CLASSIFICAÇÃO AUTOMÁTICA DE GÊNEROS MUSICAIS BASEADA EM APROXIMAÇÕES POLINOMIAIS NO DOMÍNIO DO TEMPO

Marcos José Sant'Anna Magalhães

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Luiz Wagner Pereira Biscainho

Rio de Janeiro
Setembro de 2013

CLASSIFICAÇÃO AUTOMÁTICA DE GÊNEROS MUSICAIS BASEADA EM APROXIMAÇÕES POLINOMIAIS NO DOMÍNIO DO TEMPO

Marcos José Sant'Anna Magalhães

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Luiz Wagner Pereira Biscainho, D.Sc.

Prof. José Gabriel Rodriguez Carneiro Gomes, Ph.D.

Prof. Márcio Nogueira de Souza, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2013

Magalhães, Marcos José Sant'Anna

Classificação Automática de Gêneros Musicais baseada em Aproximações Polinomiais no Domínio do Tempo/Marcos José Sant'Anna Magalhães. – Rio de Janeiro: UFRJ/COPPE, 2013.

XI, 47 p.: il.; 29, 7cm.

Orientador: Luiz Wagner Pereira Biscainho

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2013.

Referências Bibliográficas: p. 46 – 47.

1. Classificação de Gêneros Musicais. 2. Polinômios de Legendre. 3. Matrizes de Similaridade. I. Biscainho, Luiz Wagner Pereira. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

And now, the end is near
And so I face the final curtain
My friend, I'll say it clear
I'll state my case, of which I'm certain
I've lived a life that's full
I traveled each and ev'ry highway
And more, much more than this,
I did it my way

Paul Anka, My Way.

A Deus, por ter me dado forças e determinação nos momentos mais difíceis.

Aos que vieram antes de mim, por terem me dado o conhecimento necessário para ir além.

À minha família, sem a qual nada disso faria sentido. Em especial, aos meus pais, que me ofereceram as condições para seguir trabalhando.

Ao meu padrinho, Luiz Antonio, pelo apoio incondicional.

Ao meu orientador, por ser o melhor referencial que poderia ter na minha vida profissional.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

CLASSIFICAÇÃO AUTOMÁTICA DE GÊNEROS MUSICAIS BASEADA EM APROXIMAÇÕES POLINOMIAIS NO DOMÍNIO DO TEMPO

Marcos José Sant'Anna Magalhães

Setembro/2013

Orientador: Luiz Wagner Pereira Biscainho

Programa: Engenharia Elétrica

A *Music Information Retrieval* (MIR) é um campo de Processamento de Sinais que estuda formas de categorizar automaticamente músicas, usando apenas a forma de onda digitalizada, quanto a diversos aspectos: autor, intérprete, gênero, emoção passada pela música, entre outros.

Este tipo de classificação encontra aplicação em diversas atividades; alguns exemplos são: organização de serviços de comercialização de áudio digital (lojas virtuais ou *streaming*), equalizadores automáticos, criação automática de *playlists*.

Este projeto propõe uma família de descritores baseada nas Matrizes de Similaridade e duas técnicas para classificação de gêneros musicais usando esses descritores. O primeiro método, Matricial, agiliza o processo de treinamento e classificação, além de permitir fácil inserção e retirada de músicas do modelo. Já o segundo, Sequencial, tenta explorar a natureza da música para superar a interferência entre gêneros.

Usando o descritor proposto associado aos Polinômios de Legendre, o método Matricial obtém taxas de acerto de 71% em base autoral e 45% na GTZAN. Já o Sequencial obtém taxas de acerto de 63% e 40% nas mesmas bases, respectivamente.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

AUTOMATIC MUSICAL GENRE CLASSIFICATION BASED ON POLINOMIAL APPROXIMATION IN TIME DOMAIN

Marcos José Sant'Anna Magalhães

September/2013

Advisor: Luiz Wagner Pereira Biscainho

Department: Electrical Engineering

Music Information Retrieval (MIR) is a field of Signal Processing which studies how to automatically tag music (using only the digitized waveform) in several categories: composer, singer, genre, mood and others.

This kind of classification can be useful to many applications including online stores, music streaming services, automatic equalizers and playlist generation.

This work proposes a new kind of fingerprints based on Similarity Matrices and two techniques to recognize musical genres using the proposed fingerprint. The first method, matricially organized, exhibits a fast training and classification and allow for quick addition and exclusion of musics from the model. The other method, sequentially organized, tries to explore the nature of music to overcome intra-genre interference.

Using the proposed fingerprint associated with Legendre polynomials, the matricially organized technique achieves correct rates of 71% in an authoral database and 45% in GTZAN. The sequentially organized technique achieves correct rates of 63% and 40% in the same databases, respectively.

Sumário

Lista de Figuras **x**

Lista de Tabelas **xi**

1 Introdução **1**

1.1 Motivação 1

1.2 Visão Geral 2

1.3 Revisão Bibliográfica 3

2 Método Clássico **7**

2.1 Definindo o Método 7

2.1.1 Extração 9

2.1.2 Treinamento e Classificação 10

2.1.3 Resultado 12

2.2 Críticas 12

2.2.1 *Short-Time Fourier Transform* 12

2.2.2 Natureza Sequencial da Música 13

2.2.3 *Crosstalking* 13

2.2.4 Exclusividade de Pertinência 13

3 Método Matricial **15**

3.1 Extração 15

3.1.1 Matrizes de Similaridade 16

3.1.2 Descritores 19

3.1.3 A Representação Única 22

3.2 Treinamento e Classificação 26

3.2.1 Classificação por Linha 26

3.3 Considerações Finais 27

4 Método Sequencial **28**

4.1 Treinamento e Classificação 28

4.2 Extração 30

4.3 Considerações Finais 32

5 Avaliação de Desempenho **33**

5.1 Base de Dados 33

5.1.1 Base Autoral: PF 33

5.1.2 Base *Benchmark*: GTZAN 34

5.2 Metodologia 35

5.3 Resultados 37

5.3.1 Bases Completas 37

5.3.2 Variações 40

5.3.3 Outros 41

6 Conclusão **44**

6.0.4 Trabalhos Futuros 45

Referências Bibliográficas **46**

Lista de Figuras

1.1 A taxonomia da base de dados usada por Tzanetakis Essl e Cook. Extraída de [1] 4
1.2 Trabalhos (experimentais) publicados em reconhecimento de gêneros musicais por ano divididos pelo uso (ou não) da GTZAN. Adaptada de Sturm [2] 6

2.1 Diagrama de blocos de um sistema de classificação de gêneros. Adaptada de Seo e Lee [3] 8
2.2 Ilustração mostrando possível organização do vetor de características de bloco. 10
2.3 Ilustração com blocos sobrepostos extraídos depois da tranformada. Adaptada de Seyerlehner, Widmer e Pohle [4]. 11

3.1 Matriz de Similaridade calculada para o arquivo "hiphop.00001.au" extraído do gênero Hiphop da GTZAN. 16
3.2 Exemplo do esquema de sobreposições proposto por Foote [5]. 17
3.3 Matriz de Similaridade calculada para o arquivo "classical.00001.au" extraído do gênero clássico da GTZAN. 18
3.4 Matriz de Similaridade calculada para o arquivo "Aquarela_Toquinho.wav" extraído do gênero MPB da PF. 19
3.5 Aproximação de trecho de Brasileirinho pelos Polinômios de Legendre. Comportamento no tempo e na frequência. 21
3.6 Polinomios de Legendre e seus espectros. 23
3.7 Exemplo de descritores e ativações calculados para a faixa Brasileirinho. 25

4.1 Diagrama mostrando a nomenclatura e as matrizes envolvidas na descrição de sequências markovianas 29

Lista de Tabelas

5.1 Teste tamanho da janela: número de polinômios igual a 50% do tamanho da janela e tamanho do bloco aproximadamente 50ms. 35
5.2 Teste número de polinômios: tamanho do bloco aproximadamente 50ms. 36
5.3 Teste tamanho do bloco: número de polinômios igual a 50% do tamanho da janela. 37
5.4 Taxa de acerto: GTZAN e PF sem alterações. 38
5.5 Matriz de confusão: PF. 38
5.6 Matriz de confusão: GTZAN classificada pelo algoritmo matricial por blocos (M1). 38
5.7 Matriz de confusão: GTZAN classificada pelo algoritmo matricial com classificação por linhas (M2). 39
5.8 Matriz de confusão: GTZAN classificada pelo algoritmo Sequencial. . 39
5.9 Taxas de acerto: combinações de 4 gêneros da GTZAN. 40
5.10 Taxas de acerto: Variações da PF. 40
5.11 Matriz de confusão: versão de 30s da base PF. 41
5.12 Matriz de confusão: versão codificada da base PF. 41
5.13 Taxas de acerto: Classico x Popular e Cantado x Instrumental. 41
5.14 Matriz de confusão: Clássico x Popular. 42
5.15 Matriz de confusão: Cantada x Instrumental. 42

Capítulo 1

Introdução

1.1 Motivação

Este trabalho trata de classificação automática de gêneros musicais. Muito antes da era dos *gadgets*, se retornamos aos primórdios da humanidade, já existia uma relação muito próxima entre música e tecnologia: desde os artesãos primitivos[1] até a invenção dos instrumentos elétricos, o uso de DSPs (*Digital Signal Processors*) e a criação do mundo digital.

No decorrer da era digital, a forma básica de comercialização e distribuição das músicas tem sofrido transformações. O CD rapidamente se tornou anacrônico, com o nascimento do comércio eletrônico e do canal de distribuição inteiramente digital. Entre idas e vindas de um modelo que tentava resistir ao tempo, usuários criavam coleções (individuais) gigantescas de arquivos de áudio, possibilitadas pelas então novas tecnologias de compressão, mas ainda reminiscentes das antigas coleções de vinil e CDs. Mas em algum momento surgiu um novo modelo, e é nessa fase que hoje nos encontramos.

Se alguns anos atrás era comum encontrar usuários com coleções gigantescas de arquivos MP3, nos quase ultrapassados *MP3 Players*, agora essa coleção fica armazenada em serviços *online* que oferecem aos usuários o direito de acesso às faixas. Os CDs (de música), os DVDs (de vídeo) e os aplicativos (especialmente jogos e suas ferramentas de desenvolvimento) deixam de ser de prateleira e passam a ser comercializados, distribuídos e armazenados digitalmente, declarando o fim de muitas coleções de usuários comuns[2].

Essa nova forma de ter acesso às músicas traz mudanças no comportamento dos usuários. Dificilmente alguém entraria em uma loja de CDs e compraria aleatoriamente com o intuito de descobrir novos intérpretes. Tampouco era possível recortar as faixas interessantes de um CD, armazenando somente o que fosse de interesse.

[1] Flautas com mais de 35000 anos já foram encontradas [6].

[2] É provável que profissionais e colecionadores continuem mantendo suas coleções privadas.

Um outro aspecto é que essas lojas cresceram e ficou complicado separar o que é de interesse no meio de uma coleção tão vasta.

Se na era dos *MP3 Players*, das grandes coleções individuais, já era complicado achar as faixas, nesta nova fase, das grandes coleções coletivas, ficou ainda mais. Assim, é esperado que os algoritmos de MIR ganhem renovado interesse, entre eles os do tema deste trabalho: classificação automática de gêneros musicais.

O papel desses algoritmos nessa nova era seria prover automaticamente o rótulo de gênero das faixas (estamos falando de milhões de faixas), do modo mais objetivo possível (ainda que a caracterização de gêneros seja muitas vezes pessoal). Uma vez conhecido o gênero das faixas, podem ser implementados filtros, equalizadores automáticos; essa informação pode ser usada ainda em geradores de *playlists* ou na organização física dos arquivos no servidor.

Na próxima seção veremos como esses algoritmos se estruturam, bem como algumas idéias básicas para entendê-los.

1.2 Visão Geral

O objetivo desta seção é dar ao leitor apenas as informações básicas sobre como os algoritmos de classificação automática de gêneros musicais se organizam, uma vez que no Capítulo 2 trataremos desse assunto com mais detalhes.

Os métodos encontrados na literatura, em grande parte, se organizam na seguinte sequência: extração, treinamento/classificação e resultado.

A **extração** é, com o perdão da alegoria, uma forma de fazer o computador ouvir a música. Nesta fase a faixa será descrita matematicamente através de características (também chamadas de descritores ou *fingerprint*). Essas características são avaliadas em períodos curtos (10-100ms; usualmente 23ms – 1024 amostras a 44100Hz) e regulares, e posteriormente agrupadas, através das estatísticas dos descritores (frequentemente média e variância) em intervalos maiores (usualmente 1s).

Nesses métodos, CADA CONJUNTO DE ESTATÍSTICAS DAS CARACTERÍSTICAS É SUPOSTO UM REPRESENTANTE EXCLUSIVO DO GÊNERO DA FAIXA À QUAL ELE PERTENCE. Esta é uma suposição que examinaremos mais a fundo na Seção 2.2.

A fase de **treinamento/classificação** é parecida com a memória. Nesta fase são feitas associações entre os sons desconhecidos, de novas músicas, com os sons que foram "ouvidos" anteriormente. Esta associação é feita usando-se classificadores de padrões para "sugerir" em qual gênero é mais provável um determinado vetor de estatísticas ser encontrado.

A última fase, do **resultado**, vai ponderar as impressões ao longo da faixa e determinar um gênero para ela. Uma vez que a fase anterior foi feita para todos os segmentos (de 1s) da faixa que se quer classificar, sabe-se em qual gênero é

mais provável que cada vetor desta faixa seja encontrado, considerando o que já foi "ouvido". Usualmente, a contagem do gênero que aparece mais vezes determina o gênero mais provável da faixa.

Em todas as fases descritas, a importância da base de dados não pode ser desconsiderada. Esta influência cria grandes dificuldades para os algoritmos de classificação automática de gêneros musicais.

Uma grande dificuldade desses métodos está no fato de precisarem ser supervisionados, portanto capazes de classificar novas faixas apenas nos gêneros previamente treinados (e eventualmente em uma categoria NÃO SEI) e necessitar de novo treinamento a cada mudança na base de dados. Outros complicadores são a codificação da faixa, principalmente em se tratando de um ambiente com múltiplas codificações; e a quantidade de dados, já que podemos estar falando de milhões de músicas.

O objetivo desta dissertação é propor técnicas que possam evitar distorções no processo de treinamento dos gêneros. Para isso, apresentamos um método que tenta utilizar uma extração mais simples, e mais rápida, além de facilitar a inclusão e exclusão de faixas na base de dados (e nos modelos de classificação) sem grande esforço.

Entretanto, antes de começar a apresentação do presente trabalho, na seção que se segue faremos uma breve revisão de trabalhos que compõem a literatura sobre este tema.

1.3 Revisão Bibliográfica

No que se refere à classificação de gêneros, Tzanetakis, Essl e Cook [1] foram os primeiros a propor métodos para resolver o problema, ainda em 2001. Os autores examinaram a diferenciação entre música e fala (86% de acerto), entre voz masculina, feminina e transmissões esportivas (74%), entre gêneros (62%) e em subgrupos de música clássica (76%), como pode ser visto na Figura 1.1. Para a classificação entre gêneros eles usaram 2 conjuntos de características: um relacionado à forma de onda e outro ao ritmo.

No conjunto de características relacionado à forma de onda, encontramos: o centroide, o *rolloff*, o fluxo espectral, a taxa de cruzamentos por zero e o percentual de *frames* de baixa energia; e todos reapareceriam em diversos trabalhos nos anos seguintes. Já o conjunto de características rítmicas, baseado na transformada *wavelet*, não seria explorado com a mesma riqueza. Para a classificação entre subgrupos de clássico foi usada representação por MFCCs (*Mel-Frequency Cepstrum Coeficients*), que recentemente tem se tornado dominante na literatura, inclusive quanto à disponibilização de bases de dados já extraídas.

No ano seguinte, Tzanetakis e Cook [7] deram continuidade ao trabalho do ano

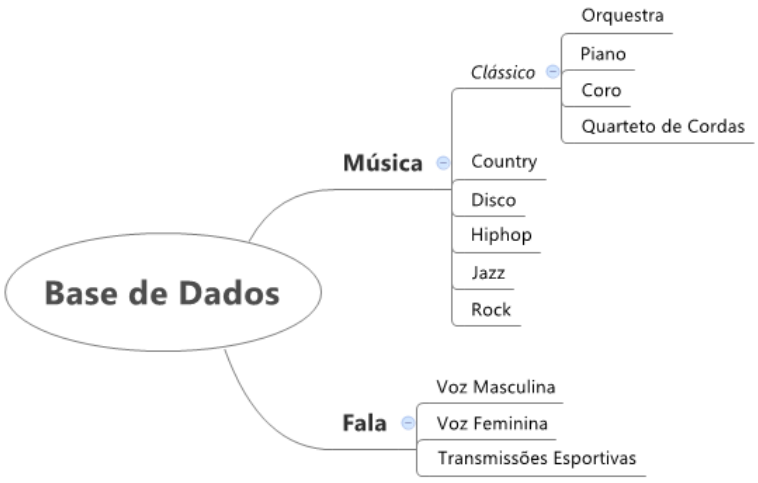

Figura 1.1: A taxonomia da base de dados usada por Tzanetakis Essl e Cook. Extraída de [1]

anterior, agora usando os MFCCs como características para a classificação de gêneros e enriquecendo a base de dados. Esta nova base de dados (GTZAN) se tornaria uma das bases de *benchmark* para a área até hoje. Na Seção 5.1.2 discutiremos as características dessa base de dados durante a avaliação de desempenho do algoritmo proposto neste trabalho.

Em 2004, West e Cox [8] fizeram amplo estudo sobre a influência de características e dos tipos de classificadores sobre a taxa de classificação, classificando 6 gêneros (*Rock*, *Classical*, *Heavy Metal*, *Drum and Bass*, *Reggae* e *Jungle*) com 150 faixas cada gênero. Cada faixa, amostrada em 22050Hz, foi transformada em mono através da média dos canais e apenas 30s escolhidos aleatoriamente (os 10s iniciais não podem ser incluídos nesses 30s) foram segmentados em intervalos de 23ms e agrupados em segmentos de análise contendo 512 desses *frames*. Eles concluíram que o uso das estatísticas (média e variância) das características produz um ganho no desempenho dos classificadores.

Shao, Xu e Kankanhalli [9] trabalharam com HMMs (*Hidden Markov Models*) para classificação "não"[3] supervisionada de gêneros musicais. Para isso eles treinaram um HMM para cada faixa e criaram uma matriz de distância entre todos os modelos das faixas. Com essa matriz, foi executado um algoritmo de clusterização até que se tivesse o número de gêneros da base (ou que se atingisse uma condição de parada no caso não supervisionado). Considerando a segmentação em intervalos regulares de 20ms, os autores obtiveram uma taxa de 75% para uma base de 4 gêneros (*Pop*, Country, Jazz e Classic), inferior ao desempenho quando usando um classificador SVM (Support Vector Machine): 93%. Os autores argumentam

[3]Embora os autores proponham uma algoritmo capaz de ser não supervisionado, os resultados apresentados são supervisionados.

que o uso de segmentos de tamanhos variáveis para cada gênero (determinado pelo intervalo entre *beats*[4] e exposto em trabalho publicado anteriormente pelos autores) melhoraria o desempenho para 89%.

E, ainda em 2004, a ISMIR (*International Society for Music Information Retrieval Conference*) hospedou uma competição de descrição de áudio que tinha entre diversas tarefas de MIR a classificação de gêneros musicais. Juntamente com a base criada por Tzanetakis, a base de dados usada no ISMIR 2004 (ISMIR2004 ou MAGNATUNE) se tornaria também um *benchmark* para a área.

Em 2006, Bergstra, Casagrande, Erhan, Eck e Kégl [10] descreveram o algoritmo que obteve os melhores resultados do Mirex 2005 (*Music Information Retrieval Evaluation eXchange*). Em vez de usar simplesmente média e variância para descrever um conjunto de trechos consecutivos, os autores propuseram usar as médias e variâncias de gaussianas ajustadas a cada uma das características. Como características os autores usaram: Coeficientes da Transformada de Fourier, Coeficientes Cepstrais Reais, Coeficientes Cepstrais na Escala Mel, Taxa de Cruzamento por Zero, medidas espectrais (espalhamento, *rolloff* e centroide) e o LPC (*Linear Prediction Coefficients*). Os autores reportam taxas de 75% na GTZAN. Assim como West e Cox, os autores concluíram que o uso das estatísticas de características produz uma melhora no desempenho dos algoritmos.

Em 2009, a tese de doutorado de Guaus [11] alcança o que é até a presente data a maior classificação na GTZAN, 99,4%. O autor relata o uso de diversos descritores e propõe uma "Transformada Rítmica" que, embora isoladamente não apresente um bom resultado, em conjunto com características espectrais alcança taxas da ordem da citada anteriormente. Como classificador ele sugeriu o uso do Simca (*Soft Independent Modelling by Class Analogy*). Este classificador inclui a PCA (*Principal Component Analysis*) mais um especialista[5] para cada classe, permitindo que uma amostra seja classificada em mais de uma classe (o que parece propenso a *overfitting* e taxas de acerto generosas). Entretanto, não ficou claro se no trabalho em questão foi feito o uso dessa possibilidade.

Em 2010, Seyerlehner, Schedl, Pohle e Knees [4] propuseram o uso de caraterísticas para a descrição de blocos maiores de sinal (de poucos segundos) em vez da descrição estatística tradicional que relatamos até aqui. Essas características de blocos seriam sumarizadas em um único vetor, através da mediana de cada característica, criando uma única representação para a música como um todo. Os autores reportam taxas próximas a 80% na GTZAN.

Em 2011, Seo e Lee [3], examinaram a importância do uso das estatísticas de ordem superior (*skewness*, curtose) no processo de descrição dos blocos de sinal. Os

[4]É a unidade básica de tempo. Frequentemente, é o ritmo que se acompanha com as mãos.
[5]Um *software* capaz de emular o comportamento de especialista humano.

autores dividem a STFT por sub-bandas geométricas e usam os momentos (média, variância, *skewness* e curtose) "intrabanda", o mínimo e a diferença máximo-minímo como caraterísticas. Sobre esses descritores os autores aplicam os mesmos momentos usados anteriormente como métodos de sumarização (ou integração temporal). Os autores concluem que a variância produz ganho superior ao da curtose, mas que o melhor resultado se obtém com o uso combinado dos dois momentos.

Em 2013, Sturm [2] faz minucioso estudo sobre a GTZAN, compilando desde informações como a frequência de uso da base de dados em trabalhos científicos (ver Figura 1.2) até estudos sobre distorções do áudio e possíveis erros de classificação. Abordaremos as considerações feitas por Sturm na Seção 5.1.2, quando discutirmos as características das base de dados usadas neste trabalho.

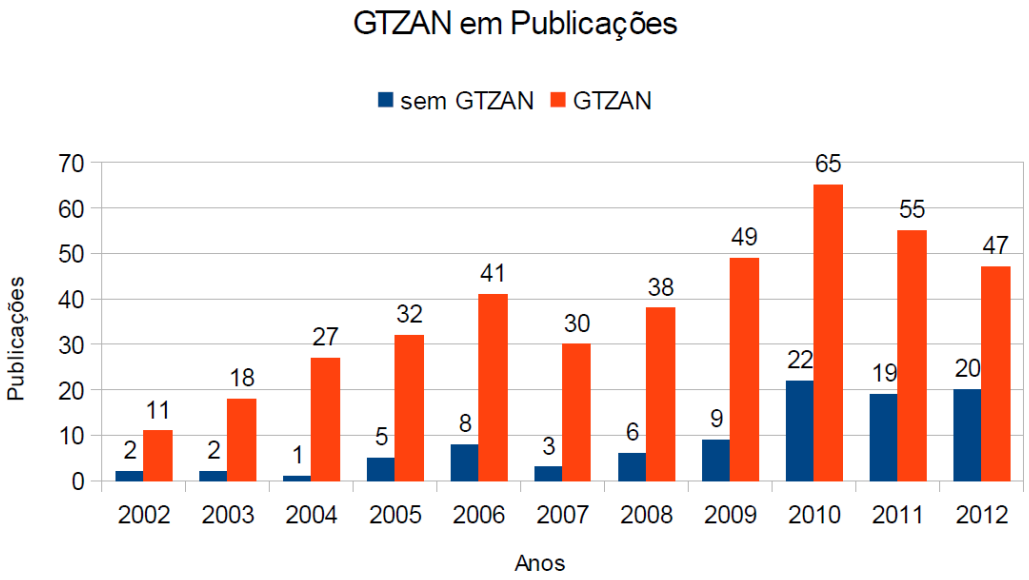

Figura 1.2: Trabalhos (experimentais) publicados em reconhecimento de gêneros musicais por ano divididos pelo uso (ou não) da GTZAN. Adaptada de Sturm [2]

No próximo capítulo analisaremos a estrutura básica mais recorrente na literatura sobre classificação de gêneros musicais.

Capítulo 2

Método Clássico

Este capítulo se inicia mostrando como grande parte dos algoritmos para classificação de gêneros musicais encontrados na literatura são estruturados, mas evita abordar especificidades de cada algoritmo. Já na Seção 2.2, apresentaremos algumas considerações sobre esta família de algoritmos que servirão como motivação para os demais capítulos.

2.1 Definindo o Método

Como vimos na Seção 1.2, de forma geral, os métodos de clássificação automática de gêneros musicais são formados por três fases: extração, treinamento/classificação e resultado. A Figura 2.1 mostra o diagrama de blocos de um sistema de classificação de gêneros musicais.

Esta figura representa dois momentos do classificador. Abaixo da linha tracejada se vê o processo de seu treinamento. Como podemos ver na imagem, nessa fase o classificador recebe como entrada um conjunto de músicas com os metadados (gênero) e cria um modelo matemático/estatístico para diferenciação entre os gêneros.

Já acima da linha tracejada está ilustrado o processo de classificação, onde podemos ver as três fases: Extração, Classificação (Classificador) e Resultado (Votação). Neste momento, já existe um modelo de gênero previamente treinado e o classificador tem como entrada um arquivo de áudio cujo gênero é a saída esperada.

Antes de aprofundar o estudo dessas fases vamos definir uma nomenclatura para ser usada no decorrer deste trabalho:

- Por **Base de Dados** (ou simplesmente **base**) entendemos uma coleção de faixas que serão usadas para treinamento e avaliação de desempenho.
- Por **faixa** (*track*) entendemos um arquivo de áudio, contendo uma única música, mas não necessariamente inteira.

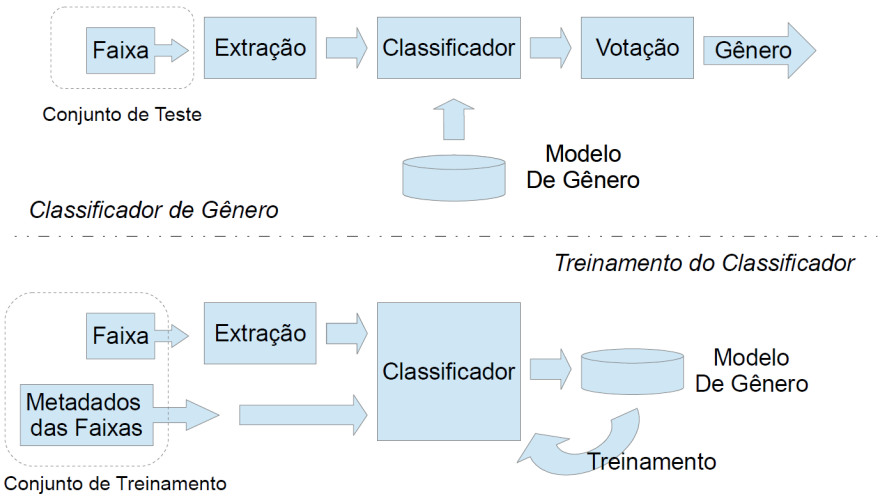

Figura 2.1: Diagrama de blocos de um sistema de classificação de gêneros. Adaptada de Seo e Lee [3]

- Uma faixa será divida em **trechos** com um número fixo de amostras podendo ser superpostos ou não.
- Um **bloco** é formado por um número n de trechos consecutivos. Em geral, todos os blocos têm o mesmo número de trechos[1] .
- Estamos tratando a classificação de gêneros, exceto quando explicitado em contrário, num contexto supervisionado[2], portanto existe um (e apenas um) gênero associado a cada faixa.

Uma vez definida a nomenclatura que será usada durante este trabalho, podemos então nos debruçar sobre a classificação de gêneros musicais.

Um bom ponto para começar esta discussão é perceber o que está acontecendo para além das definições matemáticas. A tarefa de classificar gêneros musicais se propõe um trabalho bastante desafiador, que é definir matematicamente um gênero musical.

Se colocar em palavras o que é um gênero musical já é bastante complicado, definir matematicamente algo que é, na opinião do autor, um sentimento, beira o impossível.

Mas também não é necessário. Mais importante que saber, mesmo que hipoteticamente, o que nos torna capazes de diferenciar ou assemelhar os gêneros de duas faixas, é saber qual a hipótese que assegura que depois de todos os modelos e matemática utilizados, o gênero encontrado pode estar certo.

[1]Salvo raríssimas exceções, as faixas não são compostas por um número inteiro de trechos ou blocos. A maioria dos trabalhos apenas descarta os trechos e blocos incompletos.

[2]Entende-se como contexto supervisionado quando existe a informação do rótulo das amostras de teste. Nesta aplicação, este rótulo existe e é único.

No modelo clássico, a hipótese que sustenta a matemática é:

> Se uma faixa f é dominada por blocos semelhantes aos encontrados em faixas do gênero A, então esta faixa f também é do gênero A.

Imagino que o leitor tenha em mente duas perguntas: Como saber se um bloco é semelhante aos encontrados em um determinado gênero? Qual o critério para dizer se uma faixa é dominada por um gênero?

As respostas a estas perguntas são as três etapas do algoritmo de classificação.

- Encontrar um padrão de descrição dos blocos.
- Criar um modelo de ocorrência dos blocos nos gêneros.
- Sabendo as ocorrências dos gêneros na faixa, determinar o gênero "final".

As seções que se seguem abordam cada uma dessas fases.

2.1.1 Extração

A extração é a porta de entrada dos algoritmos de classificação. Os objetivos principais aqui são:

- **Preparação** do arquivo para o processamento. Os algoritmos de classificação, salvo em casos especiais, são preparados para trabalhar com áudio que não sofreu qualquer compactação. Algumas bases de dados podem incluir arquivos que não são mono. Então, aqui também é feita a média dos canais.
- **Segmentação** da faixa em diversos trechos. Muitos trabalhos usam trechos com sobreposição, em geral, de 50%[3] e janelamento para tornar mais lenta a variação espectral, já que metade da janela também é analisada na janela anterior, bem como atenuar a distorção espectral produzida pelo retirada do trecho. Se usado o padrão de CD, isto significa uma taxa de 1,4Mb/s (2*16bits*44100Hz)[4] de dados para processamento, sem considerar que ao se usar *float* (32 bits) para representar o valor das amostras, a quantidade de dados dobra de novo.

[3] Como não há necessidade de reconstrução do sinal as taxas não precisam estar relacionadas à reconstrução perfeita.

[4] Esse 2 vem da sobreposição de 50%, que dobra a quantidade de trechos, e não do número de canais. Normalmente os arquivos são mono, ou é usada a média dos canais.

- **Extração** das características de trechos e blocos. Boa parte das características de trechos são calculadas sobre uma transformada; então, também é necessário fazer a conversão do domínio do tempo para o domínio da transformada (frequentemente pela Transformada de Fourier).

Uma vez que as características dos trechos foram extraídas, é hora de buscar a descrição dos blocos. A abordagem encontrada mais frequentemente na literatura busca descrever o bloco pelo comportamento estatístico das características dos trechos [3, 10]. Este tipo de processo aparece na literatura como *Summarizing* ou *Time Integration*. O resultado dele é um vetor organizado como na Figura 2.2.

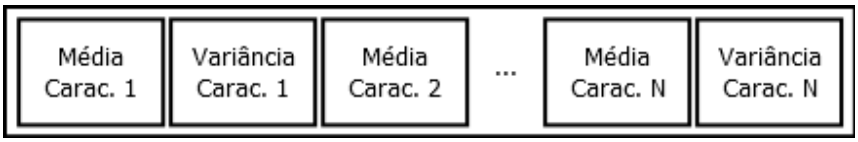

Figura 2.2: Ilustração mostrando possível organização do vetor de características de bloco.

A Figura 2.3 mostra como Seyerlehner, Widmer e Pohle [4] escolheram os blocos mantendo 50% dos trechos em comum. Esse tipo de escolha dos blocos não é frequente, mas faz sentido no contexto proposto pelos autores. A idéia do uso da sobreposição é tornar mais lenta a variação estatística de um bloco para outro.

Feita a extração completa, os dados serão enviados ao classificador para treinamento ou classificação.

2.1.2 Treinamento e Classificação

O treinamento recebe sequências de blocos (vetores), cada uma originada em uma faixa, agrupadas por gênero. Cada um desses vetores vai ajudar a compor um modelo para os blocos dos gêneros em questão.

Cada bloco da base de dados será apresentado ao classificador como um membro característico do gênero da sequência de onde foi retirado. Isto significa, para a maioria dos classificadores, que cada bloco é suposto uma realização independente de um processo estatístico. Assim, do ponto de vista teórico, estamos tentando definir a densidade de probabilidade, no espaço dos blocos, que define um gênero.

Entretanto, uma outra peculiaridade do método clássico é que, salvo duplicidade de blocos (o que virtualmente só ocorre com blocos de silêncio e causaria erros no classificador), os blocos só devem pertencer a um único gênero. Isso implica, como dissemos na Seção 1.2, que:

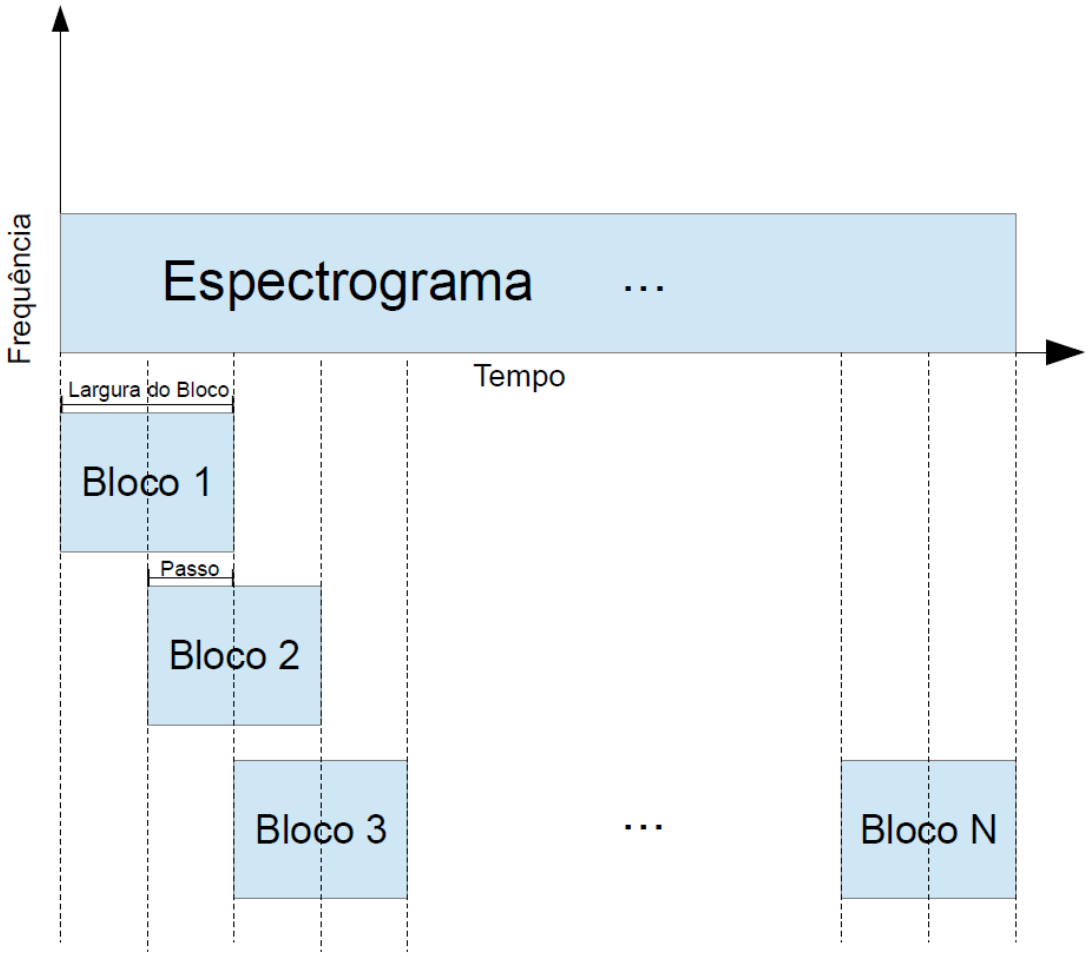

Figura 2.3: Ilustração com blocos sobrepostos extraídos depois da tranformada. Adaptada de Seyerlehner, Widmer e Pohle [4].

Cada conjunto de estatísticas das características (bloco) é suposto um representante **exclusivo** do gênero da faixa à qual ele pertence.

Isto significa que blocos próximos podem estar em lados diferentes da fronteira de classificação, tendendo a sobrepor as densidades de probabilidade de cada gênero e dificultando a classificação dos blocos das faixas que serão classificadas.

Já na fase de classificação, o algoritmo recebe um bloco como entrada e, baseado no modelo criado, tenta inferir em qual gênero teria sido mais provável que o bloco fosse encontrado durante o treinamento. Assim, ao final da classificação, para uma sequência de blocos dada como entrada existe uma sequência de gêneros associada.

Uma vez que temos uma sequência de gêneros formada, a última etapa precisa transformar esta sequência de gêneros em um único resultado.

2.1.3 Resultado

A última etapa do processo de classificação é, de certa forma, uma compilação do resultado. Esta etapa recebe como entrada uma sequência de gêneros mais prováveis, como vimos na seção anterior.

Também na seção anterior, vimos que cada bloco é um representante do gênero da faixa. Então, se por acaso, um arquivo de treinamento fosse reclassificado, a sequência de gêneros esperada seria algo do tipo: A A A A A (...) A A A A, sendo A o gênero atribuído à faixa.

Assim, quanto mais blocos de um mesmo gênero houver na faixa classificada, mais próxima das faixas de treinamento ela estará. Por isso, o critério adotado, de votação por maioria, é adequado, mas tem suas limitações.

A seção seguinte reúne comentários do autor sobre esta e outras limitações do algoritmo. E nos Capítulos 3 e 4 propomos métodos que tentam contornar estas limitações.

2.2 Críticas

Até este ponto nos esforçamos para mostrar ao leitor como os algoritmos de classificação de gêneros são, frequentemente, organizados. Nesta seção pretendemos mostrar que esta organização gera alguns efeitos indesejáveis. As seções que seguem fazem comentários sobre cada um dos pontos mais delicados.

2.2.1 *Short-Time Fourier Transform*

Não há dúvidas sobre a utilidade da STFT em diversas aplicações. Entretanto, para esta aplicação esta técnica enfrenta um problema às vezes difícil de superar.

Imaginemos uma base de dados que use áudio padrão de CD, portanto frequência de amostragem 44100 amostras/s, 2 canais e 16 bits por amostra, 3 minutos de duração por música e trechos SEM sobreposição. Nessas condições, estamos falando de 180s * 44100 a/s (8×10^6) de amostras, ou cerca de 7.751 trechos por arquivo. Uma base com 1000 arquivos, como a GTZAN, totalizaria 7.751.000 trechos e, portanto, requereria o mesmo número de STFTs. Embora a FFT (*Fast Fourier Transform*) seja *fast*, ela não é tão *fast* assim...

Uma base com 1000 arquivos é infinitesimal quando estamos falando de aplicações comerciais; por exemplo, Rdio.com®, um serviço *online* de *streaming* de áudio, anuncia 20 milhões de músicas (`http://www.rdio.com/features/`). Para essas aplicações, o uso da STFT pode ficar difícil, embora não haja a necessidade de se extrair um arquivo mais de uma vez.

Neste trabalho propomos um descritor baseado na forma de onda temporal, eliminando a STFT. Como veremos adiante, esse descritor torna todo o processo mais rápido, mas ainda é preciso provar a sua eficiência.

2.2.2 Natureza Sequencial da Música

Os classificadores estatísticos consideram que cada amostra apresentada ao classificador é independente das demais. Quando os blocos, que não são independentes entre si, são apresentados ao classificador, esta hipótese desconsidera a natureza sequencial da música.

Propomos dois caminhos diferentes para tentar explorar esta propriedade da música. No Capítulo 3, pretendemos adotar uma representação única para a faixa, o que termina com o problema das sequências. Já no capítulo 4, usamos HMMs para modelar essas sequências para tentar melhorar o desempenho de classificação.

2.2.3 *Crosstalking*

O outro ponto que queremos questionar é a hipótese inicial em si. Por hipótese, o método clássico considera que se um trecho foi ouvido em uma faixa do gênero A, então ele é um representante do gênero A.

O que derruba esta hipótese é que não precisamos nos esforçar muito para achar músicas que apresentem trechos característicos de diversos gêneros; a mistura de pop/rock, blues/jazz, rock/clássico é bastante frequente. Isto significa que ao treinar um gênero, trechos que estavam em músicas rotuladas com este gênero, mas possivelmente eram mais característicos de outros gêneros, também foram treinados, "contaminando" o que o classificador entende como este gênero.

Mas por que um bloco em um gênero seria mais representativo de outro gênero? A resposta é: porque ele é um bloco. Ao se fragmentar a faixa, não há mais garantia de que o gênero válido para todo o conjunto faixa seja também válido para todos os blocos isoladamente. Um exemplo claro é *Bohemian Rhapsody* do *Queen*, onde as partes vocais da música não são características do gênero da faixa, possivelmente Rock.

2.2.4 Exclusividade de Pertinência

A forma como o treinamento do classificador foi concebido impõe, como dissemos, um relacionamento forte entre gênero e bloco, já que um bloco necessariamente pertence a um único gênero. Mas é fácil citar contraexemplos. Muitas músicas Pop começam com solos de guitarra mais característicos de Rock, mas ainda assim podem ser classificadas como Pop.

Nossa proposta é o uso de HMMs como classificador de blocos, o que permitiria uma relação de pertinência mais flexível entre um bloco e possíveis gêneros dos quais ele fosse característico.

É importante salientar a diferença para o problema citado na seção anterior. Lá o trecho era característico de Pop, mas estava em Rock. Aqui ele é característico dos dois gêneros, e portanto sugerimos a flexibilização desse relacionamento bloco-gênero para uma descrição mais próxima da realidade.

Capítulo 3

Método Matricial

Neste capítulo, apresentaremos o primeiro método proposto por este trabalho, que surgiu ao buscarmos soluções para as críticas feitas ao Método Clássico, em especial para as dificuldades no uso da STFT e para evitar o que chamamos de *Crosstalking* usando uma única representação matricial para todo o arquivo de áudio.

Baseado nas matrizes de similaridade propostas por Jonathan Foote [5], apresentadas com mais detalhes na Seção 3.1.1, este método dá origem a um novo descritor que será usado em todo o trabalho e descrito na Seção 3.1.2. Como foi feito para o Método Clássico, vamos descrever a extração já na próxima seção, e o treinamento e a classificação na seção que segue. Este método não faz uso de mais uma fase após a classificação.

3.1 Extração

Antes de iniciar a descrição do método proposto, é importante definir o que queremos que os algoritmos façam. Assim como o Método Clássico tem uma hipótese que sustenta o modelo matemático, também precisamos de uma hipótese de sustentação.

Como vimos no capítulo anterior, é mais importante saber diferenciar os gêneros do que defini-los. Por isso, a **hipótese** que usamos é de que a forma como padrões se repetem no tempo é capaz de prover a discriminação entre gêneros.

O trabalho proposto por Jonathan Foote [5] é uma forma de tentar encontrar padrões que se repetem ao longo de um arquivo. Partindo deste trabalho, mostraremos como chegamos ao descritor proposto. Na Seção 3.1.3 utilizaremos o descritor proposto para gerar uma única representação para a faixa como um todo.

3.1.1 Matrizes de Similaridade

As matrizes de similaridade foram propostas por Jonathan Foote como uma forma de visualização da música. A idéia central é comparar todos os trechos de uma faixa de áudio dois a dois, criando um "mapa de similaridade". Assim, quando trechos parecidos (por exemplo, do refrão), ainda que de momentos diferentes da faixa, forem comparados, a tendência é que a similaridade seja alta; já no restante da música, é esperado um valor de similaridade mais baixo. A Figura 3.1 mostra um exemplo de matriz de similaridade.

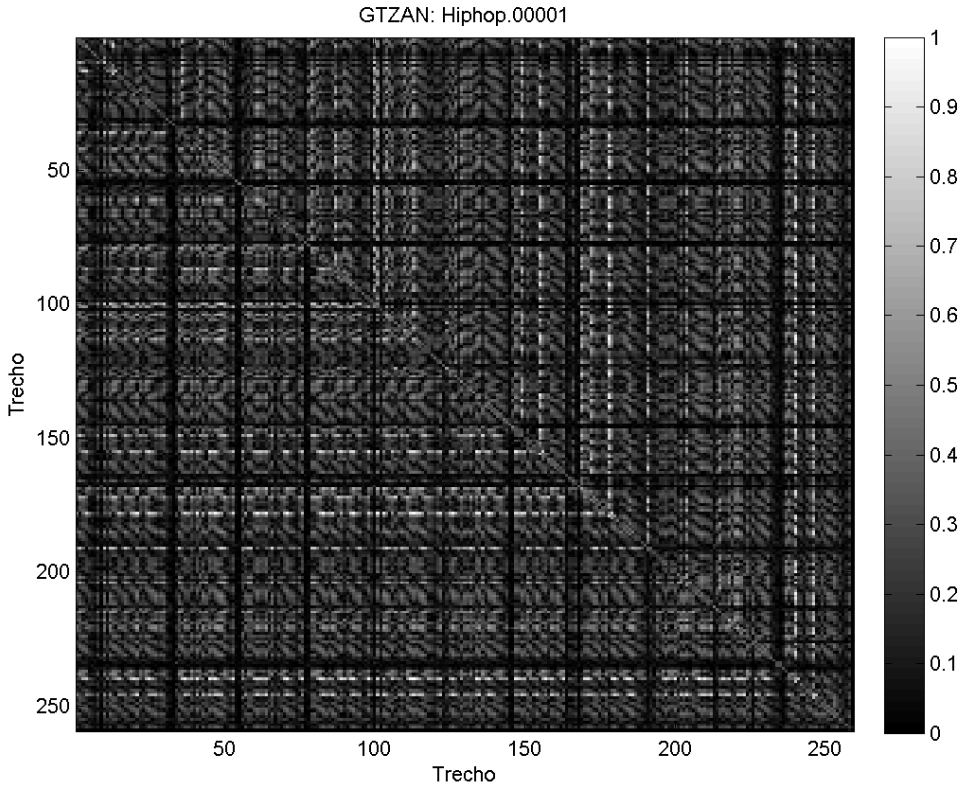

Figura 3.1: Matriz de Similaridade calculada para o arquivo "hiphop.00001.au" extraído do gênero Hiphop da GTZAN.

Foote propôs segmentar o arquivo em janelas de 25ms com passo de 1/3 de janela (o que resulta em aproximadamente 100 janelas por segundo) e usar como medida de similaridade entre dois trechos a correlação entre os 12 primeiros coeficientes cepstrais na escala Mel mais a energia. O autor sugere que seja feita a média das similaridades de janelas consecutivas para levar em conta o contexto em volta do trecho. Então, a medida de similaridade proposta por Foote assume a forma:

$$S_w(i,j) = \frac{1}{w}\sum_{k=0}^{w-1}(v_{i+k} \cdot v_{j+k}), \tag{3.1}$$

onde w é o número de janelas em cada média (sugerido entre 5 e 10), v é o vetor de característica de cada trecho e i,j são os índices dos trechos. Com $w = 10$, valor usado por Foote nos exemplos, esta fórmula faz a média no espaço de tempo equivalente de quatro vetores não sobrepostos, como mostra a Figura 3.2.

Nesta figura estão os quatro vetores (0-3, 3-6, 6-9, 9-12) com as divisões marcando 1/3 de vetor e os seis vetores criados pela sobreposição estão marcados na parte de cima da imagem. O autor argumenta que não é necessário calcular a similaridade para todos os trechos de uma faixa e que a similaridade pode ser calculada apenas para os múltiplos de w. Com efeito, no exemplo da Figura 3.2, poderíamos recomeçar o cálculo a partir do w-ésimo trecho (supondo a contagem iniciada no trecho 0).

Figura 3.2: Exemplo do esquema de sobreposições proposto por Foote [5].

As Figuras 3.1, 3.3, e 3.4 foram calculadas com $w = 5$, sem sobreposição, e a correlação foi feita diretamente entre as formas de onda no tempo. Os arquivos da GTZAN foram amostrados em 22050Hz; assim, segmentamos esses arquivos em trechos de 512 amostras. Já o arquivo "Aquarela_Toquinho.wav"[1] foi amostrado em 44100Hz, e por isso o segmentamos em 1024 amostras para manter a razão entre píxeis da matriz e tempo.

A Figura 3.1 foi calculada a partir do arquivo "hiphop.00001.au" da GTZAN. O padrão xadrez da imagem mostra que existe semelhança frequente entre os trechos próximos da faixa, característica de faixas muito marcadas como as do *Hiphop*. As linhas que se juntam na diagonal, de forma semelhante à moldura de um quadro (por exemplo, entre os trechos 150 e 200), sugerem que houve regiões similares até o trecho na diagonal, mas que os trechos futuros não serão similares.

Já na Figura 3.3, as semelhanças apresentam um padrão diferenciado. Os quadrados perto da diagonal sinalizam uma região de forte semelhança. Se essas regiões tendem a ocorrer outras vezes, o padrão vira um quadrado mais amplo, como o que pode ser visto na figura. A música clássica tende a ser mais lenta, com trechos similares mais longos, o que resulta em padrões desse tipo.

Como os arquivos da GTZAN têm duração de 30s, a matriz completa do arquivo "Aquarela_Toquinho.wav" foi recortada no mesmo tamanho das matrizes anteriores. O resultado pode ser visto na Figura 3.4, onde há uma mistura dos padrões mostrados nas figuras anteriores.

[1]Este arquivo faz parte do gênero MPB da base de dados autoral (PF) descrita na Seção 5.1.

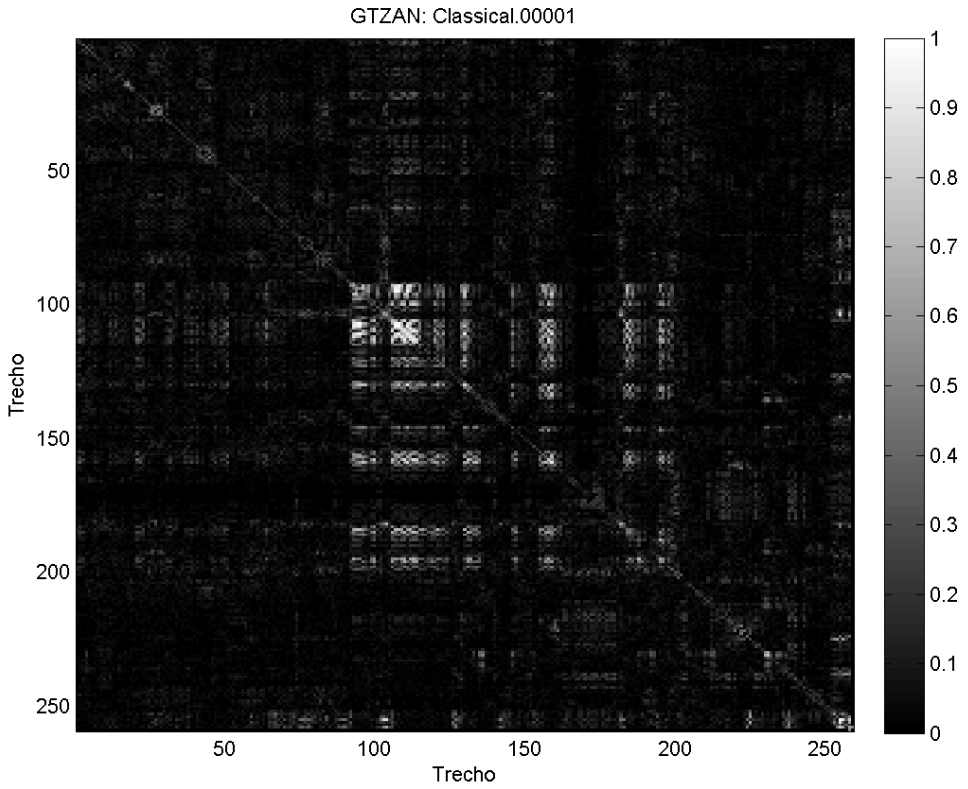

Figura 3.3: Matriz de Similaridade calculada para o arquivo "classical.00001.au" extraído do gênero clássico da GTZAN.

Essas matrizes podem ser úteis para mostrar onde há semelhança entre trechos do arquivo de áudio, mas, do ponto de vista de classificação, são difíceis de se utilizar. O principal problema é que elas têm tamanhos diferentes para cada faixa (já que a matriz tem o número de linhas e colunas relacionado ao número de trechos da faixa). Entretanto, podemos adaptar a idéia de Foote, e para isso temos duas opções: truncar o arquivo (como fizemos na imagem Aquarela) ou tentar ajustar pelo menos uma das dimensões.

Numa base como a GTZAN, em que os arquivos têm sempre **aproximadamente**[2] a mesma duração, as matrizes tendem a ser do mesmo tamanho. Entretanto, no caso prático não é essa a situação. Uma solução seria usar apenas 30s de cada faixa, mas essa solução dá início a nova discussão sobre quais os melhores 30s. Por isso, preferimos buscar uma solução que ajustasse uma das dimensões da matriz de similaridade para adequá-la mais facilmente à classificação.

Para ajustar uma das dimensões, vamos precisar mudar um pouco a ideia de Foote. Em vez de comparar trechos da faixa entre si, vamos compará-los a uma referência externa conhecida e constante. Se esta referência for a mesma para todas

[2]Como já era esperado, o número de amostras em cada faixa varia.

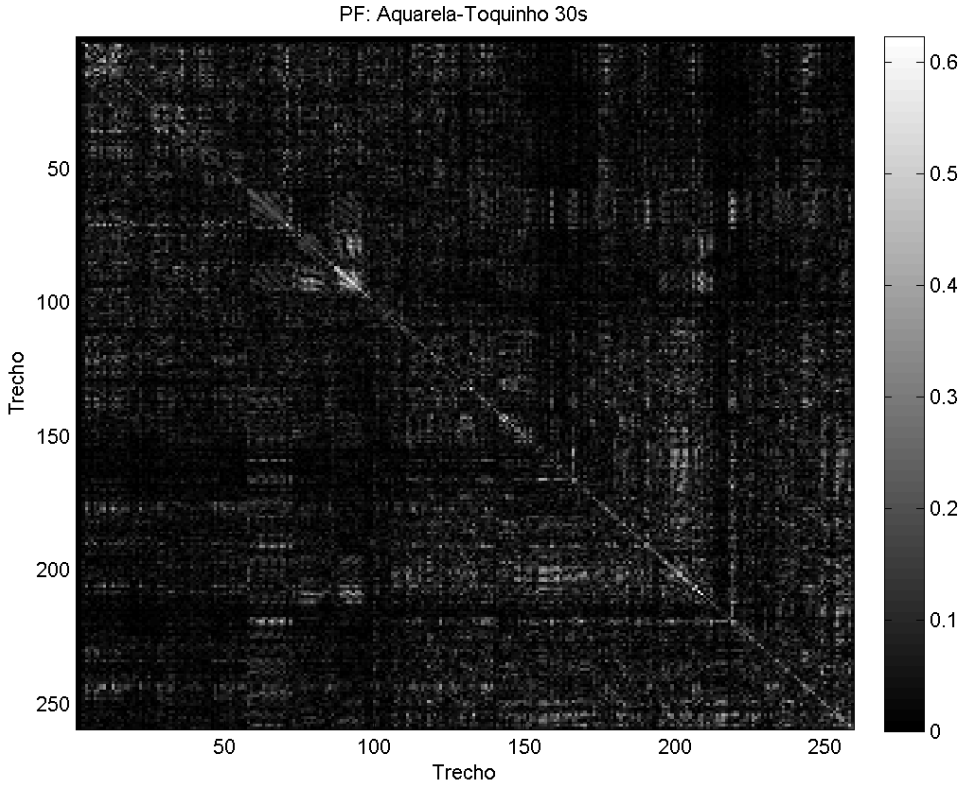

Figura 3.4: Matriz de Similaridade calculada para o arquivo "Aquarela_Toquinho.wav" extraído do gênero MPB da PF.

as faixas e for definida com um tamanho fixo, então uma das dimensões das Matrizes de Similaridade será relacionada ao tamanho da referência e a outra, ao número de trechos da faixa. O que podemos usar como referência? Qual deve ser o critério de similaridade?

A próxima seção busca responder essas perguntas e perceber que ao perseguir esta solução estamos na verdade criando uma nova família de descritores.

3.1.2 Descritores

Na seção anterior, vimos que as Matrizes de Similaridade são adequadas para revelar padrões que se repetem durante uma faixa, mas não se adaptam tão bem aos algoritmos de classificação. Dissemos que se conseguirmos fixar uma das dimensões dessas matrizes, o resultado será mais conveniente aos algoritmos de classificação. Esta seção mostra os detalhes das escolhas que fizemos para alcançar este objetivo.

Podemos comparar um trecho de áudio e uma referência (segundo um critério de similaridade) de muitas formas, mas é importante pensar no que significam essas comparações e referências. Normalmente, o objetivo da busca pela similaridade

é ressaltar **características** relevantes ou **descrever** um trecho da melhor forma possível. Isto é completamente análogo à fase de extração do método clássico, onde o par critério-referência define uma nova característica.

Neste trabalho, escolhemos o critério e a referência de modo que pudéssemos descrever o áudio da forma mais completa possível. Pensando o trecho como uma forma de onda de um espaço de formas de ondas possíveis, comum a todas as faixas (supondo o mesmo número de amostras por trecho), podemos descrever os trechos se adotarmos uma referência que seja base ortogonal desse espaço, com o bônus de minimizar a correlação entre as dimensões dos vetores características, já que a base é ortogonal.

Uma tentativa de criar uma base ortogonal nesse espaço é levar para ele uma base que já era ortogonal no espaço contínuo. A ideia é que se existe a ortogonalidade no espaço contínuo, então existe uma combinação linear capaz de, em todos os pontos de um subespaço discreto, aproximar uma função qualquer, ou a hipótese inicial de ortogonalidade seria violada; entretanto, as projeções entre vetores (após a redução de dimensionalidade) da base podem não ser nulas. Não há garantias de que ao reduzir a dimensão do espaço as direções continuem ortogonais, mas há garantia de que a aproximação é correta.

Nossa opção foi usar os Polinômios de Legendre[3] como a referência que procurávamos. Eles são ortogonais entre [-1, 1] no espaço contínuo; então, se "amostrarmos" esses polinômios e usarmos os vetores resultantes como base no espaço das formas de onda, seremos capazes de criar aproximações polinomiais das formas de onda no domínio do tempo; além disso, essa escolha leva a uma aproximação do espectro cujo conteúdo frequencial vai aumentando com a ordem da aproximação.

Na Figura 3.5, temos um trecho com 256 amostras (no tempo) da música Brasileirinho, que faz parte da base de dados PF (descrita na Seção 5.1). Em contínuo, estão os dados relativos a este trecho, do lado esquerdo a forma de onda e na direita o espectro, ambos ajustados para energia unitária; já os pontos são as aproximações feitas usando os Polinômios de Legendre. Na primeira linha, usamos o número de polinômios igual a 25% do tamanho da janela, na segunda linha 50% e na terceira linha 75% (64, 128 e 196 amostras, respectivamente). Os gráficos do espectro mostram que à medida que acrescentamos polinômios de maior grau à aproximação, o conteúdo frequencial vai progressivamente ocupando as regiões do espectro de frequência mais alta.

A Figura 3.6 explica por que isso acontece. Nela podem ser vistos a forma de onda (esquerda) e o espectro (direita) dos polinômios de graus 1, 21, 41, 61, 81 e

[3]Numericamente, o comportamento dos Polinômios de Legendre, após a redução da dimensionalidade, é manter a ortogonalidade entre pares e ímpares, o que não ocorre entre os de mesma paridade.

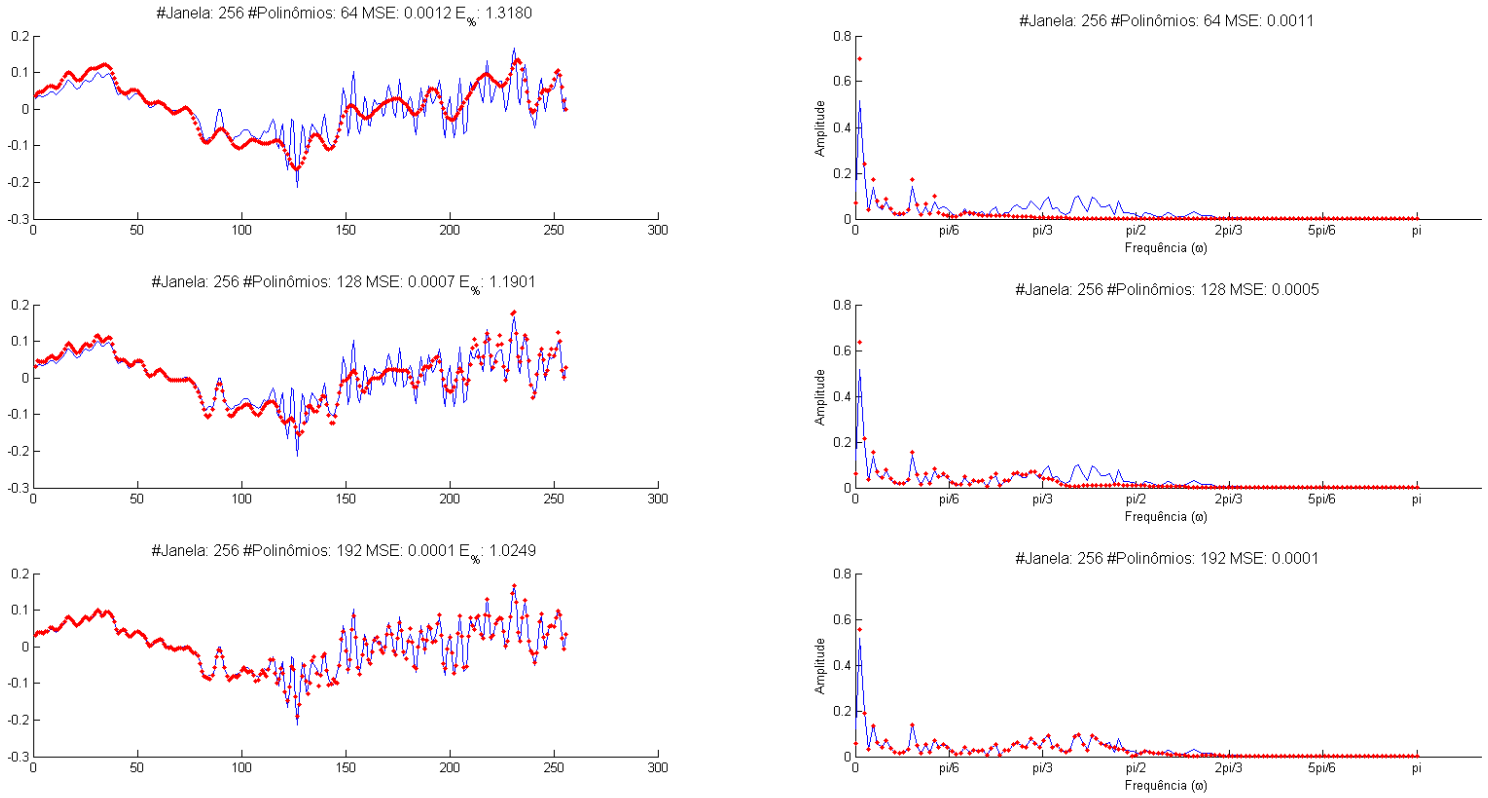

Figura 3.5: Aproximação de trecho de Brasileirinho pelos Polinômios de Legendre. Comportamento no tempo e na frequência.

101. É interessante notar que no domínio da frequência o efeito de aumentar o grau desses polinômios é criar um "passa-banda". Isso significa que o conteúdo de baixa frequências, além de concentrado apenas nos polinômios de grau baixo, também é "descrito primeiro". Com efeito, Cohen e Tan mostram em [12] que os Polinômios de Legendre convergem bem mais rapidamente que a Série de Taylor, e que para um mesmo número de componentes o erro de aproximação é uma ordem de grandeza menor.

Para completar o descritor, também precisamos de um critério de comparação da referência com os trechos de áudio. O mais simples seria usar a idéia de projeção e usar as projeções como descritor, mas, como veremos na próxima seção, este descritor não se adaptaria ao processo de classificação porque este precisa de descritores não-negativos. A solução mais óbvia é, então, usar o quadrado de cada projeção, que, como trechos e referências têm módulo (e energia) unitário, é também o percentual de energia que cabe a cada polinômio.

3.1.3 A Representação Única

Já definimos como serão descritos os trechos das faixas, mas ainda falta definir uma forma de agrupar esses trechos em blocos e depois como chegar a uma única representação para a faixa.

Para transformar os descritores de trecho em descritores de bloco, usaremos apenas a função média (sem outros momentos); entretanto, calculamos primeiro a forma de onda média (média dos trechos), e depois o descritor deste trecho médio.

Até este momento, temos um processo de extração absolutamente idêntico ao processo clássico, apenas usando um tipo de descritor diferente. Isto significa que até agora ainda sofremos dos mesmos problemas que sofríamos no método clássico, e que agora precisamos transformar este tipo de representação "por blocos" em uma representação da faixa como um todo.

Para formularmos esta transição, precisamos voltar à hipótese que adotamos para diferenciar os gêneros: a repetição de padrões espectrais ao longo do tempo é suficiente para distinguir entre gêneros musicais. Como não temos como definir que padrões seriam esses, podemos definir, por exemplo, bandas espectrais, e à medida que essas bandas espectrais forem ativadas ao longo do tempo, um índice de intensidade de ativação pode ser calculado. Como vimos, os espectros dos Polinômios de Legendre tendem a ser passa-banda; então, o descritor que temos já nos fornece a informação da ativação de cada banda ao longo da faixa.

Entretanto, existem duas questões: i) a largura da banda aumenta com o grau do polinômio; ii) será que a ativação de uma única banda revela um padrão? Os dois problemas estão relacionados, já que quanto mais larga for a banda, mais fácil

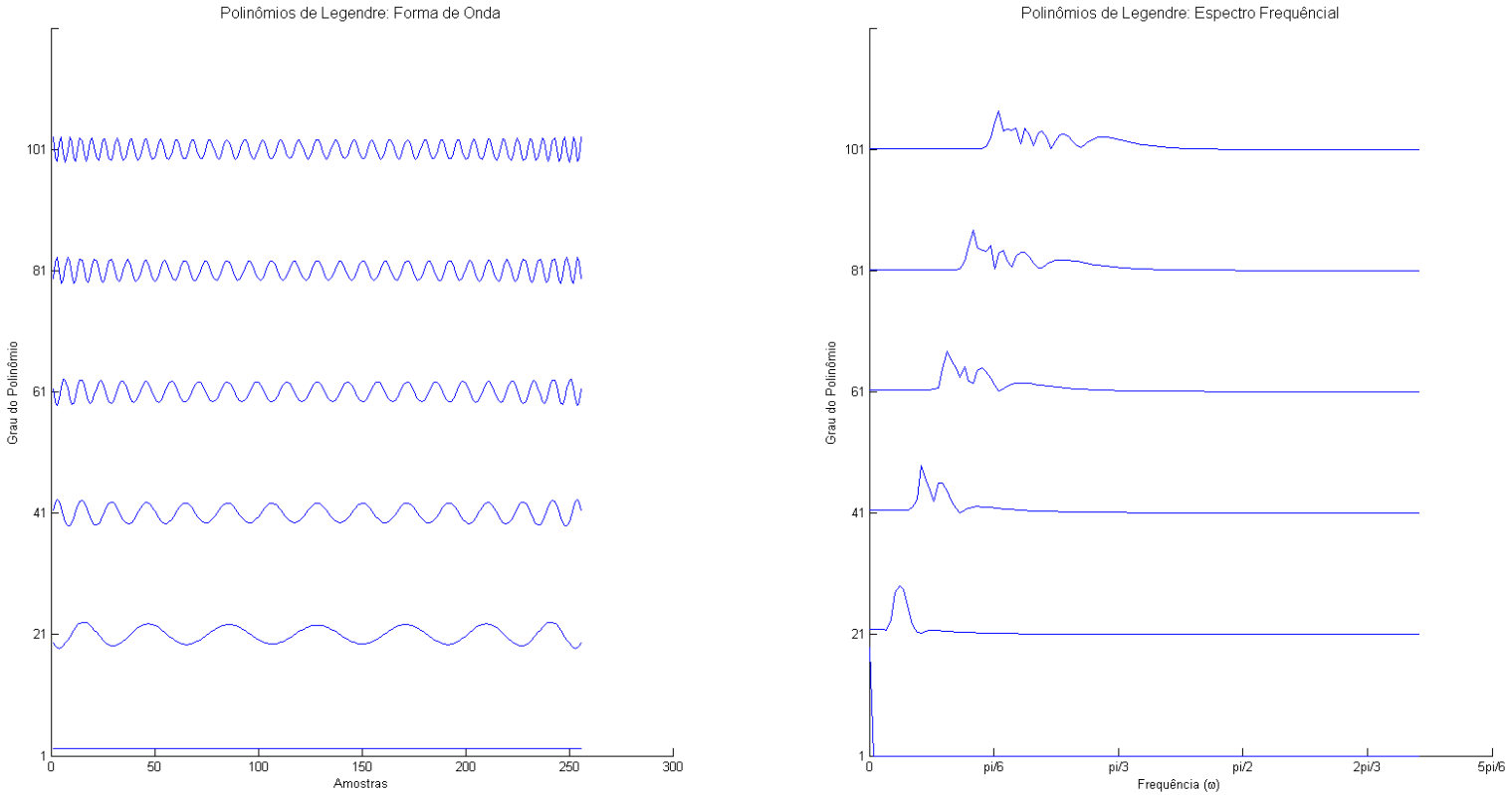

Figura 3.6: Polinomios de Legendre e seus espectros.

é responder positivamente à pergunta feita. Resolvemos adotar uma solução que avalie as ativações em uma única banda, mas que verifique se existe outra banda sendo ativada conjuntamente.

Esta verificação pode ser feita de forma computacionalmente eficiente apenas multiplicando a matriz dos descritores pela transposta. Esta operação faz com que os índices de ativação (a energia percentual) obtidos em uma banda sejam multiplicados pelos índices em outra banda (e nela mesma). Como dissemos, os índices são percentuais; portanto, variam entre 0 e 1 e são positivos.

Onde os índices são relativamente grandes nas duas bandas, o valor tende a se destacar; onde são menores, tendem a se aproximar de 0; e onde uma banda é ativada e a outra não, tende a ser consideravelmente atenuado. O somatório pode ser dividido pelo número de blocos para virar média e, então, temos uma matriz (quadrada, simétrica e cuja ordem é o número de polinômios usados) contendo números positivos entre 0 e 1, que na diagonal correspondem à ativação singular da banda correspondente e fora da diagonal correspondem à ativação conjunta (ou simultânea) das duas bandas[4].

Suponha que há um padrão se repetindo ao longo do tempo que ocupe as bandas 2, 3 e 4. O reflexo desse padrão na matriz (D) seria o aumento no valor dos elementos na diagonal ($d_{2,2}$, $d_{3,3}$ e $d_{4,4}$) e fora da diagonal ($d_{2,3}$, $d_{2,4}$ e $d_{3,4}$), bem como dos elementos simétricos ($d_{3,2}$, $d_{4,2}$ e $d_{4,3}$), por exemplo. A Figura 3.7 mostra um exemplo prático calculado a partir da faixa Brasileirinho. Nela escolhemos um trecho de 500 blocos e calculamos o descritor proposto, que pode ser visto no gráfico que ocupa a faixa à esquerda da figura (polinômios 8, 9 e 10). Os gráficos menores mostram o resultado da multiplicação dos descritores entre si, de forma que cada linha de gráficos corresponda ao descritor mais à esquerda multiplicado por si mesmo e pelos demais, e no título a média desse produto.

Ainda na mesma figura, podemos ver que os descritores obtidos pelo polinômio 10 apresentam 4 picos razoavelmente definidos, enquanto que nos demais esses picos perdem definição. Isto fica bastante claro ao observamos os gráficos das ativações singulares (diagonal) e suas médias, já que o polinômio 10 tem a maior ativação singular média. Note que a ativação conjunta dos polinômios 9 e 10 não acontece com intensidade relevante (a escala do gráfico conjunto é de apenas 0,12), levando a uma média menor que a ativação entre 8 e 10. Isto significa que o polinômio 9 parece "fora de fase" em relação aos demais, fato corroborado pelas baixas médias das ativações conjuntas com os demais polinômios cuja ativação singular é próxima à ativação do polinômio 10.

Para cada faixa na base de dados, será criada uma matriz no formato descrito.

[4]Para efeito de nomenclatura, usaremos ativação singular quando a medida se referir a um único polinômio e conjunta quando fizer referência a dois.

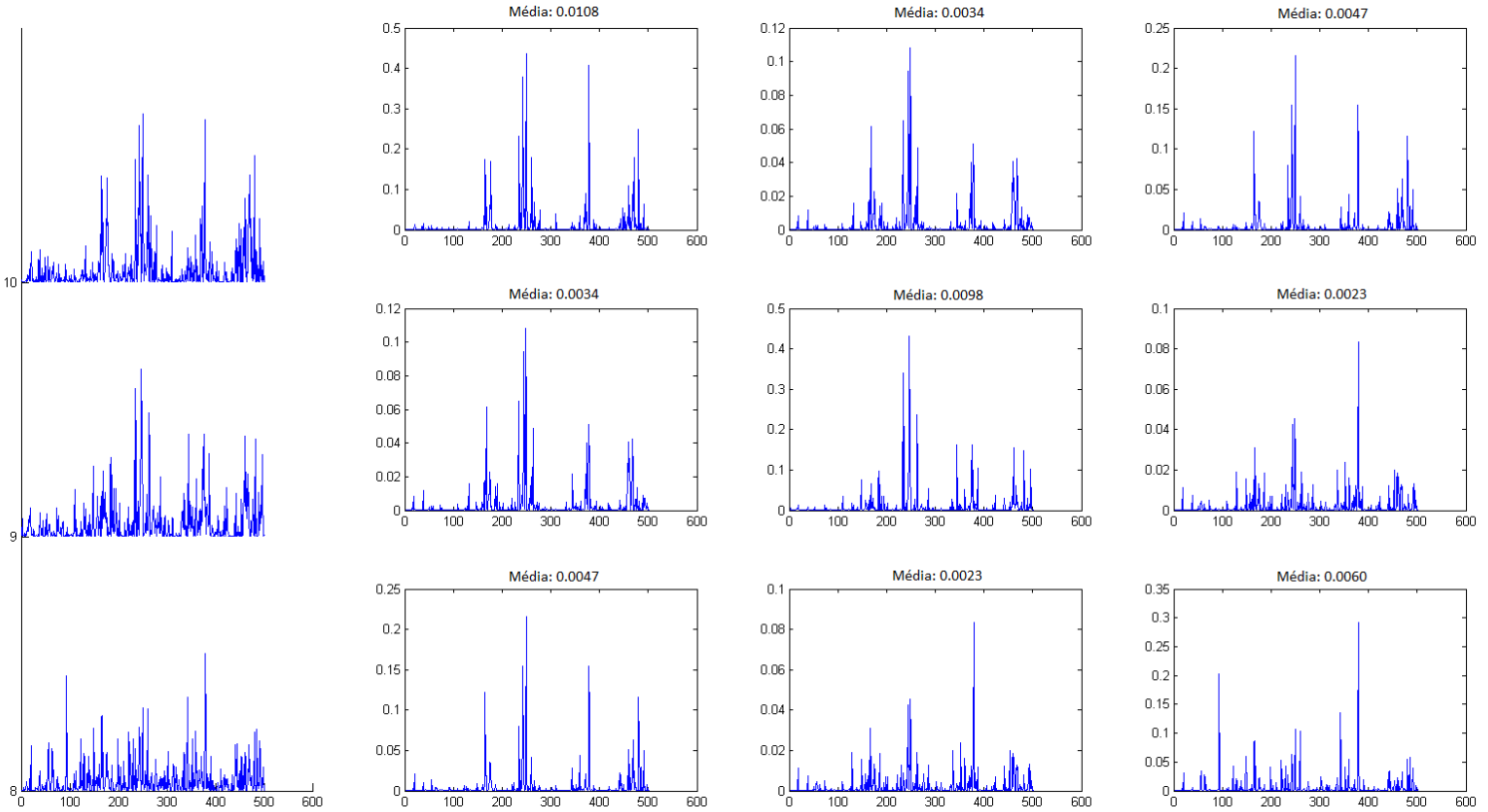

Figura 3.7: Exemplo de descritores e ativações calculados para a faixa Brasileirinho.

Essas matrizes serão enviadas ao classificador para por fim criarem um modelo de gênero ou serem classificadas.

3.2 Treinamento e Classificação

No treinamento, onde queremos criar um modelo para o gênero, optamos por apenas agrupar as matrizes nos gêneros das faixas. Este tipo de modelo indica que, no futuro, o modelo de gêneros possa ser tratado como tensor e que transformações tensoriais possam ser úteis no processo de classificação para, por exemplo, seleção de características. Esse tema foge ao escopo deste trabalho.

Uma vantagem deste modelo de gênero é ser trivial inserir e remover novas faixas no modelo com custo computacional ínfimo comparado ao da execução de todo o processo de treinamento presente no método clássico. Este tipo de operação, principalmente a inclusão, é fundamental para os serviços que disponibilizam o *streaming* de áudio pela internet, já que fazem adições diárias a suas bases de dados.

Na classificação, adotamos como classificador um KNN (*k-Nearest Neighbors*) adaptado para matrizes. Para fazer essa adaptação, calculamos uma distância entre a matriz de classificação e todas as matrizes da base, e o KNN é feito nessas distâncias buscando os valores próximos a 0. A distância calculada é simplesmente a distância *City Block* considerando a matriz como um vetor.

O uso da distância *City Block*, sem pesos, significa que para O (ordem da matriz e número de polinômios usados) termos da distância que consideram a ativação individual da banda (diagonal), $O \cdot (O - 1)$ termos fora da diagonal consideram a ativação simultânea das bandas.

3.2.1 Classificação por Linha

A seção anterior descreveu uma forma de classificarmos a matriz de ativação que aplica um único rótulo à matriz, supondo que arquivos de mesmo gênero têm comportamento similar em todas as bandas. Entretanto, isso pode não ser o mais adequado; e por isso, estamos propondo uma abordagem mais flexível, que seria classificar o comportamento de cada uma das bandas (linhas ou colunas da matriz) e usar o rótulo aplicado a cada uma delas para classificar a matriz como um todo.

O processo de classificação é simples: cada banda é classificada usando 5 técnicas diferentes – uma estatística, duas através da medida de correlação e outras duas baseadas em medidas de distância. As bandas em que um mesmo rótulo recebe 60% ou mais dos votos são consideradas "eleitores" válidos para o gênero global da matriz; as outras, não. Os votos dos eleitores válidos são contabilizados, e o rótulo mais votado é assumido como rótulo da matriz.

A técnica estatística envolve classificar a banda calculando suas distâncias às medianas dos rótulos e escolhendo o mais próximo. Para calcular as medianas são usadas apenas as bandas correspondentes; assim, se em um exemplo são usadas 100 bandas e 4 rótulos, então existem 400 medianas, uma para cada rótulo em cada banda.

As medidas de correlação e distância são calculadas de formas bastante similares entre si; elas calculam quais as bandas mais próximas (maior correlação e menor distância), mantendo a classificação somente entre as bandas originadas pelo mesmo polinômio, e ordenam essas medidas. O valor mais próximo tem direito a um voto, e o valor mais frequente entre os 10 mais próximos tem direito a outro voto.

Assim, suponha que durante a classificação de um arquivo os gêneros atribuídos aos arquivos das 10 linhas mais próximas sejam dados por BAAACAAABA. O rótulo B receberia um voto por ser o rótulo da banda mais próxima e o rótulo A receberia outro voto por ser o rótulo mais frequente dentre os mais próximos.

3.3 Considerações Finais

Para encerrar este capítulo, gostaríamos de fazer algumas observações que não foram incluídas anteriormente no texto.

O descritor que apresentamos, por ser baseado em polinômios, só precisa ter a extração feita para um determinado polinômio apenas uma vez. Suponha que os polinômios de 1 a 100 foram extraídos. Uma execução que use os primeiros 90 polinômios não precisa reextrair o arquivo, bastando carregar apenas os primeiros 90 descritores. Já uma outra execução que use 110 polinômios precisará extrair apenas os 10 polinômios (101 ao 110) que ainda não foram extraídos.

Uma característica importante do classificador baseado no KNN é que o modelo de gênero criado é o agrupamento das matrizes da representação única, o que dispensa uma fase de treinamento propriamente dita, possibilitando a inserção e remoção de arquivos do modelo. Já o classificador descrito na Seção 3.2.1, por utilizar as medianas, não tem essa mesma facilidade.

Conseguimos atingir um método que evite a STFT para ter uma fase de extração mais rápida. Também conseguimos criar uma representação compacta capaz de mostrar na média o que está acontecendo ao longo do arquivo, contornando o problema do *Cross Talking*.

Entretanto, esse método é ainda mais agressivo do que o método clássico ao desconsiderar as sequências intrínsecas à música. No capítulo seguinte, faremos uma proposta que tenta aproveitar essa característica do sinal com que estamos trabalhando.

Capítulo 4

Método Sequencial

Neste capítulo, apresentamos uma proposta de uso de HMMs para a classificação de gêneros musicais. Partindo dos descritores apresentados no capítulo anterior, tentamos contornar o *crosstalking* e fazer uso das sequências naturais da música.

Para facilitar a leitura, optamos por primeiro definir o modelo de treinamento e classificação e depois apresentar a extração, isso porque esta última foi definida para se encaixar em um modelo especifico de treinamento.

4.1 Treinamento e Classificação

Nesta seção, vamos definir como chegar a um modelo de gênero baseado em HMM partindo de uma sequência que, para cada bloco da faixa, indica de qual gênero o bloco é mais característico. A seção seguinte descreve como esta sequência foi obtida.

Os HMMs usam três matrizes para expressar o funcionamento de sequências: P_i, a matriz de probabilidade inicial; P_{se}, a matriz que mostra para cada estado s qual a probabilidade de ocorrer a emissão e; e P_{ss}, a matriz de transição de estados. Uma forma possível de interpretar o funcionamento dos HMMs é de que o estado nada mais é que um "seletor" de PDFs (*Probability Density Function*), que, por sua vez, estão descritas nas linhas da matriz P_{se}.

Pensando os HMMs dessa forma, e lembrando das sequências de gêneros do método clássico, uma possível hipótese teórica para o uso dos HMMs seria que os gêneros naturalmente compartilham trechos (aliás essa foi uma das críticas que fizemos ao método clássico); então, talvez um gênero possa ser descrito pela forma como ele compartilha seus trechos com os outros gêneros. Se dado um gênero, o compartilhamento de trechos entre ele e os demais ocorre de uma forma particular a ele, então podemos dizer que esse gênero está funcionando como um seletor de probabilidades de compartilhamento de trechos – logo, pela ótica do HMM, como

estado. Mas ele também funciona como emissão, porque também aparece para definir as probabilidades de compartilhamento.

Imaginemos que uma base seja composta por três gêneros: Rock, Pop e Jazz. Ao adotarmos os gêneros como estados e emissões, sugerimos que para cada um desses gêneros o compartilhamento de trechos entre eles ocorre com diferentes probabilidades. Assim, no estado Rock as probabilidades de se ter um trecho Rock, Pop ou Jazz podem ser, por exemplo, de 60%, 30% e 10%, respectivamente. Já no estado Pop, as probabilidades de compartilhamento podem ser totalmente diferentes.

Do ponto de vista conceitual, os gêneros como estado representam uma idealização que leva em conta o contexto; já como emissões, representam uma realização, levando em conta apenas a física das amostras e o período de duração dos blocos. No exemplo anterior, dado o contexto Rock, existe 30% de probabilidade de ser encontrado um vetor com características (físicas, espectrais, por exemplo) similares ao de Pop.

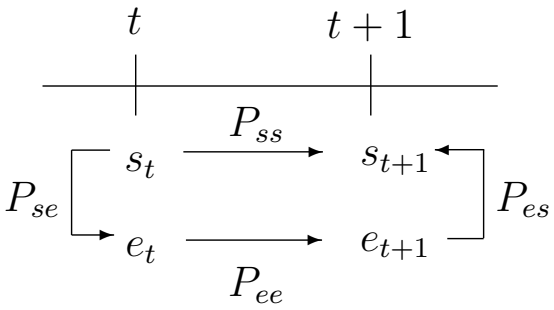

Figura 4.1: Diagrama mostrando a nomenclatura e as matrizes envolvidas na descrição de sequências markovianas

Ter os gêneros como estados desse modelo traz uma complicação considerável: como treinar os HMMs se não sabemos as sequências de estado? Sem elas não há como calcular P_{ss} ou P_{se}. Entretanto, se encontrarmos uma forma de estimar essas sequências, teremos um modelo. Na Figura 4.1, podemos ver dois instantes consecutivos t e $t+1$, além dos respectivos estados (s_t e s_{t+1}) e das emissões (e_t e e_{t+1}). A teoria de HMM diz que a probabilidade $P_{ee}(e_t, e_{t+1})$ é determinada pela probabilidade de acontecerem a emissão e_t no estado s_t, a transição para o estado s_{t+1} e a emissão de e_{t+1} dado o novo estado:

$$P_{ee}(e_t, e_{t+1}) = \sum_{s_t} P_{se}(s_t, e_t) * \sum_{s_{t+1}} P_{ss}(s_t, s_{t+1}) * P_{se}(s_{t+1}, e_{t+1}). \tag{4.1}$$

Ora, da mesma forma podemos dizer então que probabilidade de transição de estados ($P_{ss}(s_t, s_{t+1})$) pode ser dada pela probabilidade de o estado s_t emitir e_t e e_t ser sucedida por e_{t+1}, e esta última ser emitida durante um estado s_{t+1}, onde P_{es} é a matriz que diz qual a probabilidade de uma emissão ter ocorrido durante um

estado (é o dual de P_{se}). A Equação (4.2) nos dá uma forma de estimar P_{ss} ainda que não haja sequências ou que talvez tenhamos sequências pouco apropriadas:

$$P_{ss}(s_t, s_{t+1}) = \sum_{e_t} P_{se}(s_t, e_t) * \sum_{e_{t+1}} P_{ee}(e_t, e_{t+1}) * P_{es}(e_{t+1}, s_{t+1}) \tag{4.2}$$

ou, na forma matricial, simplesmente

$$P_{ss} = P_{se} * (P_{ee} * P_{es}). \tag{4.3}$$

Note que P_{ee} só precisa das emissões, e P_{es} pode ser calculada através de P_{se}, que ainda precisa ser determinada. Se arbitrarmos sequências de forma que P_{se}, e portanto as demais matrizes, possam ser calculadas e então reestimamos essas sequências, usando, por exemplo, o algoritmo de Viterbi, podemos iterativamente determinar as sequências de estado para treinamento.

Qual seria a melhor forma de iniciar as sequências? Usando o próprio gênero das faixas. Fazer a inicialização desta forma significa dizer que as faixas de um gênero devem ser dominadas pelo estado que corresponde a ele, mas também que essas faixas dão uma boa ideia de como este gênero compartilha seus trechos com os demais.

O Algoritmo 4.1 mostra a forma final do processo iterativo para estimação das sequências de estado e treinamento dos HMMs. Nele fizemos a estimação de P_{se} usando toda a base de dados (linha 6) e a estimação de P_{ee} usando a parte correspondente ao gênero em treinamento (linha 8). Com a escolha que fizemos das sequências de estado, os conjuntos de dados de cada gênero serão responsáveis por preencher as linhas correspondentes na matriz P_{se}.

Calcular esta matriz por gêneros significa viabilizar que os gêneros se comportem de maneiras diferentes dependendo do gênero da faixa, dando mais flexibilidade ao modelo, mas enfraquecendo-o, já que seriam mais situacionais. Já calcular P_{ee} por gênero significa dizer que cada gênero tem uma composição diferente e uma dinâmica diferente, como de fato é esperado.

4.2 Extração

Como dissemos anteriormente, este método faz uso do descritor apresentado no capítulo anterior. De maneira mais objetiva, são feitas médias de trechos dos sinais, e essas médias são descritas usando os Polinômios de Legendre. Entretanto, desta vez não faremos o produto pela transposta.

Na literatura, quando há a necessidade de usar um HMM com estados discretos, mas emissões contínuas, como neste caso, a saída padrão é a clusterização ou a

Algorithm 1 Treinamento dos HMMs por gênero

```
1: G ← Gêneros
2: F ← Faixas
3: S(g, f) ← SequenciaEstadoUniforme(G)
4: E(g, f) ← SequenciaEmissão(G,F)
5: while S tem mudanças do
6:    P_se, P_es ← EstimativaPse(S,E)
7:    for all g ∈ G do
8:       P_ee ← EstimativaPee(E(g, :))
9:       P_ss ← P_se * (P_ee * P_es)
10:      for all f ∈ F do
11:         S(g, f) ← Viterbi(E(g, f), P_ss, P_se)
12:      end for
13:   end for
14: end while
    P_se, P_ss ← AtualizarEstatística(S, E)
```

classificação. Existem inúmeros métodos para realizar as duas tarefas, mas, em geral, eles não se adaptam muito bem ao nosso problema, principalmente porque o nosso objectivo não é criar a melhor clusterização ou classificação possível, mas sim servir de pré-processamento para os HMMs.

Sempre com o objetivo de ter um método rápido e eficiente, preferimos adotar um método de classificação bem elementar. Para cada gênero existe uma nuvem de pontos que fica localizada em um hipercubo entre 0 e 1. O método mais trivial de classificação seria calcular a média dos dados e usá-la como referência para reclassificar os vetores que foram usados para calcular essa mesma média.

Um problema desse tipo de classificador é que ele está completamente vulnerável ao *crosstalking*, algo que, por suposição, deve se manifestar nessa nuvem como *outliers*. Uma medida que pode dar maior proteção contra esses pontos, e faz mais sentido geométrico, é a mediana. No espaço multidimensional, a mediana é o ponto que, em todas as dimensões, tem o mesmo número de valores maiores e menores, ou seja, tende ao centro.

Os vetores que estão dentro da nuvem, portanto próximos à mediana, podem ser tomados como representativos do gênero; já os que estão nos arredores, já se aproximando de outras medianas, devem ser tomados como *outliers* dessa classe. Para medir essa distância entre as medianas e os descritores, usamos uma métrica *City Block*, porque para nós o que importa é a diferença na distribuição da energia em todas as dimensões.

Poderíamos usar a métrica euclidiana, mas ela, por conta do expoente, pode ser vista como a *City Block* com pesos (ω) iguais a diferença entre os valores considerados ao invés de peso unitário, como mostram as Equações (4.4) e (4.5). Isto significa que onde a diferença é maior, tem mais peso, tendendo a espalhar o erro ao longo

de todas as dimensões; já a *City Block*, visa a diminuir a soma do erro, mesmo que isso signifique concentrar o erro.

$$CityBlock : \sum_{n=0}^{N} \underbrace{1}_{\omega} \cdot (v_{n,1} - v_{n,2}) \tag{4.4}$$

$$Euclidiana : \sum_{n=0}^{N} \underbrace{(v_1 - v_2)}_{\omega} \cdot (v_{n,1} - v_{n,2}), \tag{4.5}$$

onde $v_{n,1}$ e $v_{n,2}$ representam os valores em cada uma das N dimensões do espaço.

Considere que a diferença entre um vetor e os centróides de dois *clusters* resultou nos vetores $A : \{0{,}5; 0{,}5; 0{,}5\}$ e $B : \{0{,}1; 0{,}4; 0{,}8\}$. Usando a distância euclidiana, o vetor estaria mais próximo de A (0,75 menor que 0,81); já usando *City Block*, o vetor estaria mais próximo de B (1,3 menor que 1,5).

A fase final da extração é reclassificar as próprias sequências de treinamento para enviar aos HMMs as sequências de gêneros característicos dos blocos.

4.3 Considerações Finais

O método iterativo proposto para o treinamento dos HMMs converge rapidamente. Na implementação, calculamos o número de sequências alteradas e adotamos como critério de parada 15 iterações ou nenhuma sequência alterada. A grande maioria das partições convergem já na segunda rodada; algumas precisam de até 5 rodadas.

Análise semelhante à feita na Seção 4.1 poderia ser feita considerando as emissões *clusters* em vez de gêneros. Esta foi a nossa abordagem inicial; nela, os *clusters* funcionavam como *kernels*, tentando estimar a PDF para um dado estado. Entretanto, o custo computacional da clusterização feita repetidas vezes era proibitivo.

Este capítulo encerra a apresentação dos métodos propostos. No próximo capítulo, mostraremos os resultados obtidos por esses métodos sob diferentes condições, além da descrição das bases de dados e dos parâmetros dos modelos.

Capítulo 5

Avaliação de Desempenho

Neste capítulo mostramos os resultados obtidos pelos algoritmos propostos e os comparamos com os resultados encontrados na literatura, após descrevermos a metodologia usada nos testes.

Porém, antes de iniciar a descrição da metodologia, iremos descrever uma outra parte do *setup* de testes tão importante quanto a própria metodologia: a base de dados. A Seção 5.1 faz esta descrição, enquanto que a Seção 5.2 descreve a metodologia usada na comparação. A Seção 5.3 apresenta os resultados obtidos.

5.1 Base de Dados

Para validar o desempenho dos algoritmos usamos duas base de dados: a PF, criada anteriormente pelo orientador deste trabalho para o trabalho de projeto final de graduação do autor; e a GTZAN, como já dissemos, criada por George Tzanetakis. As seções que seguem descrevem essas bases.

5.1.1 Base Autoral: PF

Esta base é formada por 122 faixas inteiras, entre 43s e 8m45s, extraídas de CDs de áudio e codificadas mantendo o padrão de CD: 16 bits por amostras, stereo, 44.100Hz e sem compactação.

Essas faixas estão dividas em: Barroco (30 faixas), Choro (30), MPB (32) e Piano (30). Essas categorias foram escolhidas para testar a capacidade de generalização dos classificadores, e por isso, escolhidas de forma que os elos entre os arquivos fossem bastante diversificados. A categoria Barroco se refere a um período de tempo, Piano a um instrumento, MPB ao local de origem e Choro a um gênero.

Ao se escolherem as faixas, foi tomado o cuidado de evitar a interseção entre classes; portanto, não há músicas barrocas em piano, nem solo de piano em Barroco. Caso se criasse mais essa dificuldade para o classificador, ficaria difícil apurar a real

capacidade de generalização do algoritmo. Autores e intérpretes podem aparecer mais de uma vez numa mesma categoria, principalmente em Barroco.

Uma outra particularidade das classes dessa base é que Piano, Barroco e Choro são puramente instrumentais; já MPB é sempre cantado. Foi a forma de evitar interseção desta com as outras classes.

5.1.2 Base *Benchmark*: GTZAN

Esta base é formada por 1000 faixas com duração de 30 segundos extraídas de diversas fontes (CD, radio e gravação com microfone) e codificadas usando 16 bits por amostra, mono, 22050Hz de frequência de amostragem e sem compactação.

Como dissemos na Seção 1.3, Sturm [2] publicou recentemente, em junho de 2013, uma análise detalhada sobre essa base, levantando algumas questões. Como a GTZAN não foi publicada com metadados, Sturm usou uma ferramenta de identificação de músicas (Echo Nest) e, com base nos resultados desta ferramenta, analisou os rótulos do Last.fm® (`http://www.lastfm.com.br/`).

O Last.fm® se define como:

> "um serviço de descobertas de músicas que faz recomendações personalizadas com base nas músicas que você ouve".

Este serviço oferece ao usuário a possibilidade de aplicar rótulos às faixas, albuns e artistas. Ao analisar os rótulos mais frequentes, Sturm espera obter uma possível classificação "consensual" com a qual poderá validar os rótulos aplicados pela GTZAN.

A análise feita por ele mostra que todos os dez gêneros da base, Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae e Rock, cada um com 100 faixas, têm autores/intérpretes repetidos.

Outra característica que a análise feita por Sturm mostra é que as faixas parecem ter sido escolhidas de forma a borrar as fronteiras de gênero. Aproximadamente 10% das classes Metal e Rock foram marcadas com o rótulo Rock; o rótulo Pop foi aplicado a arquivos das classes Disco, Pop e Rock, e o rótulo Dance, às classes Pop e Disco.

Sturm argumenta que na base de dados existem 71 repetições, das quais 50 repetições são "exatas"[1] e 21 são repetições de gravações, além de haver uma música com duas versões; distorções como estática e saturação em 18 arquivos; e na visão do analista, 59 faixas com rótulos questionáveis.

Para obter os nossos resultados mantivemos a GTZAN sem alterações, para possibilitar comparação com outros resultados.

[1] O autor da análise considera faixas com descritores "muito similares" como repetições exatas

5.2 Metodologia

Nesta seção, vamos reportar os testes que fizemos para encontrar os valores dos três parâmetros usados pelo descritor proposto. São eles:

- **Tamanho da Janela** define a unidade básica de tempo com que o algoritmo trabalhará. Assim como na STFT, quanto maior a janela maior a precisão no domínio da frequência; entretanto, aumentar esse parâmetro significa precisar de mais polinômios para descrever um trecho.
- **Número de Polinômios** define o número de bandas que serão usados no algoritmo. É determinante na velocidade de execução do algoritmo.
- **Largura do Bloco** é a unidade fundamental de análise. Um bloco muito grande é incapaz de descrever bem ativações curtas, enquanto que um bloco muito curto é incapaz de perceber as ativações que persistem no tempo.

Para determinar os valores desses parâmetros, usamos a base PF e o classificador matricial baseado no KNN. Todos os desempenhos reportados foram apurados usando validação cruzada entre dez partições.

Uma preocupação que tem nos acompanhado ao longo do texto é com a velocidade de execução dos algoritmos. Por isso, decidimos escolher parâmetros que maximizassem a taxa de acerto, mas que possibilitassem execuções relativamente rápidas.

Teste do Tamanho da Janela foi realizado deixando-se fixo o número de polinômios em 50% do tamanho da janela e adotando-se o tamanho de bloco como 50ms.

Na Tabela 5.1, podemos ver o desempenho alcançado e o tempo de processamento necessário para o classificador processar todas as (10) partições da base PF. No tempo de processamento, medido através das funções *tic* e *toc* do Matlab®, não está incluída a extração; entretanto, o tempo necessário para carregar os dados pré-extraídos do disco foi contabilizado.

Janela	2048	**1024**	512	**256**	128	64
Acerto	65.57	**65.57%**	59.02%	**63.11%**	60.66%	52.46%
Tempo	863s	**254s**	188s	**182s**	167s	172s

Tabela 5.1: Teste tamanho da janela: número de polinômios igual a 50% do tamanho da janela e tamanho do bloco aproximadamente 50ms.

As janelas nos extremos, com 2048 e 64 amostras, foram prontamente desconsideradas, ou pelo tempo de execução bem acima das demais (1024) ou pela taxa de acerto abaixo das demais (64).

Consideramos que as janelas de tamanho 1024, 512 e 256 estão no mesmo patamar em termos de velocidade de execução e taxa de acerto. Optamos, assim, por fazer os próximos testes para 1024 e 256 para explorar a maior variação do parâmetro.

Teste do Número de Polinômios foi feito para as janelas de tamanho 1024 e 256, mantendo-se o comprimento do bloco fixo em 50ms. Na Tabela 5.2, podemos ver os desempenhos e tempos de execução obtidos.

256					
Acerto	64.75	64.75%	**62.30%**	65.57%	59.84%
Tempo	182s	177s	**173s**	170s	168s
1024					
Acerto	64.75	59.84%	**61.48%**	59.84%	59.84%
Tempo	518s	306s	**275s**	221s	217s
Janela	90%	70%	**50%**	30%	10%

Tabela 5.2: Teste número de polinômios: tamanho do bloco aproximadamente 50ms.

Não há evidências de que um valor tenha taxa de acerto melhor que as demais. Como os polinômios de menor grau representam a região de menor frequência do espectro, estes resultados sugerem que as baixas frequências carregam mais informação sobre o gênero da faixa.

Entretanto, consideramos que diminuir demais o número de descritores pode ser prejudicial para a classificação da GTZAN. Essa base foi amostrada com 22050Hz limitando, portanto, a frequência máxima a 11kHz; então, se usarmos poucos descritores desprezaremos ainda mais informação frequencial que pode ser útil para a classificação.

Aumentar o percentual de polinômios pode, nas ordens mais altas, comprometer o tempo de execução, enquanto que diminuí-lo demais pode empobrecer a descrição. Assim, optamos por manter os 50% iniciais.

Teste Tamanho do Bloco foi realizado para os tamanhos de janela 1024 e 256 mantendo fixo o número de polinômios em 50% do tamanho da janela. O tamanho do bloco foi aproximado para números inteiros de janelas.

Os resultados (Tabela 5.3) apresentam evidências de que não há diferenças significativas entre blocos de 50ms e 100ms, nem entre blocos de 250ms e 500ms, considerando taxas de acerto e tempo de processamento; entretanto, usar blocos de 50 ou 100ms parece um pouco melhor.

Optamos por usar blocos de 50ms para dar mais dados ao classificador Sequencial.

256						
Acerto	55.74	**63.93%**	60.66%	62.30%	59.02%	57.38%
Tempo	273s	**259s**	257s	259s	259s	258s
1024						
Acerto	68.85	**67.21%**	66.39%	59.02%	60.66%	58.20%
Tempo	452s	**403s**	404s	389s	389s	387s
Bloco	25ms	**50ms**	100ms	250ms	500ms	1s

Tabela 5.3: Teste tamanho do bloco: número de polinômios igual a 50% do tamanho da janela.

5.3 Resultados

Nesta seção apresentaremos e analisaremos os resultados obtidos pelos métodos propostos. Na Seção 5.3.1, discutimos os resultados das duas bases de dados quando classificadas sem alterações; na Seção 5.3.2, fazemos algumas mudanças nas bases, tentando torná-las mais semelhantes para facilitar as comparações entre elas; e finalizando o capítulo, usamos a base PF para testar os métodos propostos em situações diferentes da classificação de gêneros musicais.

Todos os resultados desta seção foram obtidos usando os classificadores definidos nos capítulos anteriores e usando os parâmetros com os valores determinados na seção anterior. São eles:

- **Matricial 1 (M1)**: Classificador baseado no KNN definido no Capítulo 3.
- **Matricial 2 (M2)**: Classificador por linhas definido na Seção 3.2.1.
- **Sequencial (S)** : Classificador sequencial descrito no Capítulo 4.

5.3.1 Bases Completas

Esta seção traz, na Tabela 5.4, os resultados obtidos ao se classificar as bases de dados usando a abordagem todos-contra-todos (isto é, quando todos os gêneros competem entre si pela indicação como gênero da faixa). Na Tabela 5.5, estão as matrizes de confusão geradas ao se classificar a base PF; já as matrizes de confusão geradas ao se classificar a GTZAN estão descritas nas Tabelas 5.6, 5.7 e 5.8. Em todas as matrizes de confusão apresentadas neste capítulo, os rótulos atribuídos pelo classificador se encontram nas colunas.

Para deixar claro a quais confusões entre gêneros estamos nos referindo, adotamos a seguinte notação: $A \rightarrow B$, se refere aos casos em que arquivos do gênero A foram rotulados com o gênero B (na tabela, linha A, coluna B). $A \leftrightarrow B$, se refere tanto a $A \rightarrow B$ como $B \rightarrow A$.

	Matricial 1	Matricial 2	Sequencial
PF	63.93%	71.31%	63.11%
GTZAN	42.40%	45.70%	40.10%

Tabela 5.4: Taxa de acerto: GTZAN e PF sem alterações.

Na Tabela 5.4, vemos que os classificadores propostos alcançam taxas ainda menores que as reportadas inicialmente por Tzanetakis [7]. Entretanto, os resultados na base PF indicam que os classificadores conseguem discernir entre os gêneros. Na Seção 5.3.2 apresentamos alguns testes feitos para tentar identificar quais as causas de tamanha diferença entre as taxas de acerto.

	Matricial 1				Matricial 2				Sequencial			
	B	C	M	P	B	C	M	P	B	C	M	P
Barroco	12	6	5	7	17	2	4	7	18	3	4	5
Choro	1	18	11	0	1	20	9	0	2	18	10	0
MPB	1	8	20	3	3	8	20	1	4	9	16	3
Piano	2	0	0	28	0	0	0	30	5	0	0	25

Tabela 5.5: Matriz de confusão: PF.

Na Tabela 5.5, vemos que os classificadores enfrentam dificuldades para separar Choro $\leftrightarrow$ MPB e fazem alguma confusão entre Barroco $\leftrightarrow$ Piano. Curiosamente, o método sequencial consegue o melhor resultado na classe Barroco e o pior na classe MPB. Acreditamos que na classe Barroco, até mesmo por conta da repetição de autor/intérprete, há menos variação, levando a um modelo que descreve a classe de forma mais precisa; já para a classe MPB, a diversidade é maior, e o modelo perde precisão.

	Matricial 1									
	B	Cl	Co	D	H	J	M	P	Re	Ro
Blues	42	2	15	6	9	1	4	10	5	6
Classical	3	69	5	4	1	5	2	6	1	4
Country	9	1	42	15	7	3	7	9	3	4
Disco	8	3	7	34	7	1	10	22	1	7
Hiphop	4	0	8	16	34	0	9	21	6	2
Jazz	8	9	16	8	7	29	7	11	3	2
Metal	2	0	6	7	11	1	69	0	1	3
Pop	3	2	10	14	8	2	1	54	3	3
Reggae	7	4	6	9	29	2	2	22	17	2
Rock	7	2	6	14	9	3	12	10	3	34

Tabela 5.6: Matriz de confusão: GTZAN classificada pelo algoritmo matricial por blocos (M1).

Na Tabela 5.6, chama atenção a confusão Reggae $\rightarrow$ Hiphop. Os dois gêneros são bastante marcados ritmicamente, o que pode ter confundido o classificador. Outras

confusões, como, Disco → Pop e Blues → Country → Jazz, já eram esperadas pela própria natureza desses gêneros.

	Matricial 2									
	B	Cl	Co	D	H	J	M	P	Re	Ro
Blues	48	12	11	7	7	1	11	2	1	0
Classical	1	91	1	1	1	0	2	0	3	0
Country	10	8	45	12	4	2	13	5	1	0
Disco	2	6	8	34	7	2	24	13	2	2
Hiphop	10	10	7	9	35	0	17	9	1	2
Jazz	11	28	7	5	4	26	9	8	2	0
Metal	2	0	3	4	3	0	83	2	0	3
Pop	4	14	2	15	7	0	3	51	4	0
Reggae	9	9	6	8	18	3	7	11	28	1
Rock	8	18	15	13	3	0	20	2	5	16

Tabela 5.7: Matriz de confusão: GTZAN classificada pelo algoritmo matricial com classificação por linhas (M2).

Na Tabela 5.7, podemos perceber que as taxas de acerto em todas as classes, exceto Jazz, Pop e Rock, melhoraram. Entretanto, as confusões mudaram bastante. A confusão entre Jazz → Classical, frequentemente reportada na literatura, ficou muito maior; entre Pop ↔ Reggae e Disc → Metal, gêneros bem diferentes, também. Acreditamos que durante o processo de classificação, algumas linhas, que eram muito disputadas, acabaram não sendo consideradas votos válidos, dando força a linhas que claramente eram semelhantes a um gênero, mesmo que esse outro gênero fosse bem diferente.

	Sequencial									
	B	Cl	Co	D	H	J	M	P	Re	Ro
Blues	39	9	5	14	4	8	9	4	8	0
Classical	4	75	1	0	0	8	0	4	2	6
Country	19	5	24	22	3	8	4	1	7	7
Disco	3	1	6	49	10	2	12	8	4	5
Hiphop	7	3	4	23	20	7	5	12	19	0
Jazz	13	19	10	9	6	16	2	9	8	8
Metal	1	0	2	5	7	0	80	0	1	4
Pop	5	8	1	18	10	10	2	21	17	8
Reggae	5	4	2	9	18	6	3	6	46	1
Rock	12	8	8	13	2	2	16	3	5	31

Tabela 5.8: Matriz de confusão: GTZAN classificada pelo algoritmo Sequencial.

Na Tabela 5.8, percebemos confusão entre gêneros similares, Country → Blues, Hiphop → Disco, Jazz → Classical; o que corrobora que o modelo consegue de fato descrever a dinâmica dos gêneros. A confusão Reggae → Hiphop também volta a aparecer.

5.3.2 Variações

Nesta seção fizemos algumas variações na GTZAN e na PF para tornar a comparação entre elas mais fácil. Quatro características principais diferenciam as duas bases; são elas: i) o número de gêneros; ii) a duração dos arquivos; iii) a fonte dos arquivos; e iv) a frequência de amostragem.

Fizemos experimentos tentando adaptar as três primeiras características para melhorar a comparação entre as bases. Na Tabela 5.9, estão os resultados da base GTZAN quando submetemos aos métodos propostos os conjuntos gerados a partir de todas as combinações quatro a quatro dos 10 gêneros da base.

	Matricial 1	Matricial 2	Sequencial
GTZAN	42.40%	45.70%	40.10%
GTZAN (4x4) Média	54.20%	64.22%	58.08%
GTZAN (4x4) Completa	34.60%	45.10%	39.10%

Tabela 5.9: Taxas de acerto: combinações de 4 gêneros da GTZAN.

Ao combinar os dez gêneros da GTZAN, 4 a 4, são 210 combinações possíveis. A média das taxas obtidas na classificação de todas essas combinações se encontra na segunda linha da tabela. A terceira linha mostra o resultado ao classificarmos a base inteira usando os resultados dessas classificações parciais. Cada parcial vota uma vez no gênero do arquivo e o gênero vencedor é o que tem mais votos. Na primeira linha, replicamos o resultado obtido na seção anterior.

Já na Tabela 5.10, estão os desempenhos obtidos: ao encurtarmos os arquivos da base PF para apenas 30s de duração; e após passar os arquivos da PF (com duração completa) pelo processo de codificação e decodificação MP3 (128kbps). Para encurtar os arquivos, usamos os primeiros 30s após o primeiro segundo, ou seja, do segundo ao trigésimo primeiro segundo.

	Matricial 1	Matricial 2	Sequencial
PF	63.93%	71.31%	63.11%
PF: 30s	47.54%	63.93%	49.18%
PF: MP3	63.11%	68.03%	64.75%

Tabela 5.10: Taxas de acerto: Variações da PF.

Quanto à quarta característica, poderíamos ter reamostrado a base PF reduzindo a frequência de amostragem. Entretanto, julgamos que o processo de reamostragem sempre deixaria a desejar e não permitiria uma boa comparação.

A Tabela 5.11 traz as matrizes de confusão obtidas ao classificar a versão de 30 segundos da base PF pelos três classificadores que definimos anteriormente. De forma geral, este experimento comprova que classificar a matriz como um bloco fixo, como é feito no classificador M1, não tem o mesmo desempenho da classificação por

	Matricial 1				Matricial 2				Sequencial			
	B	C	M	P	B	C	M	P	B	C	M	P
Barroco	13	4	4	9	18	3	2	7	15	2	5	8
Choro	4	14	10	2	5	17	7	1	4	9	15	2
MPB	4	19	7	2	4	11	13	4	6	10	13	3
Piano	4	2	0	24	0	0	0	30	6	1	0	23

Tabela 5.11: Matriz de confusão: versão de 30s da base PF.

linhas. Quanto ao classificador S, parece que a quantidade de dados foi insuficiente para treinar as matrizes do modelo de Markov.

	Matricial 1				Matricial 2				Sequencial			
	B	C	M	P	B	C	M	P	B	C	M	P
Barroco	13	4	5	8	17	3	3	7	20	3	3	4
Choro	1	17	12	0	3	17	9	1	2	16	11	1
MPB	1	9	19	3	3	9	19	1	3	9	17	3
Piano	2	0	0	28	0	0	0	30	4	0	0	26

Tabela 5.12: Matriz de confusão: versão codificada da base PF.

A Tabela 5.12 traz as matrizes de confusão obtidas ao classificar a versão da base PF submetida à codificação perceptual pelos três classificadores que definimos anteriormente. Acreditamos que a aproximação polinomial já descarta informações que a codificação descartaria e que, portanto, é "imune" aos efeitos da codificação.

5.3.3 Outros

Esta seção reporta os resultados de dois experimentos feitos usando diferentes "arrumações" da base PF. No primeiro experimento, agrupamos as classes Barroco e Piano na classe Clássico, e as classes Choro e MPB deram origem à classe Popular. No segundo experimento, agrupamos Barroco, Piano e Choro em Instrumental e MPB foi renomeada para Cantada.

Na Tabela 5.13, podemos ver os resultados obtidos nesses experimentos. Como vimos nos casos anteriores, o algoritmo encontra dificuldade para separar a classe Barroco de Piano e, principalmente, Choro de MPB. Ao agrupar essas classes, a taxa de acertos naturalmente aumenta.

	Matricial 1	Matricial 2	Sequencial
PF: Clássico x Popular	86.07%	86.07%	82.79%
PF: Cantado x Instrumental	76.23%	84.43%	68.03%

Tabela 5.13: Taxas de acerto: Classico x Popular e Cantado x Instrumental.

Entretanto, o caso mais interessante nesse experimento é o desempenho do classificador Sequencial no segundo experimento. Como vimos no Capítulo 4, esse al-

goritmo foi construído para diferenciar diferentes dinâmicas temporais ao longo da faixa, diferentemente dos classificadores Matriciais, que, apesar de levarem em conta essa dinâmica, parecem estruturalmente mais aptos a entender a diferença que existe entre essas classes (por sua forma própria de tratar a informação espectral).

Clássico x Popular

	Matricial 1		Matricial 2		Sequencial	
	Clássico	Popular	Clássico	Popular	Clássico	Popular
Clássico	47	13	50	10	51	9
Popular	4	58	7	55	12	50

Tabela 5.14: Matriz de confusão: Clássico x Popular.

A Tabela 5.14 apresenta dois resultados interessantes: i) A classificação da matriz de ativação em bloco (M1) apresentou, diferentemente do que vimos até agora, o mesmo resultado da classificação por linha (M2); ii) Mesmo com uma grande variedade, o classificador (S) conseguiu obter boa classificação.

Acreditamos que ao unir as classes que apresentavam maior dificuldade de separação, criamos duas classes claramente separadas, diante do que os classificadores M1 e S conseguem superar suas dificuldades: no caso do M1, a de precisar casar as matrizes de ativação como um todo; no caso do S, a de descrever diferentes padrões de compartilhamento.

Cantada x Instrumental

	Matricial 1		Matricial 2		Sequencial	
	Cantada	Instr.	Cantada	Instr.	Cantada	Instr.
Cantada	17	15	20	12	25	7
Instrumental	14	76	7	83	32	58

Tabela 5.15: Matriz de confusão: Cantada x Instrumental.

Na Tabela 5.15, podemos ver as matrizes de confusão para o experimento Cantado x Instrumental. Novamente é interessante ressaltar que, como percebemos anteriormente, o classificador Sequencial tem dificuldades de descrever classes com grande variedade.

Ainda nessa tabela, percebemos que o erro cometido pelos classificadores matriciais, dentro da classe Cantada, é expressivo (aproximadamente 63% de acerto, para o M2 contra 92% na outra classe). Nem o classificador matricial nem os descritores são aptos para essa tarefa; por isso, entendemos que o classificador esteja na verdade usando a informação rítmica das faixas para realizar a classificação, mesmo sem "notar" que em uma classe há a presença de voz cantada e em outra, não.

Com este experimento, encerramos a descrição dos algoritmos propostos, restando apenas as conclusões finais, expostas no capítulo seguinte.

Capítulo 6

Conclusão

Para concluir, gostaríamos de fazer um breve resumo das nossas contribuições neste trabalho.

No Capítulo 3, propusemos uma nova família de descritores, que, dependendo do par critério-referência, pode se adaptar a diferentes sinais ou aplicações. Nesse mesmo capítulo, apresentamos uma possível representação compacta do arquivo de áudio e finalizamos o capítulo apresentando duas formas de classificar essas representações.

No Capítulo 4, tentamos explorar as sequências intrínsecas às músicas utilizando modelos de Markov. Nesse processo, contornamos o problema do *Crosstalking*.

No Capítulo 5, apresentamos os resultados de alguns experimentos que tentam demonstrar as virtudes e os problemas dos classificadores propostos.

O classificador Matricial por bloco consegue descrever as dinâmicas das classes, mas é refém da necessidade de encontrar um padrão que seja mais próximo a um bloco como um todo. Já o classificador Matricial por linha, apresentou resultados superiores (ou iguais) ao classificador por bloco em todos os testes. Embora represente um ganho em termos de desempenho, esse classificador cria situações anômalas quando descarta muitas linhas dos processo de votação.

O classificador Sequencial mostrou ser bem preciso quando há pouca variação dentro de uma mesma classe, mas quando existem muitos padrões de compartilhamento ele parece não convergir para nenhum.

Uma grata surpresa que os experimentos reportados no capítulo anterior nos proporcionou foi o que parece ser a imunidade do descritor proposto quanto à codificação perceptual. Conseguir classificar arquivos que passaram pelo processo de codificação com perdas é um desafio, que cedo ou tarde, a área terá que enfrentar.

Consideramos ainda que o encurtamento dos arquivos da base PF para 30s trouxe resultados bem abaixo do esperado. Aliado à frequência de amostragem, consideramos que o tamanho reduzido das faixas pode fornecer uma pista sobre o fraco desempenho na GTZAN.

De modo geral, os classificadores demonstraram boa capacidade de generalização, mas estão ainda muito longe de conseguir resolver de forma comercial o problema da classificação de gêneros musicais.

Na próxima seção, sugeriremos algumas ações que podem melhorar os algoritmos propostos ou trazer outros resultados.

6.0.4 Trabalhos Futuros

Talvez a continuação mais trivial deste trabalho seria realizar os testes descritos usando referências e similaridades diferentes. Por exemplo, usar os polinômios de Chebyshev como referência. Uma outra possibilidade seria testar apenas a diagonal da matriz de ativação, ou ainda, tentar fazer uma espécie de *feature selection* com as matrizes.

Do ponto de vista dos classificadores matriciais, outras distâncias e classificadores podem ser explorados. A criação de um modelo de gêneros mais complexo, por exemplo, através de tensores.

Do ponto de vista do classificador sequencial, outras formas de criar as sequências de emissões podem ser exploradas, por exemplo, usando a distância de Mahalanobis, mantendo os gêneros como emissões, ou não. As sequências podem ser inicializadas de diferentes formas.

As bases usadas nesse projetos poderiam ser melhoradas. A base PF tem poucos arquivos e poderia ser estendida. Já a GTZAN poderia ser refeita levando em consideração as observações feitas por Sturm [2].

Como uma última sugestão, o classificador matricial pode ser usado como uma primeira etapa para o classificador sequencial.

Referências Bibliográficas

[1] TZANETAKIS, G., ESSL, G., COOK, P. "Automatic Musical Genre Classification Of Audio Signals". In: *2nd Annual International Symposium on Music Information Retrieval*, pp. 205–210, Bloomington, Indiana, EUA, Outubro 2001. Disponível em: <http://webhome.cs.uvic.ca/~gtzan/work/pubs/ismir01gtzan.pdf>.

[2] STURM, B. L. "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use", *CoRR – Computer Research Repository*, v. abs/1306.1461, 2013. Disponível em: <http://arxiv.org/abs/1306.1461>.

[3] SEO, J. S., LEE, S. "Higher-order moments for musical genre classification", *Signal Processing*, v. 91, n. 8, pp. 2154 – 2157, Agosto 2011. ISSN: 0165-1684. doi: DOI:10.1016/j.sigpro.2011.03.019. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0165168411000946>.

[4] SEYERLEHNER, K., SCHEDL, M., POHLE, T., et al. "Using Block-Level Features for Genre Classification, Tag Classification and Music Similarity Estimation". In: *Music Information Retrieval Evaluation eXchange*, Utrecht, Holanda, Agosto 2010. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.176.4206>.

[5] FOOTE, J. "Visualizing music and audio using self-similarity." In: *7th ACM International Conference on Multimedia*, pp. 77–80, Orlando, Florida, EUA, Novembro 1999. Disponível em: <http://rotorbrain.com/foote/papers/footeACM99.pdf>.

[6] CONARD, N. J., MALINA, M., MÜNZEL, S. C. "New flutes document the earliest musical tradition in southwestern Germany", *Nature*, v. 460, n. 7256, pp. 737–740, Jun 2009. Disponível em: <http://www.nature.com/nature/journal/v460/n7256/full/nature08169.html>.

[7] TZANETAKIS, G., COOK, P. "Musical genre classification of audio signals", *Speech and Audio Processing, IEEE Transactions on*, v. 10, n. 5,

pp. 293–302, Julho 2002. ISSN: 1063-6676. doi: 10.1109/TSA.2002.800560. Disponível em: <http://webhome.cs.uvic.ca/~gtzan/work/pubs/tsap02gtzan.pdf>.

[8] WEST, K., COX, S. "Features and classifiers for the automatic classification of musical audio signals". In: *5th Annual International Symposium on Music Information Retrieval*, Barcelona, Espanha, Outubro 2004. Disponível em: <http://ismir2004.ismir.net/proceedings/p096-page-531-paper115.pdf>.

[9] SHAO, X., XU, C., KANKANHALLI, M. S. "Unsupervised classification of music genre using hidden markov model". In: *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 2023–2026, Taipei, Taiwan, Junho 2004. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.4854>.

[10] BERGSTRA, J., CASAGRANDE, N., ERHAN, D., et al. "Aggregate features and AdaBoost for music classification", *Machine Learning*, v. 65, n. 2, pp. 2–3, Dezembro 2006. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.640>.

[11] GUAUS, E. *Audio content processing for automatic music genre classification: descriptors, databases, and classifiers*. Tese de Doutorado, Department of Information and Communication Technologies, Universitat Pompeu Fabra, 2009. Disponível em: <http://www.dtic.upf.edu/~eguaus/phd/eguaus_phd_2009_genre_classification_A4.pdf>.

[12] COHEN, M. A., TAN, C. O. "A Polynomial Approximation for Arbitrary Functions", *ArXiv e-prints*, Agosto 2011. Disponível em: <http://adsabs.harvard.edu/abs/2011arXiv1108.0608C>.