26s
Hip Hop	500	332	29h 13m 42s
Jazz	500	314	39h 12m 16s
Latinas	500	354	33h 08m 41s
Metal	500	500	39h 07m 00s
Pop	500	434	31h 13m 07s
Reggae	500	333	33h 18m 05s
Rock	500	388	33h 32m 51s

BASE GTZAN

Para este trabalho também foi utilizada a base de dados GTZAN Genre Collection[2], que foi criada e usada no trabalho de Tzanetakis e Cook (2002). Esta base possui 1000 músicas, divididas em 10 gêneros (Blues, Clássica, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae e Rock) com 100 músicas cada. Os arquivos estão em formato AU, com frequência de 22.050 Hz Mono 16-bit, e cada música possui 30 segundos de duração, segmento este retirado do meio da música. Também foi escolhido o uso desta base de dados por ser uma das mais conhecidas e mais utilizadas em diversos trabalhos, como nos trabalhos de Haggblade *et al.* (2012), Lidy e Rauber (2005), Lee *et al.* (2009), Benetos e Kotropoulos (2010), Ariyaratne e Zhang (2012), entre outros. Com isto, foi possível testar o desempenho do sistema proposto com outra base de músicas, e comparar os resultados obtidos com outros métodos já publicados.

5.2.2 PARÂMETROS E CONFIGURAÇÕES DO SISTEMA

O sistema proposto foi totalmente desenvolvido no software MATLAB. A segmentação dos arquivos, extração de características, classificação, regras de combinação e modificação das probabilidades foram todas implementadas utilizando as ferramentas nativas do software MATLAB. Somente para o MFCC foi utilizado o algoritmo desenvolvido por Ellis (2005).

A seguir são especificados os parâmetros e configurações utilizadas para as rotinas de treinamento e teste no sistema, extração por segmentos, extração de características, e classificação.

TREINAMENTO E TESTE

Para a realização dos treinamentos e testes entre as bases de dados, foi utilizado *N-fold cross-validation*. Neste método, a base é dividida em *N* partições de igual

[2] http://marsyas.info/download/data_sets/

tamanho, e o treinamento é realizado em $N - 1$ partições e testado na partição remanescente. O processo é repetido N vezes, alternando a partição que está sendo testada (WEBB, 2002). Esta forma garante melhor qualidade e confiabilidade nos resultados, pois os treinamentos/testes são realizados em toda a base. Foi escolhida a divisão de cada base em 10 partições (*10-fold cross-validation*), que é a quantidade mais utilizada em trabalhos de reconhecimento de gêneros musicais (TZANETAKIS *et al.*, 2001; POHLE *et al.*, 2004; SILLA *et al.*, 2004; LIDY e RAUBER, 2005; LEE *et al.*, 2009; BENETOS e KOTROPOULOS, 2010), e que garante bastante amostras para treinamento em cada iteração.

Foram realizadas 4 abordagens de treinamento/teste entre as bases de dados:

- Treinamento e teste na base experimental utilizando *10-fold cross-validation*;
- Treinamento na base experimental, teste na base GTZAN: o objetivo desta abordagem é verificar se um grande volume de músicas para treinamento contribui para um melhor reconhecimento; e também analisar se é possível integrar bases de músicas diferentes. Devido a incompatibilidade dos gêneros presentes nas bases, para esta abordagem foram utilizados 9 gêneros musicais que são comuns em ambas as bases (Clássica, Country, Eletrônica, Hip Hop, Jazz, Metal, Pop, Reggae, Rock), ficando de fora o gênero Latinas da base experimental e o gênero Blues da base GTZAN. Para o treinamento, foram utilizadas todas as 4500 músicas da base experimental pertinentes aos 9 gêneros, e o teste foi realizado nas 900 músicas dos 9 gêneros envolvidos da base GTZAN. Nesta abordagem, devido à incompatibilidade de frequência das músicas das bases (44.100 Hz da base experimental contra 22.050 Hz da base GTZAN), as músicas da base GTZAN tiveram suas frequências alteradas para 44.100 Hz. Como a taxa de frequência das músicas em ambas as bases é alta, a alteração destas na base GTZAN ou na Experimental não faria muita diferença. Assim, foi optado pela alteração na GTZAN por ser uma base menor, resultando em menos custo de processamento;
- Treinamento e teste na base GTZAN com 9 gêneros: foi utilizado somente os 9 gêneros usados na abordagem anterior, para que seja possível comparar as taxas de acerto e verificar se há mudanças quando o treinamento é feito

em uma base diferente. O mesmo conceito de *10-fold cross validation* foi utilizado nesta abordagem.

- Treinamento e teste na base GTZAN com 10 gêneros: utilizando todos os gêneros da base GTZAN, esta abordagem foi utilizada para comparar o desempenho do sistema com outros trabalhos já publicados. Foi utilizado *10-fold cross validation* para a partição da base em conjuntos de treinamento/teste.

SEGMENTAÇÃO

A segmentação foi definida conforme a base de dados. Para a base Experimental, o tamanho do segmento é de 30 segundos, totalizando 90 segundos a ser analisado por música. Os 20 segundos iniciais e finais da música são descartados, pois geralmente estes trechos apresentam partes silenciosas (*fade in* e *fade out*) e não contém todos os instrumentos da canção, podendo não representar corretamente o gênero. O segmento inicial é analisado no trecho logo após os 20 segundos iniciais. Já para o segmento intermediário é calculado o ponto que representa a metade da música, e a partir deste ponto, são analisados os 15 segundos anteriores e posteriores a este ponto. O segmento final é extraído do trecho de 30 segundos anterior aos 20 segundos finais da música. Caso a duração da canção seja inferior a 130 segundos (os 90 segundos dos segmentos mais os 20 segundos iniciais e finais que são descartados), os segmentos são parcialmente sobrepostos. Isto garante que em qualquer música, independente do seu tamanho, o volume de dados analisados seja sempre o mesmo.

Para a base de dados GTZAN, o cálculo dos segmentos é diferente. Como as músicas desta base possuem apenas 30 segundos de duração, toda a música é analisada, e os segmentos são de 10 segundos cada. O segmento de início é dos segundos 0 a 9, o segmento intermediário pertence à faixa dos 10 a 19 segundos e o segmento final da música é dos segundos 20 a 29.

CARACTERÍSTICAS

Para o Grupo 1 de características (relacionadas ao timbre e MFCC), foram analisadas janelas com duração de 100 ms com deslocamento de 50 ms, totalizando 600 vetores de características por segmento de música na base experimental, e 200 vetores na base GTZAN. Para o *spectral centroid*, *spectral rolloff* e *spectral flux* foi montada a transformada de Fourier do sinal com amplitude normalizada entre 0 e 1. Para o MFCC, foram utilizados 13 coeficientes, que é a quantidade de coeficientes mais utilizada nos trabalhos de reconhecimento de gêneros. Ao final, este grupo forma um vetor de 17 características (13 do MFCC e 4 de informações relacionadas ao timbre).

Já o Grupo 2, com características temporais, é formado pela média e variância das características do Grupo 1, além do *low energy*. Este grupo totaliza um vetor com 35 dimensões.

Para o Grupo 3, constituído por características rítmicas, foi utilizada a técnica de Tzanetakis e Cook (2002), explicada na Seção 3.3.1. Optou-se por janelas de 4 segundos com deslocamento de 2 segundos, filtro α igual a 0.99 na filtragem passa-baixa e diminuição da taxa de amostragem com fator igual a 16, que são os mesmos valores utilizados por Tzanetakis e Cook (2002). Na construção do histograma, foi realizada a soma das forças das batidas das janelas. A captura dos maiores picos não se mostrou interessante devido ao tamanho menor de música analisada (segmentos de 10 e 30 segundos), o que resultava em um histograma com poucos dados. O histograma abrange batidas de 40 a 300 BPM, com isso permitindo a captura de batidas mais rápidas. É extraído um total de 20 características do histograma:

- Soma total do histograma, indicando a força da batida;
- Variância do histograma, mostrando se o ritmo possui variações ao longo da música;
- Amplitude dos dois maiores picos, indicando a força e presença das duas principais batidas;
- Período dos dois maiores picos, que apresenta a velocidade em BPM da batida principal e auxiliar;
- Largura dos dois maiores picos, indicando a variação em velocidade das duas batidas principais;

- Distância entre os dois maiores picos, resultando na velocidade média da música em BPM;
- Quantidade total de picos, mostrando a quantidade de diferentes batidas durante a música;
- Quantidade de picos com amplitude acima da média, indicando a quantidade de batidas com maior presença na canção;
- Distância média entre os picos, indicando a variação de velocidade média da música;
- Média das amplitudes, que calcula de outra forma a força da batida;
- Quantidade de posições sem picos, indicando se a música apresenta diversas batidas rítmicas diferentes;
- Diferença da amplitude entre os dois maiores picos, apresentando a diferença da força da batida entre a batida principal e auxiliar;
- Quantidade de ilhas (grupos fechados de picos contínuos), que também é uma forma de calcular a variação rítmica nas mais diferentes velocidades;
- Amplitude relativa dos dois maiores picos em relação à soma do histograma, medindo o quão distintas são as batidas em relação ao resto do sinal;
- Diferença relativa entre as amplitudes dos dois maiores picos, expressando a relação entre a batida principal e a auxiliar;
- Diferença relativa das distâncias dos dois maiores picos, indicando a relação de velocidade entre as duas batidas principais.

Já para o Grupo 4 de características (pitch), foi construído um histograma do pitch utilizando a técnica de Tolonen e Karjalainen (2000), explicada na Seção 3.3.2. Para este trabalho, foi optado por janelas de 100 ms de duração, com deslocamento de 100 ms. Para a filtragem de suavização do sinal, foram utilizados filtros de ordem 12; e para a compressão da magnitude do espectro foi utilizado um fator de 0.67. Estes valores para os parâmetros foram os valores sugeridos por Tolonen e Karjalainen (2000) para o uso deste modelo de extração de pitch. A construção do histograma foi feita somando-se as forças do pitch das janelas, devido aos mesmos motivos apontados na construção do histograma rítmico. Após a construção do histograma, este é analisado e é montado um vetor por segmento (totalizando 20 dimensões), com as mesmas

características que são extraídas do histograma rítmico, porém com finalidades diferentes:

- Soma total do histograma, indicando a força do pitch;
- Variância do histograma, mostrando se a melodia/harmonia possui variações ao longo da música;
- Amplitude dos dois maiores picos, indicando a força e presença das duas tonalidades principais;
- Período dos dois maiores picos, que apresenta as duas tonalidades predominantes na canção;
- Largura dos dois maiores picos, indicando a presença ou não de tons próximos das duas tonalidades principais;
- Distância entre os dois maiores picos, resultando no intervalo tonal da música;
- Quantidade total de picos, mostrando a quantidade de tonalidades diferentes durante a música;
- Quantidade de picos com amplitude acima da média, indicando a quantidade de tons principais;
- Distância média entre os picos, indicando a variação média da tonalidade da canção;
- Média das amplitudes, que calcula de outra forma a força do pitch;
- Quantidade de posições sem picos, indicando a quantidade de tons não explorados na música;
- Diferença da amplitude entre os dois maiores picos, apresentando a diferença da força da tonalidade principal e seu intervalo;
- Quantidade de ilhas (grupos fechados de picos contínuos), que é uma forma de calcular a quantidade de grupos fechados de tons;
- Amplitude relativa dos dois maiores picos em relação à soma do histograma, medindo o quão distintas são as tonalidades principais em relação ao resto da música;
- Diferença relativa entre as amplitudes dos dois maiores picos, expressando a relação entre a tonalidade principal e a auxiliar;

- Diferença relativa das distâncias dos dois maiores picos, indicando a relação de tempo entre as duas tonalidades principais.

CLASSIFICADORES

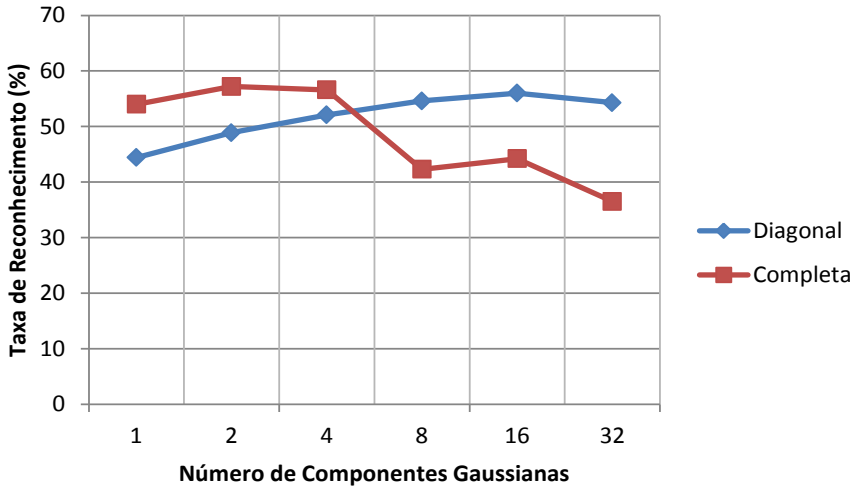
Nos dois classificadores utilizados (GMM e SVM) foi necessário definir os parâmetros de cada um deles. Para isto, foram realizados diversos testes, definindo assim as melhores configurações para cada classificador.

Para o GMM, os parâmetros necessários para a construção dos modelos foram:

- Número de componentes gaussianas;
- Tipo da matriz de covariância: completa ou diagonal;

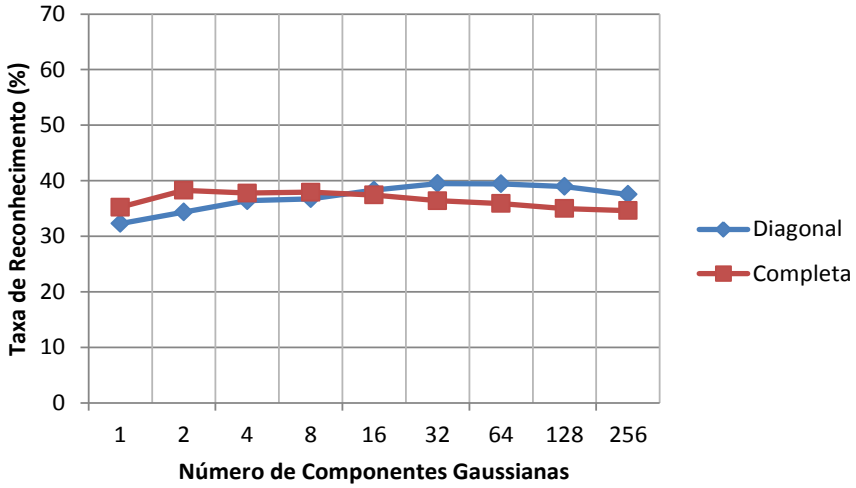
Testes foram realizados em ambas as bases com a combinação dos dois parâmetros. Nestes testes, a base foi particionada utilizando *3-fold cross validation*. A quantidade de componentes gaussianas foi testada em potências de 2. Para a base GTZAN, foram efetuados testes com até 32 componentes gaussianas, já que a base é menor e possui menos músicas por gênero, contendo, portanto, menos amostras para treinar o modelo. Já para a base experimental, com mais músicas de cada gênero por treinamento, foram testados modelos de GMM com até 256 componentes gaussianas. Todos os testes utilizaram as mesmas características e a mesma regra de decisão do gênero (regra da soma), e os modelos de GMM foram configurados para testar 3 diferentes parâmetros de início para o algoritmo EM, escolhendo a configuração que obteve o melhor desempenho. Os resultados mostrados são a taxa de acerto geral entre os 10 gêneros das respectivas bases. Os resultados obtidos encontram-se na Figura 5.5 e Figura 5.6.

Figura 5.5 - Teste de parâmetros do GMM na base GTZAN.



Fonte: Autor.

Figura 5.6 - Teste de parâmetros do GMM na base Experimental.



Fonte: Autor.

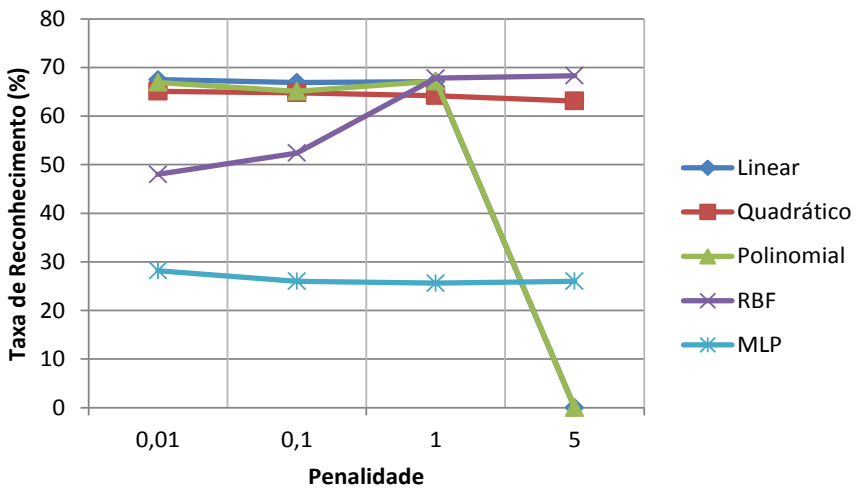
A partir dos resultados, foram obtidos diferentes valores para melhor configurar o GMM em cada base de dados. Para a base GTZAN, a melhor configuração foi com a matriz de covariância completa e 2 componentes gaussianas, obtendo taxa de reconhecimento de 57,2% contra 56% obtido com matriz de covariância diagonal e 16 componentes gaussianas. Já para a base Experimental, os resultados foram diferentes, e a melhor configuração foi obtida com matriz de covariância diagonal e 32 componentes gaussianas, com 39,5% de reconhecimento. Estas melhores configurações foram utilizadas durante o restante dos testes do sistema utilizando o GMM.

Para o SVM, os parâmetros necessários para a construção dos classificadores foram:

- Função do kernel, onde foram testados os seguintes algoritmos: linear, polinomial, quadrático, *radial basis function* (RBF) e *multilayer perceptron* (MLP);
- Parâmetros do kernel, quando necessário. Foram testados diferentes ordens do polinômio (para o algoritmo polinomial) e diferentes valores para o sigma no kernel RBF.
- Penalidade de erro, onde foram testados os seguintes valores: 0.01, 0.1, 1 e 5.

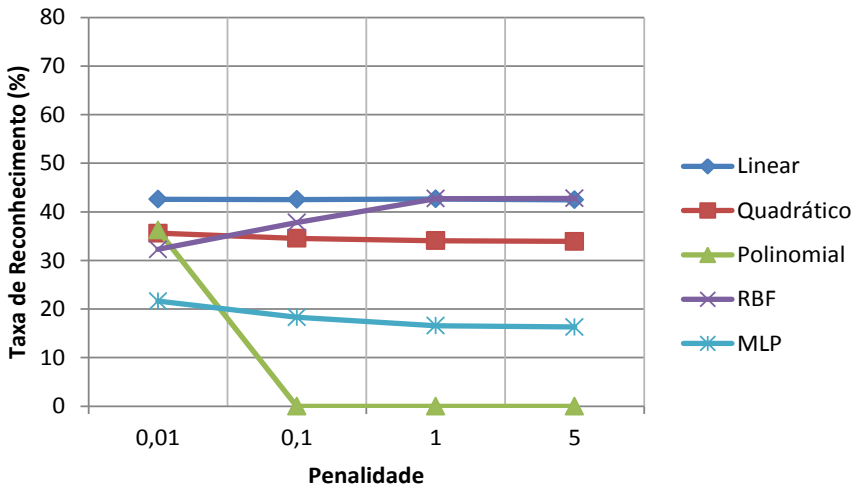
Testes foram realizados em ambas as bases com a combinação dos dois parâmetros. Nestes testes, a base foi particionada utilizando *3-fold cross validation*. Para os parâmetros específicos do kernel (necessários no algoritmo polinomial e RBF), foram testados diversas configurações, e os melhores parâmetros obtidos foram: polinômio de ordem 3 para o kernel polinomial e valor sigma igual a 15 para o RBF. Todos os testes utilizaram as mesmas características e a mesma regra de decisão do gênero (regra da soma). Os resultados mostrados são a taxa de acerto geral entre os 10 gêneros das bases. Taxas de acerto igual a zero significam que o algoritmo não convergiu com a configuração apresentada. Os resultados obtidos encontram-se na Figura 5.7 e na Figura 5.8.

Figura 5.7 - Teste de parâmetros do SVM na base GTZAN.



Fonte: Autor.

Figura 5.8 - Teste de parâmetros do SVM na base Experimental.



Fonte: Autor.

Através dos resultados, foram obtidas as seguintes conclusões:

- Tanto o kernel MLP como o quadrático obtiveram desempenhos inferiores quando comparado aos demais. O pior deles foi o MLP, com resultados entre 26% e 28% para a base GTZAN e 16% e 21% para a base Experimental;
- O kernel polinomial apresentou dificuldades para convergir quando usado com o parâmetro de penalidade muito alto;
- O kernel linear apresentou bons resultados quando comparado aos demais (aproximadamente 67% para a base GTZAN e 42% para a base Experimental) e praticamente não sofreu influência quando variado a penalidade de erro;
- O melhor kernel foi o RBF, principalmente quando utilizado com uma penalidade grande, alcançando 68,3% de acerto para a base GTZAN e 42,8% para a base Experimental (com penalidade igual a 5).

A partir disto, foi escolhido como melhor configuração o kernel RBF com penalidade de erro igual a 5 e valor sigma igual a 15, e esta configuração foi usada no restante dos testes do sistema utilizando o SVM.

Os parâmetros do GMM utilizados na construção dos supervetores foram determinados através dos resultados dos testes explanados anteriormente (apresentados

nas Figura 5.5 e Figura 5.6). Foram testados os classificadores GMM com cada tipo de matriz de covariância (diagonal e completa), e a quantidade de componentes gaussianas que obteve o melhor resultado nos testes em cada base de dados e em cada tipo de matriz. Estes testes foram realizados, pois a quantidade de componentes gaussianas influencia consideravelmente no reconhecimento, pois define a quantidade de dimensões dos dados de saída. Após os testes, a melhor configuração para o GMM em ambas as bases de dados foi matriz de covariância completa e 2 componentes gaussianas.

5.3 RESULTADOS

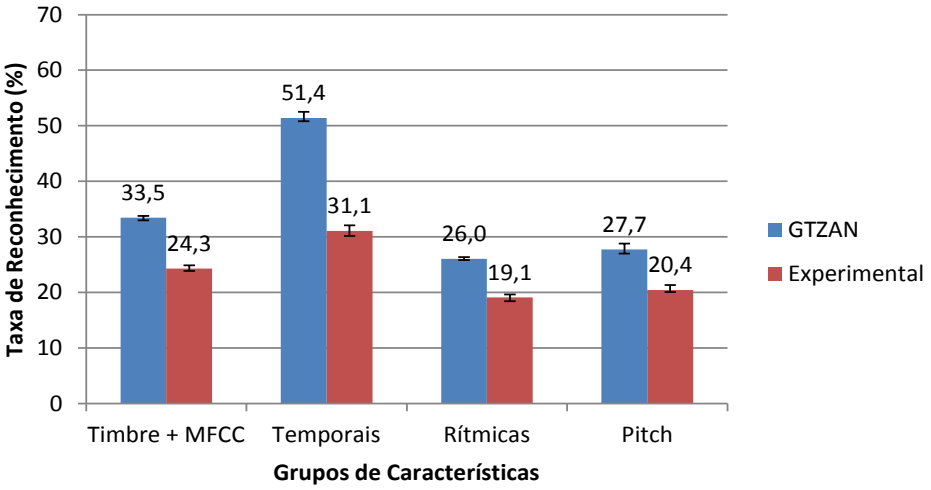
Para os resultados, foi reportado o desempenho individual de cada um dos 15 classificadores utilizados no sistema (os 4 específicos para cada grupo de características, em cada um dos 3 segmentos, e o classificador principal de cada segmento), além do resultado final do sistema, após as combinações dos resultados. Assim, é possível identificar, por exemplo, qual grupo de características mais ajudou no reconhecimento, se a combinação dos classificadores aumentou a taxa de acerto, entre outras questões.

As seções a seguir mostram os resultados obtidos nos diversos testes realizados no sistema entre as bases de dados. O teste binomial para diferenças de proporção apresentado no trabalho de Gillick e Cox (1989) é usado para testar se as diferenças são estatisticamente significantes. Quando não especificado, o nível de significância α é igual a 0,05.

5.3.1 GRUPOS DE CARACTERÍSTICAS

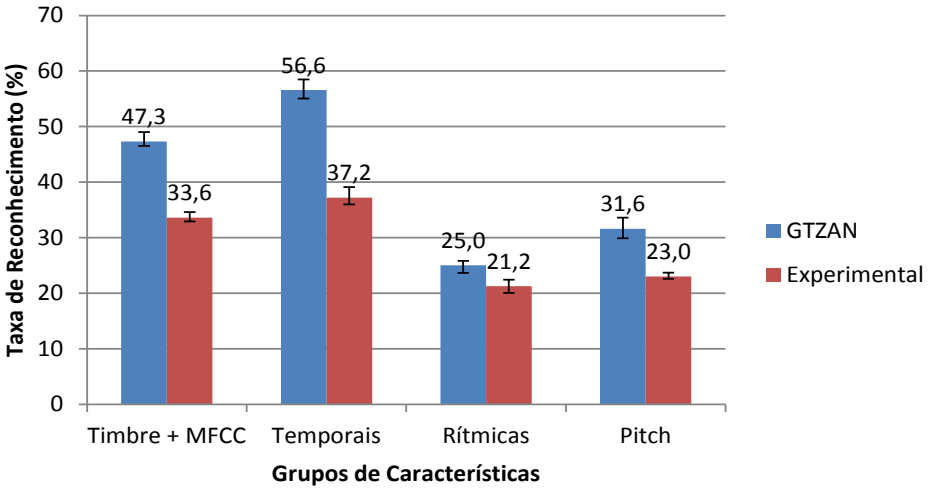
Com o objetivo de identificar quais características mais ajudam no reconhecimento dos gêneros, foi verificada a taxa de acerto de cada grupo de características, quando utilizadas individualmente. Para isto, foi analisado o desempenho de cada grupo de classificadores em cada grupo de características, para cada base de dados. Os valores encontram-se na Figura 5.9 e na Figura 5.10, e apresentam as melhores taxas de reconhecimento obtidas nos testes para 10 gêneros.

Figura 5.9 - Resultados obtidos nos grupos de características utilizando GMM.



Fonte: Autor.

Figura 5.10 - Resultados obtidos nos grupos de características utilizando SVM.



Fonte: Autor.

Em ambas as bases e em ambos os classificadores, nota-se as diferenças de qualidade dos grupos de características. As características temporais (média e variância do timbre e MFCC, além do *low energy*), obtiveram os melhores desempenhos individuais, com 51,4% e 56,6% para a base GTZAN (para GMM e SVM respectivamente) e 31,1% e 37,2% para a base Experimental. Em seguida, as características de timbre e MFCC para cada janela obtiveram os segundos melhores desempenhos. Por fim, as características de pitch e rítmicas obtiveram piores

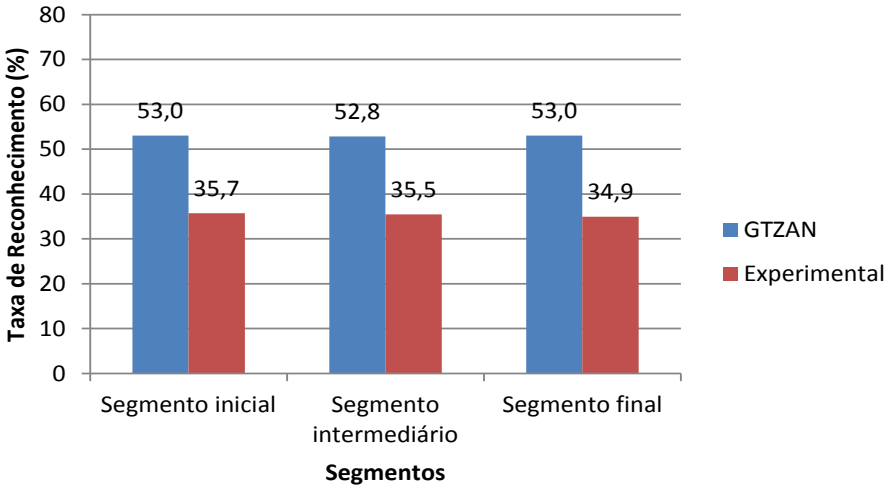
reconhecimentos quando utilizadas sozinhas. Utilizando GMM, as diferenças de reconhecimento usando pitch e rítmicas não foram estatisticamente significantes, porém, usando SVM, o pitch apresenta resultados estatisticamente superiores do que as informações rítmicas.

Os resultados obtidos corroboram com os apresentados em outros trabalhos, mostrados na Tabela 3.1 e na Tabela 3.2, nas Seções 3.2.3 e 3.3.4, respectivamente. Nesta comparação, reforça-se a ideia de que as características de baixo nível (timbre, MFCC e temporais) apresentam melhores taxas de reconhecimento quando comparadas as características de médio nível (rítmicas e pitch). Diferenças similares às apontadas entre os grupos de características são encontradas nos trabalhos de Tzanetakis e Cook (2002), Li *et al.* (2003) e Li e Ogihara (2006). Isto provavelmente se deve a maior complexidade das características de médio nível, que procuram representar informações abstratas do áudio como, por exemplo, um ritmo.

5.3.2 SEGMENTOS

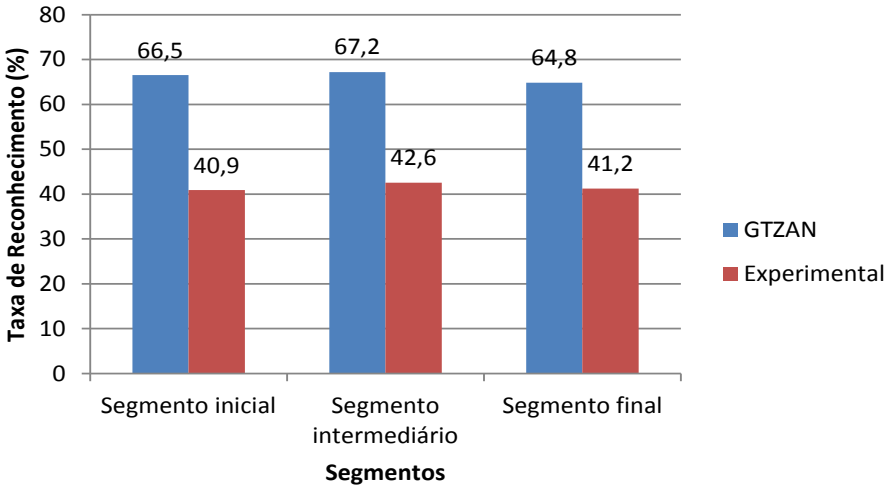
Para verificar se um determinado segmento da música apresenta um conteúdo musical mais propício para o reconhecimento de gêneros, foi analisado o desempenho de cada um dos três segmentos em cada base de dados. A Figura 5.11 e a Figura 5.12 apresentam as melhores taxas de reconhecimento obtidas nos testes para 10 gêneros.

Figura 5.11 - Resultados obtidos nos segmentos utilizando GMM.



Fonte: Autor.

Figura 5.12 - Resultados obtidos nos segmentos utilizando SVM.



Fonte: Autor.

A taxa de reconhecimento entre as análises de diferentes segmentos da música mostrou certas diferenças entre elas. Para a base GTZAN é difícil tirar conclusões sobre esta informação, visto que as músicas da base são de apenas 30 segundos (retiradas do meio da música) e os segmentos dela são contínuos. Já para a base Experimental, que possui músicas completas, é possível avaliar esta diferença entre os segmentos. Tanto com o uso do GMM como do SVM, em ambas as bases, as diferenças entre as taxas de reconhecimento dos segmentos não foram estatisticamente significantes. Assim, para a base Experimental e GTZAN, não é possível afirmar qual segmento apresenta o melhor conteúdo musical para a definição do gênero da música.

A fusão das probabilidades da primeira fase do sistema (com classificadores para cada grupo de características) para a segunda etapa (com classificador para cada segmento) mostrou resultados interessantes de como unir as diversas informações obtidas, aumentando a taxa de reconhecimento atual. Em cada classificador, em cada base, após a união das probabilidades, a taxa de acerto aumentou quando comparado aos resultados individuais, como verificado na Tabela 5.2. Para cada caso, a tabela mostra a média dos resultados de cada grupo de classificadores nos 3 segmentos. A última coluna mostra a média de reconhecimento dos classificadores que receberam as probabilidades dos classificadores anteriores em cada segmento. Após a fusão, a taxa de reconhecimento aumentou entre 1,5% e 9,6% na base GTZAN, e 4,3% na base

Experimental quando comparado ao melhor resultado individual. Os resultados da fusão reafirmam o que foi apresentado em outros trabalhos (Tabela 3.1 e Tabela 3.2, nas Seções 3.2.3 e 3.3.4, respectivamente), mostrando que o uso combinado de diversos tipos de características aumenta a taxa de reconhecimento.

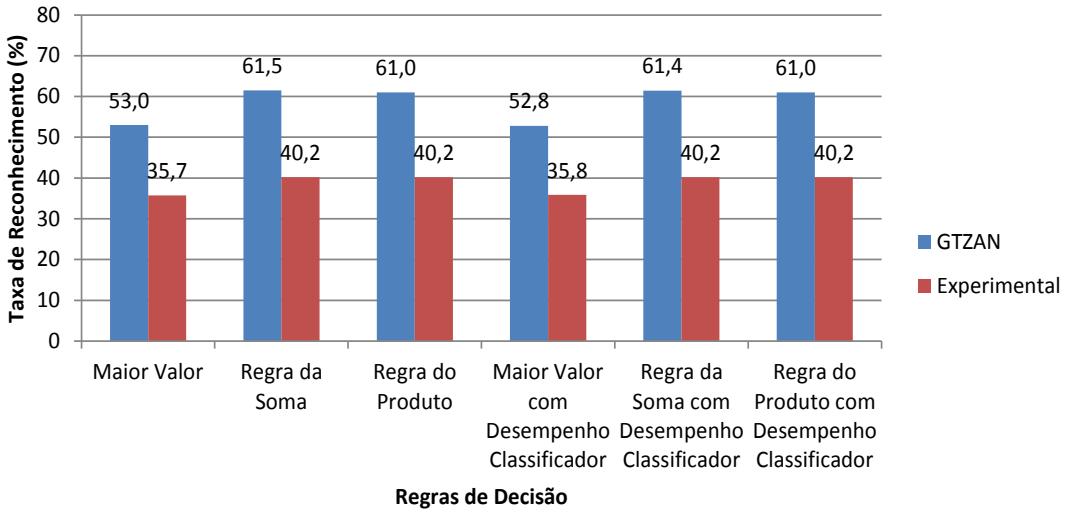
Tabela 5.2 - Taxa de reconhecimento entre os grupos de classificadores.

CLASSIFICADOR	TIMBRE + MFCC	TEMPORAIS	RÍTMICAS	PITCH	FUSÃO
Base GTZAN					
GMM	33,5%	51,4%	26%	27,7%	52,9%
SVM	47,3%	56,6%	25%	31,6%	66,2%
Base Experimental					
GMM	24,3%	31,1%	19,1%	20,4%	35,4%
SVM	33,6%	37,2%	21,2%	23%	41,5%

5.3.3 DECISÃO FINAL

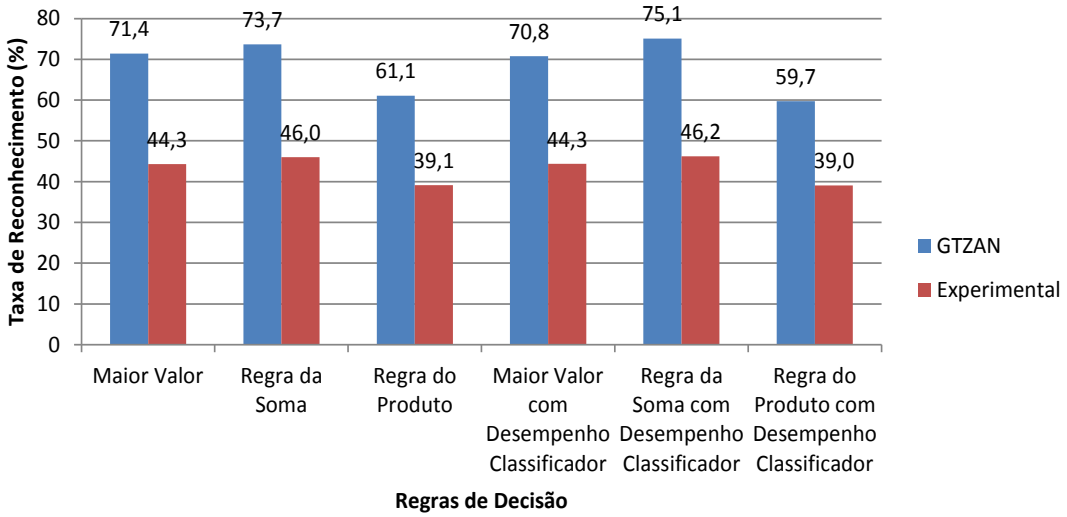
Para combinar os resultados de diversos classificadores, várias são as regras de decisão que podem ser utilizadas. Para verificar qual regra de decisão obtém os melhores resultados, foram testadas as regras do maior valor, da soma e do produto, para cada classificador em cada base de dados. Também foi verificado o uso das regras com alteração de probabilidades pelo desempenho dos classificadores de cada segmento. A Figura 5.13 e a Figura 5.14 apresentam os melhores resultados obtidos para cada uma das regras de decisão aplicadas.

Figura 5.13 - Resultados obtidos nas regras de decisão utilizando GMM.



Fonte: Autor.

Figura 5.14 - Resultados obtidos nas regras de decisão utilizando SVM.



Fonte: Autor.

Os resultados encontrados variaram para cada classificador utilizado. Usando GMM, a regra do maior valor apresentou resultados estatisticamente inferiores aos demais. Já entre a regra da soma e a do produto, as diferenças não foram estatisticamente significantes em nenhuma das bases. Utilizando SVM, os resultados foram um pouco diferentes. Em ambas as bases de dados, a regra do produto mostrou resultados estatisticamente inferiores em comparação às outras regras. Entre a regra da

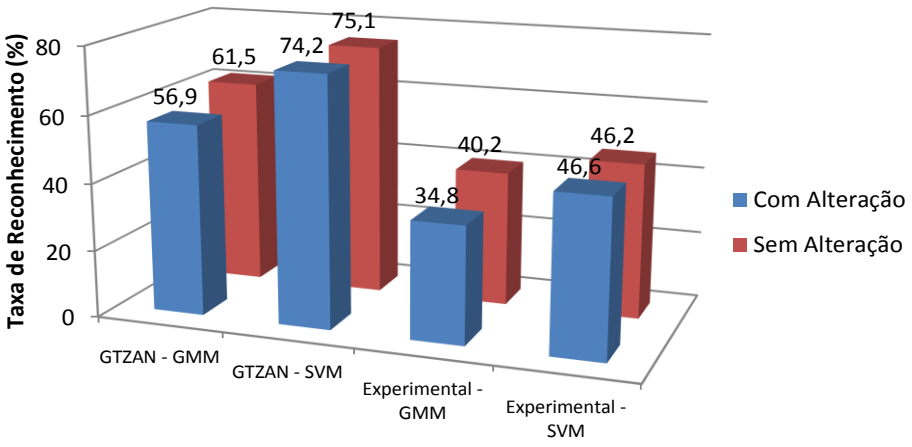
soma e a do maior valor, tanto na base Experimental como na GTZAN, as diferenças entre elas não foram estatisticamente significantes.

Por fim, o uso das regras de decisão com alteração das probabilidades pelo desempenho dos classificadores mostrou poucas diferenças. Em todas as regras de decisão aplicadas, em ambas as bases e classificadores, as diferenças encontradas com a alteração não foram estatisticamente significantes. Logo, não é possível afirmar se a alteração das probabilidades pelo desempenho dos classificadores melhora ou piora a taxa de reconhecimento.

5.3.4 PROBABILIDADE MODIFICADA PELO ACERTO NOS GÊNEROS

Outra técnica utilizada no reconhecimento de gêneros é a alteração de probabilidades baseado no desempenho dos classificadores em cada gênero, método este explicado na Seção 5.1.3. A ideia é de que as características que reconheçam melhor certos gêneros tenham mais poder de decisão na escolha destes. Assim, foi verificado se esta técnica melhora a taxa de reconhecimento do sistema. Para este teste, foram utilizadas as melhores regras de decisão verificadas nos testes anteriores, para cada base de músicas em cada classificador. A Figura 5.15 mostra as taxas de reconhecimento obtidas.

Figura 5.15 - Taxas de reconhecimento obtidas com a modificação de probabilidades pelo desempenho dos classificadores nos gêneros em ambas as bases e classificadores.



Fonte: Autor.

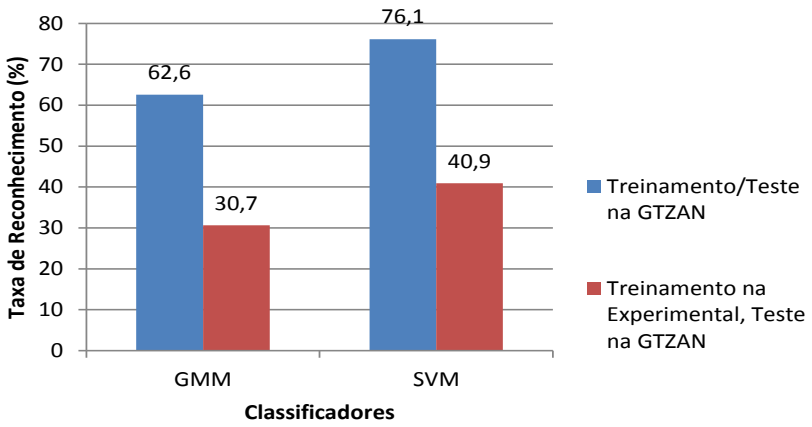
Conforme apontado pelos resultados, a alteração das probabilidades pelo desempenho dos classificadores nos gêneros foi bastante influenciada pelo tipo do classificador utilizado. Com o GMM, as taxas obtidas foram significantemente inferiores, com redução de 4,6% na base GTZAN e 5,4% na base Experimental. Já com o SVM, não houve diferenças estatisticamente significantes entre as taxas de acerto com e sem a alteração.

Os resultados mostram que a forma de cálculo das probabilidades influencia consideravelmente no desempenho desta técnica. No GMM foi observado que, utilizando a alteração das probabilidades, o sistema reconheceu demais certos gêneros, mas muito pouco os outros, resultando no final uma taxa média de reconhecimento inferior. Outro fator que pode ter influenciado para o desempenho desta técnica no GMM foi a grande diferença de reconhecimento deste classificador nos mais diferentes gêneros. A Seção 5.4.1 mostra o reconhecimento do sistema em cada gênero, onde é possível verificar que, utilizando GMM, a variação de reconhecimento entre os gêneros foi muito maior do que utilizando SVM. Isto pode ter influenciado, fazendo com que a técnica de alteração aumentasse ainda mais a divergência no reconhecimento entre os gêneros.

5.3.5 TREINAMENTO E TESTE EM DIFERENTES BASES

Outro teste realizado foi com relação ao treinamento/teste em diferentes bases, com o intuito de verificar se é possível trabalhar com diversas bases de dados e se um treinamento com um grande número de amostras influencia no reconhecimento do sistema. Foi comparado o desempenho do sistema ao ser treinado com a base Experimental e testado na base GTZAN, e ao ser treinado/testado na mesma base de dados (GTZAN). Devido a incompatibilidade de gêneros das bases, estes testes utilizaram apenas 9 gêneros musicais. As configurações de cada classificador e as regras de decisão aplicadas foram as que, nos testes anteriores, obtiveram as melhores taxas de acerto em cada base de dados. Os resultados obtidos são apresentados na Figura 5.16.

Figura 5.16 - Resultados obtidos no treinamento/teste em diferentes bases.



Fonte: Autor.

Através dos resultados obtidos, se percebe a enorme diferença de reconhecimento quando utilizado diferentes bases para treinamento e teste. Os resultados mostram uma redução relativa de desempenho de aproximadamente 51% para o GMM e de aproximadamente 46% para o SVM. No teste utilizando as duas bases de dados (treinamento na Experimental, teste na GTZAN), alguns gêneros como Clássica, Eletrônica, Hip-Hop e Metal obtiveram altas taxas de reconhecimento (geralmente acima dos 70%), porém para os outros gêneros (Country, Jazz, Pop, Reggae e Rock) os desempenhos foram muito baixos, muitas vezes não alcançando nem 10% de reconhecimento.

Por estas diferenças, nota-se que as baixas taxas de reconhecimento utilizando as duas bases se deram justamente pela diferença de conteúdo entre elas. Pela própria diferença do ano de construção das bases, as músicas se transformaram bastante, e o conceito de alguns gêneros mudou consideravelmente nos últimos anos (mais detalhes são explicados na Seção 5.4.1). Isto mostra que outro desafio a ser enfrentado para uma usabilidade real dos sistemas de reconhecimento de gêneros é a evolução e adaptação destes para as constantes mudanças no mundo da música.

5.4 DISCUSSÃO DOS RESULTADOS

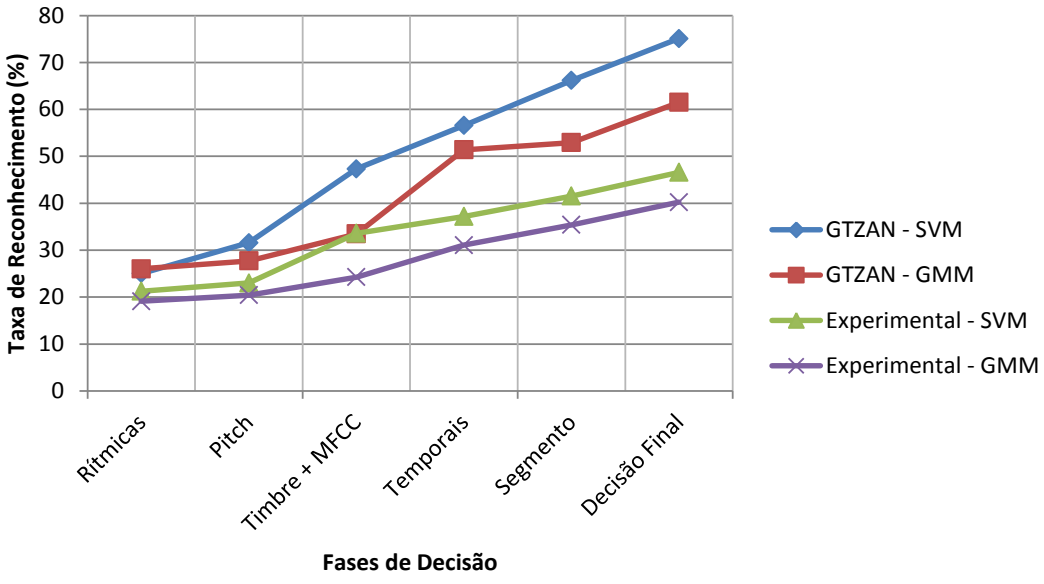
Após a realização de todos os testes anteriores, alcançaram-se as melhores taxas de reconhecimento em cada base de dados, em cada classificador:

- 61,5% de reconhecimento com GMM na base GTZAN;
- 75,1% de reconhecimento com SVM na base GTZAN;
- 40,2% de reconhecimento com GMM na base Experimental;
- 46,6% de reconhecimento com SVM na base Experimental.

Em ambas as bases de dados, o SVM mostrou resultados estatisticamente superiores. Na base GTZAN, a diferença absoluta chegou a 13,6% e na base Experimental, o SVM apresentou taxa de reconhecimento 6,4% maior que o GMM. Os resultados corroboram com os apresentados em outros trabalhos, mostrados na Tabela 4.1 na Seção 4.4. Através destes resultados, reforça-se a ideia de que o SVM apresenta melhores resultados em comparação ao GMM para o reconhecimento de gêneros.

Um fato a ser considerado sobre o sistema proposto foi a evolução da taxa de reconhecimento ao longo de todas as etapas do sistema. O sistema proposto possui a divisão em três etapas principais: classificadores por grupo (separado por grupo de características), classificadores por segmento (com a junção dos classificadores por grupo de cada segmento) e decisão final (com a análise das probabilidades dos três classificadores por segmento). Em todas as bases, em ambos os classificadores, o sistema mostrou evolução na taxa de reconhecimento em cada etapa, conforme mostrado na Figura 5.17. Isto mostra que a junção das saídas dos classificadores por grupo e as regras de decisão surtiram um bom efeito no reconhecimento dos gêneros.

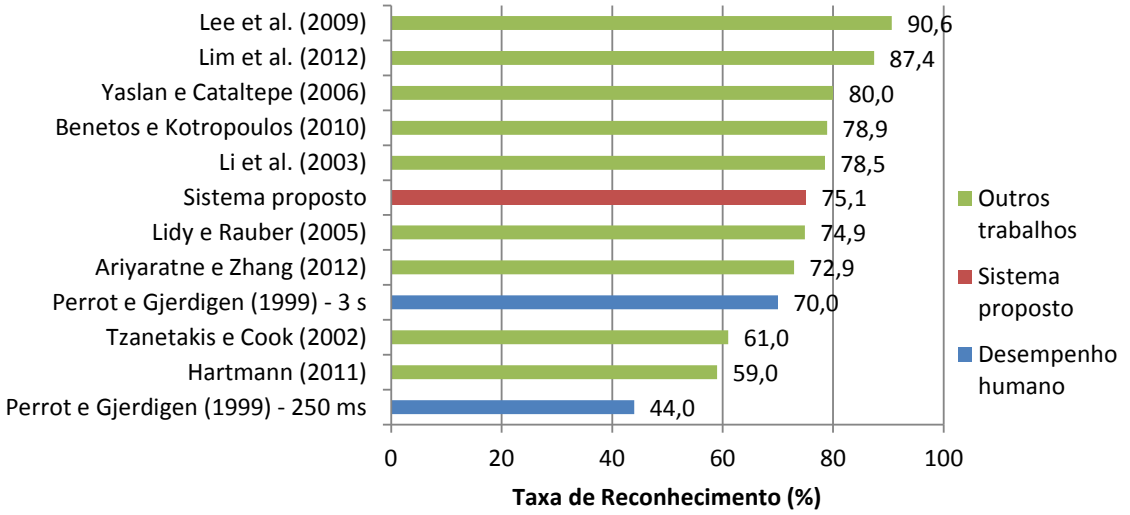
Figura 5.17 - Evolução da taxa de reconhecimento ao longo das etapas do sistema.



Fonte: Autor.

A Figura 5.18 mostra os melhores resultados obtidos em outros trabalhos de reconhecimento de gêneros musicais utilizando todas as músicas da base de dados GTZAN, com *10-fold cross-validation*. Também é mostrado o desempenho humano em reconhecer gêneros musicais, conforme reportado no trabalho de Perrot e Gjerdigen (1999). Todos os resultados apresentados representam a melhor taxa média de reconhecimento em 10 gêneros musicais.

Figura 5.18 - Taxa de reconhecimento de gêneros de trabalhos correlatos e do sistema proposto.



Fonte: Autor.

Conforme verificado na Figura 5.18, o sistema proposto apresentou resultados muito próximos ao desempenho humano, e encontra-se na faixa média de resultados encontrados em outros trabalhos. Para os trabalhos que obtiveram taxas de reconhecimento inferiores ao sistema proposto, o classificador e as características utilizadas podem ter sido a causa destes resultados. No trabalho de Hartmann (2011), um conjunto menor de características foi utilizado (somente informações espectrais), o que pode ter gerado pouca informação para a distinção dos gêneros. Já no trabalho de Tzanetakis e Cook (2002), a escolha de classificadores provavelmente menos aptos para o reconhecimento de gêneros, como o GMM e o kNN, pode ter sido a causa de uma taxa de reconhecimento menor.

Já para os trabalhos que obtiveram resultados superiores ao sistema proposto, foi verificada a existência de diversas técnicas para aumentar o reconhecimento, seja na realização de variações na arquitetura padrão de reconhecimento de gêneros, ou na criação de novas características e classificadores. Como principais técnicas, pode-se citar:

- Uso de características pouco exploradas, como o caso do contraste espectral baseado em oitavas nos trabalhos de Lee *et al.* (2009) e Lim *et al.* (2012), envelope espectral normalizado do áudio no trabalho de Lee *et al.* (2009) e histograma dos coeficientes da transformada Wavelet de *Daubechies* no trabalho de Li *et al.* (2003). Estas características são mais complexas de serem obtidas, e é necessária uma série de cálculos para extraí-las do áudio;
- Criação de um classificador (*Non-Negative Tensor Factorization*), conforme reportado no trabalho de Benetos e Kotropoulos (2010);
- Utilização de algoritmos de seleção de características nos trabalhos de Yaslan e Cataltepe (2006), Benetos e Kotropoulos (2010) e Lim *et al.* (2012). Estes algoritmos objetivam reduzir o número de informações e retirar as características que não contribuem para o reconhecimento.

A diferença significativa da taxa de reconhecimento do sistema entre as duas bases de dados é uma questão a ser observada. Conforme verificado, os resultados do sistema na base GTZAN foram muito maiores do que na base Experimental (diferença absoluta de 21,3% utilizando GMM e 28,5% com o SVM). Algumas possíveis razões para esta diferença de reconhecimento são:

- Gêneros envolvidos entre as bases: apesar da semelhança de 9 gêneros, a base Experimental contém o gênero Latinas, que é um grupo de músicas difícil de ser classificado corretamente, pois envolve diversos subgêneros bastante divergentes entre eles (como Cuba, Salsa, Tango, Axé, etc.). Maiores explicações sobre a divergência dos gêneros é verificada na Seção 5.4.1;
- Formato dos arquivos de áudio: enquanto a base GTZAN é constituída por arquivos de áudio no formato AU, um formato em que os dados não são comprimidos, a base Experimental contém músicas no formato MP3, que utiliza compactação, o que reduz a qualidade dos dados analisados;

- Segmento da música: enquanto a base Experimental apresenta músicas completas (podendo conter partes com silêncios ou que não representem fielmente o gênero da música), a base GTZAN contém apenas 30 segundos de cada música, retiradas do meio, o que garante maior qualidade no conteúdo musical lido;
- Diversidade das bases: conforme Sturm (2013), a base GTZAN apresenta muitas músicas repetidas, e inclusive dos mesmos artistas, o que restringe bastante à definição do gênero, facilitando sua identificação. Já para a base Experimental, um dos objetivos durante sua construção foi justamente diversificar ao máximo as músicas, incluindo canções dos mais diversos subgêneros. Isto, sem dúvida, dificulta o reconhecimento, mas deixa a base mais próxima de uma possível situação real de atuação de um sistema de reconhecimento de gêneros musicais;
- Diferença do ano de construção das bases: durante os últimos anos, alguns gêneros musicais passaram por diversas "atualizações" e "modificações". Gêneros como Eletrônica, Hip-Hop, Metal e Pop, por exemplo, tiveram novos subgêneros criados e a abrangência do gênero tornou-se muito mais ampla do que quando comparado aos anos 2000, ano de construção da base GTZAN. Certamente hoje a definição de alguns gêneros é mais complicada devido as constantes evoluções no mundo da música. Maiores explicações sobre a evolução dos gêneros ao longo dos anos são apresentadas na Seção 5.4.1.

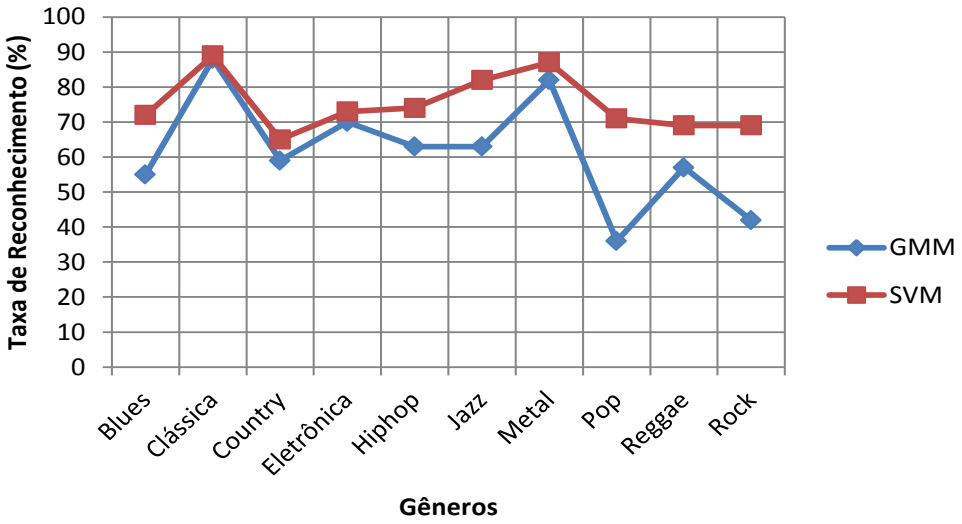
Os resultados completos do melhor teste com cada base de dados em cada classificador são encontrados ao final deste trabalho, na Seção de Apêndices.

5.4.1 RECONHECIMENTO NOS GÊNEROS

Complementando os resultados obtidos e apresentados anteriormente, também se buscou analisar o reconhecimento do sistema nos diferentes gêneros musicais envolvidos. Assim, é possível identificar se alguns gêneros são mais fáceis de serem identificados do que outros. Os valores obtidos foram capturados do teste que obteve a

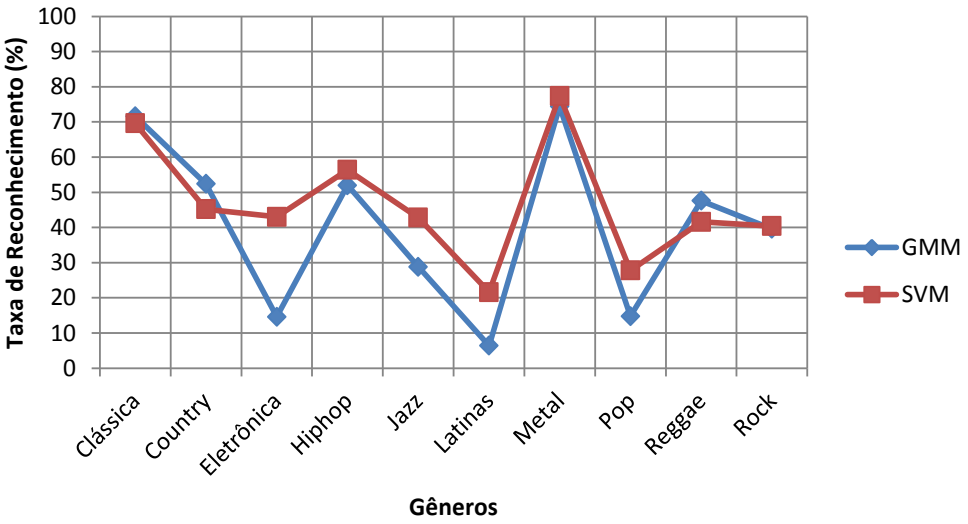
melhor taxa de reconhecimento geral em cada base de músicas. Os resultados encontram-se na Figura 5.19 e na Figura 5.20.

Figura 5.19 - Resultados obtidos em cada gênero musical da base GTZAN.



Fonte: Autor.

Figura 5.20 - Resultados obtidos em cada gênero musical da base Experimental.



Fonte: Autor.

Nota-se que os resultados para os gêneros foram similares para ambos os classificadores, mostrando que as dificuldades e facilidades encontradas para distinguir cada gênero são as mesmas. Percebe-se que alguns gêneros são naturalmente mais

distinguíveis do que outros, como é o caso da música Clássica e de Metal, que apresentaram as maiores taxas de reconhecimento em ambas às bases. De fato, estes são dois gêneros com características bem peculiares. A música Clássica apresenta batidas leves e instrumentos que dificilmente aparecem em outros gêneros, como piano, violino, órgão, etc. Já o Metal apresenta ritmos mais rápidos, e com uma sonoridade mais grave, o que não é encontrada em outros gêneros (KOSINA, 2002). O mesmo já não acontece, por exemplo, com o gênero Rock, o qual apresentou baixas taxas de reconhecimento em ambas às bases. O Rock é um gênero que apresenta uma sonoridade mais comum, e geralmente é mesclado com outros gêneros musicais, formando, por exemplo, o Blues Rock, Country Rock ou mesmo o Pop Rock, tornando a distinção deste gênero mais complicada (CHEN e RAMADGE, 2013). Verificando outros trabalhos de reconhecimento de gêneros, percebe-se que de fato alguns gêneros são melhores distinguidos do que outros. A Tabela 5.3 mostra alguns trabalhos que utilizaram a base GTZAN, indicando quais foram os 2 gêneros mais bem reconhecidos pelo sistema, e os 2 gêneros que obtiveram menor taxa de reconhecimento. Nestes exemplos, fica claro que a música Clássica e Metal são gêneros mais fáceis de serem reconhecidos, enquanto que Blues, Country, Reggae e Rock são mais complicados.

Tabela 5.3 - Gêneros mais e menos reconhecidos nos sistemas de reconhecimento de gêneros.

TRABALHO	MELHOR TAXA DE RECONHECIMENTO	PIOR TAXA DE RECONHECIMENTO
Sistema proposto	Clássica (89%) e Metal (87%)	Country (65%) e Reggae/Rock (69%)
Tzanetakis e Cook (2002)	Jazz (75%) e Clássica (69%)	Rock (40%) e Blues (43%)
Yaslan e Cataltepe (2006)	Clássica (96%) e Hip-Hop (86%)	Reggae (78%) e Rock (81%)
Bagci e Erzin (2007)	Clássica (94%) e Metal (81%)	Rock (44%) e Blues (60%)
Benetos e Kotropoulos (2010)	Metal (92%) e Clássica (88%)	Hip-Hop (67%) e Rock (71%)
Ariyaratne e Zhang (2012)	Clássica (86%) e Metal (82%)	Rock (58%) e Country (58%)
Chathuranga e Jayaratne (2013)	Clássica (98%) e Metal (86%)	Rock (63%) e Reggae (65%)

Uma observação é necessária sobre o resultado do gênero Latinas da base Experimental. O gênero Latinas foi pouco reconhecido pelo sistema (não alcançou nem 22%), e isto provavelmente aconteceu devido a uma dificuldade própria deste gênero. Originalmente, o gênero Latinas envolve músicas de diversos gêneros que tiveram

origem na América Latina, como Cuba, Flamenco, Salsa, Tango, etc. Estes gêneros, por si só, já são bem diferentes (por exemplo, o Tango faz uso de elementos da música Clássica e possui um ritmo mais lento, enquanto que a Salsa é um gênero mais animado e que utiliza diferentes instrumentos, como percussão, castanholas, etc.), o que abrange demais a definição deste gênero, dificultando o reconhecimento. Além disso, uma pesquisa mais aprofundada no site Last.fm mostrou divergências em algumas músicas classificadas pelo site como Latinas. Algumas canções que seriam mais próximas de outros gêneros são classificadas como Latinas simplesmente pelo fato da banda ser de origem Latina, levando em conta a classificação apenas pela localização geográfica do grupo ao invés da sonoridade da música. Estes fatores podem ter contribuído consideravelmente para o baixo reconhecimento deste gênero.

Outra observação importante se diz respeito ao gênero Pop, que foi altamente reconhecido na base GTZAN, mas o mesmo não aconteceu na base Experimental. Um provável motivo para esta diferença pode ter sido o ano de construção das bases. A base GTZAN, por exemplo, foi construída por volta dos anos 2000, onde o gênero Pop era bem definido, com músicas de ritmo dançante e com uma batida mais constante, sem muitas variações rítmicas. Dos anos 2000 para os dias de hoje, o gênero Pop passou por inúmeras mudanças, abrangendo características de diversos gêneros. Atualmente, grande parte das músicas Pop faz um uso bem mais significativo de efeitos eletrônicos; muitos vocalistas cantam de forma similar aos cantores de Rap ou Hip-Hop e as músicas chamadas de acústicas (voz e violão) são muitas vezes classificadas como Pop. A mudança do gênero ao longo dos anos é grande, e a conceituação do Pop atualmente é mais complicada do que à de alguns anos atrás. Logo, torna-se um gênero mais difícil de ser diferenciado dos demais. O mesmo acontece com outros gêneros como, por exemplo, a música Eletrônica. Na base GTZAN, inclusive, a música Eletrônica é referenciada apenas como “Disco”, envolvendo músicas principalmente dos anos 80 e 90, com um ritmo dançante similar ao Rock, bastante tocadas em discotecas. Hoje, o conceito de música Eletrônica é muito maior, envolvendo, além destas músicas, canções com muitos efeitos eletrônicos e batidas mais fortes. Os próprios instrumentos e o som também são diferentes, abrangendo atualmente uma sonoridade mais futurista e/ou psicodélica. A mistura e globalização da música, com diversas canções envolvendo características de vários gêneros torna a categorização destas músicas mais complicada.

A matriz de confusão é uma forma interessante de visualizar como o sistema classificou as músicas, permitindo assim descobrir quais gêneros foram confundidos com outros. A Tabela 5.4 mostra a matriz de confusão para os gêneros da base GTZAN, e a Tabela 5.5 contém a matriz de confusão do melhor resultado da base Experimental, onde ambas as bases envolvem os seguintes gêneros: BLU – Blues, CLA – Clássica, COU – Country, ELE – Eletrônica, HIP – Hip-Hop, JAZ – Jazz, LAT – Latinas, MET – Metal, POP – Pop, REG – Reggae, ROC – Rock. As células coloridas (diagonal principal da matriz) representam a quantidade de amostras classificadas corretamente pelo sistema (no total são 100 amostras de cada gênero para a base GTZAN e 500 para a base Experimental). Cada linha representa as amostras daquele gênero e onde elas foram classificadas. Já cada coluna indica quantas músicas de cada gênero foram classificadas no gênero representado por aquela coluna.

Tabela 5.4 - Matriz de confusão para a base GTZAN.

	BLU	**CLA**	**COU**	**ELE**	**HIP**	**JAZ**	**MET**	**POP**	**REG**	**ROC**
BLU	**72**	0	8	2	0	3	5	0	3	7
CLA	2	**89**	2	0	0	2	0	0	0	5
COU	9	0	**65**	6	0	0	3	3	4	10
ELE	3	1	5	**73**	2	0	3	7	1	5
HIP	2	0	1	4	**74**	0	3	4	12	0
JAZ	4	6	2	0	1	**82**	1	2	0	2
MET	1	0	0	6	0	0	**87**	0	0	6
POP	0	1	5	4	5	1	1	**71**	2	10
REG	5	1	6	2	7	1	0	6	**69**	3
ROC	4	1	6	8	1	3	2	0	6	**69**

Tabela 5.5 - Matriz de confusão para a base Experimental.

	CLA	**COU**	**ELE**	**HIP**	**JAZ**	**LAT**	**MET**	**POP**	**REG**	**ROC**
CLA	**348**	30	23	2	52	15	9	12	3	6
COU	16	**226**	15	11	40	40	9	57	35	51
ELE	41	19	**215**	33	27	29	16	65	31	24
HIP	0	20	41	**282**	17	30	5	20	81	4
JAZ	54	35	26	20	**214**	73	5	29	17	27
LAT	32	72	39	37	61	**108**	12	51	62	26
MET	3	11	11	0	3	2	**387**	8	1	74
POP	17	67	70	18	27	40	12	**139**	30	80
REG	5	31	40	77	18	41	12	29	**208**	39
ROC	7	72	21	5	15	12	71	63	32	**202**

As matrizes de confusão apresentadas mostram que o sistema esteve no caminho correto para a classificação das músicas, e grande parte das confusões foram entre gêneros que de fato são muito parecidos. A música Clássica, por exemplo, possui certas semelhanças com o Jazz (ambas possuem ritmos mais lentos e uma sonoridade mais suave), e grandes partes das confusões da música Clássica foram com o Jazz e vice-versa. O Rock, que também aparece em subgêneros de outros gêneros musicais, como Blues Rock, Country Rock e Pop Rock, também foi bastante confundido entre estes gêneros, além de também ter certas semelhanças com o Metal. Já a música Eletrônica e o Pop também possuem alguns elementos musicais próximos, e por isso foram, na maioria dos casos, confundidos entre eles. Por fim, a grande maioria dos erros de classificação das músicas de Hip-Hop foi com o Reggae e vice-versa. Aparentemente, são gêneros bem diferentes, porém alguns subgêneros do Reggae, como o Ska, apresentam uma batida rítmica bem semelhante ao Hip-Hop. A forma de cantar dos dois gêneros também é parecida, onde a voz é mais falada do que gritada. Estas semelhanças provavelmente fizeram estes gêneros serem mais próximos e, portanto, mais confundidos. Estas confusões entre os gêneros foram similares às encontradas nos trabalhos de Tzanetakis e Cook (2002), Benetos e Kotropoulos (2010) e Chathuranga e Jayaratne (2013).

As matrizes de confusão mostram que o sistema cometeu erros que provavelmente qualquer pessoa cometeria. A grande maioria das confusões se deu entre gêneros que são parecidos, mostrando que o sistema está conseguindo encontrar características diferenciais entre os gêneros, mas é afetado pelo fato de que algumas destas informações são comuns entre as classes. A grande variedade de canções dentro de um gênero abrange muito o conceito do grupo, e alguns subgêneros são originalmente difíceis de serem classificados. O que se percebe é que o sistema consegue diferenciar aquelas músicas que servem “de base” para a definição de um gênero musical e que possuem as características mais marcantes de cada gênero. Já casos de canções que apresentam elementos musicais de vários gêneros tornam o reconhecimento muito mais complicado para um sistema computacional.

6. CONSIDERAÇÕES FINAIS

O reconhecimento de gêneros musicais é uma tarefa complexa e requer um alto conhecimento em música, processamento de sinais e reconhecimento de padrões. Este trabalho procurou explicar os principais conceitos e técnicas envolvidas no reconhecimento de gênero. Até hoje, várias técnicas foram desenvolvidas, envolvendo os mais diversos tipos de características, classificadores e arquiteturas de sistema. Porém, ainda não foi desenvolvido um sistema apto a ser utilizado em uma situação real.

O sistema proposto neste trabalho está focado na modificação da arquitetura padrão dos sistemas de reconhecimento de gêneros, dividindo a tarefa de identificar um gênero em várias etapas. Primeiramente as músicas são analisadas em 3 segmentos (início, meio e fim da música). Para cada segmento, características são extraídas e unidas em 4 grupos principais, conforme o tipo destas características. Classificadores são utilizados para cada um destes grupos e em seguida, as probabilidades dos classificadores dos grupos de cada segmento são unificadas e utilizadas como novas características em classificadores para cada segmento. Após, as probabilidades de cada gênero em cada segmento são analisadas e é decidido o gênero da música. É sabido que, por ser um sistema demasiadamente grande, provavelmente este sistema não terá um tempo de execução aceitável para uma situação real, devido aos vários classificadores e a grande quantidade de dados envolvidos. Como obter uma alta taxa de reconhecimento já é um desafio muito grande, optou-se por priorizar o reconhecimento, sem se preocupar muito com o tempo de execução.

A partir do desenvolvimento deste trabalho, várias informações foram obtidas quanto ao reconhecimento de gêneros musicais:

- A divisão da música em segmentos mostra-se uma boa forma de analisar o gênero de uma música. A análise de um segmento não necessariamente aumenta a taxa de reconhecimento, mas evita o processamento completo de uma música (o que envolve muito tempo de execução) sem diminuir o reconhecimento. Para uma aplicação em tempo real, a segmentação torna-se muito interessante;

- A separação de características por grupos através do seu tipo ajuda no reconhecimento. Conforme reportado, a análise da junção das saídas dos classificadores de cada grupo mostrou resultados superiores do que os resultados individuais de cada grupo. Percebe-se que a análise separada de cada grupo e união posterior destas informações aumenta o reconhecimento;
- Não foi determinado que algumas características fossem melhores para reconhecer certos gêneros, e piores para outros. Porém, conclui-se que alguns gêneros são mais facilmente reconhecidos do que outros, independentemente das informações utilizadas;
- Todas as características utilizadas auxiliaram no reconhecimento dos gêneros. Porém, algumas destas informações obtiveram melhores resultados, como as informações de timbre e MFCC, especialmente quando utilizadas de forma temporal (com momentos estatísticos). Já as informações rítmicas e de pitch apresentam resultados inferiores, e fizeram leves contribuições no reconhecimento dos gêneros. Isto se deve a enorme complexidade de conseguir capturar a estrutura rítmica de uma música, e de conseguir obter informações sobre as notas fundamentais de uma canção frente a todos os instrumentos e fontes sonoras utilizadas em uma música;
- Na classificação, a escolha do algoritmo tem um papel importante no reconhecimento do sistema. Neste trabalho, o classificador não paramétrico (SVM) apresentou melhores taxas de reconhecimento do que o classificador paramétrico (GMM);
- A combinação de classificadores aumenta consideravelmente a taxa de acerto. A análise de probabilidades de vários classificadores e a decisão a partir destes valores contribuem para um melhor reconhecimento. Nas regras de combinação utilizadas, a regra da soma com alteração de probabilidades pelo desempenho dos classificadores se mostrou a melhor alternativa para unir as informações de diferentes classificadores.

Conclui-se que o reconhecimento de gêneros por parte de computadores é uma tarefa complexa e que ainda está longe de ser utilizada no dia-a-dia. Para serem utilizados em situações reais, os sistemas de reconhecimento de gêneros precisam, além de conseguirem distinguir os mais diversos tipos de música, serem continuamente atualizados e adaptados às constantes mudanças que ocorrem no mundo da música.

O objetivo deste trabalho, além de desenvolver um sistema de reconhecimento de gêneros musicais, foi de prover informações e estatísticas para futuros trabalhos. O intuito foi de apresentar dados que pudessem evidenciar quais caminhos devem ser seguidos para melhorar as taxas de acerto (e quais não devem).

Para futuros trabalhos, será continuado o uso da arquitetura proposta, com algumas modificações. O SVM continuará a ser utilizado, e poderá ser feito uso de outro classificador não paramétrico. As características permanecerão sendo usadas em grupos, porém novas informações do áudio poderão ser utilizadas. Novos algoritmos de detecção rítmica e de pitch serão analisados para verificar se proveem informações mais precisas sobre o conteúdo musical processado. Além disso, continuamente novas características são lançadas em diversas publicações. Estas informações poderão ser analisadas para verificar a eficácia destas no reconhecimento de gêneros musicais. Também será analisado o uso de algoritmos de seleção de características, com o objetivo de reduzir a quantidade de informações utilizadas sem prejudicar o reconhecimento.

REFERÊNCIAS

ALDER, Michael, **An Introduction to Pattern Recognition**. HeavenForBooks.com, 2001.

ALMEIDA, P. R. L., BRITTO, A. S., SILVA, E. J., OLIVEIRA, L. E. S., CELINSKI, T. M., KOERICH A. L., **Music Genre Classification using Dynamic Selection of Ensemble of Classifiers**. 2012 IEEE International Conference on Systems, Man, and Cybernetics, COEX, Seoul, Korea, 2012.

ARYAFAR, K., JAFARPOUR, S., SHOKOUFANDEH, A., **Music Genre Classification Using Sparsity-Eager Support Vector Machines**. Drexel University, Philadelphia, USA, 2012.

ARIYARATNE, H. B., ZHANG, D., **A Novel Automatic Hierarchical Approach to Music Genre Classification**. IEEE International Conference on Multimedia and Expo Workshops, 2012.

AUCOUNTURIER, J. J., PACHET, F., **Representing Musical Genre: A State of the Art**. Journal of New Music Research, p. 83-93, 2005.

BAGCI, U., ERZIN, E., **Automatic Classification of Musical Genres Using Inter-Genre Similarity**. IEEE Signal Processing Letters, v. 14, n. 8, p. 521-524 2007.

BALTI, H., FRIGUI, H., **Feature Mapping and Fusion for Music Genre Classification**. IEEE, 11th International Conference on Machine Learning and Applications, p. 306-310, 2012.

BANITALEBI-DEHKORDI, M., BANITALEBI, A., **Music Genre Classification Using Spectral Analysis and Sparse Representation of the Signals**. Springer Science, New York, 2012.

BARREIRA, F. M., Luís, **Unsupervised Automatic Music Genre Classification**. Dissertação (Mestrado em Engenharia Informática), Universidade Nova de Lisboa, Lisboa, 2010.

BASILI, R., SERAFINI, A., STELLATO, A., **Classification of Musical Genre: A Machine Learning Approach**. Universitat Pompeu Fabra, 2004.

BENETOS, E., KOTROPOULOS, C., **Non-Negative Tensor Factorization Applied to Music Genre Classification**. IEEE Transactions on Audio, Speech, and Language Processing, v. 18, n. 8, p. 1955-1967, 2010.

BERGSTRA, J., CASAGRANDE, N., ERHAN, D., ECK, D., KÉGL, B., **Aggregate features and AdaBoost for music classification**. Springer Science, v. 65, p. 473-484, 2006.

BIRD, S., KLEIN, E., LOPER, E., **Natural Language Processing with Python**. 2009.

BISHOP, M. Christopher, **Pattern Recognition and Machine Learning**. New York: Springer Science + Business Media, 2006.

BOSI, M., GOLDBERG, R. E., **Introduction to Digital Audio Coding and Standards**. The Netherlands: Kluwer Academic Publishers, 2003.

BURRED, J. J., LERCH, A., **A Hierarchical Approach to Automatic Music Genre Classification**. 6th International Conference on Digital Audio Effects (DAFx-03), London, 2003.

CAMPBELL, W. M., STURIM, D. E., REYNOLDS, D. A., SOLOMONOFF, A., **SVM Based Speaker Verification Using a GMM Supervector Kernel and NAP Variability Compensation**. Technical Report, MIT Lincoln Laboratory, 2005.

CHATHURANGA, Y. M. D., JAYARATNE, K. L., **Automatic Music Genre Classification of Audio Signals with Machine Learning Approaches**. GSTF International Journal on Computing (JoC), v. 3, n. 2, 2013.

CHEN, X., RAMADGE, P. J., **Music Genre Classification Using Multiscale Scattering and Sparse Representations**. IEEE, v. 13, p. 978-983, 2013.

CHU, Mei-Lan, **Automatic Music Genre Classification**. Technical Report, Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taiwan, 2009.

COSTA, C. H. L., VALLE, J. D., KOERICH, A. L., **Automatic Classification of Audio Data**. IEEE International Conference on Systems, Man and Cybernetics, 2004.

COSTA, Y. M. G., OLIVEIRA, L. S., KOERICH, A. L., GOUYON, F., **Music Genre Recognition Using Spectrograms**. Technical Report, 2011.

COSTA, Y. M. G., OLIVEIRA, L. S., KOERICH, A. L., GOUYON, F., MARTINS, J. G., **Music genre classification using LBP textural features**. Elsevier, Signal Processing, n. 92, p. 2723-2737, 2012.

DAVIS, S. B., MERMELSTEIN, P., **Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences**. IEEE Transaction on Acoustics, Speech and Signal Processing, v. 28, n. 4, p. 357-366, 1980.

DOUGHERTY, Geoff., **Pattern Recognition and Classification**. New York: Springer Science + Business Media, 2013.

DUDA, R. O., HART, P. E., STORK, D. G., **Pattern Classification**. New York: Wiley Interscience, Second Edition, ISBN: 0-471-05669-3, 2001.

ELLIS, Daniel P. W., **PLP and RASTA (and MFCC, and inversion) in Matlab**, online resource, url: <http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/>, 2005. Disponível em: <http://labrosa.ee.columbia.edu/matlab/rastamat/>. Acesso em: 30 de mar. 2014.

FABBRI, Franco, **A Theory of Musical Genres: Two Applications**. First International Conference on Popular Music Studies, Amsterdam, 1981.

FINGERHUT, M., DONIN, N., **Filling Gaps Between Current Musicological Practice and Computer Technology at Ircam**. Technical Report, IRCAM, 2006.

FU, Z., LU, G., TING K. M., ZHANG, D., **A Survey of Audio-Based Music Classification and Annotation**. IEEE Transactions on Multimedia, v. 13, n. 2, p. 303-319, 2011.

FUKUNAGA, Keinosuke, **Introduction to statistical pattern recognition**. San Diego: Academic Press, Second Edition, 1990.

GILLICK, L., COX, S. J., **Some statistical issues in the comparison of speech recognition algorithms**. Acoustics, Speech, and Signal Processing, 1989, ICASSP-89., 1989 International Conference on, v. 1, p. 532-535, 1989.

GOLD, B., MORGAN, N., ELLIS, D., **Speech and Audio Signal Processing: Processing and Perception of Speech and Music**. John Wiley & Sons, Second Edition, 2011.

GRIMALDI, M., CUNNINGHAM, P., KOKARAM, A., **Discrete wavelet packet transform and ensembles of lazy and eager learners for music genre classification**. Springer, Multimedia Systems, p. 422-437, 2006.

HAGGBLADE, M., HONG, Y., KAO, K., **Music Genre Classification**. Technical Report, 2012.

HARTMANN, A. Martin, **Testing a Spectral-Based Feature Set for Audio Genre Classification**. Dissertação (Mestrado), University of Jyväskylä, 2011.

HASTIE, S. L., TIBSHIRANI, D. M, **Gaussian mixture models**. Technical Report, 2008.

HERRERA-BOYER, P., GOUYON, F., **MIRrors: Music Information Research reflects on its futures**. Springer Science, New York, 2013.

HUANG, X., ACERO, A., HON, H., **Spoken Language Processing: A Guide to Theory, Algorithm, and System Development**. Prentice Hall, 2001.

JOTHILAKSHMI, S., KATHIRESAN, N., **Automatic Music Genre Classification for Indian Music**. 2012 International Conference on Software and Computer Applications, v. 41, 2012.

KASSLER, M., **Toward Musical Information Retrieval**. Perspectives of New Music, v. 4, p. 59-67, 1966.

KIL, D. H., SHIN, F. B., **Pattern Recognition and Prediction with Application to Signal Characterization**. New York: American Institute of Physics, 1996.

KOERICH, A. L., POITEVIN, C., **Combination of Homogeneous Classifiers for Musical Genre Classification**. Technical Report, 2005.

KOSINA, Karin, **Music Genre Recognition**. Trabalho de Conclusão (Mídia e Tecnologia), Technical College of Hagenberg, Austria, Hagenberg, 2002.

LAMPROPOULOS, A. S., LAMPROPOULOU, P. S., TSIHRINTZIS, G. A, **Musical Genre Classification Enhanced by Improved Source Separation Techniques**. University of London, 2005.

LAMYA, F., HOUACINE A., **Artificial Neural Network genre classification of musical signals**. 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications – SETIT, Tunisia, 2007.

LEE, C. H., SHIH, J. L., YU, K. M., LIN, H. S., **Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features**. IEEE Transactions on Multimedia, v. 11, n. 4, p. 670-681, 2009.

LEON, F., MARTINEZ, K., **Towards Efficient Music Genre Classification Using Fastmap**. 15th International Conference on Digital Audio Effects (DAFx-12), York, 2012.

LI, T., OGIHARA, M., LI, Q., **A Comparative Study on Content-Based Music Genre Classification**. ACM 1-58113-646-3, Toronto, Canada, 2003.

LI, T., OGIHARA, M., **Toward Intelligent Music Information Retrieval**. IEEE Transactions on Multimedia, v. 8, n. 3, p. 564-574, 2006.

LIDY, T., RAUBER, A., **Evaluation of Feature Extractors and Psycho-Acoustic Transformations for Music Genre Classification**. Vienna University of Technology, Austria, 2005.

LIM, S. C., LEE, J. S., JANG, S. J., LEE, S. P., KIM, M. Y., **Music-Genre Classification System based on Spectro-Temporal Features and Feature Selection**. IEEE Transactions on Consumer Electronics, v. 58, n. 4, p. 1262-1268, 2012.

LINCOLN, H., **Some Criteria And Techniques For Developing Computerized Thematic Indeces**. Electronishe Datenverarbeitung in der Musikwissenschaft. Regensburg: Gustave Bosse Verlag, 1967.

MADEVSKA-BOGDANOVA, A., NIKOLIK, D., CURFS, L., **Probabilistic SVM outputs for pattern recognition using analytical geometry**. Elsevier, Neurocomputing, v. 62, p. 293-303, 2004.

MADJAROV, G., PESANSKI, G., SPASOVSKI, D., GJORGJEVIKJ, D., **Automatic Music Classification into Genres**. ICT Innovations 2012 Web Proceedings, ISSN 1857-7288, p. 623-632, 2012.

MALHEIRO, R., PAIVA, R. P., MENDES, A. J., MENDES, T., CARDOSO, A., **A Prototype for Classification of Classical Music Using Neural Networks**. Eighth

IASTED International Conference of Artificial Intelligence and Soft Computing, Marbella, Spain, p. 294-299, 2004.

MCKAY, C., FUJINAGA, I., **Automatic Genre Classification Using Large High-Level Musical Feature Sets**. Universitat Pompeu Fabra, Canada, 2004.

MCKINNEY, M. F., BREEBAART, J., **Features for Audio Music Classification**. Johns Hopkins University, 2003.

MENG, A., SHAWE-TAYLOR, J., **An Investigation of Feature Models for Music Genre Classification Using the Support Vector Classifiers**. Queen Mary, University of London, 2005.

MENG, A., AHRENDT, P., LARSEN, J., HANSEN, L. K., **Temporal Feature Integration for Music Genre Classification**. IEEE Transactions on Audio Speech, and Language Processing, v. 15, n. 5, p. 1654-1664, 2007.

MOORE, Allan F., **Categorical conventions in music discourse: style and genre**. Music and Letters, v. 82, n. 3, p. 432-442, 2001.

NOROWI, N. M., DORAISAMY, S., WIRZA, R., **Factors Affecting Automatic Genre Classification: An Investigation Incorporating Non-Western Musical Forms**. University Putra Malaysia, Faculty of Computer Science and Information Technology, Malaysia, 2005.

OPPENHEIM, A. V., SCHAFER, R. W., **Processamento em Tempo Discreto de Sinais**. São Paulo: Editora Pearson Education do Brasil, Terceira Edição, 2012. Tradução: Vieira, D.

PANAGAKIS, I., BENETOS, E., KOTROPOULOS, C., **Music Genre Classification: A Multilinear Approach**. International Symposium Music Information Retrieval, 14 – 18 September, USA, Philadelphia, 2008.

PARADZINETS, A., HARB, H., CHEN, L., **MultiExpert System For Automatic Music Genre Classification**. Research Report, Ecole Centrale de Lyon, 2009.

PERROT, D., GJERDIGEN, R. O., **Scanning the dial: An exploration of factors in the identification of musical style**. Proceedings of the Society for Music Perception and Cognition, pp. 88, 1999.

PLUMBEY, M., DIXON, S., **Tutorial: Music Signal Processing**. IMA Conference Mathematics in Signal Processing, 2012.

POHLE T., PAMPALK, E., WIDMER, G., **Evaluation of Frequently Used Audio Features for Classification of Music Into Perceptual Categories**. Technical Report, 2004.

REED, J., LEE, C.H., **A Study on Music Genre Classification Based on Universal Acoustic Models**. University of Victoria, 2006.

REYNOLDS, Douglas, **Gaussian Mixture Models**. Technical Report, MIT Lincoln Laboratory, 2001.

SCARINGELLA, N., ZOIA, G., MLYNEK, D., **Automatic Genre Classification of Music Content: A Survey**. IEEE Signal Processing Magazine, v. 23, n. 2, p. 133-141, 2006.

SCHAFER, R. W., Homomorphic Systems and Cepstrum Analysis of Speech. In: BENESTY, J., SONDHI, M. M., HUANG, Y., **Springer Handbook of Speech Processing**. Springer Science, capítulo 9, Parte B|9, p. 161-180, 2008.

SILLA, C. N., KAESTNER, C. A. A., KOERICH, A, L., **Classificação Automática de Gêneros Musicais Utilizando Métodos de Bagging e Boosting**. Pontifícia Universidade Católica do Paraná, Programa de Pós-Graduação em Informática Aplicada, Brasil, Curitiba, 2004.

SILLA, C. N., KAESTNER, C. A. A., KOERICH, A, L., **Automatic Music Genre Classification Using Ensemble of Classifiers**. IEEE, n. 7, p. 1687-1692, 2007.

STURM, L. Bob, **A Survey Of Evaluation in Music Genre Recognition**. Audio Analysis Lab, Aalborg University Copenhagen, Denmark, 2012.

STURM, L. Bob, **Classification accuracy is not enough on the evaluation of music genre recognition systems**. Springer Science, 2013.

TERMENS, G. Enric, **Audio Content Processing for Automatic Music Genre Classification: Descriptors, Databases, and Classifiers**. Dissertação (Doutorado em Ciência da Computação e Comunicação Digital), Universitat Pompeu Fabra, Barcelona, 2009.

THEODORIDIS, S., KOUTROUMBAS, K., **Pattern Recognition**. Greece: Elsevier Academic Press, Second Edition, 2003.

TOLONEN, T., KARJALAINEN, **A Computationally Efficient Multipitch Analysis Model**. IEEE Transactions on Speech and Audio Processing, v. 8, n. 6, p. 708-716, 2000.

TZANETAKIS, G., ESSL, G., COOK, P., **Automatic Musical Genre Classification of Audio Signals**. Technical Report, 2001.

TZANETAKIS, G., COOK, P., **Musical Genre Classification of Audio Signals**. IEEE Transactions on Speech and Audio Processing, v. 10, n. 5, p. 293-302, 2002.

WEBB, R. Andrew, **Statistical Pattern Recognition**. England: John Wiley & Sons Ltd., Second Edition, 2002.

WEST, K., COX, S., **Features and Classifiers for the Automatic Classification of Musical Audio Signals**. School of Computing Sciences, University of East Anglia, 2004.

WÜLFING, J., RIEDMILLER, M., **Unsupervised Learning of Local Features for Music Classification**. 13th International Society for Music Information Retrieval Conference – ISMIR, p. 139-144, 2012.

XU, C., MADDAGE, N. C., SHAO, X., **Automatic Music Classification and Summarization. IEEE Transactions on Speech and Audio Processing**, v. 13, n. 3, p. 441-450, 2005.

YASLAN, Y., CATALTEPE, Z., **Audio Music Genre Classification Using Different Classifiers and Feature Selection Methods**. The 18th International Conference on Pattern Recognition, IEEE, 2006.

ZHOU, X., YU, K., ZHANG T., HUANG, T. S., **Image Classification using Super-Vector Coding of Local Image Descriptors**. ECCV-10 submission ID 453, 2009.

APÊNDICE A - Resultados completos do teste na base Experimental utilizando GMM

A Tabela A.1 apresenta os resultados do melhor teste obtido na base Experimental utilizando GMM. As colunas representam os gêneros envolvidos, sendo que a última coluna apresenta a taxa média, enquanto que as linhas representam o grupo de classificadores.

Tabela A.1 - Resultados do teste na base Experimental com GMM.

	CLA	COU	ELE	HIP	JAZ	LAT	MET	POP	REG	ROC	GER
Grupo 1 – Segmento 1	51,20	28,60	8,40	16,80	16,60	7,40	72,00	8,20	12,00	17,80	23,90
Grupo 1 – Segmento 2	51,40	25,60	12,60	17,00	18,60	7,60	75,00	9,20	11,40	20,20	24,86
Grupo 1 – Segmento 3	49,20	25,40	10,60	18,40	15,00	7,60	72,00	11,80	10,80	19,40	24,02
Grupo 2 – Segmento 1	55,60	37,60	14,80	38,40	21,00	9,80	69,40	18,00	27,80	16,60	30,90
Grupo 2 – Segmento 2	56,80	34,00	14,60	41,80	25,60	12,00	68,00	14,80	26,60	26,80	32,10
Grupo 2 – Segmento 3	53,60	39,20	13,20	35,80	18,20	11,80	68,00	14,20	27,60	20,40	30,20
Grupo 3 – Segmento 1	22,20	17,20	13,20	28,20	20,00	8,40	50,20	5,20	16,80	11,00	19,24
Grupo 3 – Segmento 2	31,80	16,20	9,40	22,20	14,20	11,00	43,60	8,20	15,00	12,60	18,42
Grupo 3 – Segmento 3	20,00	20,80	12,60	25,80	17,80	7,20	49,20	11,20	15,00	16,80	19,64
Grupo 4 – Segmento 1	41,00	21,60	9,40	17,20	15,80	5,80	36,40	10,40	18,20	24,40	20,02
Grupo 4 – Segmento 2	45,00	28,40	9,20	14,40	10,80	10,80	40,20	9,20	21,60	23,00	21,26
Grupo 4 – Segmento 3	38,00	18,00	9,20	22,60	13,00	5,60	36,80	6,60	20,00	30,00	19,98
Segmento 1	71,80	45,60	8,80	47,40	20,80	5,00	70,20	15,00	44,20	28,60	35,74
Segmento 2	68,80	39,00	10,60	48,80	25,40	6,60	74,00	9,00	41,20	31,20	35,46
Segmento 3	67,00	39,60	11,60	53,40	20,80	7,40	71,60	13,40	37,00	27,60	34,94
Decisão Final	71,60	52,40	14,60	52,00	28,80	6,40	74,40	14,80	47,60	39,60	40,22

APÊNDICE B - Resultados completos do teste na base Experimental utilizando SVM

A Tabela B.1 apresenta os resultados do melhor teste obtido na base Experimental utilizando SVM. As colunas representam os gêneros envolvidos, sendo que a última coluna apresenta a taxa média, enquanto que as linhas representam o grupo de classificadores.

Tabela B.1 - Resultados do teste na base Experimental com SVM.

	CLA	COU	ELE	HIP	JAZ	LAT	MET	POP	REG	ROC	GER
Grupo 1 – Segmento 1	60,20	32,20	19,40	34,80	29,40	15,40	59,40	19,80	28,00	34,60	33,32
Grupo 1 – Segmento 2	57,60	32,20	23,60	34,60	29,20	17,00	66,60	23,20	26,00	35,80	34,58
Grupo 1 – Segmento 3	56,80	34,40	21,60	32,80	23,80	16,00	61,40	25,40	23,40	33,40	32,90
Grupo 2 – Segmento 1	47,60	44,80	21,80	43,80	29,00	15,80	61,80	21,80	33,80	40,00	36,02
Grupo 2 – Segmento 2	58,00	44,60	27,80	45,60	31,00	19,40	69,80	25,40	32,80	37,00	39,14
Grupo 2 – Segmento 3	51,20	36,00	22,60	43,20	27,40	15,60	62,00	34,40	33,00	38,40	36,38
Grupo 3 – Segmento 1	26,20	21,60	13,80	23,60	19,80	12,80	44,80	7,40	33,00	21,40	22,44
Grupo 3 – Segmento 2	23,80	21,20	17,60	21,80	12,40	6,00	38,40	21,60	22,60	15,00	20,04
Grupo 3 – Segmento 3	28,60	25,60	14,60	23,60	15,00	6,00	42,20	9,20	19,80	27,80	21,24
Grupo 4 – Segmento 1	33,40	22,00	16,60	21,20	15,20	5,60	37,60	15,60	26,00	33,20	22,64
Grupo 4 – Segmento 2	31,80	33,00	13,20	21,80	13,00	3,40	38,80	18,80	32,20	21,00	22,70
Grupo 4 – Segmento 3	34,80	30,60	19,00	28,00	14,00	7,20	37,40	14,60	26,60	24,40	23,66
Segmento 1	63,20	35,40	33,60	51,60	33,20	20,20	70,60	24,80	40,00	36,20	40,88
Segmento 2	62,60	41,00	36,80	52,80	37,60	25,00	71,60	24,40	37,80	36,00	42,56
Segmento 3	63,20	37,60	37,40	49,40	35,00	19,20	67,60	28,00	36,60	38,00	41,20
Decisão Final	69,60	45,20	43,00	56,40	42,80	21,60	77,40	27,80	41,60	40,40	46,58

APÊNDICE C - Resultados completos do teste na base GTZAN utilizando GMM

A Tabela C.1 apresenta os resultados do melhor teste obtido na base GTZAN utilizando GMM. As colunas representam os gêneros envolvidos, sendo que a última coluna apresenta a taxa média, enquanto que as linhas representam o grupo de classificadores.

Tabela C.1 - Resultados do teste na base GTZAN com GMM.

	BLU	**CLA**	**COU**	**ELE**	**HIP**	**JAZ**	**MET**	**POP**	**REG**	**ROC**	**GER**
Grupo 1 – Segmento 1	31,00	66,00	31,00	16,00	16,00	34,00	76,00	20,00	18,00	22,00	33,00
Grupo 1 – Segmento 2	30,00	65,00	34,00	18,00	21,00	27,00	76,00	22,00	19,00	24,00	33,60
Grupo 1 – Segmento 3	36,00	67,00	32,00	18,00	25,00	29,00	74,00	22,00	16,00	19,00	33,80
Grupo 2 – Segmento 1	57,00	80,00	46,00	26,00	41,00	66,00	62,00	41,00	56,00	33,00	50,80
Grupo 2 – Segmento 2	51,00	84,00	47,00	29,00	54,00	61,00	62,00	44,00	47,00	30,00	50,90
Grupo 2 – Segmento 3	56,00	86,00	43,00	34,00	46,00	62,00	61,00	51,00	60,00	26,00	52,50
Grupo 3 – Segmento 1	22,00	65,00	9,00	48,00	17,00	29,00	39,00	10,00	11,00	13,00	26,30
Grupo 3 – Segmento 2	16,00	38,00	19,00	52,00	17,00	22,00	52,00	10,00	16,00	16,00	25,80
Grupo 3 – Segmento 3	6,00	61,00	22,00	50,00	15,00	24,00	45,00	4,00	14,00	19,00	26,00
Grupo 4 – Segmento 1	17,00	71,00	22,00	13,00	13,00	57,00	33,00	13,00	19,00	12,00	27,00
Grupo 4 – Segmento 2	10,00	66,00	35,00	10,00	27,00	41,00	45,00	16,00	23,00	15,00	28,80
Grupo 4 – Segmento 3	25,00	74,00	38,00	14,00	29,00	30,00	20,00	10,00	21,00	12,00	27,30
Segmento 1	39,00	88,00	46,00	63,00	46,00	61,00	78,00	32,00	38,00	39,00	53,00
Segmento 2	35,00	89,00	48,00	70,00	51,00	49,00	81,00	25,00	45,00	35,00	52,80
Segmento 3	41,00	87,00	50,00	66,00	57,00	43,00	77,00	30,00	46,00	33,00	53,00
Decisão Final	55,00	88,00	59,00	70,00	63,00	63,00	82,00	36,00	57,00	42,00	61,50

APÊNDICE D - Resultados completos do teste na base GTZAN utilizando SVM

A Tabela D.1 apresenta os resultados do melhor teste obtido na base GTZAN utilizando SVM. As colunas representam os gêneros envolvidos, sendo que a última coluna apresenta a taxa média, enquanto que as linhas representam o grupo de classificadores.

Tabela D.1 - Resultados do teste na base GTZAN com SVM.

	BLU	**CLA**	**COU**	**ELE**	**HIP**	**JAZ**	**MET**	**POP**	**REG**	**ROC**	**GER**
Grupo 1 – Segmento 1	42,00	69,00	44,00	45,00	38,00	40,00	71,00	58,00	41,00	42,00	49,00
Grupo 1 – Segmento 2	43,00	68,00	37,00	42,00	45,00	38,00	63,00	56,00	34,00	40,00	46,60
Grupo 1 – Segmento 3	35,00	67,00	36,00	42,00	48,00	39,00	72,00	52,00	43,00	31,00	46,50
Grupo 2 – Segmento 1	49,00	81,00	49,00	49,00	54,00	65,00	77,00	57,00	49,00	55,00	58,50
Grupo 2 – Segmento 2	44,00	76,00	50,00	51,00	50,00	63,00	70,00	54,00	54,00	49,00	56,10
Grupo 2 – Segmento 3	51,00	70,00	45,00	62,00	51,00	54,00	77,00	58,00	48,00	35,00	55,10
Grupo 3 – Segmento 1	16,00	25,00	14,00	42,00	24,00	24,00	28,00	22,00	21,00	20,00	23,60
Grupo 3 – Segmento 2	19,00	38,00	17,00	51,00	22,00	28,00	26,00	15,00	16,00	25,00	25,70
Grupo 3 – Segmento 3	18,00	38,00	12,00	50,00	21,00	31,00	35,00	10,00	22,00	21,00	25,80
Grupo 4 – Segmento 1	11,00	64,00	27,00	28,00	32,00	45,00	31,00	21,00	37,00	17,00	31,30
Grupo 4 – Segmento 2	13,00	51,00	32,00	33,00	30,00	45,00	22,00	24,00	30,00	19,00	29,90
Grupo 4 – Segmento 3	20,00	58,00	33,00	31,00	36,00	50,00	31,00	19,00	34,00	24,00	33,60
Segmento 1	58,00	90,00	59,00	67,00	59,00	72,00	77,00	65,00	58,00	60,00	66,50
Segmento 2	64,00	81,00	57,00	67,00	66,00	78,00	75,00	65,00	63,00	56,00	67,20
Segmento 3	57,00	82,00	52,00	65,00	64,00	70,00	82,00	66,00	60,00	50,00	64,80
Decisão Final	72,00	89,00	65,00	73,00	74,00	82,00	87,00	71,00	69,00	69,00	75,10