UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

RAFAEL HENRIQUE DALEGRAVE ZOTTESSO

Identificação de espécies de pássaros utilizando espectrogramas e dissimilaridade

Maringá
2017

RAFAEL HENRIQUE DALEGRAVE ZOTTESSO

Identificação de espécies de pássaros utilizando espectrogramas e dissimilaridade

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Yandre Maldonado e Gomes da Costa
Coorientador: Prof. Dr. Diego Bertolini Gonçalves

Maringá
2017

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá – PR, Brasil)

Z89i

Zottesso, Rafael Henrique Dalegrave
Identificação de espécies de pássaros utilizando espectrogramas e dissimilaridade / Rafael Henrique Dalegrave Zottesso. -- Maringá, PR, 2017.
70 f.: il., color., tabs.

Orientador: Prof. Dr. Yandre Maldonado e Gomes da Costa.
Coorientador: Prof. Dr. Diego Bertolini Gonçalves.
Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Tecnologia, Departamento de Informática, Programa de Pós-Graduação em Ciência da Computação, 2017.

1. Sistemas de reconhecimento de padrões - Pássaros - Identificação de espécies. 2. Pássaros - Padrões (Informática) - Sistemas de reconhecimento. 3. Espectrogramas. 4. Processamento de sinais. I. Costa, Yandre Maldonado e Gomes, orient. II. Gonçalves, Diego Bertolini, orient. III. Universidade Estadual de Maringá. Centro de Tecnologia. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. IV. Título.

CDD 23.ed. 006.45

Márcia Regina Paiva de Brito – CRB-9/1267

FOLHA DE APROVAÇÃO

RAFAEL HENRIQUE DALEGRAVE ZOTTESSO

Identificação de espécies de pássaros utilizando espectogramas e dissimilaridade

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação pela Banca Examinadora composta pelos membros:

BANCA EXAMINADORA

Prof. Dr. Diego Bertolini Gonçalves
Universidade Tecnológica Federal do Paraná – DACOM/UTFPR-CM

Profa. Dra. Valéria Delisandra Feltrim
Universidade Estadual de Maringá – DIN/UEM

Prof. Ph.D. Luiz Eduardo Soares de Oliveira
Universidade Federal do Paraná – DInf/UFPR

Aprovada em: 27 de setembro de 2017.
Local da defesa: Sala 101, Bloco C56, *campus* da Universidade Estadual de Maringá.

AGRADECIMENTOS

Em primeiro lugar, gostaria de agradecer à Deus por me dar esta vida plena de verdade e alegria. Também, por todas as bênçãos que já recebi e irei receber, pois sei que nunca me abandonarás.

À minha esposa Mayra, que sempre foi companheira e presente. Em nossas vidas, nunca mediu forças para me apoiar e incentivar.

Aos meus pais Cláudio e Lindamir, que me apoiaram incondicionalmente. Presentes direta ou indiretamente, nunca mediram esforços para que tudo em minha vida se tornasse possível.

Ao meu orientador professor Dr. Yandre Maldonado e Gomes da Costa, que sempre esteve presente quando precisei e foi muito compreensivo e paciente nesse período. Sempre com sábias palavras, comentários e sugestões que foram fundamentais para o desenvolvimento deste projeto.

Ao meu coorientador professor Dr. Diego Bertolini Gonçalves, que sempre me deu dicas (e códigos) essenciais à este trabalho. Me incentivou a ingressar no mestrado e forneceu todo suporte pra concluir com sucesso.

Ao Departamento de Informática da Universidade Estadual de Maringá e ao Instituto Federal do Paraná - Campus Paranavaí, por fornecer as condições necessárias para o desenvolvimento de um trabalho de mestrado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro concedido a este trabalho.

Aos meus amigos Plínio, Thiago e Bruna que sempre foram companhias presentes para me apoiar e tornar os momentos difíceis mais felizes.

Enfim, à todos que passaram algum tempo comigo durante esta etapa e que direta ou indiretamente foram companheiros e contribuíram de alguma forma para a conclusão deste trabalho.

Identificação de espécies de pássaros utilizando espectrogramas e dissimilaridade

RESUMO

Este trabalho tem por finalidade apresentar uma proposta para a identificação de espécies de pássaros utilizando espectrogramas e a abordagem de dissimilaridade, em uma base de dados com alta quantidade de espécies (classes). A base de dados é composta por sinais de áudio disponibilizados pelo projeto Xeno-canto, pré-selecionados pelo *LifeClef* 2015 *Bird Task*. Para este trabalho, oito subconjuntos balanceados de dados foram criados a partir dessa base, a fim de variar a quantidade de espécies disponíveis e o tempo de duração dos sinais de áudio nos testes, selecionando somente vocalizações do tipo canto e descartando os chamados. Todos os sinais de áudio utilizados foram pré-processados para reduzir o impacto dos ruídos, removendo outras fontes de sons, e para detectar trechos de interesse que possuem maior relevância. Depois, para representar os sinais de áudio no domínio de imagens houve a geração de espectrogramas, que passaram pelo processo de zoneamento com o intuito de destacar informações específicas de cada região criada. Três descritores de textura foram utilizados para extrair características das regiões dos espectrogramas: *Local Binary Pattern* (LBP), *Local Phase Quantization* (LPQ) e *Robust Local Binary Pattern* (RLBP). Na abordagem dependente de modelo essas características foram diretamente classificadas. Na abordagem de dissimilaridade foi necessário computar os vetores de dissimilaridade (positivos e negativos), para então aplicar o esquema de classificação. Ambos os casos empregaram a classificação por meio do SVM, permitindo a aplicação de regras de combinação para se chegar às decisões finais. Depois de uma série de experimentos, percebeu-se que a abordagem de dissimilaridade apresentou resultados superiores em relação a abordagem dependente de modelo e a literatura.

Palavras-chave: Identificação de Espécies de Pássaros. Dissimilaridade. Reconhecimento de Padrões. Espectrograma a Textura. Processamento de Sinais.

Bird species identification using spectrograms and dissimilarity

ABSTRACT

This work presents a proposal for bird species identification using spectrograms and dissimilarity approach, in a database with a high number of species. The database is composed by audio recordings pre-selected by the LifeClef 2015 Bird Task that can be easily found on Xeno-canto website. In this work, eight subsets of data were created from this database, in order to diversify the amount of species and the duration of the audio samples in our tests, selecting only bird songs and discarding the bird calls. All audio samples used were preprocessed to reduce the impact of noise, removing other sources of sounds, and to detect points of interest with greatest relevance. Then, to transform the audio samples in images, there was a task to generate spectrograms, which went through the zoning process in order to enhance local information from each region created. Three texture descriptors were used to perform feature extraction: Local Binary Pattern (LBP), Local Phase Quantization (LPQ) and Robust Local Binary Pattern (RLBP). In the model-dependent approach these features were directly classified. In the dissimilarity approach it was needed to compute dissimilarity vectors (positive and negative), to further apply the classification scheme. Both cases used a classification through the SVM, allowing the application of combination rules to reach a final decision. After a series of experiments, it was perceived that the dissimilarity approach presented superior results in relation to a model-dependent approach and the literature.

Keywords: Bird Species Identification. Dissimilarity. Pattern Recognition. Spectrogram and Texture. Signal Processing.

LISTA DE FIGURAS

Figura 1.1 Espectrograma com muito ruído. 16
Figura 1.2 Espectrograma gerado de uma amostra que possui a voz de uma pessoa como ruído. 17

Figura 3.1 Variações de `P` e `R` para o LBP. Adaptada de (Mäenpää, 2003). . . 26
Figura 3.2 Cálculo do código LBP original (Mäenpää, 2003). 27
Figura 3.3 Padrões uniformes de diferentes texturas detectados pelo LBP. Adaptada de (Mäenpää, 2003). 27
Figura 3.4 Exemplo do funcionamento do RLBP (Chen et al., 2013). 28
Figura 3.5 Representação do espaço de características de um problema multiclasse, com amostras divididas em cinco classes distintas. 31
Figura 3.6 Representação do espaço de dissimilaridade. 32

Figura 4.1 Etapas que compõe a classificação das espécies de pássaros. 35
Figura 4.2 Etapas sequenciais que fazem parte do pré-processamento. 38
Figura 4.3 A Figura (a) representa um espectrograma de um sinal de áudio antes da redução de ruídos. A Figura (b) representa um espectrograma do mesmo sinal de áudio depois do processamento de redução de ruídos com o SOX. 39
Figura 4.4 Sinal de áudio antes (entrada) e depois (saída) da segmentação automática. 40
Figura 4.5 Espectrogramas de uma mesma amostra da base gerado com diferentes valores para o limite inferior de amplitude. Na Figura (a) o limite inferior da amplitude igual a -60 `dBFS`. Na Figura (b) o limite inferior da amplitude igual a -100 `dBFS`. 42
Figura 4.6 Espectrograma original extraído pelo SOX com amplitude de sinal -60 `dBFS` pronto para utilização. 42
Figura 4.7 Espectrograma dividido em três zonas verticais. 44
Figura 4.8 Espectrograma dividido em três zonas lineares. 45
Figura 4.9 Espectrograma dividido proporcionalmente conforme a escala Mel. 46
Figura 4.10 Metodologia para geração de vetores positivos e negativos na abordagem da dissimilaridade. 48
Figura 4.11 Cálculo dos vetores de dissimilaridade positivos e negativos. . . . 48
Figura 4.12 Esquema de zoneamento de um espectrograma com três zonas verticais e três zonas lineares. 49

Figura 4.13 Criação e combinação de classificadores para as características extraídas de cada zona. Adaptada de (Costa et al., 2013). 51

Figura 5.1 Desempenho das abordagens de dissimilaridade e classe dependente. 62

LISTA DE TABELAS

Tabela 2.1 Melhores resultados obtidos nos trabalhos relacionados. 23

Tabela 4.1 Subconjuntos de dados propostos neste trabalho. 36
Tabela 4.2 Composição de cada subconjunto em relação a espécies. 37
Tabela 4.3 Dados sobre os descritores de textura utilizados 46

Tabela 5.1 Divisão dos subconjuntos para as etapas de treinamento e teste. . 54
Tabela 5.2 Valores de C e *Gamma* (G) encontrados por *grid-search* no LIBSVM. 55
Tabela 5.3 Taxas de reconhecimento (%) obtidas com um *kernel* RBF e *grid-search* para C e *Gamma* (G). 55
Tabela 5.4 Taxas de reconhecimento (%) com *kernel* RBF, $C = 8$ e $Gamma = 2$. 55
Tabela 5.5 Taxas de reconhecimento (%) com *kernel* linear, $C = 8$ e $Gamma = 2$. 56
Tabela 5.6 Taxas de reconhecimento (%) com *kernel* RBF, $C = 32$ e $Gamma = 2$. 56
Tabela 5.7 Taxas de reconhecimento (%) dos descritores de textura LBP, RLBP e LPQ no subconjunto #3. 57
Tabela 5.8 Taxas de reconhecimento (%) obtidas a partir da variação das zonas verticais e horizontais. 57
Tabela 5.9 Taxas de reconhecimento (%) obtidos variando as zonas horizontais. 58
Tabela 5.10 Taxas de reconhecimento média (%) utilizando dissimilaridade em subconjuntos com diferentes quantidades de classes. 59
Tabela 5.11 Taxas de reconhecimento (%) médias obtidas ao utilizar diversos modelos para classificar subconjuntos diferentes. 59
Tabela 5.12 Taxas de reconhecimento (%) utilizando três zonas verticais e 15 zonas horizontais. 60
Tabela 5.13 Taxas de reconhecimento (%) obtidas na abordagem classe dependente em diversos subconjuntos. 61
Tabela 5.14 Melhores resultados de trabalhos relacionados e da abordagem de dissimilaridade deste trabalho. 63

LISTA DE SIGLAS E ABREVIATURAS

AR: *Autoregressive*
DCTMFCC: *2-D Discrete Cosine Transform in Mel-Frequency Cepstral Coefficients*
DTW: *Dynamic Time Warping*
HMM: *Hidden Markov Model*
IOIHC: *Inter-Onset Interval Histogram*
LBP: *Local Binary Pattern*
LIBSVM: *A Library for Support Vector Machines*
LPQ: *Local Phase Quantization*
MFCC: *Mel-Frequency Cepstral Coefficients*
MIR: *Music Information Retrieval*
MVD: *Modulation Frequency Variance Descriptor*
RBF: *Radial Basis Function*
RH: *Rhythm Histogram*
RLBP: *Robust Local Binary Pattern*
RP: *Rythm Patterns*
SOX: *Sound Exchange*
SSD: *Statistical Spectrum Descriptor*
STFT: *Short-time Fourier Transform*
SVM: *Support Vector Machine*
TDCMFCC: *Two-Dimensional Cepstrum in Mel-Frequency Cepstral Coefficients*

SUMÁRIO

1 Introdução **12**
1.1 Motivação . . . 14
1.2 Desafios . . . 14
1.3 Objetivos . . . 16
1.4 Contribuições . . . 18
1.5 Organização . . . 18

2 Revisão Bibliográfica **19**
2.1 Trabalhos Relacionados . . . 19
2.2 Considerações Finais . . . 21

3 Fundamentação Teórica **24**
3.1 Base de Dados . . . 24
3.2 Representação da Textura . . . 25
3.2.1 LBP . . . 26
3.2.2 RLBP . . . 27
3.2.3 LPQ . . . 29
3.3 Dissimilaridade . . . 30
3.4 Regras de Combinação . . . 33

4 Metodologia Proposta **35**
4.1 Base de Dados . . . 36
4.2 Pré-processamento . . . 37
4.2.1 Redução de Ruídos . . . 38
4.2.2 Segmentação do Sinal de Áudio . . . 39
4.2.3 Geração de Espectrogramas . . . 41
4.2.4 Zoneamento do Espectrograma . . . 41
4.3 Extração de Características . . . 45
4.4 Dissimilaridade . . . 47
4.5 Classificação . . . 50
4.6 Combinação de Classificadores . . . 51
4.7 Avaliação dos Resultados . . . 52

5 Resultados Experimentais **53**
5.1 Abordagem de Dissimilaridade . . . 53

5.1.1 Avaliação de Parâmetros do SVM, Descritores de Textura e Zoneamento do Espectrograma 54
5.1.2 Avaliação da Abordagem de Dissimilaridade em Diferentes Subconjuntos 58
5.2 Abordagem Classe Dependente 59
5.3 Discussão dos Resultados 61

6 Considerações Finais **64**
6.1 Trabalhos Futuros 65

REFERÊNCIAS **66**

1

Introdução

O monitoramento de espécies de pássaros é de suma importância para o controle de fluxo migratório e identificação de espécies. Sobre o fluxo migratório, Negret (1988) explica que diferentes espécies de aves migram conforme as estações do ano, o que dificulta a identificação delas. Faria et al. (2006) relatam diversos métodos para o monitoramento de regiões a fim de identificar espécies de pássaros existentes: observação direta ao longo de "transectos", captura com redes, pontos de escuta e identificação a partir do uso de vocalizações.

Entre as técnicas utilizadas para a identificação das espécies desses animais, é muito comum a montagem de redes de neblina (Faria et al., 2006). Essa prática se baseia em suspender uma rede, feita normalmente de nylon, entre dois pontos como se fosse uma rede de vôlei, possibilitando a captura dos pássaros que se prendem a ela. Todavia, práticas desse tipo colocam em risco a integridade das aves de uma região que, em muitas vezes, acabam se ferindo ao colidir com a rede, podendo até morrer enroscadas. Sendo assim, devido à preocupação com o bem-estar das aves, especialistas sugerem que técnicas não invasivas devam ser utilizadas desde a coleta de dados até o reconhecimento. Além disso, é muito difícil que todas as espécies sobrevoem a exata área na qual a armadilha foi montada, dificultando a identificação.

Com o desenvolvimento tecnológico, diversos dispositivos de gravação de áudio passaram a ser frequentemente utilizados (Conway, 2011; Faria et al., 2006; Schuchmann et al., 2014), o que possibilitou que sistemas de monitoramento de pássaros gravassem seus cantos e chamados de forma menos invasiva, sem a necessidade do contato direto com os animais e seu *habitat*.

As vocalizações podem ser entendidas como a produção ou emissão de som em certos animais. Catchpole e Slater (2003) explicam os que sons emitidos por pássaros podem ser classificados, basicamente, como cantos ou chamados. Segundo eles, o canto dos pássaros tende a ser mais longo e completo, aparecendo de forma espontânea e, muitas vezes, produzido em longos intervalos durante o dia. Já sobre o chamado, tende a ser mais curto, simples, geralmente relacionados com brigas, ameaças, alarmes e outros tipos de comportamento. A partir dos cantos e chamados é possível identificar pulsos, que de acordo com Lopes et al. (2011a), são pequenos intervalos de som que possuem altas amplitudes.

A identificação de espécies a partir do uso dessas vocalizações envolve horas de trabalho humano, divididas entre preparar o equipamento, gravar o som e anotar dados enquanto o som é gravado (Conway, 2011). No Brasil, por exemplo, há um projeto de monitoramento acústico no Pantanal que inclui espécies de pássaros (Schuchmann et al., 2014). Por ser uma região grande, é necessário o auxílio da tecnologia para automatizar e facilitar a coleta e análise de dados.

Mesmo com a dificuldade para gravar esse tipo de áudio, o acesso a bases de pássaros ficou mais fácil para a comunidade científica, possibilitando novos estudos relacionados à identificação de espécies utilizando vocalizações. Um exemplo disso é o projeto Xeno-canto[1], que possui um site dedicado ao compartilhamento de sons de pássaros, no qual colaboradores podem obter e enviar gravações, além de ajudar na identificação de espécies. Diversos trabalhos têm utilizado amostras selecionadas do Xeno-canto (Lopes et al., 2011a,b; Lucio e Costa, 2015; Marini et al., 2015; Zottesso et al., 2016).

A identificação automática de espécies de pássaros é um problema típico de reconhecimento de padrões e boa parte dos estudos incluem as etapas de aquisição, pré-processamento, segmentação, extração de características e classificação (Fagerlund, 2007). Alguns dos primeiros trabalhos de classificação utilizando vocalizações datam o final da década de 90 (Anderson et al., 1996; Kogan e Margoliash, 1998).

Com técnicas mais recentes, muitos trabalhos propõem diversas abordagens para o reconhecimento de espécies de pássaros (Briggs et al., 2009; Cai et al., 2007; Lee et al., 2008; Lopes et al., 2011a; Lucio e Costa, 2015; Marini et al., 2015; Zottesso et al., 2016). Porém, todos os trabalhos utilizaram até 50 espécies para aplicar suas técnicas. Apenas Chou e Liu (2009) realizam testes com centenas de espécies e percebe-se que quando o número de classes aumenta, as taxas de reconhecimento diminuem.

[1] `http://www.xeno-canto.org/`

1.1 Motivação

Neste trabalho, utilizamos a abordagem de dissimilaridade em uma base com um número grande de espécies. Primeiro, pelo fato da dissimilaridade não ter sido explorada na identificação de espécies de pássaros e, também, porque é uma abordagem que, de acordo com Bertolini et al. (2013), não tem a necessidade de retreinar o modelo sempre que novas classes são adicionadas ao sistema de classificação. Segundo, por não existir muitos trabalhos que utilizem centenas de espécies de pássaros no esquema de classificação, realizando uma variação em relação ao tempo do áudio disponível para avaliar as taxas de reconhecimento. Assim, será possível testar uma abordagem nova (dissimilaridade) na identificação de espécies de pássaros e verificar o desempenho deste sistema de classificação à medida que o número de espécies aumenta e chega a centenas.

1.2 Desafios

A base de dados utilizada para a realização deste trabalho é formada por sons de pássaros disponibilizados pelo Xeno-canto e selecionados pelo *LifeClef* 2015 *Bird Task*. O *LifeClef Bird Task* é uma competição anual na qual as equipes participantes criam esquemas para a identificação de espécies de pássaros.

Na base de dados completa disponibilizada pelo *LifeClef* 2015 *Bird Task* é comum perceber o uso de diferentes dispositivos de gravação, muita variação no tempo de duração dos sinais áudio, gravações em diferentes tipos de ambiente e a disparidade na quantidade de amostras por espécies. Isso dificulta a tarefa de encontrar um padrão entre amostras intraclasse, podendo confundir o sistema de identificação. Ademais, a base completa conta com 999 espécies, o que a torna muito mais desafiadora.

Mais especificamente, alguns desafios encontrados nessa base de dados são:

- Ruídos de ambiente contendo barulho de diferentes animais, rios, pessoas, carros e outras fontes;
- A não uniformidade entre sons dos pássaros da mesma espécie, além da presença de sons de outras espécies de pássaros;
- Períodos silenciosos em que não há canto de pássaros nos sinais de áudio;
- Espécies com apenas um canto de pássaro, enquanto outras passam de 50;

- Em uma mesma espécie, o tamanho dos arquivos pode variar entre 119 KB e 17,8 MB;
- Há muitas amostras com duração de um segundo distribuídas pela base;

Este trabalho segue as abordagens propostas em Costa et al. (2011); Lucio e Costa (2015); Zottesso et al. (2016), que utilizam espectrogramas para representar e extrair características de sinais de áudio. Desta forma, os problemas citados influenciam diretamente na textura encontrada no espectrograma e, consequentemente, no resultado da classificação das amostras, pois a presença de ruídos pode trazer uma certa poluição nos espectrogramas.

A sobreposição de diferentes sons nos sinais de áudio pode ser facilmente detectada por um ser humano ao ouvi-lo. Nos espectrogramas, qualquer tipo de som existente em um sinal de áudio é representado no espaço de tempo por frequência. Além disso, a intensidade da cor é diferente quando um som está mais presente do que outro. Se há sobreposição de sons no mesmo tempo e frequência, se sobressai aquele com maior nível de amplitude. Por isso, ter muito ruído nos sinais de áudio pode fazer com que os cantos dos pássaros não sejam corretamente representados nos espectrogramas.

A Figura 1.1 e a Figura 1.2 ilustram dois exemplos ruins de espectrogramas gerados a partir de amostras da base que possuem muito ruído. Na primeira, é possível notar o canto do pássaro na região demarcada (entre um e cinco segundos) e em volta uma grande quantidade de ruído presente no sinal de áudio, causando até uma poluição visual no espectrograma. Na segunda, há o som de uma pessoa conversando (presente entre os tempos de dois a 10 segundos aproximadamente) que começa nas baixas frequências e se estende até as altas, se misturando com o som do pássaro na região demarcada, gerando informações no espectrogramas que serão entendidas, de alguma forma, como características daquela espécie.

Ambas as imagens apresentadas possuem muito ruído tanto na região destacada com o canto dos pássaros quanto no restante do espectrograma. Utilizar espectrogramas muito diferentes da mesma espécie pode fazer com que um modelo de aprendizagem que fique mal treinado devido a constante presença de ruídos. Além disso, na abordagem de dissimilaridade proposta neste trabalho, é necessário que os espectrogramas intraclasse tenham um conjunto de características em comum e os interclasse tenham conjuntos de características diferentes.

Além disso, ao se utilizar espécies com uma única amostra, não é possível ter mais do que um subconjunto de dados. A amostra pode estar presente somente na etapa de treino ou de teste, além do fato de que com um único sinal de áudio é muito difícil generalizar

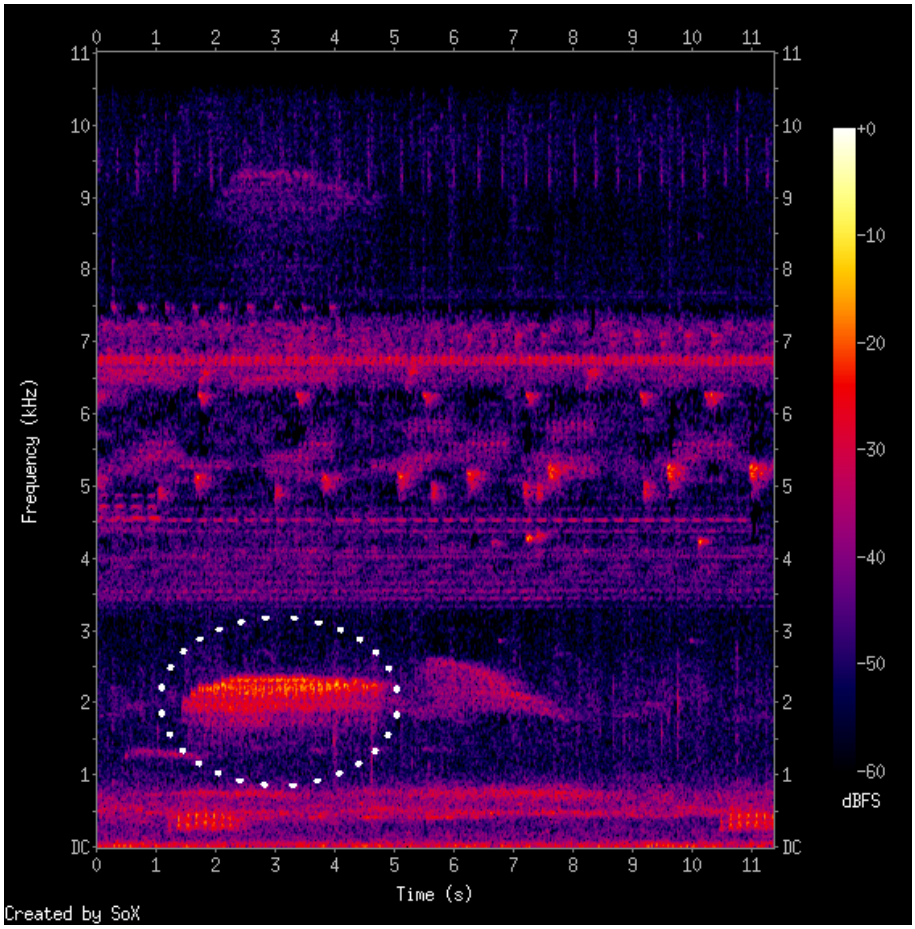


Figura 1.1: Espectrograma com muito ruído.

uma espécie inteira de pássaros. Esse fato se agrava com a presença de amostras muito pequenas, que também podem não representar corretamente uma espécie de pássaro porque não há como garantir um som característico de sua espécie. Nos dois casos, essas amostras podem ter sons de brigas, ameaças, alarmes e outros tipos de comportamento rapidamente expressados pelos pássaros que não são exclusivos de uma única espécie.

1.3 Objetivos

O principal objetivo deste trabalho foi construir um esquema de identificação de espécies de pássaros utilizando espectrogramas e a abordagem de dissimilaridade, em uma base disponibilizada pelo *LifeClef* 2015 *Bird Task* com centenas de classes, pois comparamos o desempenho do classificador, com e sem o uso de dissimilaridade, conforme aumentou-se a quantidade de espécies envolvidas no problema.

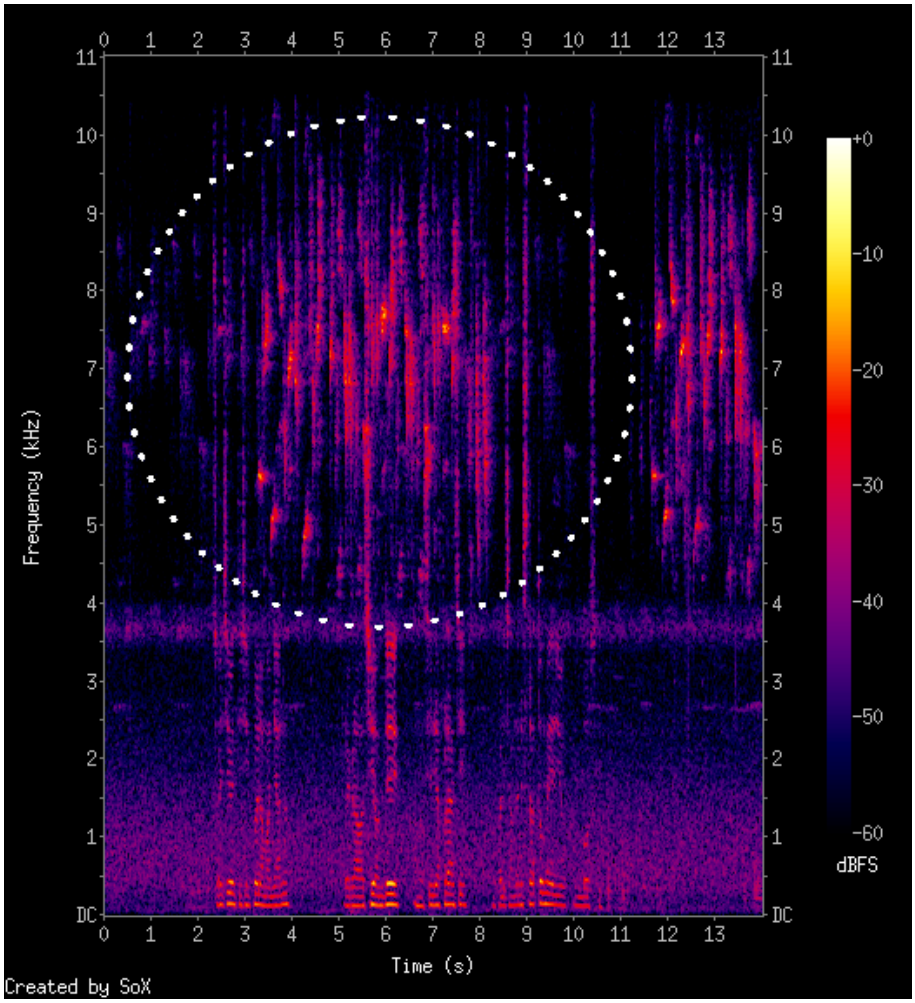


Figura 1.2: Espectrograma gerado de uma amostra que possui a voz de uma pessoa como ruído.

Para atingir esse objetivo principal, pode-se ainda destacar alguns objetivos específicos:

- Avaliar o impacto do tempo de áudio disponível em relação às taxas de reconhecimento;
- Avaliar o desempenho do sistema ao empregar modelos treinados com diferentes números de classes;
- Aplicar o método proposto em uma base de dados com centenas de espécies de pássaros, utilizando a abordagem de dissimilaridade;
- Investigar maneiras de eliminar ruídos e buscar destacar o canto de pássaros em sinais de áudio;

- Avaliar o zoneamento de imagens em espectrogramas com limites equivalentes a bandas de frequência;
- Avaliar diferentes descritores de textura;
- Comparar a abordagem proposta com a literatura.

1.4 Contribuições

O desenvolvimento deste trabalho proporcionou algumas importantes contribuições no contexto da identificação de espécies de pássaros.

As primeiras contribuições, apresentadas em Zottesso et al. (2016), mostraram que é possível encontrar automaticamente partes dos sinais de áudio que melhor caracterizam a vocalização dos pássaros, evitando uma quantidade significativa de trabalho manual e mantendo boas taxas de reconhecimento em problemas de classificação.

Destaca-se também a proposta de uma abordagem nova para a identificação de espécies de pássaros com base na dissimilaridade, que alcançou taxas maiores do que os melhores resultados já apresentados na literatura. Além disso, a abordagem possibilitou a utilização de espécies diferentes nos conjuntos de treino e de teste, não sendo necessário retreinar um modelo toda vez que novas espécies forem inseridas no sistema de identificação.

1.5 Organização

Este trabalho se desenvolve ao longo de cinco tópicos. Este tópico contém algumas informações iniciais sobre a identificação de espécies de pássaros, além de uma contextualização com esta dissertação. O tópico 2 traz uma relação de trabalhos semelhantes a este. O tópico 3 apresenta a revisão dos principais conceitos utilizados. No tópico 4, a metodologia proposta para o desenvolvimento deste trabalho é explicada. O quinto tópico apresenta os resultados experimentais obtidos ao longo desta pesquisa. Por fim, o tópico 6 descreve as considerações finais e trabalhos futuros.

2

Revisão Bibliográfica

Esta seção apresenta resumos de trabalhos relacionados à identificação de espécies de pássaros. Em geral, são citadas as principais informações de cada trabalho, como: origem do conjunto de características (descritores), classificador, origem da base de dados, quantidade de espécies, medidas de avaliação e principais resultados. Por fim, uma tabela foi criada para facilitar a observação de todos os trabalhos citados, apresentando suas principais características.

2.1 Trabalhos Relacionados

Entre os primeiros trabalhos em classificação de espécies de pássaros utilizando sons estão o de Anderson et al. (1996), classificando com um algoritmo *Dynamic Time Warping* (DTW), e o de Kogan e Margoliash (1998) com uma comparação entre DTW e *Hidden Markov Models* (HMMs). Nos dois casos, apenas duas espécies de pássaros estavam presentes. As acurácias foram superiores a 90%.

Chou et al. (2007) propuseram um trabalho com 420 espécies de pássaros com vocalizações, obtidas de uma fonte comercial, segmentadas em sílabas. Nos experimentos, dois terços de cada vocalização foram selecionados aleatoriamente para treinamento e um terço para classificação. Cada conjunto de sílabas foi modelado por um HMM para representar suas características. Os autores utilizaram o algoritmo de Viterbi para classificar o conjunto de teste. A melhor acurácia foi de 78,3%.

Cai et al. (2007) apresentaram um trabalho usando características *Mel-Frequency Cepstral Coefficients* (MFCC) extraídas de sinais de áudio do tipo "chamado", pré-processados

com algoritmos para a redução de ruídos. Redes Neurais foram utilizadas para realizar a classificação das amostras e as acurácias foram de 98,7% e 86,8% usando quatro e 14 espécies, respectivamente.

Com a combinação de classificadores gerados pelos descritores *Two-Dimensional Cepstrum in Mel-Frequency Cepstral Coefficients* (TDMFCC) e *2-D Discrete Cosine Transform in Mel-Frequency Cepstral Coefficients* (DTDMFCC) ao analisar as sílabas dos cantos, Lee et al. (2008) obtiveram a taxa de 84,06% como melhor resultado utilizando o classificador baseado no *nearest neighbor*. A base utilizada era composta por 28 espécies de pássaros.

Com classificadores semelhantes ao de Lee et al. (2008), Briggs et al. (2009) realizaram experimentos em uma base com seis espécies e com vários tipos de características. Para utilizar a densidade do espectro como característica, os autores consideraram que o espectro de magnitude de um *frame* pode ser normalizado para formar uma distribuição de probabilidade. A melhor acurácia foi de 92,10%.

Lopes et al. (2011b) realizaram vários experimentos variando tipos de características e classificadores. A base utilizada era composta por sons de três espécies que foram divididas em cinco *folds* para realizar a validação cruzada. As melhores taxas (*F-measure*) computadas foram de 79,2% para o áudio inteiro e 99,7% para pulsos, ambos obtidos utilizando Redes Neurais *Multilayer Perceptron* (MLP) e características extraídas com o *framework* MARSYAS[1] baseadas no timbre, incluindo MFCC.

Em seguida, Lopes et al. (2011a) apresentaram um trabalho semelhante a Lopes et al. (2011b), porém com mais espécies de pássaros. Ao utilizar os sinais de áudio inteiros, destacam-se resultados para 12 e 20 espécies, com as taxas (*F-Measure*) de 48,8% e 47,4% respectivamente. Nesses experimentos, um classificador baseado em Redes Neurais MLP também classificou as características extraídas com o *framework* MARSYAS baseadas no timbre, incluindo MFCC.

Marini et al. (2015) propuseram a classificação com SVM de 50 espécies com 422 amostras de áudio divididas em cinco *folds*. A acurácia foi calculada conforme um esquema de "TOP N melhores hipóteses" entre 1 e 10, resultando em 45,97% (TOP 1) e 86,97% (TOP 10). Os sinais de áudio foram pré-processados para remover os espaços silenciosos entre cantos.

Lucio e Costa (2015) apresentaram a classificação de espécies de pássaros utilizando espectrogramas gerados a partir dos sons disponibilizados pelo Xeno-Canto. Os descritores de textura LBP, LPQ e Filtros de Gabor foram utilizados para extrair características do sinal de áudio do canto de 46 espécies, cujas amostras foram divididas em 10 *folds*. A

[1] `http://marsyas.info/`

melhor acurácia encontrada foi de 77,65% usando Filtros de Gabor e SVM para realizar a classificação. Porém, todos os sinais de áudio utilizados foram segmentados manualmente a fim de encontrar as regiões de interesse com cantos de pássaros e descartar ruídos externos.

Em seguida, Zottesso et al. (2016) apresentaram um trabalho semelhante ao de Lucio e Costa (2015). A ideia era propor um esquema automático para segmentar os sinais de áudios, removendo partes não interessantes para a identificação das espécies. Os mesmos descritores foram utilizados: LBP, LPQ e Filtros de Gabor, com a melhor acurácia de 78,97% no classificação com SVM e características dos Filtros de Gabor. O único ponto diferente nos trabalhos foi a quantidade de espécies utilizadas, pois o segmentador automático gerou amostras para 45 espécies contra 46 da segmentação manual.

O trabalho de Albornoz et al. (2017) utilizou 25 espécies de uma família que habita a mesma região na América do Sul, com parte das amostras obtidas do Xeno-canto. Os sinais de áudio foram pré-processados com o Filtro de Wiener para redução de ruídos e aplicou-se uma técnica para detectar atividade acústica, baseada no método de Rabiner e Schafer, a fim de identificar os sons dos pássaros. As principais características presentes na literatura foram extraídas com a ferramenta *openSMILE toolkit*. Vários esquemas de classificação foram utilizados e novamente a combinação de Redes Neurais MLP com características MFCC alcançou a maior acurácia, que foi de 89,32%.

Zhao et al. (2017) realizaram um trabalho utilizando 11 espécies com amostras disponibilizadas pelo Xeno-canto, duas delas compostas por vocalizações do tipo canto e nove por chamados. Os autores segmentaram os sinais de áudio com uma técnica baseada no *Gaussian Mixture Model* (GMM) para selecionar eventos acústicos mais representativos. Os espectrogramas desses eventos foram submetidos a um filtro MFCC e depois parametrizado por um modelo denominado autorregressivo (AR). Por fim, o SVM classificou as amostras tendo como medidas de desempenho 93,3% para *Precision* e 91,7% para *Recall*.

2.2 Considerações Finais

Considerando os últimos 10 anos, a maioria dos trabalhos relacionados à identificação de espécies de pássaros fazem o uso de características extraídas diretamente do sinal de áudio. Dentre elas, a *Mel-Frequency Cepstral Coefficients* (MFCC) é uma das mais utilizadas. Outro aspecto comum entre os trabalhos mais recentes é o uso de sons obtidos do projeto Xeno-canto.

Em suma, a avaliação de trabalhos publicados ao longo dos anos contribui vigorosamente para a elaboração desta dissertação, pois é possível perceber a evolução dos sistemas de identificação. Porém, é difícil fazer uma comparação de desempenho devido ao uso de bases com origens diferentes e com muita variação no número de classes utilizadas.

A Tabela 2.1 apresenta uma síntese dos trabalhos descritos nesta seção, bem como as técnicas utilizadas para extração de características e classificação, além da quantidade de espécies, a medida de desempenho e o melhor resultado encontrado.

Tabela 2.1: Melhores resultados obtidos nos trabalhos relacionados.

Trabalho	**Características**	**Classificador**	**Esp.**	**Medida**	**Desempenho (%)**
Anderson et al. (1996)		DTW	2	Acurácia	98,1
Kogan e Margoliash (1998)		DTW e HMM	2	Acurácia	92,5
Chou et al. (2007)	HMM	Algoritmo de Viterbi	420	Acurácia	78,3
Cai et al. (2007)	MFCC	Redes Neurais	4	Acurácia	98,7
		Redes Neurais	14	Acurácia	86,8
Lee et al. (2008)	Combinação de TDMFCC e DTDMFCC	*Nearest Neighbor*	28	Acurácia	84,06
Briggs et al. (2009)	Densidade do espectro	*Nearest Neighbor*	6	Acurácia	92,1
Lopes et al. (2011b)	MARSYAS	Redes Neurais MLP	3	*F-measure*	79,2
			3		99,7
Lopes et al. (2011a)	MARSYAS	Redes Neurais MLP	12	*F-measure*	48,8
			20		47,4
Marini et al. (2015)	MFCC	SVM	50	Acurácia	45,90 a 86,97
Lucio e Costa (2015)	Filtros de Gabor	SVM	46	Acurácia	77,65
Zottesso et al. (2016)	Filtros de Gabor	SVM	45	Acurácia	78,97
Albornoz et al. (2017)	MFCC	Redes Neurais MLP	25	Acurácia	89,32
Zhao et al. (2017)	MFCC/AR	SVM	11	*Precision*	93,3
				Recall	a 91,7

3

Fundamentação Teórica

Nesta seção são apresentadas técnicas computacionais que sustentam esta proposta de dissertação, com o objetivo de dar suporte ao leitor, contribuindo para um melhor entendimento dos métodos empregados. Uma maior quantidade de detalhes pode ser encontrada nas referências aqui citadas. A subseção 3.1 faz uma breve descrição sobre a origem da base de dados utilizada neste trabalho. A subseção 3.2 aborda a textura e o funcionamento de alguns descritores de textura. Na subseção 3.3 há uma breve descrição sobre a abordagem da dissimilaridade. Por fim, a subseção 3.4 aponta as regras de combinação usadas para combinar classificadores.

3.1 Base de Dados

O Xeno-canto é um site dedicado ao compartilhamento de sons de pássaros ao redor do mundo. É também um projeto colaborativo no qual as pessoas podem enviar suas gravações de sons de pássaros e ajudar na identificação de espécies. Além disso, tem como objetivos popularizar as gravações com sons de pássaros, melhorar a acessibilidade aos cantos e contribuir com o conhecimento sobre canto de pássaros.

Com a grande diversidade de sons disponibilizados pelo projeto Xeno-Canto, o *LifeClef* 2015 *Bird Task*, uma competição anual de identificação de espécies de pássaros pelo som, criou uma base de dados com sons de pássaros, distribuídos em 999 espécies, seguindo alguns requisitos importantes para que a tarefa de classificação tivesse condições de chegar o mais perto possível de aplicações do mundo real:

- As amostras de áudio de uma mesma espécie foram obtidas de pássaros distintos presentes em diferentes regiões;
- Os sons foram gravados por diversos usuários que podem não ter utilizado a mesma combinação de microfone e dispositivo de gravação;
- Os sinais de áudio foram obtidos de gravações feitas em vários períodos do ano e em diferentes horários do dia, além de possuir uma variedade de ruídos no ambiente (outros pássaros, zumbido de insetos, etc).

Além dos sinais de áudio, foram disponibilizados alguns dados sobre as amostras. Entre eles, podemos destacar a espécie do pássaro, que será utilizada na classificação das amostras, e o tipo do sinal, que pode ser canto ou chamado. Catchpole e Slater (2003) explicam as diferenças entre cantos e chamados. Segundo eles, o canto dos pássaros tende a ser mais longo, complexo e, geralmente, produzido por machos. Além disso, aparece de forma espontânea e muitas vezes é produzido em longos intervalos durante o dia, com mais frequência em algumas épocas do ano. Já o chamado, tende a ser mais curto, simples e produzido por ambos os sexos durante o ano todo. Usualmente, está relacionado com funções específicas como: brigas, ameaças, alarmes e outros tipos de comportamento.

Desta forma, as amostras do tipo "chamado" foram descartadas neste trabalho por não serem tão típicas de uma espécie como é caso do canto. Assim, os sinais de áudio selecionados a partir da base completa do LifeClef 2015 *Bird Task* são amostras de cantos de espécies de pássaros da América do Sul, totalizando 12.623 cantos distribuídos em 988 espécies. Os sinais de áudio foram normalizados em 44.1 kHz no formato *wav* de 16 bits.

3.2 Representação da Textura

Os sinais de áudio da base de dados utilizada neste trabalho foram convertidos em espectrogramas, para que fosse possível trabalhar no domínio de imagens. Os espectrogramas são representações dos sinais de áudio em forma de imagem, apresentando a densidade espectral de energia em uma relação de tempo por frequência. Assim, o principal atributo a ser explorado neste trabalho é a textura. A textura dos espectrogramas já foi explorada em trabalhos que têm como base dados sinais de áudio (Costa et al., 2011; Lucio e Costa, 2015; Zottesso et al., 2016) que apresentaram bons resultados se comparados com outros trabalhos que utilizam características extraídas diretamente do sinais de áudio.

A textura é facilmente percebida por um observador humano. Na forma digital, essa tarefa é mais difícil porque as imagens são interpretadas por dispositivos eletrônicos como

um composto numérico. Segundo Gonzalez e Woods (2010), a textura pode ser um conjunto de características estatísticas ou outras propriedades locais da imagem que sejam constantes, com pouca variação ou aproximadamente periódica.

As próximas subseções abordam os descritores textura utilizados neste trabalho.

3.2.1 LBP

De acordo com Ojala et al. (2002), o *Local Binary Pattern* (LBP) opera sobre os pixels de uma imagem e seus pixels adjacentes para encontrar um histograma dos padrões binários locais, que pode ser utilizado como descritor de texturas. Para ser capaz de operar texturas de diferentes escalas, pode-se criar padrões LBP estabelecendo-se diferentes quantidades de pixels vizinhos para o seu funcionamento. Tais variações são identificadas por $\text{LBP}_{P,R}$, em que `P` é a quantidade de pixels vizinhos existentes em um círculo de raio `R` ao redor do pixel central. A Figura 3.1 demonstra algumas possibilidades de valores para `P` e `R`.

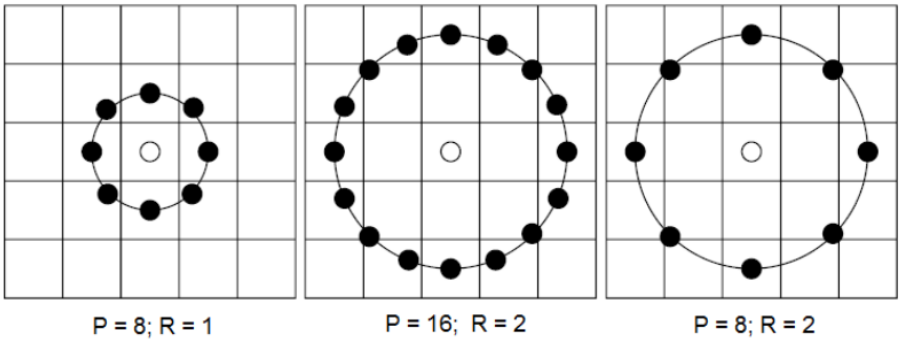


Figura 3.1: Variações de `P` e `R` para o LBP. Adaptada de (Mäenpää, 2003).

Na busca do padrão LBP para um pixel, os valores são computados ao se comparar a intensidade do pixel central com seus vizinhos. Mäenpää (2003) explica que funciona como um limiar: o valor 1 é tomado se a intensidade do pixel vizinho é igual ou maior que o pixel central e 0 se for menor. Depois, esses valores são multiplicados pelos pesos dos pixels vizinhos que correspondem a potências com base dois e com o expoente igual ao seu número de ordem na sequência de vizinhos, sendo o primeiro igual a zero. Por fim, é atribuído ao pixel central o somatório deste cálculo.

A Figura 3.2 apresentada por Mäenpää (2003) ilustra um exemplo seguindo as operações originalmente propostas com o LBP. O primeiro quadrado representa alguns pixels de uma imagem, que tem o valor do pixel central igual a 3. No segundo quadrado, é atribuído 1 onde os valores dos pixels adjacentes são maiores ou iguais a 3 (mais escuros), e 0 quando for menor (mais claros). No terceiro, estão presentes os valores da potência

de 2 elevado ao índice de cada pixel, que começa em 0. No último quadrado, há somente os valores das potências de 2 conforme o índice dos pixels vizinhos que tem o valor maior ou igual a 3. O somatório dos valores deste quadrado representa o padrão binário local (LBP) do pixel central.

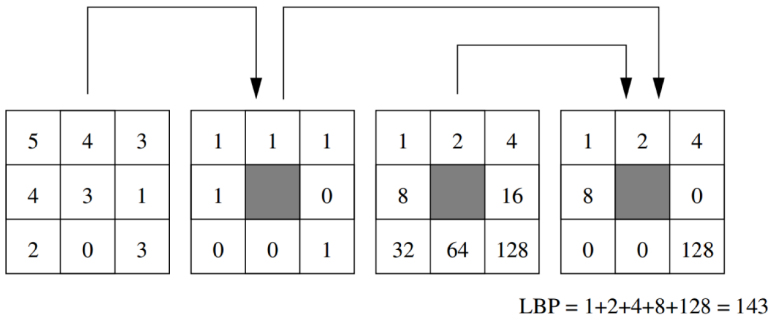


Figura 3.2: Cálculo do código LBP original (Mäenpää, 2003).

O total de padrões LBP que podem ser calculados depende da variação do número de pixels adjacentes em relação ao pixel central. Por exemplo, no $LBP_{8,2}$ existem 2^8 padrões possíveis. Porém, nem todos os padrões computados para cada pixel central são sempre utilizados como características. Mäenpää (2003) relata que alguns padrões são chamados "uniformes" porque possuem no máximo duas transições zero-para-um ou um-para-zero no código binário. Por exemplo: a Figura 3.2 apresenta um padrão binário `11101001` que não é considerado uniforme por ter duas ocorrências `01` e duas `10`; já a Figura 3.3 ilustra cinco exemplos de padrões uniformes representados por cores.

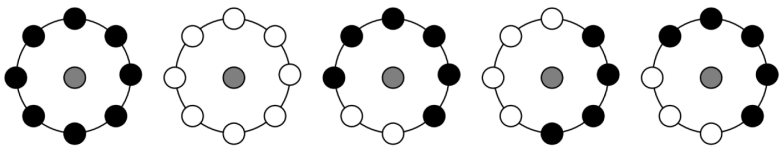


Figura 3.3: Padrões uniformes de diferentes texturas detectados pelo LBP. Adaptada de (Mäenpää, 2003).

Neste trabalho foi utilizado o $LBP_{8,2}$ que, tendo em conta apenas os padrões ditos uniformes, possui um histograma com 59 valores normalizados. Estes valores são tratados como as características extraídas de uma imagem.

3.2.2 RLBP

O *Robust Local Binary Pattern* (RLBP) tem como proposta suprir uma possível deficiência do LBP em relação a robustez à eventuais ruídos presentes nas imagens quando as

mudanças de nível de cinza resultantes do ruído não são monotônicas, mesmo que as mudanças não sejam significativas (Chen et al., 2013). Desta forma, a ideia do RLBP é encontrar um bit no padrão binário LBP que, possivelmente, foi alterado por influência do ruído e então tratá-lo.

O funcionamento do RLBP é ilustrado pela Figura 3.4. Seguindo a ideia do LBP para computar o padrão binário do pixel x_c, seria possível extrair o padrão `11010011`. No entanto, o valor do pixel $x_2 = 124$ tem uma alta probabilidade de ser ruidoso porque possui uma *substring* `(101)`. Alterando o bit correspondente a x_2 de `0` para `1`, o padrão binário passa a ser `11110011`. Este novo padrão passa a ser mais significante tanto para a representação da textura quanto para a classificação (Chen et al., 2013).

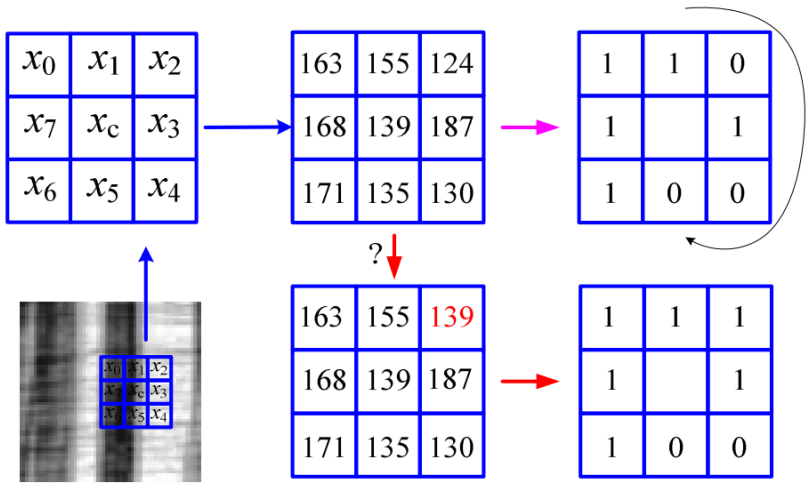


Figura 3.4: Exemplo do funcionamento do RLBP (Chen et al., 2013).

Considerando a ideia do $\text{LBP}_{8,1}$ em que um padrão binário possui 8 bits, uma *substring* de 3 bits pode assumir um dos valores: `000`, `001`, `010`, `011`, `100`, `101`, `110`, `111`. Seguindo a ideia do exemplo na Figura 3.4, assume-se que a ocorrência de uma *substring* `101` ou `010` é ruidosa e, então, acontece a troca para uma nova *substring* `111` ou `000`, respectivamente.

Devido ao fato da base de dados deste trabalho possuir muito ruído e o principal objetivo do descritor ser atuar nessas circunstâncias, alguns experimentos foram realizados para verificar sua eficiência no domínio de identificação de espécies de pássaros com espectrogramas.

Assim como no $\text{LBP}_{8,2}$, o padrão $\text{RLBP}_{8,2}$ também gera um histograma com 59 valores normalizados, considerados como características.

3.2.3 LPQ

Proposto por Ojansivu e Heikkilä (2008), o *Local Phase Quantization* (LPQ) tem como diferencial a robustez para analisar textura insensível ao borramento ou afetada por iluminação não uniforme. Embora o LPQ tenha como diferencial operar sobre imagens borradas, ele também produz bons resultados em situações que não apresentam esse problema.

Esse descritor de textura se baseia na propriedade de invariância ao borramento do espectro de fase de Fourier. Para cada *pixel* da imagem, o LPQ utiliza uma vizinhança retangular de tamanho $N \times N$, chamada de janela local, para extrair informações de fase local empregando a Transformada Discreta de Fourier 2D. A informação da fase local de uma imagem é dada pela *Short-time Fourier Transform* (STFT) descrita na equação 3.1, sendo o filtro ϕ_{u_i} dado pela equação 3.2, na qual $r = (m-1)/2$, m é o tamanho da janela local e u_i é um vetor de frequências 2D.

$$\hat{f}_{u_i}(x) = (f \times \phi_{u_i})x \tag{3.1}$$

$$\phi_{u_i} = e^{-j2\pi u_i^T} | y \in \mathbb{Z}^2 || y || \infty \leq r \tag{3.2}$$

No LPQ, apenas quatro coeficientes complexos são considerados para corresponder às frequências 2D: $u_1 = [a, 0]^T$, $u_2 = [0, a]^T$, $u_3 = [a, a]^T$, $u_4 = [a, -a]^T$, considerando $a = 1/m$. A STFT é expressa através do vetor de notação conforme a equação 3.3, de modo que uma matriz $m^2 \times N^2$ pode ser evidenciada por $F = [f(x_1), f(x_2), \cdots, f(x_{x^2})]$ compreendendo a vizinhança de todos os *pixels* da imagem e $w = [w_R, w_I]$, na qual $w_R = Re[W_u1, W_u2, W_u3, W_u4]$ e $w_I = Im[W_u1, W_u2, W_u3, W_u4]$. Um número complexo tem suas partes reais representadas por $Re[...]$ e as imaginárias por $Im[...]$. $\hat{F} = wF$ é dado para denotar uma matriz de transformação $(8 \times N^2)$.

$$\hat{f}_{u_i}(x) = w_{u_i}^T f(x) \tag{3.3}$$

De acordo com Ojansivu e Heikkilä (2008), a função $f(x)$ de uma imagem é resultado de um processo de primeira ordem de Markov, em que o coeficiente de correlação entre dois *pixels* x_i e x_j é relacionada exponencialmente com a sua distância L^2. Uma matriz de covariância C com tamanho $m^2 \times m^2$ pode ser definida para o vetor f, com base na equação 3.4. A matriz de covariância dos coeficientes de Fourier pode ser obtida por $DwCw^T$. Considerando que D não é uma matriz diagonal, os coeficientes são correlatos

e podem deixar de ser através de $E = C^T \hat{F}$, sendo V uma matriz ortogonal derivada do valor de decomposição singular da matriz D, com $D' = V^T DV$.

$$C_{i,j} = \sigma^{||x_i - x_j||} \tag{3.4}$$

A equação 3.5 é utilizada para quantizar os coeficientes, considerando que e_{ij} são componentes de E. A equação 3.6 representa a transformação dos elementos binários para decimal, representados por números inteiros presentes no intervalo de 0 a 255. Assim, com o histograma do LPQ pode-se formar um vetor de 256 valores, considerando todas as posições da imagem.

$$q_{ij} = \begin{cases} 1 & \text{se } e_{ij} \geq 0 \\ 0 & \text{caso contrário} \end{cases} \tag{3.5}$$

$$b_j = \sum_{i=0}^{7} q_{ij} 2^i \tag{3.6}$$

Mesmo sendo proposto para lidar bem com imagens afetadas por borramento, este descritor tem apresentado bons resultados em situações que não há esse tipo de problema. Desta forma, alguns experimentos também foram realizados com este descritor.

No geral, as características foram extraídas utilizando uma janela de tamanho 3×3, o coeficiente de correlação 0,90 e STFT (*Short-Term Fourier Transform*) com uma janela uniforme. Desta forma, o vetor final de características LPQ corresponde ao histograma construído e possui um total de 256 características.

3.3 Dissimilaridade

Neste trabalho foi adotado o esquema proposto por Pekalska e Duin (2000), baseado na ideia da dicotomia (Cha e Srihari, 2002), chamado de abordagem de dissimilaridade. Essa abordagem tem sido empregada com sucesso em problemas de verificação. Nesse contexto, Jain et al. (2006) descrevem brevemente as finalidades de "verificação" e "identificação":

- Verificação: consiste, basicamente, em averiguar a autenticidade de uma amostra. Isso pode ser feito ao verificar se duas amostras pertencem a uma mesma classe ou não. Por exemplo, queremos saber se dois pássaros pertencem a uma mesma espécie ou não, independente de quais são suas espécies. Dessa forma, acontece uma comparação um pra um $(1:1)$.

- Identificação: Dado uma amostra, a identificação determina qual classe ela pertence com base em um conjunto de dados conhecido. Por exemplo, dado o canto de um pássaro queremos saber qual espécie ele pertence. Assim, acontece uma busca de um para muitos $(1 : N)$.

Um ponto interessante nessa abordagem é a possibilidade de transformar um problema de reconhecimento de padrões multiclasse em um problema binário. A identificação de espécies de pássaros utilizando espectrogramas é um problema multiclasse que também pode ser explorado utilizando essa abordagem.

Um exemplo dessa transformação é ilustrado pela Figura 3.5 e pela Figura 3.6. A primeira apresenta várias amostras de cinco classes diferentes distribuídas em um espaço bidimensional, em que cada amostra é representada por um vetor de características $(f1, f2)$. Na segunda, ocorre a transformação ao computar a dissimilaridade entre as características de cada par de amostras para formar vetores $(f1, f2)$. Esses vetores são chamados de vetores de dissimilaridade.

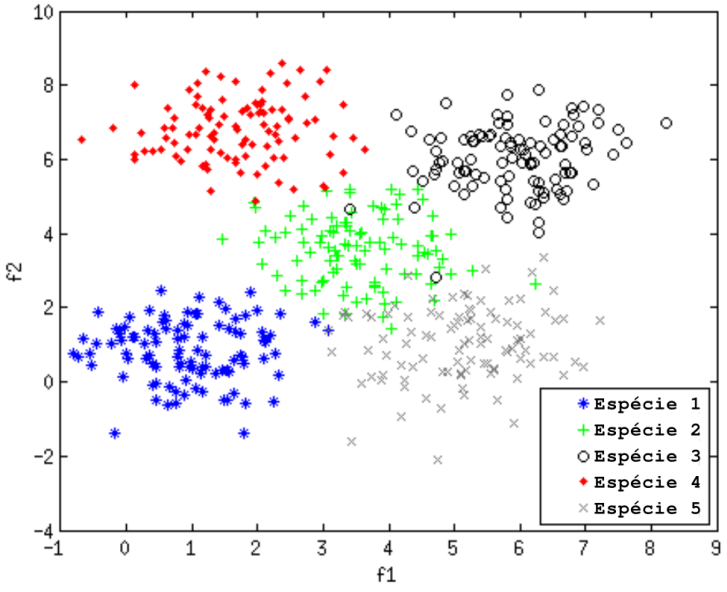


Figura 3.5: Representação do espaço de características de um problema multiclasse, com amostras divididas em cinco classes distintas.

Na Figura 3.6 é possível ver a representação das amostras na abordagem de dissimilaridade evidenciadas apenas por duas classes: positivo `(+)` e negativo `(*)`. O número de amostras é maior porque elas são geradas a partir de cada par dos vetores de características. Se ambos os vetores forem de amostras da mesma classe, então o

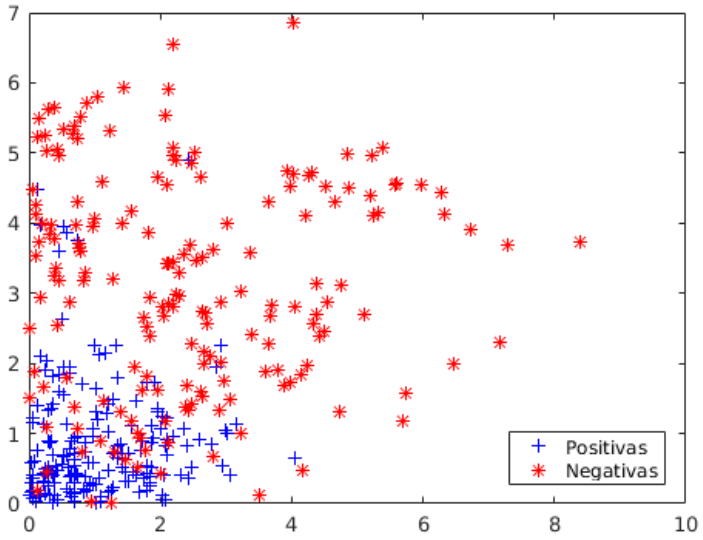


Figura 3.6: Representação do espaço de dissimilaridade.

vetor de dissimilaridade é positivo, e se forem de amostras de classes distintas, o vetor de dissimilaridade é negativo (Bertolini et al., 2013).

> Seja R um conjunto de representação composta de n objetos. Um conjunto de treinamento T com m objetos é representado por $D(T, R)$ por uma matriz de dissimilaridade m. Através da abordagem de representação da dissimilaridade, a classificação de um novo objeto x representado por $D(x, R)$ é realizada usando a regra do vizinho mais próximo. Ao objeto x é atribuído a classe de seu vizinho mais próximo, sendo a classe de representação do objeto r_j dada por $d(x, r_j) = min_{r \in R} D(x, R)$. O ponto chave, aqui, é que as diferenças devem ser pequenas para objetos semelhantes (pertencentes à mesma classe) e, grande para objetos distintos. (Bertolini, 2014)

Deste modo, a proposta deste trabalho é extrair vetores de características de espectrogramas utilizando descritores de textura para, então, computar os vetores de dissimilaridade. A partir de duas amostras da mesma classe, distâncias intraclasse são primeiramente computadas. Se ambos os vetores forem de amostras da mesma classe, então o vetor de dissimilaridade resultante deve ter seus componentes próximos de zero porque evidenciam atributos semelhantes. Em contrapartida, se forem de amostras de classes distintas, deve ter seus componentes longe de zero, pois classes diferentes têm características diferentes (Bertolini et al., 2013).

Outro ponto importante citado por Pekalska e Duin (2000), é que até mesmo classes que não foram utilizadas no conjunto de treinamento podem fazer parte do sistema de

identificação. Isso quer dizer que a transformação para um problema de duas classes possibilita a inserção de novas espécies no esquema de classificação sem a necessidade de treinar o modelo novamente, pois o modelo criado é treinado para distinguir se duas amostras pertencem à uma mesma classe ou não, independentemente de qual seja a classe. Neste trabalho, alguns experimentos exploram essa característica utilizando conjuntos disjuntos de espécies para treinamento e teste.

3.4 Regras de Combinação

Quando há mais de um classificador para classificar uma única amostra, pode-se utilizar algumas técnicas para combinar as saídas de tais classificadores e chegar a uma decisão final. A combinação de classificadores pode ser realizada quando as saídas dos classificadores (predições) apresentam, para cada amostra, uma estimativa de probabilidade para cada classe existente no esquema de classificação. As regras propostas por Kittler et al. (1998) podem ser usadas para combinar as predições dos classificadores para gerar uma decisão final.

Regra do produto

De acordo com Kittler et al. (1998), esta regra calcula o produto dos valores das predições de cada classe presente nos classificadores, e escolhe a classe com maior valor. O resultado da combinação é dado pela equação 3.7:

$$pr(x) = arg\,max_{k=1}^{c} \prod_{i=1}^{n} P(\omega_k|y_i(x)) \tag{3.7}$$

Em que x é a amostra a ser classificada, n é a quantidade de classificadores selecionados para a combinação, y_i o rótulo de saída do i-ésimo classificador em um problema com as classes $\Omega = \omega_1, \omega_2, ..., \omega_c$ e $P(\omega_k|y_i(x))$ a probabilidade de que a amostra x seja da classe ω_k encontrada pelo i-ésimo classificador.

A regra do produto pode ser considerada a mais severa, pois se um dos classificadores apresentar uma probabilidade com valor baixo para alguma classe, essa classe pode ter um valor resultante final baixo. Esta regra é indicada para casos em que não há tolerância para erro dos classificadores (Kittler et al., 1998).

Regra da soma

Esta regra calcula o somatório dos valores das predições encontradas para cada classe em todos os classificadores. É dada pela equação 3.8 (Kittler et al., 1998):

$$sr(x) = arg\, max_{k=1}^{c} \sum_{i=1}^{n} P(\omega_k|y_i(x)) \tag{3.8}$$

Onde x é a amostra que será classificada, n é a quantidade de classificadores relacionados na combinação, y_i o rótulo de saída do i-ésimo classificador em um problema com as classes $\Omega = \omega_1, \omega_2, ..., \omega_c$ e $P(\omega_k|y_i(x))$ a probabilidade de que a amostra x esteja vinculada à classe ω_k encontrada pelo classificador i.

Regra do máximo

A regra do máximo utiliza a maior probabilidade dentre as classes entre todos os classificadores. Ou seja, todos os classificadores são verificados, mas apenas um deles é utilizado na decisão final. É dada pela equação 3.9 (Kittler et al., 1998):

$$max(x) = arg\, max_{k=1}^{c}\, max_{i=1}^{n} P(\omega_k|y_i(x)) \tag{3.9}$$

Na qual x é a amostra a ser classificada, n é a quantidade de classificadores selecionados para a combinação, y_i o rótulo de saída do i-ésimo classificador em um problema com as classes $\Omega = \omega_1, \omega_2, ..., \omega_c$ e $P(\omega_k|y_i(x))$ a probabilidade de que a amostra x esteja vinculada à classe ω_k encontrada pelo classificador i.

Esta regra tende a ser menos severa porque utiliza apenas a predição daquele classificador com a maior probabilidade para as classes. Assim, basta que um dos classificadores seja bom para ter um alto valor de probabilidade.

4

Metodologia Proposta

Esta seção apresenta a metodologia seguida para realização deste trabalho. A Figura 4.1 ilustra, de forma geral, as etapas existentes no esquema de identificação de espécies de pássaros utilizando espectrogramas e dissimilaridade. A subseção 4.1 especifica os subconjuntos de dados utilizados neste trabalho. A subseção 4.2 descreve em detalhes as etapas que fazem parte do pré-processamento. A extração de característica é feita pelos descritores de textura apresentados na subseção 4.3. A dissimilaridade é uma etapa proposta que está localizada entre a extração de características e a classificação, pois manipula vetores de características antes de serem classificados. A classificação é realizada com SVM utilizando *kernel* RBF e parâmetros C e $Gamma$ otimizados por *grid-search* em alguns experimentos. O processo de combinação, *Late-Fusion*, consiste em utilizar as regras de combinação para combinar as predições geradas na etapa de classificação.

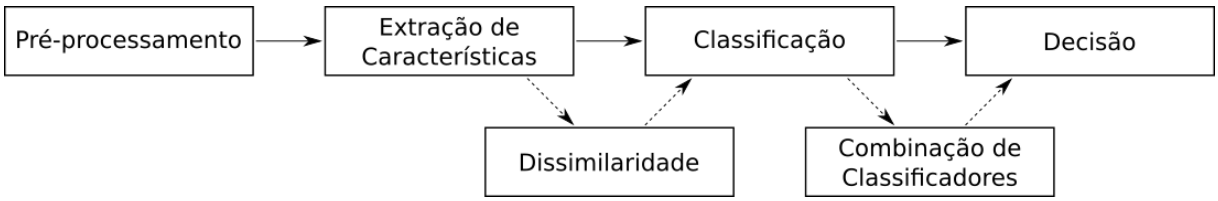


Figura 4.1: Etapas que compõe a classificação das espécies de pássaros.

4.1 Base de Dados

Devido a grande variedade no tempo de duração e na quantidade das amostras disponíveis para cada espécie na base proposta no *LifeClef* 2015 *Bird Task*, vários subconjuntos foram definidos empiricamente conforme a Tabela 4.1:

Tabela 4.1: Subconjuntos de dados propostos neste trabalho.

Subconjunto	Duração mínima (s)	Amostras por espécie	Espécies encontradas	Total de amostras
#1	30	10	23	230
#2	20	10	48	480
#3	15	10	88	880
#4	10	10	180	1800
#5	05	10	349	3490
#6	05	06	614	3684
#7	05	04	772	3088
#8	05	02	915	1830

Para criar o subconjunto #1, realizamos uma busca na base do *LifeClef* 2015 *Bird Task* filtrando somente vocalizações do tipo "canto" para encontrar espécies que continham, pelo menos, 10 amostras de áudio com no mínimo 30 segundos de duração. No total, 23 espécies foram encontradas. Em seguida, buscamos espécies com pelo menos 10 amostras de áudio que tinham duração de 20 segundos ou mais, criando o subconjunto #2 com 48 espécies. O mesmo processo foi realizado repetidamente com os requisitos de duração mínima das amostras de áudio e a quantidade das mesmas por espécie para criar os oito subconjuntos, apresentados na Tabela 4.1.

Com os critérios apresentados na Tabela 4.1, as classes que possuíam somente uma amostra foram descartadas, não sendo possível utilizar as 988 classes da base de cantos.

Com esse processo de criação de subconjuntos, as espécies presentes no #1 também compõe o #2, pois o fato de ter 20 segundos ou mais também inclui os casos que têm 30 segundos ou mais. Porém, as amostras não são necessariamente as mesmas, pois houve uma seleção aleatória de amostras. Deste modo, as espécies do #2 estão presentes no #3, e assim por diante conforme a Tabela 4.2.

Estes subconjuntos de dados foram divididos igualmente em *folds*, cada um contendo uma única amostra por espécie e a seleção da amostra ocorreu de forma aleatória. Assim,

Tabela 4.2: Composição de cada subconjunto em relação a espécies.

Subconjunto	Composição das espécies	Total de espécies
#1	23	23
#2	espécies de #1 + 25	48
#3	espécies de #2 + 40	88
#4	espécies de #3 + 92	180
#5	espécies de #4 + 169	349
#6	espécies de #5 + 265	614
#7	espécies de #6 + 158	772
#8	espécies de #7 + 143	915

os conjuntos de treinamento e teste ficaram balanceados e distintos, evitando que o modelo treinado tivesse mais habilidade para classificar algumas espécies do que outras.

4.2 Pré-processamento

Assim como foi dito na Introdução e na subseção 3.1, a base de dados é formada, em sua grande maioria, por sinais de áudio captados diretamente da natureza e enviados por pessoas ao redor do mundo e, por isso, algumas amostras não possuem somente o canto do pássaro. Há também o intervalo entre cantos de um mesmo pássaro e o som emitido por outras espécies de pássaros, animais, rios, pessoas, carros e outras fontes. Depois de obter a base de dados, o pré-processamento tem algumas etapas que ocorrem de maneira sequencial, como pode ser visto na Figura 4.2. A ideia dessa sequência é:

1. Primeiramente, é necessário realizar a coleta de uma amostra do sinal original completo a fim de identificar um perfil do ruído presente no mesmo. Uma vez definido o perfil do ruído do sinal, ocorre a redução do ruído do sinal como um todo;

2. Em seguida, acontece a remoção de partes do sinal que não tem canto de pássaro. O objetivo é detectar segmentos de interesse para que seja possível utilizar a parte do sinal mais representativa, descartando o restante. Assim, pode acontecer uma redução no tamanho de cada sinal de áudio;

3. A criação dos subconjuntos propostos neste trabalho é realizada depois das duas etapas anteriores, pois foram utilizados dois requisitos para a seleção de amostras: a duração mínima dos sinais de áudio e a quantidade de amostras por espécies;

4. Por fim, é gerado um espectrograma para cada amostra de áudio, que será dividido em zonas para a extração de informações locais em cada região (banda de frequência).

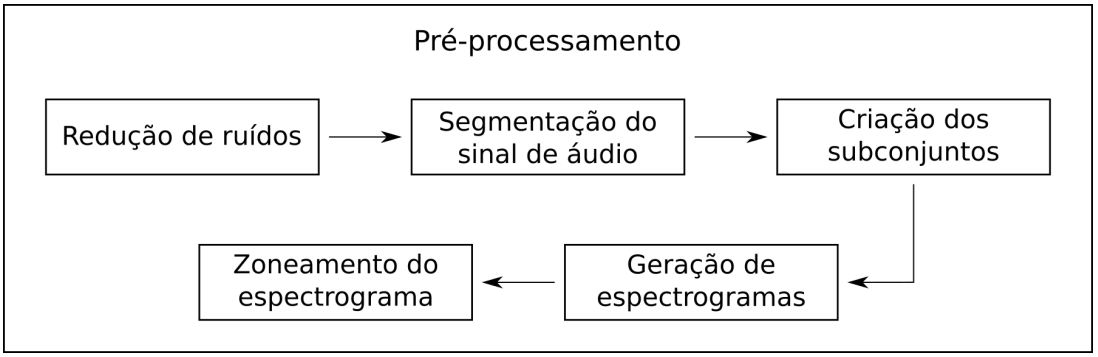


Figura 4.2: Etapas sequenciais que fazem parte do pré-processamento.

Espectrogramas foram utilizados com sucesso em várias tarefas de classificação de áudio com gêneros musicais (Costa et al., 2011), espécies de pássaros (Lucio e Costa, 2015) e língua falada (Montalvo et al., 2015). Por isso, essa abordagem foi escolhida para a representação de características neste trabalho.

O zoneamento do espectrograma é realizada por meio de zoneamentos verticais para viabilizar o uso da abordagem da dissimilaridade, e horizontais para extrair características específicas de diferentes regiões do espectrograma.

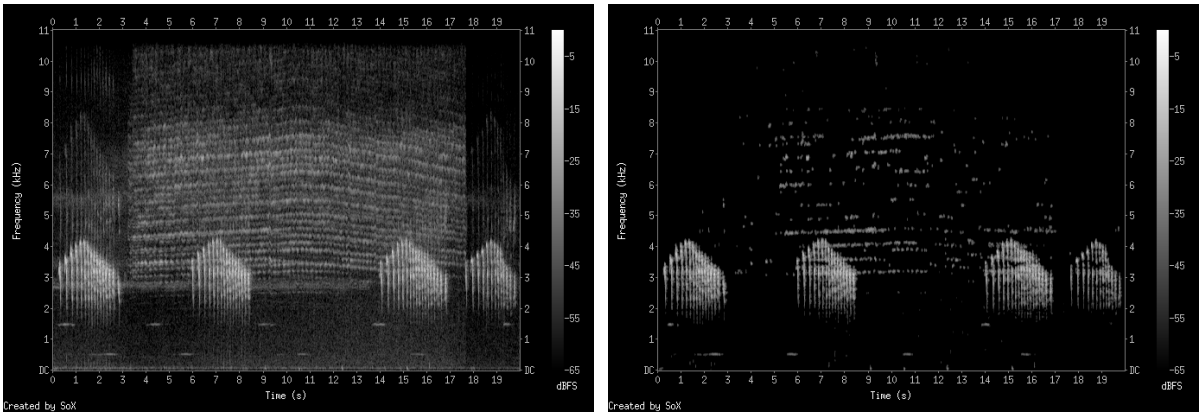
4.2.1 Redução de Ruídos

Ao se trabalhar com bases originadas do Xeno-canto, é comum notar que os sinais de áudio disponibilizados não têm um padrão de gravação para o ambiente em que são gravadas ou para os dispositivos utilizados. Há gravações de áudio tanto em regiões pouco ou não habitadas quanto em cidades ou perto da civilização. Dessa forma, nota-se junto com o canto dos pássaros outras fontes de sons, como: ventos, cachoeiras, riachos, sobreposição do áudio com outros animais ou insetos, carros, pessoas e outras mais.

Para minimizar a presença de ruídos e destacar o som dos pássaros, foi utilizada neste trabalho uma estratégia para a redução de ruídos presentes nos sinais de áudio, igual à Zottesso et al. (2016). Primeiramente, é realizada a coleta de uma amostra do sinal a fim de identificar um perfil do ruído presente no mesmo. Essa amostragem é feita com base nos primeiros 400 milissegundos do sinal de áudio (tamanho estabelecido empiricamente). Uma vez definido o perfil do ruído do sinal, ocorre a redução do ruído do sinal como um todo. Essa redução é baseada na subtração do perfil ruidoso identificado do sinal original.

Para realizar esta etapa, utilizamos a ferramenta de remoção de ruídos disponibilizada junto ao software *Sound eXchange* (SOX) versão 14.4.1.

A Figura 4.3 ilustra os espectrogramas gerados de um mesmo sinal de áudio antes e depois do processo de redução de ruídos com o SOX. Visualmente, o espectrograma gerado depois de aplicar o processo do SOX está menos poluído com ruídos.



(a) (b)

Figura 4.3: A Figura (a) representa um espectrograma de um sinal de áudio antes da redução de ruídos. A Figura (b) representa um espectrograma do mesmo sinal de áudio depois do processamento de redução de ruídos com o SOX.

Neste trabalho, todos os subconjuntos definidos na subseção anterior passaram por uma etapa de redução de ruídos.

4.2.2 Segmentação do Sinal de Áudio

Além da presença de ruídos, grande parte dos sinais disponibilizados possuem trechos em que não há o som dos pássaros. Sendo assim, a aplicação de um método para detecção dos segmentos de interesses se torna de extrema importância para a obtenção de melhores resultados, pois, de acordo com Evangelista et al. (2014), para obter melhores resultados na etapa de classificação, é necessário utilizar a parte do sinal de áudio mais representativa.

A fim de extrair estes segmentos considerados mais importantes, a técnica de segmentação proposta por Zottesso et al. (2016) foi aplicada em todas as amostras dos subconjuntos utilizados neste trabalho. De acordo com os autores, o processo consiste, basicamente, em:

- Extrair duas sequências com características do sinal de áudio, uma com base na Energia do Sinal (*Signal Energy*) e outra no Centroide Espectral (*Spectral Centroid*);

- Para cada sequência, dois limiares são estimados dinamicamente utilizando o histograma dos valores da sequência e os máximos locais;
- Um critério de limiar é aplicado nas sequências para separar os segmentos que possuem som dos segmentos com pouco ou nenhum som;
- Juntar os segmentos identificados no passo anterior.

A Figura 4.4 apresenta o exemplo de um sinal de áudio antes e depois do processo de segmentação.

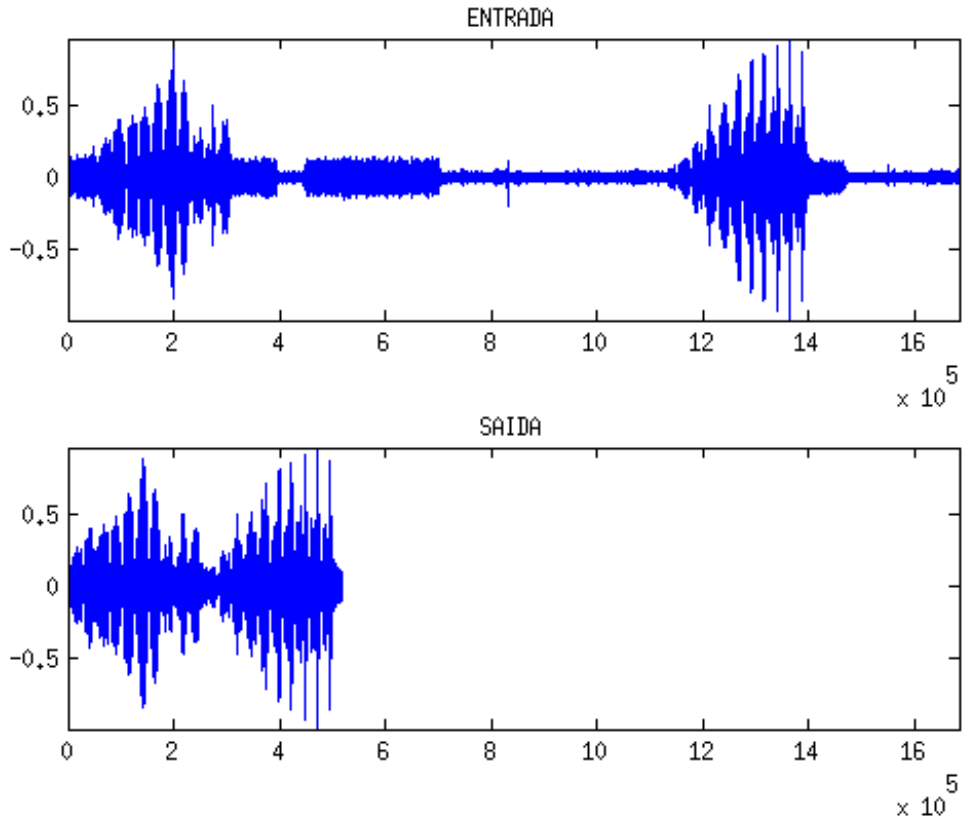


Figura 4.4: Sinal de áudio antes (entrada) e depois (saída) da segmentação automática.

Neste trabalho, ao referenciar que uma base da dados é "segmentada", significa dizer que as amostras de áudio que fazem parte dela passaram por este procedimento de segmentação.

4.2.3 Geração de Espectrogramas

Para poder representar um sinal de áudio no domínio das imagens, uma metodologia semelhante à utilizada em Costa et al. (2011), Lucio e Costa (2015) e Zottesso et al. (2016) foi aplicada. Ela consiste em gerar espectrogramas a partir dos sinais de áudio usando o software *Sound eXchange* (SOX), versão 14.4.1. Esses espectrogramas gerados representam o tempo no eixo horizontal, a frequência no eixo vertical e a amplitude do sinal na intensidade das cores dos pixels.

Nesta etapa, é possível variar uma série de parâmetros que influenciam diretamente na textura das imagens e no resultado final da classificação, como: altura e largura da imagem, densidade de pixels por segundo, limite de frequência até onde há representação do sinal no espectrograma gerado, amplitude do sinal, entre outros.

Para este trabalho, alguns parâmetros foram estabelecidos conforme Costa et al. (2011), Lucio e Costa (2015) e Zottesso et al. (2016), exceto a amplitude do sinal. Ela foi definida empiricamente com base nas taxas de reconhecimento dos experimentos iniciais deste trabalho, pois pode variar muito conforme a origem e o tipo dos sinais de áudio.

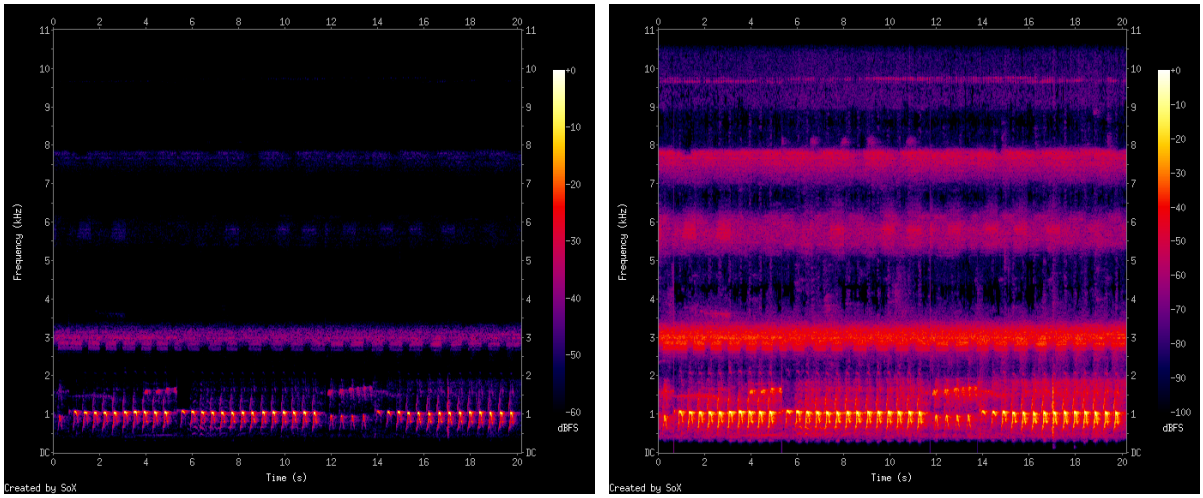
A amplitude é medida em `dBFS` e o seu limite inferior influencia diretamente na textura porque altera o contraste da imagem. Quanto menor o limite inferior, mais informação é representada no espectrograma, diminuindo o contraste entre as informações representadas na imagem. A Figura 4.5 ilustra dois espectrogramas com limites inferiores de amplitude diferentes. O primeiro foi gerado utilizando -60 `dBFS` (Figura 4.5(a)) e o segundo, -100 `dBFS` (Figura 4.5(b)).

Os dois exemplos de espectrogramas apresentados na Figura 4.5 foram gerados com representação no espaço RGB. Porém, para se adequarem corretamente às etapas e técnicas presentes neste trabalho, todos os espectrogramas foram automaticamente gerados em escala de cinza e sem rótulos nos eixos, conforme demonstrado pela Figura 4.6.

É importante ressaltar que as técnicas de processamento de imagens aqui utilizadas para a representação da textura operam eficientemente sobre imagens em níveis de cinza. Além disso, a representação da imagem em escala de cinza preserva a intensidade de energia do sinal representada no espectrograma e não compromete a tarefa de classificação aqui investigada.

4.2.4 Zoneamento do Espectrograma

Durante a realização dos experimentos desenvolvidos neste trabalho, observou-se que a textura presente nos espectrogramas dos cantos de pássaros não apresenta um conteúdo



(a) (b)

Figura 4.5: Espectrogramas de uma mesma amostra da base gerado com diferentes valores para o limite inferior de amplitude. Na Figura (a) o limite inferior da amplitude igual a -60 `dBFS`. Na Figura (b) o limite inferior da amplitude igual a -100 `dBFS`.

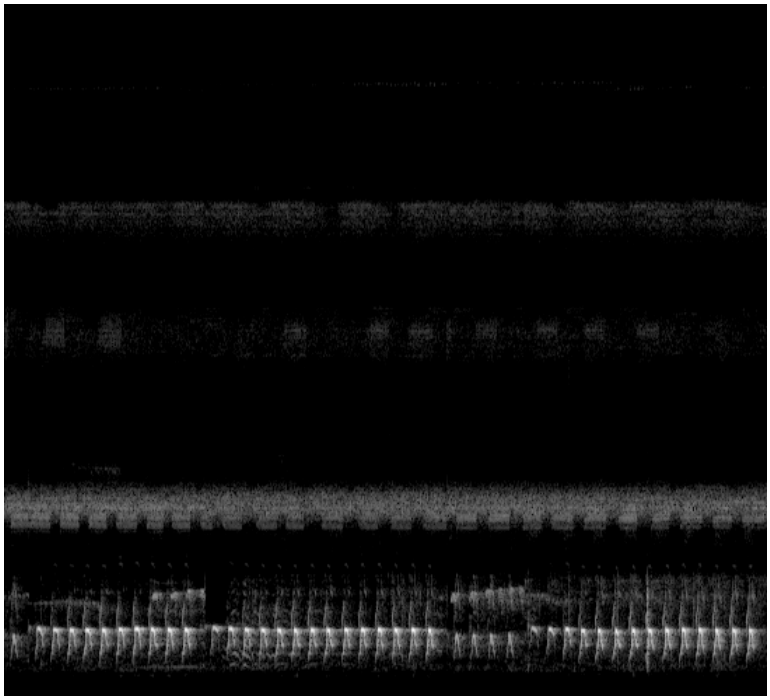


Figura 4.6: Espectrograma original extraído pelo SOX com amplitude de sinal -60 `dBFS` pronto para utilização.

padrão. Assim, foi proposta uma estratégia para dividir os espectrogramas em zonas para que fosse possível destacar informações em regiões específicas do espectrograma.

A ideia do zoneamento da imagem consiste em extrair informações locais de cada região e tentar destacar as especifidades das diferentes faixas de frequência em cada amostra (Costa et al., 2011). Isso quer dizer que cada região criada pelo zoneamento linear tem potencial para dar origem à um vetor de características específico e, consequentemente, alimentar um classificador específico associado àquela região (Costa et al., 2012a). Os classificadores criados para cada zona podem ser combinados com base em algumas regras de fusão como às propostas por Kittler et al. (1998).

Dois tipos de zoneamento foram abordados neste trabalho: vertical e horizontal. O zoneamento vertical tem como objetivo proporcionar a variação da quantidade de vetores de dissimilaridade positivos e negativos que podem ser gerados. O zoneamento horizontal possibilita extrair características de acordo com as especificidades de cada banda de frequência, gerando novos classificadores.

Zonas Verticais

Com a divisão vertical, são estabelecidas zonas de mesmo tamanho na imagem do espectrograma que correspondem a períodos de tempo com a mesma duração. O tamanho de cada zona depende da duração de cada sinal de áudio e da quantidade de zonas verticais estabelecidas.

A fim de viabilizar a criação de mais vetores descritores a partir de cada amostra, foram estabelecidas divisões verticais para o zoneamento dos espectrogramas. Essa estratégia foi utilizada para melhor viabilizar o uso da abordagem de dissimilaridade, que está descrita na subseção 4.4.

Baseando-se no número de referências utilizadas por Bertolini et al. (2013), os espectrogramas foram divididos em três, cinco ou nove partes verticais iguais possibilitando a criação de diferentes quantidades de vetores de dissimilaridade. A Figura 4.7 ilustra uma divisão em três zonas verticais.

Zonas Horizontais

O principal objetivo das zonas horizontais é possibilitar a extração de características específicas de regiões que são limitadas por bandas de frequência, gerando novos classificadores. Estas zonas podem ser definidas por regiões que têm sempre a mesma medida de altura e dividem a imagem em partes iguais em tamanho (zonas lineares) ou de acordo

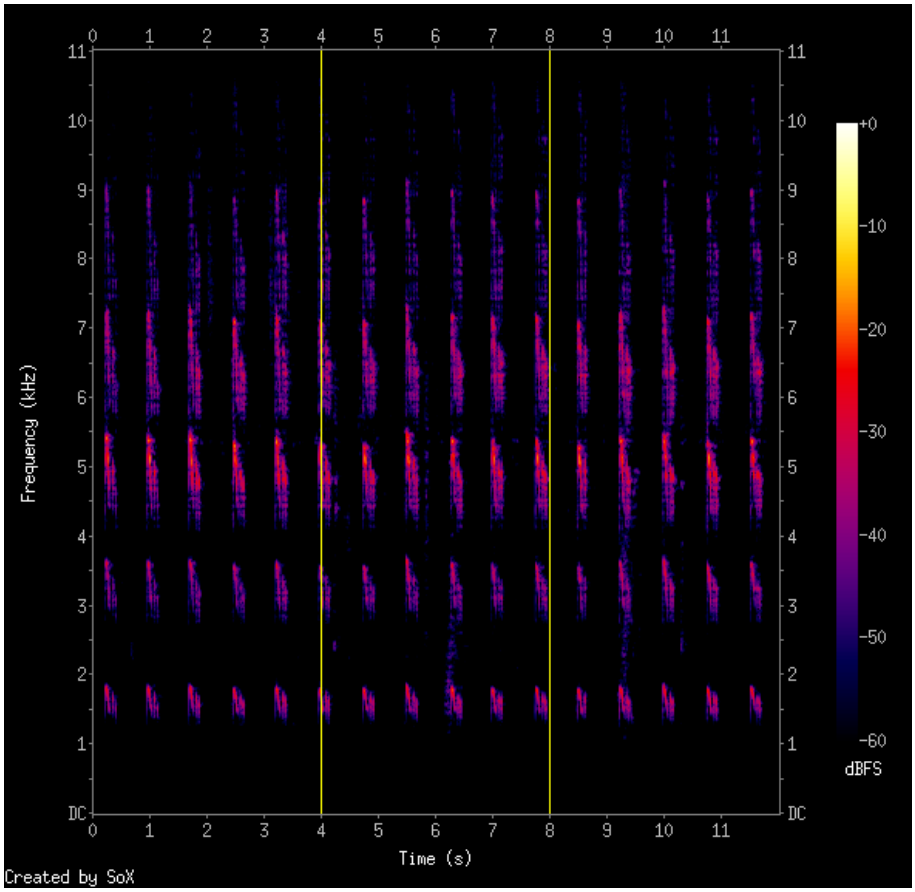


Figura 4.7: Espectrograma dividido em três zonas verticais.

com bandas de frequência pré-definidas que geram regiões dinamicamente (por exemplo, Escala Mel).

Zonas Lineares

As zonas lineares dividem a imagem em regiões de tamanhos iguais que têm como limite de altura uma frequência. Os limites de frequência dependem da quantidade de zonas que são criadas. A Figura 4.8 apresenta o zoneamento do espectrograma em três zonas lineares.

Algumas quantidades de zonas lineares foram definidas empiricamente neste trabalho para buscar melhores resultados, são elas: um, três, cinco e dez.

Escala Mel

Costa et al. (2012a) utilizaram espectrogramas para classificar gêneros musicais e o zoneamento seguindo a escala Mel apresentou um bom desempenho em relação a zoneamentos

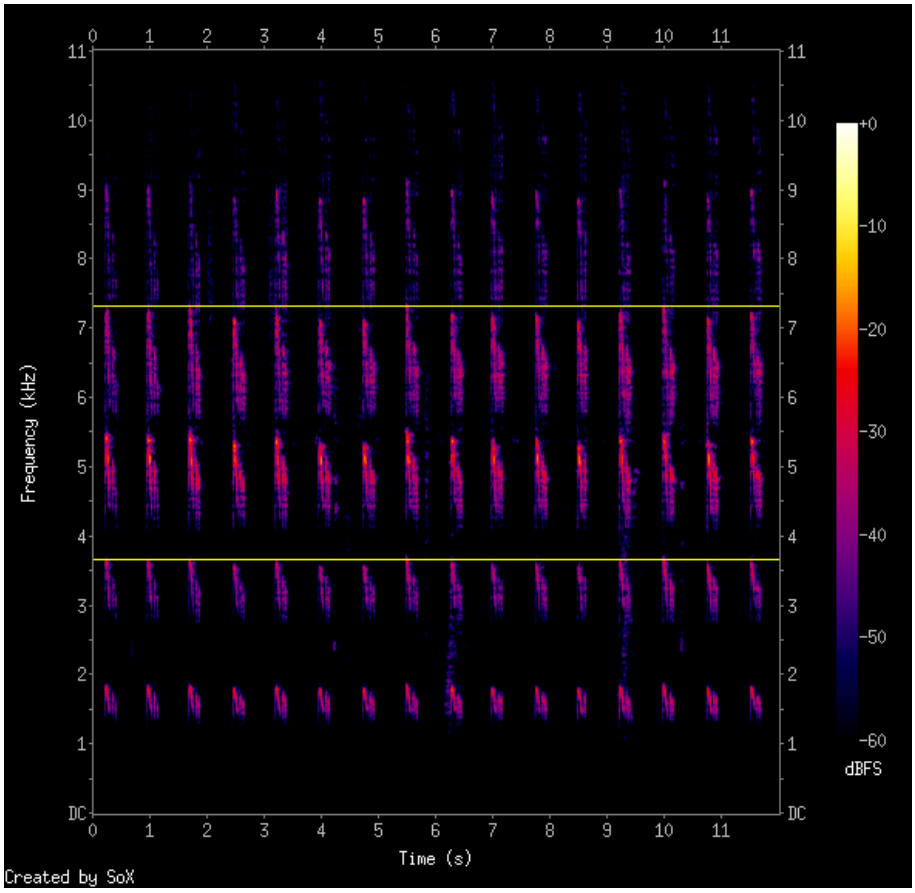


Figura 4.8: Espectrograma dividido em três zonas lineares.

lineares. Assim, alguns experimentos seguindo essa escala também foram realizados neste trabalho.

Na escala Mel, as divisões representam bandas de frequência que estão diretamente relacionadas com as frequências percebidas por humanos. Ao todo são 15 bandas (regiões) e cada uma delas tem seus limites, que em Hz são: 40, 161, 200, 404, 693, 867, 1.000, 2.022, 3.000, 3.393, 4.109, 5.526, 6.500, 7.743 e 12.000 (Umesh et al., 1999). O maior limite no zoneamento da imagem depende do limite de frequência definido na geração do espectrograma a partir do sinal de áudio. A Figura 4.9 exemplifica um espectrograma com limite de frequência de 11.000Hz e a criação de regiões conforme a divisão pela escala Mel.

4.3 Extração de Características

Considerando a textura como o principal atributo visual em imagens de espectrogramas, neste trabalho foram utilizados alguns descritores de textura apresentados na literatura

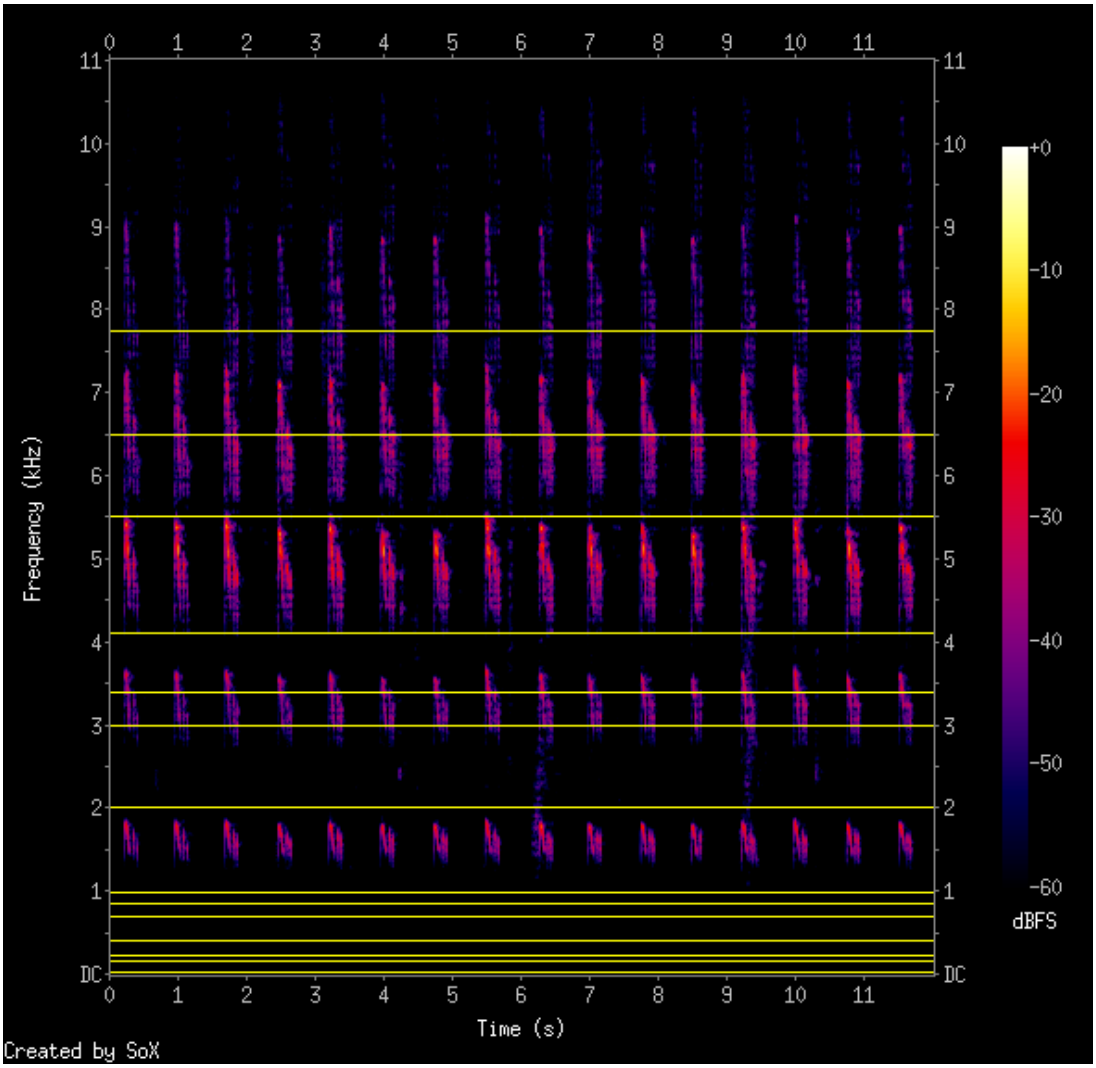


Figura 4.9: Espectrograma dividido proporcionalmente conforme a escala Mel.

para descrever o conteúdo dos espectrogramas para posteriormente utilizá-los por classificadores. A Tabela 4.3 mostra os descritores utilizados neste trabalho e a dimensão do vetor de características para cada descritor de textura.

Tabela 4.3: Dados sobre os descritores de textura utilizados

Descritor	Tamanho do vetor de características
$LBP_{8,2}$	59
$RLBP_{8,2}$	59
LPQ com janela 3×3	256

Assim como no trabalho de Costa (2013), foi utilizado um mecanismo de normalização segundo o qual os dados são mapeados para o intervalo $[-1, 1]$. O valor normalizado para uma característica x é encontrado conforme descrito na equação 4.1.

$$x_{norm} = \frac{2(x - m_i)}{(M_i - m_i) - 1} \tag{4.1}$$

Na qual x é o valor da característica antes da normalização, M_i é o maior valor de característica encontrado no conjunto de dados e m_i é o menor valor de característica encontrado no conjunto de dados.

4.4 Dissimilaridade

Neste trabalho, utilizou-se o conceito de dissimilaridade introduzido na subseção 3.3. A Figura 4.10 representa o esquema geral usado na abordagem da dissimilaridade para a geração de vetores positivos e negativos.

Inicialmente, os sinais de áudio com cantos de pássaros, depois da redução de ruídos e segmentação, passam pela geração dos espectrogramas e zoneamento. Os vetores de características extraídos das mesmas zonas são utilizados para computar os vetores de dissimilaridade positivos quando são da mesma espécie, e negativos quando vem de espécies diferentes. Em seguida, esses vetores de dissimilaridade são enviados para um classificador *Support Vector Machine* (SVM) para geração do modelo e posteriormente para classificação. Por fim, algumas regras de combinação são aplicadas para a decisão final da classificação.

Nesta abordagem, o número de vetores de dissimilaridade depende totalmente da quantidade das zonas verticais e horizontais estabelecidas, pois cada vetor de dissimilaridade é computado a partir de um par de vetores de características.

A Figura 4.11 exemplifica o cálculo de vetores de dissimilaridade positivos e negativos em um espectrograma sem zoneamento. Considerando que V_1 e V_2 são dois vetores de características e $V_{1,2}$ é um vetor de dissimilaridade, o vetor resultante dado por $V_{1,2} = |V_1 - V_2|$ será um vetor positivo se V_1 e V_2 forem vetores de características extraídos de amostras da mesma classe. Por outro lado, se V_1 e V_2 forem vetores de características de classes diferentes, o vetor resultante $V_{1,2}$ será um vetor negativo.

Com a proposta de zoneamento do espectrograma (subseção 4.2.4), cada região criada na imagem é utilizada para extração um vetor de características que, posteriormente, será utilizado para computar os vetores de dissimilaridade, seguindo a ideia apresentada

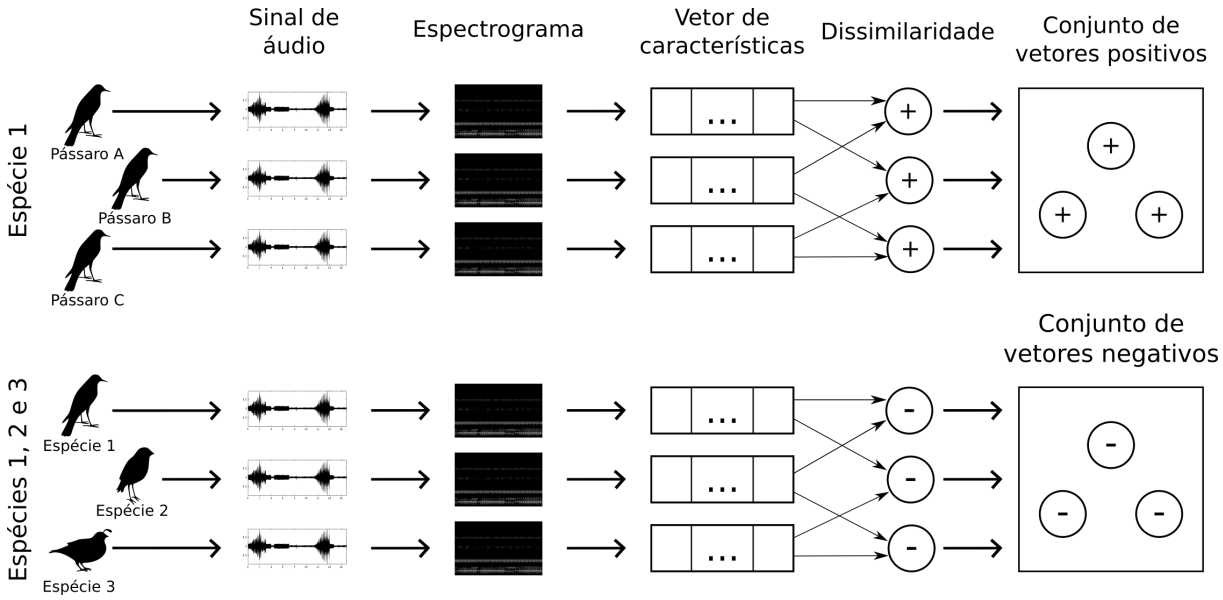


Figura 4.10: Metodologia para geração de vetores positivos e negativos na abordagem da dissimilaridade.

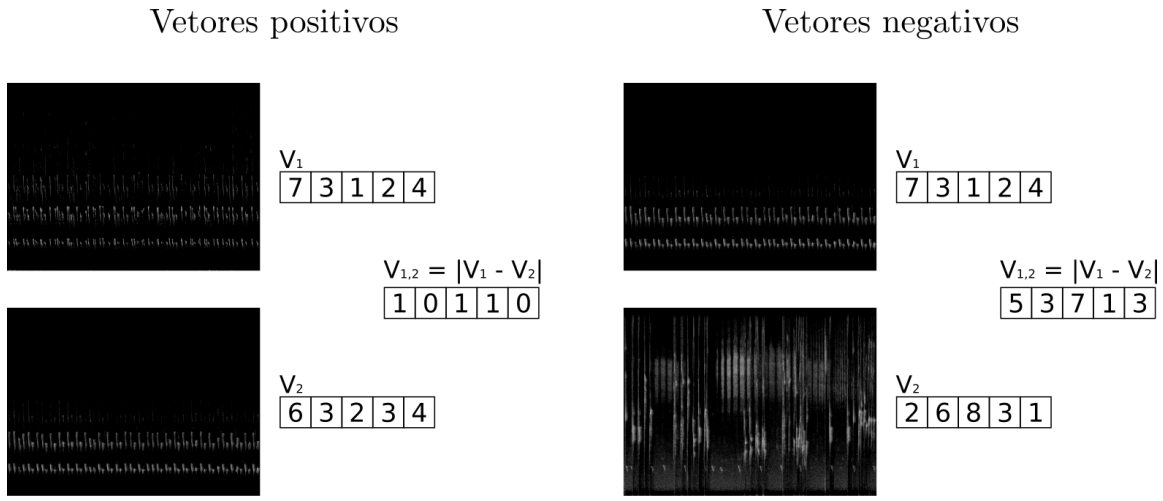


Figura 4.11: Cálculo dos vetores de dissimilaridade positivos e negativos.

na Figura 4.11, desde que ambos os vetores de características sejam extraídos da mesma banda de frequência.

Para detalhar o procedimento utilizado na geração de vetores de dissimilaridade, a Figura 4.12, que ilustra um exemplo de espectrograma com três zonas verticais e três zonas horizontais lineares, será utilizada como referência de mapeamento para explicar como os vetores positivos e negativos foram computados nas etapas de treinamento e de

teste. As regiões horizontais foram rotuladas em A, B e C, enquanto as verticais em *1*, *2* e *3*.

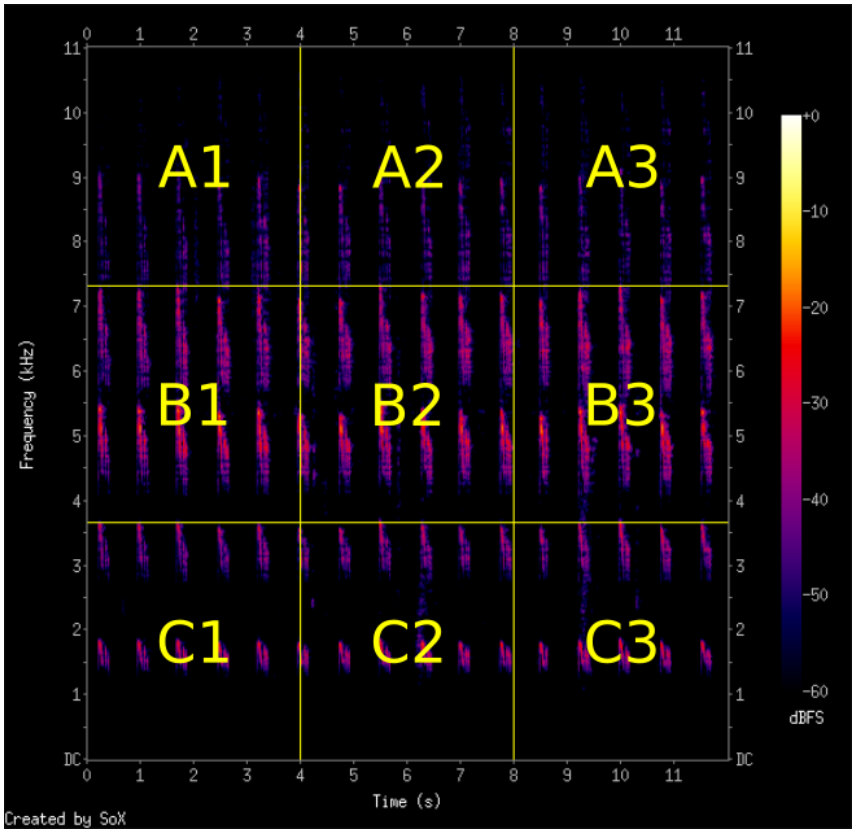


Figura 4.12: Esquema de zoneamento de um espectrograma com três zonas verticais e três zonas lineares.

Na etapa de treinamento, todos os vetores de dissimilaridade são gerados a partir de um par de vetor de características extraídas de regiões referentes à mesma banda de frequência, de modo que os vetores positivos são computados a partir de características de espectrogramas de pássaros da mesma espécie, enquanto os negativos de espécies diferentes. Cada vetor de característica é comparado aleatoriamente com vetores da mesma espécie e de espécies diferentes, de modo que a quantidade de vetores positivos e negativos é sempre igual.

Considerando a Figura 4.12, um vetor positivo V_p é dado por $V_p = |X_{A1} - Y_{A1}|$, na qual X e Y representam dois vetores de características de amostras distintas da mesma espécie extraídos da região $A1$. Esse processo se repete para todas as regiões definidas com o zoneamento vertical e horizontal. Os vetores negativos são computados do mesmo modo, porém X e Y devem ser dois vetores de características de amostras de espécies diferentes.

Já na etapa de teste, as amostras de cada subconjunto são divididas em *folds*, sendo que cada *fold* possui somente uma amostra de cada espécie. Por exemplo: um subconjunto com n amostras por espécie para teste tem n *folds*. Assim, não é possível adotar a mesma metodologia que a utilizada na etapa de treinamento para geração de vetores de dissimilaridade. A taxa de reconhecimento final é calculada pela média das taxas de reconhecimento obtidas na classificação das amostras de cada *fold*.

No processo de identificação, uma amostra X é comparada com ela mesma, por meio das zonas verticais diferentes presentes na mesma banda de frequência, e com todas as outras amostras de espécies distintas daquele *fold*. O esquema de classificação considera como classe da amostra X a mesma classe daquela amostra utilizada para computar os vetores de dissimilaridade que possuem maior probabilidade de serem positivos.

4.5 Classificação

Para realizar as tarefas de classificação, o *Support Vector Machine* (SVM), introduzido por (Vapnik, 1995), foi escolhido por ser amplamente utilizado e bem sucedido em diferentes domínios de aplicações: reconhecimento de gêneros musicais (Costa et al., 2011), identificação de escritores (Bertolini et al., 2013) e classificação de espécies de pássaros (Lucio e Costa, 2015). Além disso, o SVM foi proposto para problemas binários, sendo interessante para dissimilaridade. É um classificador robusto em tarefas de classificação e reconhecimento de padrões e tem apresentado resultados competitivos.

A implementação do SVM utilizada neste trabalho é a *Library for Support Vector Machines* (LIBSVM[1]), que foi proposta por Chang e Lin (2011) e está disponível gratuitamente.

Nos maiores subconjuntos de dados propostos neste trabalho, o tempo para execução do LIBSVM com *grid-search* inviabilizaria a realização de alguns experimentos. Assim, aplicou-se algumas rodadas de aprendizagem e classificação em subconjuntos menores para obter bons valores nos parâmetros C e *Gamma* (γ), a fim de se utilizar como padrão para a fluência do trabalho. De maneira geral, o *kernel Radial Basis Function* (RBF) foi amplamente utilizado.

Os resultados experimentais são apresentados com as configurações do LIBSVM que foram utilizadas para computar as taxas encontradas.

É importante ressaltar que, antes de computar os vetores de dissimilaridade, os dados foram devidamente normalizados, conforme descrito na seção 4.3.

[1] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`

4.6 Combinação de Classificadores

As regras de combinação citadas na subseção 3.4 foram usadas para combinar as saídas de classificadores das zonas correspondentes às diferentes faixas de frequência. Os resultados obtidos com as regras do mínimo e mediana estiveram sempre abaixo dos resultados das regras do máximo, produto e soma, e, por isso, não estão presentes nos resultados experimentais deste trabalho. As regras selecionadas também foram superiores nos trabalhos (Costa et al., 2013, 2011, 2012b; Martins et al., 2011; Nanni et al., 2016).

A Figura 4.13 ilustra o esquema da combinação de classificadores. Todas as zonas da imagem de um espectrograma são tratadas de maneira independente como um classificador até a geração das predições pelo LIBSVM, que depois são combinadas para se chegar a uma decisão final.

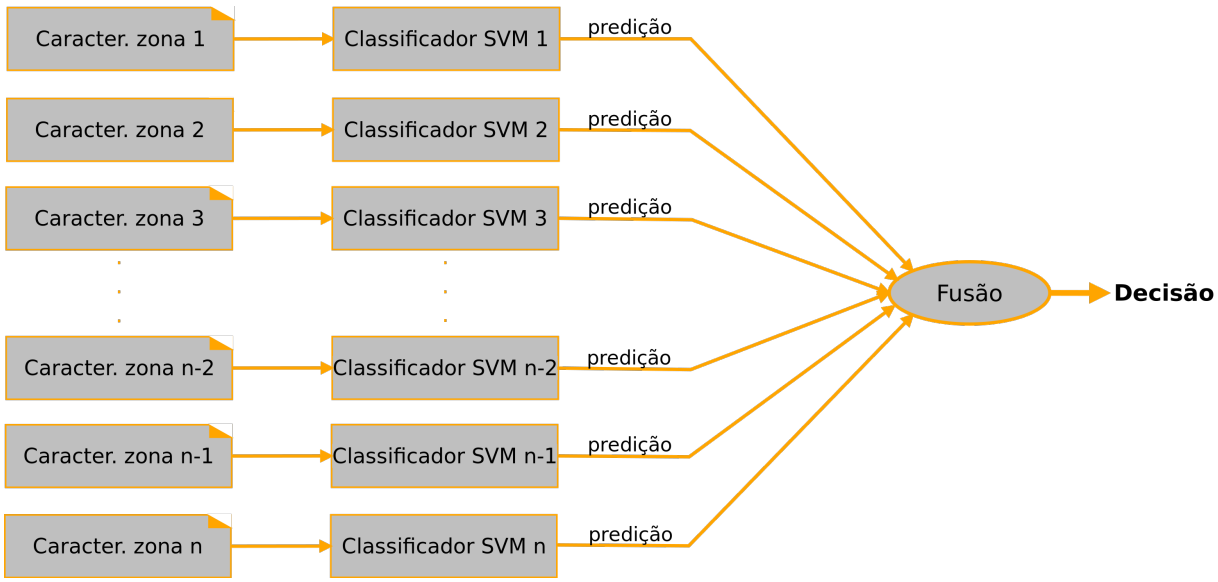


Figura 4.13: Criação e combinação de classificadores para as características extraídas de cada zona. Adaptada de (Costa et al., 2013).

Na abordagem de dissimilaridade proposta neste trabalho, para decidir a classe de uma amostra é necessário que a combinação das probabilidades dos vetores positivos sejam maiores do que as dos vetores negativos gerados a partir daquela amostra.

4.7 Avaliação dos Resultados

Considerando que a identificação de espécies de pássaros abordada neste trabalho é um problema multi-classe e que todos os subconjuntos possuem a mesma quantidade de amostras para cada classe, tanto no conjunto de treino quanto no de teste.

A partir dela, a taxa de reconhecimento geral, ou acurácia, é a medida utilizada para verificar o desempenho dos classificadores. Ela é dada pela equação 4.2, na qual n corresponde ao número de classes envolvidas na classificação, c_i é o número de instâncias corretamente classificadas pertencentes à classe i, e t_i corresponde ao número total de instâncias pertencentes à classe i.

$$Taxa\ de\ reconhecimento\ geral = \frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} t_i} \tag{4.2}$$

5

Resultados Experimentais

Esta seção apresenta os resultados obtidos dos experimentos realizados de acordo com a proposta deste trabalho. A primeira subseção descreve os principais resultados alcançados ao explorar a abordagem de dissimilaridade. Em seguida, são listados os resultados da abordagem classe dependente, e, por fim, há uma discussão acerca dos principais resultados deste trabalho.

A Tabela 5.1 detalha a utilização dos subconjuntos neste trabalho. Em todos os experimentos, independente do subconjunto ou da abordagem avaliada, metade das amostras foram utilizadas para treinamento e a outra metade para teste. Na etapa de teste, as amostras foram divididas em *folds* (uma por espécie), e as taxas de reconhecimento apresentadas nas próximas subseções foram calculadas a partir da média aritmética das taxas de reconhecimento de cada *fold*.

Mais detalhes sobre os oito subconjuntos propostos neste trabalho estão presentes na subseção 4.1.

5.1 Abordagem de Dissimilaridade

Primeiramente, são apresentados os resultados obtidos ao avaliar uma série de parâmetros para otimizar o esquema de identificação de espécies de pássaros na abordagem de dissimilaridade. Em seguida, listamos os resultados alcançados ao aplicar a abordagem de dissimilaridade em todos os subconjuntos propostos neste trabalho.

Para computar as taxas de reconhecimento "TOP N" utilizou-se sempre a regra da soma, pois ao analisar os resultados de outras regras, constatou-se que a regra da soma estava presente na maioria das combinações com taxas mais altas.

Tabela 5.1: Divisão dos subconjuntos para as etapas de treinamento e teste.

Subconjunto	Espécies	Amostras por espécie no treino	Amostras por espécie no teste	Total de amostras por espécie
#1	23	05	05	10
#2	48	05	05	10
#3	88	05	05	10
#4	180	05	05	10
#5	349	05	05	10
#6	614	03	03	06
#7	772	02	02	04
#8	915	01	01	02

5.1.1 Avaliação de Parâmetros do SVM, Descritores de Textura e Zoneamento do Espectrograma

Inicialmente, vários experimentos foram realizados para encontrar bons parâmetros de otimização no LIBSVM com a abordagem de dissimilaridade, explorando, também, as variações de zoneamento do espectrograma e descritores de textura. O objetivo dessa exploração é encontrar boas configurações de testes para maximizar as taxas de reconhecimento. Nos primeiros experimentos, os subconjuntos #1, #2 e #3 foram mais utilizados porque são menores e demandam menos tempo de processamento.

A Tabela 5.2 lista os resultados obtidos para os parâmetros C e *Gamma* do SVM configurado para executar com *grid-search* e *kernel* RBF. Nestes experimentos foram utilizadas três zonas verticais e várias configurações para zonas horizontais, todos com características extraídas do LBP. A ideia deste experimento é encontrar bons parâmetros para melhorar o tempo de execução do LIBSVM durante as tarefas de treino e teste. O valor 2 para *Gamma* foi encontrado na maioria dos testes, enquanto os valores de C tem uma maior ocorrência de 8 e 32.

Em seguida, alguns testes foram realizados para verificar o tempo de processamento[1] na etapa de aprendizagem variando algumas configurações do SVM, como: *grid-search*, C e *Gamma* fixos, *kernel* RBF e *kernel* Linear. Os experimentos foram realizados utilizando os principais valores de C e *Gamma* encontrados na Tabela 5.2, o descritor de textura LBP e um zoneamento de 10 regiões horizontais e três verticais.

[1]Tempo de processamento, no formato hh:mm:ss, em um computador com processador AMD A8-5500B 3.2 GHz, 8 GB RAM 1666 MHz e HD 500GB 7200 RPM.

Tabela 5.2: Valores de *C* e *Gamma* (G) encontrados por *grid-search* no LIBSVM.

	Sem zonas		**03 zonas**		**05 zonas**		**10 zonas**		**15 zonas (Mel)**	
Subconjunto	**C**	**G**	**C**	**G**	**C**	**G**	**C**	**G**	**C**	**G**
#1 (23)	8	2	8	0,5	8	2	512	2	128	2
#2 (48)	32	2	32	2	8	2	8	2	128	2
#3 (88)	8	2	8	2	8	2	8	2	32	2
#4 (180)	32	2	32	2	32	2	-	-	128	0,5
#5 (349)	8	2	32	2	-	-	-	-	-	-

A Tabela 5.3 lista os resultados obtidos com o LIBSVM executando com *kernel* RBF e *grid-search*, apresentando os valores encontrados para *C* e o *Gamma* e tempo de execução em cada subconjunto.

Tabela 5.3: Taxas de reconhecimento (%) obtidas com um *kernel* RBF e *grid-search* para *C* e *Gamma* (G).

Subconjunto	**Top 01**	**Top 05**	**Top 10**	**C**	**G**	**Tempo**
#1 (23)	0,817	0,974	1,000	512	2	00:45:58
#2 (48)	0,867	0,954	0,979	8	2	06:37:15
#3 (88)	0,902	0,966	0,977	8	2	31:01:22

Em seguida, a Tabela 5.4 apresenta os resultados obtidos com um *kernel* RBF e valores de *C* e *Gamma* fixos em 8 e 2, respectivamente, além do tempo de processamento para treinamento. Os resultados da Tabela 5.5 foram alcançados ao alterar o *kernel* para linear, mantendo a mesma configuração para *C* e *Gamma*.

Tabela 5.4: Taxas de reconhecimento (%) com *kernel* RBF, $C = 8$ e $Gamma = 2$.

Subconjunto	**Top 01**	**Top 05**	**Top 10**	**Tempo**
#1 (23)	0,904	0,991	1,000	00:00:16
#2 (48)	0,867	0,954	0,979	00:01:55
#3 (88)	0,902	0,966	0,977	00:07:37

Por fim, como o *kernel* RBF apresentou resultados melhores do que o *kernel* linear, fixou-se os valores de *C* em 32 e *Gamma* em 2. Os resultados estão listados na Tabela 5.6.

Tabela 5.5: Taxas de reconhecimento (%) com *kernel* linear, $C = 8$ e $Gamma = 2$.

Subconjunto	Top 01	Top 05	Top 10	Tempo
#1 (23)	0,861	0,974	1,000	00:00:23
#2 (48)	0,833	0,942	0,975	00:01:50
#3 (88)	0,843	0,952	0,964	00:05:30

Tabela 5.6: Taxas de reconhecimento (%) com *kernel* RBF, $C = 32$ e $Gamma = 2$.

Subconjunto	Top 01	Top 05	Top 10	Tempo
#1 (23)	0,896	0,983	1,000	00:00:19
#2 (48)	0,850	0,950	0,983	00:02:47
#3 (88)	0,882	0,966	0,977	00:12:38

Depois de comparar os experimentos realizados ao variar os valores de C, $Gamma$ e o tipo do *kernel*, pode-se observar que as melhores taxas de reconhecimento encontradas estão presentes na Tabela 5.4, que apresentou os resultados obtidos ao se utilizar os parâmetros $C = 8$ e $Gamma = 2$ fixos no LIBSVM com *kernel* RBF. Além disso, no treinamento o tempo de processamento com parâmetros fixos é muito menor em relação ao *grid-search*, principalmente quando o número de classes aumenta. Desta forma, tomamos como base esses parâmetros para os próximos experimentos.

Variar descritores de textura também podem auxiliar na busca por melhores resultados. Neste trabalho alguns experimentos foram realizados para comparar o LBP, RLBP e LPQ. A Tabela 5.7 lista os resultados obtidos para cada descritor. Neste experimento, utilizou-se o subconjunto #3 que possui 88 espécies e espectrogramas com três zonas verticais e escala Mel (15 horizontais), além de configurar o LIBSVM para executar com os valores de $C = 8$ e $Gamma = 2$ em um *kernel* RBF. Podemos observar que o LBP apresentou resultados levemente superiores em relação ao RLBP e LPQ. Além disso, o número de características extraídas pelo LBP é igual ao RLBP e menor em relação ao LPQ, implicando também em menor tempo de processamento. Assim, definiu-se o LBP como descritor de texturas padrão para a extração de características nos experimentos.

A variação de zonas verticais influencia na quantidade de vetores de dissimilaridade positivos e negativos que podem ser calculados. A Tabela 5.8 lista os resultados obtidos ao variar a quantidade de zonas verticais entre três, cinco e nove e as horizontais entre nenhuma ou escala Mel. Para isso, as características do subconjunto #3 (88 espécies) extraídas com o descritor de textura LBP foram classificadas com os parâmetros $C = 8$ e $Gamma = 2$ definidos no LIBSVM com *kernel* RBF. Os melhores resultados foram

Tabela 5.7: Taxas de reconhecimento (%) dos descritores de textura LBP, RLBP e LPQ no subconjunto #3.

Descritor	Top 01	Top 05	Top 10
LBP	0,914	0,982	0,986
RLBP	0,905	0,980	0,986
LPQ	0,911	0,977	0,982

obtidos ao ser utilizar três zonas verticais e os experimentos com cinco e nove zonas não apresentaram resultados satisfatórios.

Tabela 5.8: Taxas de reconhecimento (%) obtidas a partir da variação das zonas verticais e horizontais.

Verticais	Horizontais	Top 01	Top 05	Top 10
03	Nenhuma	0,570	0,857	0,914
03	15 (Mel)	0,914	0,982	0,986
05	Nenhuma	0,041	0,077	0,116
05	15 (Mel)	0,018	0,061	0,148
09	Nenhuma	0,032	0,068	0,127
09	15 (Mel)	0,025	0,059	0,157

As zonas horizontais tem como propósito destacar as especifidades de cada região de faixa de frequência e também a criação de novos classificadores. Na Tabela 5.8, além da variação das zonas verticais foram realizados testes com duas configurações de zonas horizontais (nenhuma e 15), e não houve uma delas que se sobressaiu nos resultados. Assim, se fez necessário testar quantidades diferentes para as zonas horizontais.

A Tabela 5.9 elenca os resultados obtidos ao se variar o número de zonas horizontais, mantendo sempre três zonas verticais. Foi utilizado como base o subconjunto #3 (88 espécies) com o descritor de textura LBP e o LIBSVM configurado com $C = 8$, $Gamma = 2$ e *kernel* RBF. Os melhores resultados foram obtidos utilizando a escala Mel, que cria bandas de frequência, de tamanhos diferentes, relacionadas com às percebidas por humanos. Tanto na classificação de gêneros musicais (Costa et al., 2012a) quanto na identificação de espécies de pássaros, a criação de zonas horizontais pode melhorar as taxas de reconhecimento.

Até este ponto, os experimentos foram realizados com a finalidade de encontrar bons parâmetros para configuração do esquema de classificação, considerando também o tempo de execução em subconjuntos maiores. Depois de realizar vários testes e analisar os

Tabela 5.9: Taxas de reconhecimento (%) obtidos variando as zonas horizontais.

Zonas	Top 01	Top 05	Top 10
Nenhuma	0,570	0,857	0,914
03	0,755	0,911	0,950
05	0,852	0,941	0,968
10	0,902	0,966	0,977
15 (Mel)	0,914	0,982	0,986

resultados listados nas Tabelas anteriores desta subseção, foi definido um conjunto da parâmetros: três zonas verticais, 15 zonas horizontais (escala Mel), descritor de textura LBP, $C = 8$, $Gamma = 2$ e *kernel* RBF.

5.1.2 Avaliação da Abordagem de Dissimilaridade em Diferentes Subconjuntos

Depois de definir um conjunto de parâmetros otimizados para a identificação de espécies de pássaros na abordagem de dissimilaridade, vamos avaliar o impacto do tamanho da amostra de áudio e da quantidade de classes, além de utilizar um mesmo modelo para classificar subconjuntos que possuem classes diferentes no conjunto de treinamento e de teste.

A Tabela 5.10 lista os resultados ao utilizar os oito subconjuntos criados neste trabalho. Nota-se que com o uso da dissimilaridade é possível alcançar boas taxas de reconhecimento mesmo com um significativo aumento da quantidade de classes envolvidas no problema. Mesmo com a diminuição da duração dos sinais de áudio para cinco segundos (#5 ao #8) e o aumento significativo na quantidade de classes (de 23 para 915), a abordagem da dissimilaridade manteve boas taxas de reconhecimento.

Uma das motivações para utilizar a dissimilaridade é o fato de não ser necessário retreinar o modelo de aprendizagem sempre que novas classes são adicionadas ao sistema de classificação. Desta forma, a Tabela 5.11 lista os resultados obtidos ao se utilizar um mesmo modelo para classificar todos os outros subconjuntos deste trabalho.

Podemos ver que na abordagem de dissimilaridade nem sempre um modelo de aprendizagem com mais classes apresenta os melhores resultados, pois a grande quantidade de classes pode fazer com que o modelo comece a ficar confuso. Outro ponto interessante é que, ao utilizar o mesmo subconjunto para treino e teste, somente o #6 obteve a taxa mais alta. Nos outros casos, as melhores taxas de reconhecimento sempre foram obtidas

Tabela 5.10: Taxas de reconhecimento média (%) utilizando dissimilaridade em subconjuntos com diferentes quantidades de classes.

Treino	Teste	Classes	Top 01	Top 05	Top 10
#1	#1	23	$0{,}896 \pm 0{,}059$	0,991	1,000
#2	#2	48	$0{,}875 \pm 0{,}029$	0,975	0,992
#3	#3	88	$0{,}920 \pm 0{,}036$	0,982	0,991
#4	#4	180	$0{,}849 \pm 0{,}011$	0,936	0,954
#5	#5	349	$0{,}794 \pm 0{,}027$	0,900	0,929
#6	#6	614	$0{,}749 \pm 0{,}012$	0,872	0,902
#7	#7	772	$0{,}722 \pm 0{,}005$	0,859	0,896
#8	#8	915	$0{,}702 \pm 0{,}000$	0,824	0,866

Tabela 5.11: Taxas de reconhecimento (%) médias obtidas ao utilizar diversos modelos para classificar subconjuntos diferentes.

Treino \ Teste	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**
#1 (23)	0,896	0,863	0,845	0,723	0,520	0,394	0,381	0,325
#2 (48)	**0,930**	0,875	0,911	0,752	0,588	0,446	0,493	0,443
#3 (88)	0,913	0,883	0,920	0,816	0,708	0,530	0,596	0,478
#4 (180)	0,904	**0,892**	**0,925**	0,849	0,779	0,673	0,678	0,632
#5 (349)	0,904	0,879	0,895	0,846	0,794	0,695	0,712	0,663
#6 (614)	0,861	0,875	0,895	**0,852**	0,804	**0,749**	**0,737**	**0,711**
#7 (772)	0,904	0,888	0,911	0,851	**0,806**	0,739	0,722	0,711
#8 (915)	0,870	0,863	0,902	0,846	0,789	0,723	0,716	0,702

ao se utilizar um subconjunto para treino e subconjunto outro para teste. Além disso, ter um modelo com uma diversidade maior de classes parece ter mais impacto do que ter sinais de áudio com tempo de duração maior.

5.2 Abordagem Classe Dependente

Esta subseção apresenta os resultados obtidos a partir da abordagem classe dependente (sem dissimilaridade). Para ter uma comparação justa entre abordagens, as mesmas configurações de zoneamento, descritor de textura, algoritmo de classificação e regras de combinação foram utilizadas nos próximos experimentos.

Assim como na abordagem de dissimilaridade, alguns experimentos foram realizados para encontrar bons valores de C e *Gamma* a fim de otimizar a velocidade de execução do LIBSVM. Para isso, utilizou-se a divisão das amostras dos subconjuntos conforme a Tabela 5.1. Como o LBP obteve as melhores taxas na abordagem da dissimilaridade, os experimentos desta subseção também utilizaram o LBP.

A Tabela 5.12 mostra os resultados obtidos ao utilizar o LIBSVM com *grid-search* e depois com valores de C e *Gamma* mais comuns obtidos. Assim como na dissimilaridade, o zoneamento utilizado foi de três zonas verticais e 15 zonas horizontais (escala Mel). Tanto nos resultados utilizando o subconjunto #3 quanto no #4, os melhores resultados foram obtidos ao se utilizar *C=32* e *Gamma=0,125*.

Tabela 5.12: Taxas de reconhecimento (%) utilizando três zonas verticais e 15 zonas horizontais.

Base	**C**	***Gamma***	**Máximo**	**Soma**	**Produto**
#3	*Grid-search*		0,307	0,375	0,375
#4	*Grid-search*		0,256	0,300	0,283
#3	32	0,125	0,352	0,398	0,386
#4	32	0,125	0,267	0,300	0,300
#3	32	0,5	0,261	0,284	0,284
#4	32	0,5	0,206	0,261	0,261

Para computar os resultados listados na Tabela 5.13, os experimentos foram conduzidos com parâmetros semelhantes aos da dissimilaridade, três zonas verticais e 15 zonas horizontais (escala Mel), para ter uma comparação mais justa possível, com exceção dos valores de C e *Gamma* que foram otimizados no LIBSVM devido o conjunto de características extraídas diretamente do LBP. Com estes parâmetros, 45 vetores de características foram utilizados para representar cada um dos espectrogramas e o resultado final é resultante da combinação dos 45 classificadores gerados. As regras do mínimo e mediana não apresentaram resultados satisfatórios e, por isso, foram descartadas.

Pode-se observar na Tabela 5.13 que conforme o número de classes aumenta, as taxas de reconhecimento diminuem mais rapidamente do que na abordagem de dissimilaridade. Outro ponto interessante, é que a quantidade de classes teve um impacto maior nas taxas de reconhecimento do que o tempo de duração dos sinais de áudio.

Tabela 5.13: Taxas de reconhecimento (%) obtidas na abordagem classe dependente em diversos subconjuntos.

Treino	Teste	Classes	Máximo	Soma	Produto
#1	#1	23	0,470	0,443	0,452
#2	#2	48	0,408	0,450	0,433
#3	#3	88	0,314	0,339	0,334
#4	#4	180	0,272	0,298	0,300
#5	#5	349	0,231	0,251	0,252
#6	#6	614	0,138	0,160	0,158
#7	#7	772	0,094	0,102	0,100
#8	#8	915	0,052	0,052	0,052

5.3 Discussão dos Resultados

Os primeiros experimentos deste trabalho mostraram que o fato de utilizar valores pré-definidos para C e *Gamma* em um classificador baseado no SVM, pode representar até 0,41% do tempo de processamento computacional em relação ao uso do *grid-search* na etapa de treinamento. Esta redução considerável de tempo tornou viável a execução de experimentos com a dissimilaridade nos maiores subconjuntos presentes neste trabalho.

Outro ponto interessante a ser observado é a variação na quantidade de zonas verticais. Os resultados com três zonas foram muito superiores em relação a cinco e nove zonas. A diferença pode estar relacionada com o fato de não possuir tanta informação no espectrograma representada em vetores de algumas regiões (em baixas ou altas frequências e no início ou final do espectrograma). Desta forma, os vetores positivos e negativos podem ser bem parecidos porque não há o que representar nessas regiões. Assim, os vetores de dissimilaridade acabam sendo bem parecidos nos dois casos. Isso influencia o modelo de aprendizagem de máquina e pode deixar o sistema de classificação confuso.

A variação na quantidade de zonas horizontais também possibilitou aumentar as taxas de reconhecimento. Assim como em alguns resultados dos experimentos dos trabalhos de Costa et al. (2013, 2012a), a escala Mel esteve presente nas configurações das maiores taxas de reconhecimento obtidas. Essa escala tem apresentado resultados competitivos quando é utilizada com espectrogramas em diferentes aplicações.

A Figura 5.1 apresenta os resultados listados na Tabela 5.10 e Tabela 5.13, que foram computados nas abordagens de dissimilaridade e classe dependente, respectivamente. As duas abordagens utilizaram as mesmas configurações e parâmetros para o zoneamento

e extração de características. É notável que a dissimilaridade manteve boas taxas de reconhecimento a medida que o número de classes aumentava em cada subconjunto, o que não ocorreu na abordagem classe dependente. Este fato pode ser explicado porque na dissimilaridade o sistema de classificação é binário.

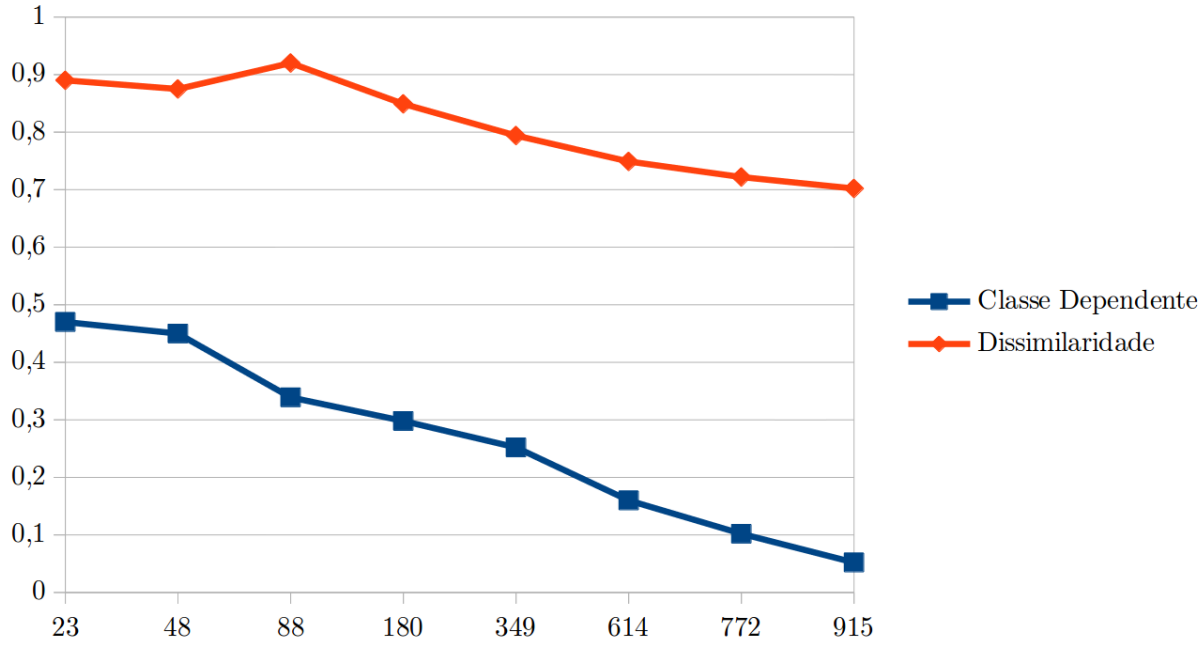


Figura 5.1: Desempenho das abordagens de dissimilaridade e classe dependente.

Outra vantagem na dissimilaridade é o fato de não ser necessário retreinar um modelo sempre que novas classes forem inseridas no sistema de classificação. Os resultados da Tabela 5.11 mostram que nem sempre um modelo com mais classes ou com maior duração dos sinais de áudio garante as melhores taxas de reconhecimento. O modelo gerado a partir do subconjunto #6, que possui três amostras com o mínimo de cinco segundos de duração e 614 espécies, foi utilizado para obter as maiores taxas de reconhecimento quando as amostras dos subconjuntos #4, #6, #7 e #8 foram classificadas.

A Tabela 5.14 lista o desempenho de trabalhos relacionados a identificação de espécies de pássaros que utilizaram uma quantidade maior de classes. Podemos observar que o trabalho de Marini et al. (2015) efetuou testes com 50 espécies e alcançou a taxa de reconhecimento de 45,9%. Chou et al. (2007) conseguiram uma taxa mais alta com uma quantidade maior de espécies, porém obtiveram a base de dados de um CD comercial.

Este trabalho, com a abordagem da dissimilaridade, alcançou taxas de reconhecimento mais altas do que vários trabalhos relacionados, mesmo utilizando subconjuntos com muito mais classes. Conforme a Tabela 5.14, a taxa de reconhecimento mais alta foi de 92,5%

Tabela 5.14: Melhores resultados de trabalhos relacionados e da abordagem de dissimilaridade deste trabalho.

Trabalho	**Espécies**	**Desempenho (%)**
Cai et al. (2007)	14	86,8
Albornoz et al. (2017)	25	89,3
Lee et al. (2008)	28	84,1
Zottesso et al. (2016)	45	79,0
Lucio e Costa (2015)	46	77,6
Marini et al. (2015)	50	45,9
Abordagem proposta	**88**	**92,5**
Abordagem proposta	**349**	**80,6**
Chou et al. (2007)	420	78,3
Abordagem proposta	**772**	**73,7**
Abordagem proposta	**915**	**71,1**

em um subconjunto com 88 espécies, quase o dobro de classes que Zottesso et al. (2016), superando trabalhos com 14, 25 e 28 espécies. Destacamos também que os experimentos com 349, 772 e 915 espécies mantiveram as taxas acima de 70%, próxima de trabalhos que utilizaram menos do que 50 espécies, mesmo com um número muito maior de classes.

6

Considerações Finais

Neste trabalho, apresentou-se uma proposta para a identificação de espécies de pássaros utilizando espectrogramas e dissimilaridade em vários subconjuntos com centenas de espécies (classes). Com a observação dos resultados experimentais, é notável que a abordagem de dissimilaridade se mostrou muito superior em relação a abordagem dependente de modelo. As taxas de reconhecimento sempre foram superiores e conforme o número de classes aumentava, a diferença entre as duas abordagens também aumentava.

A abordagem de dissimilaridade permitiu utilizar um modelo já treinado para identificar espécies de pássaros que não existiam nesse modelo, possibilitando a utilização de classes diferentes nas etapas de treinamento e de teste. Assim, não há necessidade de retreinar um modelo sempre que novas classes forem inseridas no sistema de classificação.

Os experimentos mostraram que na dissimilaridade nem sempre um modelo treinado com mais classes ou com maior tempo de duração dos sinais de áudio pode trazer melhores resultados. Além disso, conforme o número de classes aumenta, a taxa de desempenho diminui, mas muito mais lentamente do que quando a abordagem dependente de modelo é utilizada.

O fato de realizar uma busca inicial para definir alguns parâmetros fixos no LIBSVM, como C e *Gamma*, viabilizou a execução de vários experimentos em relação ao tempo. Quando um classificador SVM é executado com otimização de parâmetros por *grid-search*, ele leva muito tempo para encontrar valores otimizados para esses parâmetros, podendo levar semanas de processamento, como em alguns casos deste trabalho. Isso fica ainda pior com o zoneamento da imagem porque o processo de otimização é executado em cada classificador criado para cada uma das regiões.

6.1 Trabalhos Futuros

Em trabalhos futuros, pretende-se aprimorar a redução de ruídos presentes nos sinais de áudio diretamente no domínio visual (espectrogramas). Muitos ruídos são representados por pixels, que podem ser removidos com técnicas de processamento de imagens, como a erosão.

Outro ponto a ser investigado é a utilização de características que podem ser extraídas diretamente dos sinais de áudio por descritores acústicos, como o *Statistical Spectrum Descriptor* (SSD), *Modulation Frequency Variance Descriptor* (MVD), *Rhythm Patterns* (RP), entre outros, possibilitando a criação de novos classificadores que podem ser utilizados para a combinação. Desta forma, seria possível analisar informações diferentes que podem ser complementares, conforme os experimentos iniciais deste trabalho apresentados na qualificação. Além disso, pretendemos realizar alguns experimentos com outros classificadores, como o algoritmo de classificação *Randon Forest*.

REFERÊNCIAS

Albornoz, E. M.; Vignolo, L. D.; Sarquis, J. A.; Leon, E. Automatic classification of furnariidae species from the paranaense littoral region using speech-related features and machine learning. *Ecological Informatics*, v. 38, p. 39–49, 2017.

Anderson, S. E.; Dave, A. S.; Margoliash, D. Template-based automatic recognition of birdsong syllables from continuous recordings. *The Journal of the Acoustical Society of America*, v. 100, n. 2, p. 1209–1219, 1996.

Bertolini, D. *Identificação e verificação de escritores usando características texturais e dissimilaridade*. Tese de Doutoramento, Universidade Federal do Paraná, 2014.

Bertolini, D.; Oliveira, L. S.; Justino, E.; Sabourin, R. Texture-based descriptors for writer identification and verification. *Expert Systems with Applications*, v. 40, n. 6, p. 2069–2080, 2013.

Briggs, F.; Raich, R.; Fern, X. Z. Audio classification of bird species: a statistical manifold approach. In: *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, IEEE, 2009, p. 51–60.

Cai, J.; Ee, D.; Pham, B.; Roe, P.; Zhang, J. Sensor network for the monitoring of ecosystem: Bird species recognition. In: *Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on*, IEEE, 2007, p. 293–298.

Catchpole, C. K.; Slater, P. J. *Bird song: biological themes and variations*. Cambridge university press, 2003.

Cha, S.-H.; Srihari, S. N. On measuring the distance between histograms. *Pattern Recognition*, v. 35, n. 6, p. 1355–1370, 2002.

Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, v. 2, p. 27:1–27:27, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 2011.

Chen, J.; Kellokumpu, V.; Zhao, G.; Pietikäinen, M. RLBP: Robust local binary pattern. In: *Proceedings of the British Machine Vision Conference*, 2013.

Chou, C.-H.; Lee, C.-H.; Ni, H.-W. Bird species recognition by comparing the hmms of the syllables. In: *Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference on*, IEEE, 2007, p. 143–143.

Chou, C.-H.; Liu, P.-H. Bird species recognition by wavelet transformation of a section of birdsong. In: *Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing*, IEEE, 2009, p. 189–193.

Conway, C. J. Standardized north american marsh bird monitoring protocol. *Waterbirds*, v. 34, n. 3, p. 319–346, 2011.

Costa, Y. M. G. *Reconhecimento de gêneros musicais utilizando espectrogramas com combinação de classificadores.* Tese de Doutoramento, Universidade Federal do Paraná, 2013.

Costa, Y. M. G.; Oliveira, L.; Koerich, A.; Gouyon, F. Music genre recognition based on visual features with dynamic ensemble of classifiers selection. In: *Systems, Signals and Image Processing (IWSSIP), 2013 20th International Conference on*, IEEE, 2013, p. 55–58.

Costa, Y. M. G.; Oliveira, L.; Koerich, A. L.; Gouyon, F.; Martins, J. Music genre classification using LBP textural features. *Signal Processing*, v. 92, n. 11, p. 2723–2737, 2012a.

Costa, Y. M. G.; Oliveira, L. S.; Koericb, A.; Gouyon, F. Music genre recognition using spectrograms. In: *Systems, Signals and Image Processing (IWSSIP), 2011 18th International Conference on*, IEEE, 2011, p. 1–4.

Costa, Y. M. G.; Oliveira, L. S.; Koerich, A. L.; Gouyon, F. Comparing textural features for music genre classification. In: *Neural Networks (IJCNN), The 2012 International Joint Conference on*, IEEE, 2012b, p. 1–6.

Evangelista, T. L.; Priolli, T. M.; Silla, C. N.; Angelico, B. A.; Kaestner, C. A. Automatic segmentation of audio signals for bird species identification. In: *Multimedia (ISM), 2014 IEEE International Symposium on*, IEEE, 2014, p. 223–228.

Fagerlund, S. Bird species recognition using support vector machines. *EURASIP J. Appl. Signal Process.*, v. 2007, n. 1, p. 64–64, 2007.
Disponível em `http://dx.doi.org/10.1155/2007/38637`

Faria, C. M.; Rodrigues, M.; do Amaral, F. Q.; Módena, É.; Fernandes, A. M. Aves de um fragmento de mata atlântica no alto rio doce, minas ger minas gerais: colonização e extinção. *Revista Brasileira de Zoologia*, v. 23, n. 4, p. 1217–1230, 2006.

Gonzalez, R. C.; Woods, R. E. *Processamento digital de imagens.* Pearson Prentice Hall, São Paulo, 2010.

Jain, A. K.; Ross, A.; Pankanti, S. Biometrics: a tool for information security. *IEEE transactions on information forensics and security*, v. 1, n. 2, p. 125–143, 2006.

Kittler, J.; Hatef, M.; Duin, R. P.; Matas, J. On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 20, n. 3, p. 226–239, 1998.

Kogan, J. A.; Margoliash, D. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: A comparative study. *The Journal of the Acoustical Society of America*, v. 103, n. 4, p. 2185–2196, 1998.

Lee, C.-H.; Han, C.-C.; Chuang, C.-C. Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients. *Audio, Speech, and Language Processing, IEEE Transactions on*, v. 16, n. 8, p. 1541–1550, 2008.

Lopes, M. T.; Gioppo, L. L.; Higushi, T. T.; Kaestner, C. A.; Silla Jr, C. N.; Koerich, A. L. Automatic bird species identification for large number of species. In: *Multimedia (ISM), 2011 IEEE International Symposium on*, IEEE, 2011a, p. 117–122.

Lopes, M. T.; Koerich, A. L.; Nascimento Silla, C.; Kaestner, C. A. A. Feature set comparison for automatic bird species identification. In: *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, IEEE, 2011b, p. 965–970.

Lucio, D. R.; Costa, Y. M. G. Bird species classification using spectrograms. In: *Computing Conference (CLEI), 2015 Latin American*, IEEE, 2015, p. 1–11.

MÄENPÄÄ, T. *The local binary pattern approach to texture analysis - extensions and applications.* Tese de Doutoramento, dissertation. Acta Univ Oul C 187, 78 p + App., 2003.
Disponível em `http://herkules.oulu.fi/isbn9514270762/`

MARINI, A.; TURATTI, A.; BRITTO, A.; KOERICH, A. Visual and acoustic identification of bird species. In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, IEEE, 2015, p. 2309–2313.

MARTINS, J. G.; COSTA, Y. M. G.; BERTOLINI, D.; OLIVEIRA, L. Uso de descritores de textura extraıdos de glcm para o reconhecimento de padroes em diferentes domınios de aplicaçao. In: *XXXVII Conferencia Latinoamericana de Informática*, 2011, p. 637–652.

MONTALVO, A.; COSTA, Y. M. G.; CALVO, J. R. Language identification using spectrogram texture. In: *Iberoamerican Congress on Pattern Recognition*, Springer, 2015, p. 543–550.

NANNI, L.; COSTA, Y. M. G.; LUMINI, A.; KIM, M. Y.; BAEK, S. R. Combining visual and acoustic features for music genre classification. *Expert Systems with Applications*, v. 45, p. 108–117, 2016.

NEGRET, Á. Fluxos migratórios na avifauna da reserva ecológica do ibge, brasília, df, brasil. *Revista Brasileira de Zoologia*, v. 5, n. 2, p. 209–214, 1988.

OJALA, T.; PIETIKAINEN, M.; MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 24, n. 7, p. 971–987, 2002.

OJANSIVU, V.; HEIKKILÄ, J. Blur insensitive texture classification using local phase quantization. In: *Image and signal processing*, Springer, p. 236–243, 2008.

PEKALSKA, E.; DUIN, R. P. Classifiers for dissimilarity-based pattern recognition. In: *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, IEEE, 2000, p. 12–16.

SCHUCHMANN, K.-L.; MARQUES, M. I.; JAHN, O.; GANCHEV, T.; FIGUEIREDO, J. Os sons do pantanal: Um projeto de monitoramento acústico automatizado da biodiversidade. *Boletim Informativo Sociedade Brasileira de Zoologia*, v. 108, p. 11–12, 2014.

Umesh, S.; Cohen, L.; Nelson, D. Fitting the mel scale. In: *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, IEEE, 1999, p. 217–220.

Vapnik, V. *The nature of statistical learning theory*. Springer-Verlag, 1995.

Zhao, Z.; Zhang, S.-h.; Xu, Z.-y.; Bellisario, K.; Dai, N.-h.; Omrani, H.; Pijanowski, B. C. Automated bird acoustic event detection and robust species classification. *Ecological Informatics*, v. 39, p. 99–108, 2017.

Zottesso, R. H.; Matsushita, G. H.; Lucio, D. R.; Costa, Y. M. G. Automatic segmentation of audio signal in bird species identification. In: *Computer Science Society (SCCC), 2016 35th International Conference of the Chilean*, IEEE, 2016, p. 1–11.