UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UMA ABORDAGEM PARA MONITORAMENTO DE ANUROS BASEADA EM PROCESSAMENTO DIGITAL DE SINAIS BIOACÚSTICOS

JUAN GABRIEL COLONNA

UMA ABORDAGEM PARA MONITORAMENTO DE ANUROS BASEADA EM PROCESSAMENTO DIGITAL DE SINAIS BIOACÚSTICOS

Tese apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas, Campus Universitário Senador Arthur Virgílio Filho, como requisito parcial para a obtenção do grau de Doutor em Informática.

Orientador: Eduardo Freire Nakamura
Co-orientador: Marco A. Pinheiro Cristo

Manaus - AM
Setembro de 2017

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Colonna, Juan Gabriel
C719u Uma abordagem para monitoramento de anuros baseada em processamento digital de sinais bioacústicos / Juan Gabriel Colonna. 2017
287 f.: il. color; 31 cm.

Orientador: Eduardo Freire Nakamura
Coorientador: Marco Antônio Pinheiro de Cristo
Tese (Doutorado em Informática) - Universidade Federal do Amazonas.

1. Redes de sensores sem fio. 2. Aprendizagem de máquina. 3. Monitoramento ambiental. 4. Teoria da informação. 5. Processamento de sinais bioacústicos. I. Nakamura, Eduardo Freire II. Universidade Federal do Amazonas III. Título

PODER EXECUTIVO
MINISTÉRIO DA EDUCAÇÃO
INSTITUTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UFAM

FOLHA DE APROVAÇÃO

"UMA ABORDAGEM PARA MONITORAMENTO DE ANUROS BASEADA EM PROCESSAMENTO DIGITAL DE SINAIS BIOACÚSTICOS"

JUAN GABRIEL COLONNA

Tese de Doutorado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Marco Antonio Pinheiro de Cristo - PRESIDENTE

Prof. Carlos Maurício Seródio Figueiredo - MEMBRO INTERNO

Prof. José Reginaldo Hughes Carvalho - MEMBRO INTERNO

Prof. Alejandro Cesar Frery Orgambide - MEMBRO EXTERNO

Prof. Antonio Alfredo Ferreira Loureiro - MEMBRO EXTERNO

Manaus, 15 de Setembro de 2017

Av. Rodrigo Otávio, 6.200 - Campus Universitário Senador Arthur Virgílio Filho - CEP 69077-000 - Manaus, AM, Brasil
Tel. (092) 3305 1193 E-mail: secretariappgi@icomp.ufam.edu.br www.ppgi.ufam.edu.br

Agradecimentos

Uma tese com agradecimentos conta duas histórias, uma que começa no primeiro capítulo e outra que começou muito antes de se passarem as páginas. Por este motivo, simplesmente fica difícil listar todas as pessoas que me ajudaram a passar por esta jornada.

Agradeço principalmente à mulher da minha vida Jézika pelo apoio incondicional em todos os momentos, principalmente nos de incerteza, muito comuns para quem tenta trilhar novos caminhos. Sem você nenhuma conquista valeria a pena.

Agradeço aos meus pais Susana e Daniel, que dignamente me ensinaram à importância da família, da honestidade e da persistência. Agradeço também profundamente a toda minha família e a família da Jézika pelo afeto e por acreditarem em mim.

Agradeço ao professor Eduardo Nakamura por ter me recebido e ensinado a ser um pesquisador me dando o privilégio de aprender ao seu lado. Ao professora Marco Cristo pelo aprendizado, pela paciência e por acreditar neste trabalho. Ao professor Alejandro Freire pela confiança que me permitiu estar aqui.

A meus amigos e colegas do laboratório pela amizade e por tornar esta jornada mais agradável. Agradeço também a todos os amigos da minha terra pela amizade que transcende o tempo e os limites geográficos.

Agradeço também ao CNPq pelo suporte através de bolsa de doutorado.

Finalmente, agradeço a todos cuja confiança, apoio, companheirismo e carinho, deram-me forças para enfrentar as dificuldades da elaboração deste trabalho.

"Nothing is too wonderful to be true, if it be consistent with the laws of nature; and in such things as these, experiment is the best test of such consistency."
(The Life and Letters of Michael Faraday)

Resumo

O monitoramento de animais silvestres em seu habitat natural é objeto de estudo de biólogos e ecólogos que coletam informações ambientais para inferir o estado das populações animais e suas variações ao longo do tempo. Um objetivo específico desses estudos é identificar problemas ecológicos em estágios iniciais. No entanto, a coleta das informações é um trabalho manual que deve ser realizado por um grupo de especialistas em áreas de difícil acesso durante períodos de tempo prolongados. Neste contexto, as Redes de Sensores Sem Fio (RSSF) são uma alternativa viável ao monitoramento manual. Estas redes são constituídas por pequenos sensores com capacidade de transmissão, armazenamento e processamento local. Isto possibilita que métodos bioacústicos para reconhecimento automático de espécies sejam embarcados nos nós sensores para automatizar e simplificar a tarefa de monitoramento. Como os sons produzidos pelos animais oferecem uma impressão digital bioacústica, esta pode ser usada para identificar a presença ou ausência de uma espécie particular em uma região. Neste trabalho, apresentamos uma abordagem que utiliza aprendizagem de máquina, RSSF e processamento digital de sinais bioacústicos para reconhecer espécies animais com base em suas vocalizações. Como prova-de-conceito, aplicamos nossa solução para identificar de forma automática diferentes espécies de anuros. Escolhemos anuros uma vez que são utilizados como indicadores precoces de estresse ecológico, pelo fato de serem sensíveis às mudanças do habitat e oferecerem informações sobre os ecossistemas terrestre e aquático. Nossa abordagem integra quatro operações fundamentais: filtragem de ruídos e aprimoramento dos sinais acústicos, segmentação automática desses sinais, extração de descritores acústicos e classificação. Além disso, nossa solução considera as limitações de RSSF, buscando reduzir a carga de processamento e comunicação para prolongar o tempo de vida dos sensores. Portanto, representamos os sinais por um conjunto de descritores acústicos de baixo nível (*Low-Level Acoustic Descriptors* - LLDs) conhecidos como *Mel Frequency Cepstral Coefficients* (MFCCs). A técnica escolhida para filtrar os ruídos ambientais foi o *Singular Spectrum Analysis* (SSA), esta escolha foi baseada nas diversas comparações que fizemos com outros métodos de filtragem. Além disso, o SSA é não paramétrico, se adapta ao coaxar de cada es-

pécie e possui um esquema equivalente na teoria de filtros FIR, o que possibilita ter uma implementação com complexidade computacional constante. Ainda no método de filtragem, desenvolvemos uma versão robusta do SSA. Esta nova versão é mais tolerante aos diferentes ruídos ambientais, sejam estes Gaussianos ou não. A robustez também permitiu identificar os componentes acústicos causados pelos ruídos ambientais associados com as baixas frequências. No que diz respeito à segmentação, primeiro realizamos uma comparação entre diferentes LLDs baseados na teoria da informação. Nesta etapa, desenvolvemos um método não supervisionado capaz de se adaptar às diferentes condições de ruídos ambientais, sejam estes branco ou coloridos. Na segunda etapa, adaptamos dois dos LLDs comparados para funcionamento incremental. Assim, foi possível definir uma metodologia para segmentar os sinais acústicos em tempo real com custo de memória constante, ideal para ser embarcado em um nó sensor de baixo custo e obter as porções dos áudios que possuem as informações relevantes para o reconhecimento das espécies. Finalmente, avaliamos diferentes estratégias de classificação e propusemos uma nova forma de validação cruzada para avaliar a capacidade de generalização do método. Portanto, a validação cruzada tradicional de sílaba-por-sílaba foi substituída por uma validação cruzada que separa diferentes indivíduos nos conjuntos de teste e treinamento. Isto viabilizou uma avaliação mais justa e permitiu estimar o comportamento final que o método de classificação embarcado no nó sensor teria em uma situação real. Dentre os métodos de classificação planos comparados descobrimos que SVM e kNN são os mais promissores. Todavia, propomos e desenvolvemos uma estratégia de classificação hierárquica multirótulo para decompor e simplificar o espaço de decisões do classificador e simultaneamente reconhecer a família, o gênero e a espécie de cada amostra. Isto nos permite concluir que nossa abordagem é flexível o suficiente para se adaptar aos diferentes cenários monitorados, sem deixar de otimizar a relação custo-benefício da solução de monitoramento proposta.

Palavras-chave: Redes de sensores sem fio, aprendizagem de máquina, monitoramento ambiental, classificação de anuros, filtros de sinais bioacústicos, segmentação não supervisionada de sinais, teoria da informação..

Abstract

Wildlife monitoring is often used by biologists and ecologists to acquire information about animals and their natural habitats. In survey programs, specialists collect environmental information to infer about animal population status and their variations over time. The main goal of such programs is to identify environmental problems in early stages. However, acquiring the necessary data for this purpose is a manual work and must be carried out by groups of experts in areas of difficult access during long periods of time. In this context, Wireless Sensor Networks (WSNs) are useful alternatives to alleviate the manual work. Such networks are made up of small sensors with transmission, storage, and local processing capabilities. These networks enable bioacoustic methods for automatic species recognition to be embedded in the sensor nodes in order to automate and simplify the monitoring task. Since animal sounds usually provide a species fingerprint, it can be used to recognize the presence or absence of a target species in a site. Accordingly, in this thesis, we present an approach that combines machine learning methods, WSNs and bioacoustic signal processing techniques for wildlife monitoring based on animal calls. As a proof-of-concept, we choose anurans as the target animals. The reason is that anurans are already used by biologists as an early indicator of ecological stress, since they provide relevant information about terrestrian and aquatic ecosystems. Our solution integrates four fundamental steps: noise filtering and bioacoustic signal enhancement, automatic signal segmentation, acoustic features extraction, and classification. We also consider the WSNs limitations, trying to reduce the communication and processing load to extend the sensors' lifetime. To accomplish with the restriction imposed by the hardware, we represent the acoustic signals by a set of low-level acoustic descriptors (LLDs or features). This representation allows us to identify specific signal patterns of each species, reducing the amount of information necessary to classify it. The adverse environmental conditions of the rainforest pose additional challenges, such as noise filtering. We developed a filtering method based on Singular Spectrum Analysis (SSA). This choice was based on several comparisons with other filtering methods. The SSA method has additional advantages: it is non-parametric, it adapts to the different input signals, and it has an equivalent

in the FIR filter theory that allows the implementation of the filter with low computational complexity. In addition, we develop a robust variant of SSA (RSSA) tolerant to Gaussian and non-Gaussian noises which is able to identify the principal components of the signals related to low-frequency environmental noises. Concerning the unsupervised signal segmentation step, we first carry out a comparison between different LLDs from information theory. In this step, we developed an unsupervised method capable of adapting to different environmental noise conditions, whether white or colored noise. In the second step, we reformulated two of the LLDs for incremental computation. With this, it was possible to define a new methodology to segment an acoustic signal in real time with constant computational complexity and reduced storage requirements, thus, being ideal for embedding in low-cost sensor motes. Finally, we evaluate different flat classification strategies and we proposed a new cross-validation procedure to evaluate the generalization capabilities of the sensor. Therefore, the traditional cross-validation (syllable-by-syllable) was replaced by a cross-validation by individuals in order to perform a fairer evaluation and be able to estimate the final performance of our method in a real situation. Among the compared methods, we emphasize that kNN and SVM are the most promising. However, we went beyond the flat classification proposing a multi-label hierarchical classification strategy to decompose and simplify the classifier's decision space while simultaneously recognizing the family, gender, and species of each sample. This allows us to conclude that our approach is flexible enough to adapt to the different monitoring scenarios while optimizing the cost-benefit ratio of the proposed monitoring solution.

Keywords: Wireless Acoustic Sensor Networks, Machine Learning, Enviromental Monitoring, Anuran Classification, Digital Signal Processing, Information Theory, Singular Spectrum Analysis.

Lista de Figuras

Lista de Tabelas

Acrônimos

1A1 Um contra um (*One-against-One*)

1AA Um contra todos (*Ones-against-All*)

ACR Sistema Automático de Reconhecimento de Chamadas (*Automatic Calls Recognition System*)

ADC Conversor analógico Digital (*Analog - to - Digital Converter*)

AEER Taxa de erro de eventos acústicos (*Acoustic Event Error Rate*)

ALE Intensificador de linha adaptativo (*Adaptive Line Enhancer*)

aMFCC Coeficientes Mel com escala exponencial (*Antimel Frequency Cepstral Coefficient*)

ANN Redes Neurais Artificiais (*Artificial Neural Networks*)

ASR Reconhecimento automático de fala (*Automatic Speech Recognition*)

AUC Área sob a curva ROC (*Area Under the Curve ROC*)

BN Ruído do ambiente (*Background Noise*)

C4.5 Árvore de Decisão C4.5 (*Decision Tree C4.5*)

CWT Transformada Wavelet Continua (*Continuous Wavelet Transform*)

DCT Transformada discreta do cosseno (*Discrete Cosine Transform*)

DT Árvore de Decisão (*Decision Tree* - DT)

DWT Transformada Wavelet Discreta (*Discrete Wavelet Transform*)

E Energia do sinal

EMD Decomposição de modos empíricos (*empirical mode decomposition*)

FFT Transformada rápida de Fourier (*Fast Fourier Transform*)

FIR Filtro de resposta ao impulso finita (*Finite impulse response filter*)

FPGA *Field-Programmable Gate Array*

GMM Modelos de misturas Gaussianas (*Gaussian Mixture Models*)

GT Referencia considerada verdadeira realizada por um especialista (*Ground truth*)

HLD Descritores acústicos de alto nível (*High-level Descriptors*)

HMM Modelos ocultos de Markov (*Hidden Markov Model*)

HxC Plano Entropia-Complexidade (*Entropy-Complexity Plane*)

ID Número identificador da espécie dentro da base de dados

IG Ganho da Informação (*Information Gain*)

IMF Funções intrínsecas (*Intrinsic Mode Functions*)

IUCN União Internacional para Conservação da Natureza (*International Union for Conservation of Nature*)

***k*-CV** Validação Cruzado com k subconjuntos (*k-folds Cross-Validation*)

KLT Transformada Karhunen-Loeve (*Karhunen-Loeve Transform*)

kNN k Vizinhos mais próximos (*k-Nearest Neighbors*)

LCPL Um classificador por nível (*One Classifier per Parent Level*)

LCPN Um classificador por nó pai (*One Classifier per Parent Node*)

LDA Análise discriminante linear (*Linear Discriminant Analysis*)

LFCC Coeficientes Mel com escala linear (*Linear Mel Frequency Cepstral Coefficient*)

LLD Descritores acústicos de baixo nível (*Low-level Descriptors*)

LOOCV Validação Cruzado separando somente uma instância para teste (*Leave-one-Out Cross-Validation*)

LS Mínimos quadrados (*Least Squares*)

LTI Sistema linear é invariante no tempo (*Linear Time Invariant System*)

MFCCs Coeficientes Mel com escala logarítmica (*Mel Frequency Cepstral Coefficients*)

ML Aprendizagem de Máquina (*Machine Learning*)

MSE Erro quadrático médio (*Mean Squared Error*)

MV Regra de votação maioritária (*Majority Voting Rule*)

NB Naive Bayes

PC Componentes Principais (*Principal Components*)

PCA Análise de componentes principais (*Principal Components Analysis*)

PDF Função densidade de probabilidade (*Probability Density Function*)

PE Entropia das Permutações (*Permutation Entropy*)

PME Entropia das permutações mínima (*Permutation Min-Entropy*)

PSD Densidade espectral de potencia (*Power Spectral Density*)

QDA Análise discriminante quadrático (*Quadratic Discriminant Analysis*)

RBF Kernel Gaussiano (*Radial-basis function kernel*)

ROC Característica de Operação do Receptor (*Receiver Operating Characteristic*)

RSSA Robust Singular Spectrum Analysis

RSSF Redes de Sensores Sem Fio

SDR Taxe de distorção (*Signal-to-Distortion Rate*)

sink Nó sensor sorvedouro

SMS Subtração espectral média (*Spectral Mean Subtraction*)

SNR Relação sinal-ruído (*Signal-to-Noise Ratio*)

SSA Singular Spectrum Analysis

SURE Critério de minimização do risco empírico

SVD Decomposição em valores singulares (*Singular Value Decomposition*)

SVM Máquinas de vetores de suporte (*Support Vector Machine*)

WPE Entropia das permutações ponderada (*Weighted PE*)

ZCR Taxa de cruzamento por zero do sinal (*Zero Crossing Rate*)

Nomenclatura

$\mathbf{C}_s$ Complexidade estatística (*Statistical Complexity*)

$\mathbf{D}$ Conjunto de *frames* representados pelos LLDs

d_δ Densidade do ruído impulsivo

F1 F-Score

fn Falsos negativos (*False Negative*)

FNR Taxa de falsos negativos (*False Negative Rate*)

fn Falsos positivos (*False Positive*)

FPR Taxa de falsos positivos (*False Positive Rate*)

f_s Frequência de amostragem (*Sampling Frequency*)

χ Função analítica

$\mathbf{H}_a$ Entropia acústica

$\mathbf{H}_f$ Entropia espectral (*Spectral Entropy*)

$\mathcal{H}$ Transformada de *Hilbert*

$\mathbf{H}_s$ Entropia das permutações (PE)

$\mathbf{H}_t$ Entropia temporal

$\delta(\cdot)$ Impulso unitário com valores ± 1

$\mathbf{Q}$ Divergência de Jensen-Shannon (*Jensen-Shannon Divergency*)

m Parâmetro *time-embedding* no contexto da PE

MCC Coeficiente de correlação de *Matthews*

M_r Saída dos filtros triangulares utilizados para o cálculo dos coeficientes MFCCs

n Tamanho máximo do sinal (*signal lenght*)

N Tamanho da janela deslizante (*frame lenght*)

P_e Histograma de referência para o calculo da divergência

Π Conjunto de padrões ordinais da PE

π_i Padrão ordinal da PE

Prec Precisão

Rec Revocação

r_Q Coeficiente de autocorrelação robusto

ξ Ruído aleatório

r_{xx} Coeficiente de autocorrelação

σ_ξ Variância do ruído

σ_x Variância do sinal

t Índice temporal de *frame*

T Conjunto de *frames*

τ Parâmetro *time-lag* no contexto da PE

tn Verdadeiros negativos (*True Negative*)

TNR Taxa de verdadeiros negativos (*True Negative Rate*)

tp Verdadeiros positivos (*True Positive*)

TPR Taxa de verdadeiros positivos (*True Positive Rate*)

$\mathcal{W}$ Transformada Wavelet

W Total de classes do classificador

w_i i-ésima classe do classificador

$\mathbf{X}_f$ Valores da transformada discreta de Fourier aplicando a FFT

x_i Valores dos pontos da série temporal x

x Série temporal x (áudio ou gravação)

Sumário

1

Introdução

Técnicas para monitoramento bioacústico são úteis para criar inventários sobre a biodiversidade, obter estimativas globais das espécies e planejar estratégias de conservação ambiental (Obrist et al., 2010, Bardeli et al., 2010, Laiolo, 2010).

Por serem sensíveis às alterações do ecossistema, o estudo e interpretação das mudanças nas populações de anfíbios, e principalmente dos anuros, teve um aumento significativo nas últimas décadas (Gibbs et al., 2005, Buckley and Jetz, 2008, Curado et al., 2011, Arntzen et al., 2017).

Através do monitoramento bioacústico de anuros é possível perceber alterações precoces nas populações e identificar: contaminações por poluentes tais como os agrotóxicos, alterações do habitat por migração de outros animais ou mudanças na vegetação devido ao desmatamento. Além disso, monitorando as variações da densidade populacional dos anuros seria possível quantificar o dano ambiental causado por alguma anomalia, como por exemplo, um incêndio, um desmatamento ou a introdução de espécies animais e vegetais exóticos (Carey et al., 2001, Collins and Storfer, 2003, Hu et al., 2009, Curado et al., 2011). Porém, para verificar estas hipóteses, é necessário criar e utilizar um conjunto de técnicas que viabilize o monitoramento ambiental bioacústico.

As técnicas de monitoramento ambiental incluem uma etapa de coleta e uma de análises dos dados. O tipo de dado coletado depende da espécie alvo, podendo incluir áudio, fotografia, observações sobre a presença ou ausência dos indivíduos e inclusive a medição de variáveis ambientais. No caso de sons bioacústicos, é utilizado um gravador (ou sensor acústico) e um software para filtrar e separar os padrões que correspondam às diferentes espécies. A análise destas gravações permite confeccionar inventários biológicos (Swiston and Mennill, 2009, Depraetere et al., 2012). Entretanto

em ambas etapas, coleta e análise dos dados, existem tarefas que devem ser executadas por um especialista, estando sujeitas a erros humanos (Bridges and Dorcas, 2000).

O monitoramento ambiental através de sinais acústicos implica no desafio de desenvolver métodos preferencialmente não supervisionados, capazes de executar as tarefas de monitoramento com a mínima intervenção humana possível. Neste contexto, a coleta de som pode ser realizada com pequenos sensores espalhados na floresta, principalmente nas áreas de interesse ecológico, formando uma Rede de Sensores Sem Fio (RSSF). A identificação das espécies presentes pode ser realizada com técnicas automáticas de classificação de séries temporais (Sueur et al., 2012, Ribas et al., 2012). Embora já existam algumas soluções tecnológicas de baixo custo para a coletas dos áudios, as limitações do hardware, as severas condições ambientais e o difícil acesso aos locais de monitoramento, como a Floresta Amazônica, acrescentam dificuldades ao processo (Ingelrest et al., 2010, Lambert and McDonald, 2014).

Portanto, em nossa hipótese de pesquisa avaliamos a possibilidade de combinar técnicas de processamento e análise de sinais bioacústicos, conceitos de teoria da informação e métodos de aprendizagem de máquina para criar uma abordagem bioacústica para reconhecimento de anuros, que possa ser embarcado nos nós sensores de monitoramento ambiental. Assim, ao longo desta tese discutiremos problemas e soluções para as diferentes partes dessa abordagem, focando principalmente nas técnicas de processamento digital de sinais acústicos discretos e na utilização e combinação de técnicas de classificação automática.

1.1 Motivação ambiental

Os desequilíbrios ecológicos, tais como mudanças no clima, desmatamento, aquecimento global, contaminação por agrotóxicos e impacto da urbanização, são as principais causas de extinção das espécies animais e vegetais, contribuindo a um processo de degradação irreversível do habitat (Williams et al., 2003, Thuiller, 2004, Bernarde and Macedo, 2008, Sueur et al., 2012). A União Internacional para a Conservação da Natureza (IUCN) elabora anualmente uma lista de espécies ameaçadas, e aponta os anfíbios como as espécies com maior perigo de extinção (Vié et al., 2009). A situação dos anfíbios no Brasil não é uma excepção (Eterovick et al., 2005).

A diminuição das populações de anfíbios segundo Collins and Storfer (2003), pode ser explicada por seis diferentes hipóteses. As três primeiras hipóteses estão relacionadas à exploração abusiva e à mudança do uso da terra. Uma quarta hipótese à mudança global, que inclui um aumento da radiação ultra-violeta e o efeito estufa. A

quinta hipótese à utilização crescente de pesticidas e outros químicos tóxicos. A sexta, são as doenças infecciosas emergentes.

Os anuros, por serem anfíbios, tem sensibilidade aos fatores ambientas. Estes possuem uma pele semi-permeável que os torna sensíveis ao mesmo tempo às condições aquáticas e terrestres (King, 1969). Assim, cada indivíduo (sapo ou rã) de uma determinada população pode ser considerado um sensor biológico complexo que capta informações do meio ambiente. Os anuros são capazes de reagir às mudanças das variáveis ambientais, ou a combinações mais complexas dessas variáveis, que seriam impossíveis de detectar através de sensores físicos ou elétricos. Por exemplo, detectar a interação entre indivíduos de diferentes espécies, ou identificar um desequilibro causado pela introdução de espécies não nativas.

Cole et al. (2014) relacionaram características espaciais e temporais dos distúrbios florestais com as dinâmicas nas variações das populações de anuros. Entretanto, entender a interação dos fatores que causam variações nas populações é uma tarefa complexa (Williams et al., 2003). Consequentemente, gerar um modelo que explique tais variações ajudaria a: entender as dinâmicas das populações e transferir esse conhecimento a outras espécies animais, avaliar problemas ecológicos ainda em estágio inicial, e estabelecer estratégias de conservação da biodiversidade (Williams, 2001). Deste modo, desenvolver técnicas de monitoramento bioacústico permite viabilizar estudos ambientais mesurando o dano causado a um habitat.

1.2 Diferentes abordagens de monitoramento ambiental

Através do monitoramento ambiental de diferentes espécies animais que habitam uma região, é possível inferir informações relevantes sobre o habitat. Essas informações nos permitem entender, estudar e interpretar as variações das condições locais do meio ambiente que podem ser a causa dos fenômenos de extinção, migração ou colonização (MacKenzie et al., 2002, Carey and Alexander, 2003, Cole et al., 2014, Kalan et al., 2015).

Para tal fim, diferentes estratégias podem ser utilizadas. Os satélites de sensoriamento remoto e as aeronaves científicas (*drones*) fornecem uma grande quantidade de dados de monitoramento em escala global para observar mudanças na cobertura florestal e no uso da terra, mas com resolução espacial e temporal relativamente baixa e sem a capacidade de monitorar a biodiversidade sob a copa das árvores. Consequentemente, sua utilidade é limitada. Ao mesmo tempo, métodos manuais para monitorar

a biodiversidade abaixo do dossel, em grandes áreas, usando protocolos padronizados e por longos períodos de tempo ainda são logisticamente e financeiramente proibitivos.

Em primeira instância, a aplicação de *surveys* acústicos é a opção mais frequente (MacKenzie et al., 2003). Neste caso, é necessário que um especialista percorra as áreas monitoradas para detectar a presença ou ausência das espécies de interesse. Desta forma, o especialista simplesmente "ouve" se a espécie de interesse encontra-se no lugar. Em alguns casos a presença desta pode ser também confirmada visualmente. Este procedimento deve ser repetido durante vários dias, enquanto o estudo durar, podendo levar anos (Shannon et al., 2014).

Para cobrir áreas extensas são formados grupos de especialistas que viajam para realizar as observações. Quando é necessário aumentar o tamanho da área monitorada, deve-se também incrementar o número de pessoas para poder realizar o monitoramento, incluindo colaboradores com nível de experiência relativamente menor. Consequentemente, a qualidade dos resultados está sujeita a fatores como: valores perdidos devido a imprevistos nos traslados ou confusões no reconhecimento de espécies similares, quando o conjunto de diferentes espécies monitoradas aumenta. Estes problemas causam falsos positivos ou negativos diminuindo a qualidade das estimativas (Royle and Link, 2006). O tempo de monitoramento gasto em cada lugar também é diminuído proporcionalmente ao aumento de pontos geográficos dispersos e à frequência utilizada para realizar a amostragem. Assim, a principal desvantagem desta abordagem é a qualidade dos dados obtidos e os custos operacionais.

Uma alternativa interessante consiste em utilizar sensores acústicos (ou gravadores) para estimar a biodiversidade. Logo, pode ser utilizado um método probabilístico para estimar a ocupação das espécies. Desta forma, a biodiversidade pode ser estimada calculando diferentes índices de riqueza acústica (Acevedo and Villanueva-Rivera, 2006, Sueur et al., 2008, Wimmer et al., 2013, Towsey et al., 2014a). Com estes índices não é necessário identificar cada espécie individualmente, para isso, é necessário simplesmente calcular a entropia acústica do lugar (Lellouch et al., 2014). Utilizando estes índices é possível comparar diferentes lugares pelo grau de atividade acústica (Lattanzi et al., 2016). Assim, com a utilização de sensores acústicos é possível automatizar as coletas, manter cópia dos registros e diminuir os custos operacionais. Contudo, através do índice de riqueza acústica não é possível identificar espécies individualmente, ou eliminar o viés causado pelo aumento dos ruídos ambientais. Para solucionar este problema, é necessário que um especialista processe manualmente várias horas de gravação.

A opção mais eficaz, do ponto de vista das informações obtidas, é o monitoramento ambiental bioacústico utilizando sensores acústicos com processamento local (Selavo et al., 2007, Ribas et al., 2012, Colonna et al., 2014a). A capacidade de proces-

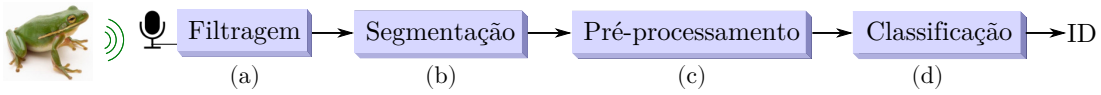


Figura 1.1. Sistema bioacústico para o reconhecimento das espécies de anuros através das vocalizações (ACR).

samento permite: selecionar minutos específicos dos áudios para evitar armazenar ou transmitir dados desnecessariamente; identificar horários de pico da atividade acústica para maximizar a probabilidade de detecção; coletar variáveis ambientais tais como: umidade, temperatura e luminosidade; reconhecer espécies automaticamente; e filtrar ruídos ambientais. A capacidade de processamento, quando combinada com a interface sem fio, possibilita também a troca de informações entre os sensores para melhorar o reconhecimento, escolher o áudio com melhor relação sinal-ruído (SNR), identificar a coordenada geográfica da fonte sonora, ou simplesmente transmitir os dados até um nó central (*sink*) (Bal et al., 2009, Bertrand, 2011). Além dessas vantagens, a transmissão dos dados evita custos operacionais e torna a abordagem de monitoramento bioacústica menos intrusiva.

1.3 Abordagem de monitoramento bioacústico

Os sinais acústicos emitidos pelos animais são captados pelo microfone e processados no nó sensor. As abordagens tradicionais de monitoramento ambiental bioacústico incluem somente três etapas: detecção do evento acústico, pré-processamento para extrair as características acústicas discriminantes, e classificação (figura 1.1). Tais abordagens são chamadas de *Automatic Calls Recognition System* (ACR) e foram originalmente baseadas nas abordagens de reconhecimento de fala humana (*Automatic Speech Recognition* - ASR) (Mammone et al., 1996). Em cada etapa do ACR existem requisitos que devem ser cumpridos, sendo que, o desempenho final do sistema depende da eficácia de cada uma das partes que o integram. No entanto, a maioria dos estudos focam em resolver cada uma destas partes de forma isolada, deixando lacunas entre a integração dos elementos.

As vocalizações dos anuros são compostas de uma sucessão de unidades menores (figura 1.2). Assim, podemos definir tais unidades como “sílabas” e “chamadas”, da forma:

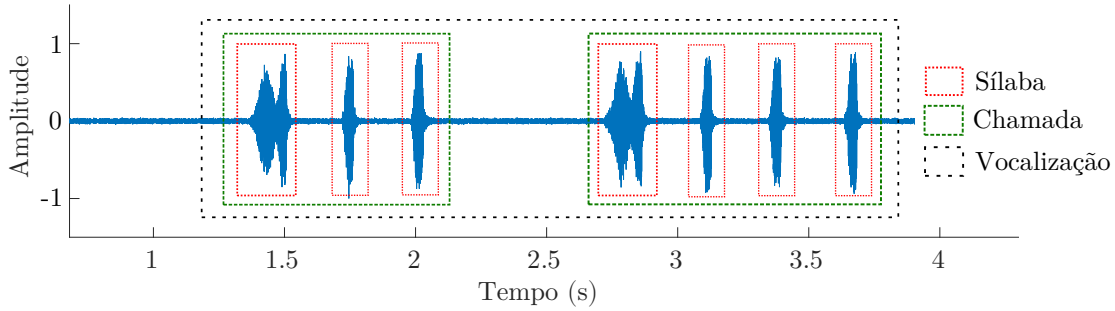


Figura 1.2. Gravação da espécie Hyla minuta. Descrição gráfica da diferença entre sílabas, chamadas e vocalização.

Sílaba: *menor padrão de sinal contínuo, contido entre dois intervalos de ruído ambiental;*

Chamada: *canto composto por uma sucessão de sílabas com alguma finalidade específica, i.e., delimitar território ou atrair as fêmeas.*

Em nossa abordagem adicionamos uma etapa de filtragem (1.1a) dos sinais úteis para melhorar a relação sinal-ruído (SNR). O objetivo principal do filtro é remover sons ambientais da floresta sem causar distorções prejudiciais às vocalizações para a qualidade das gravações. Esta etapa é negligenciada na maioria das abordagens bioacústicas, confiando que a robustez dos descritores acústicos (LLD) e a capacidade do classificador são suficientes para obter um desempenho aceitável de reconhecimento. Porém, a riqueza acústica dos cenários da floresta pode causar uma degradação do desempenho. Esta hipótese é abordada no capítulo 6.

A detecção do evento acústico (1.1b) é equivalente a segmentar os áudios em sílabas (figura 1.2) ou em chamadas completas. Este passo precede a classificação e pode ser aplicado a um conjunto de gravações (Jaafar and Ramli, 2013) ou a um fluxo contínuo de áudio em tempo real (como sendo uma abordagem incremental) (Colonna et al., 2015). Em alguns sistemas ACR as operações de segmentação e classificação são integradas na mesma etapa (Xie et al., 2015b), enquanto que, em outros, a segmentação é utilizada para selecionar os dados de entrada para o classificador (Huang et al., 2009, Colonna et al., 2012). A segmentação deve ser precisa para não descartar segmentos das sílabas que possam ser chaves no reconhecimento dos eventos acústicos. O impacto causado pela etapa de segmentação é avaliado no capítulo 4.

O pré-processamento (1.1c) pode ser aplicado com diferentes propósitos, por exemplo, comprimir os dados ou reduzir as sílabas a conjuntos de características acústicas. As características (ou *features*) são os valores dos "descritores acústicos" (LLDs)

que o classificador utiliza para separar e reconhecer as espécies (seção 2.2). A independência destes descritores em relação à duração ou amplitude das sílabas, a robustez aos ruídos, às diferentes frequências de amostragem, ou aos diferentes níveis de quantização, são qualidades convenientes na maioria das aplicações práticas (uma comparação de diferentes descritores acústicos pode ser encontrada em Colonna et al. (2012)). No capítulo 5 avaliamos o conjunto de LLDs considerados como estado da arte na maioria das abordagens ACR.

Por último, a classificação (1.1d) é a etapa de reconhecimento das espécies. Neste passo, o classificador relaciona os valores dos descritores acústicos que representam as sílabas com o identificador da espécie correspondente (ID). Diferentes técnicas de classificação podem ser utilizadas (Colonna et al., 2012, Jaafar et al., 2014). Cada uma destas tem uma forma particular de criar as funções matemáticas que separam as classes (ou espécies), obtendo mais ou menos sucesso no reconhecimento (seção 2.5). Em nossa abordagem nós estendemos o conceito de classificação plana, com um único rótulo, para um método multirótulo capaz de reconhecer a família e o genero de cada espécie.

Na literatura podem ser encontradas diferentes abordagens, por exemplo: Cai et al. (2007) incluíram uma primeira etapa de filtragem, mas utilizaram uma segmentação fixa por *frames* dos sinais; Diaz et al. (2012) utilizaram uma técnica de compressão para diminuir a quantidade de dados transmitidos; Oliveira et al. (2015) aplicaram uma técnica de processamento de imagens baseada na morfologia do espectro de frequências; Xie (2017) abordou o problema de reconhecimento utilizando técnicas de processamento de imagens com Redes Neurais Convolucionais e extração automática de *features*; Ribas et al. (2012) agruparam sensores acústicos para realizar fusão dos áudios; e Colonna et al. (2014a) realizaram fusão das classificações entre os nós sensores da rede.

Nesta tese vamos explorar cada etapa do ACR ilustrado na figura 1.1 com o propósito de (1) definir uma abordagem de monitoramento bioacústico geral, (2) contribuir com um método de processamento de sinais bioacústicos que permita interpretar, além das vocalizações, os sons ambientais, e (3) fornecer uma solução que integre a segmentação automática, a filtragem de ruídos e a classificação, com o objetivo de definir uma solução completa, robusta, autônoma e adequada às tecnologias de monitoramento atuais.

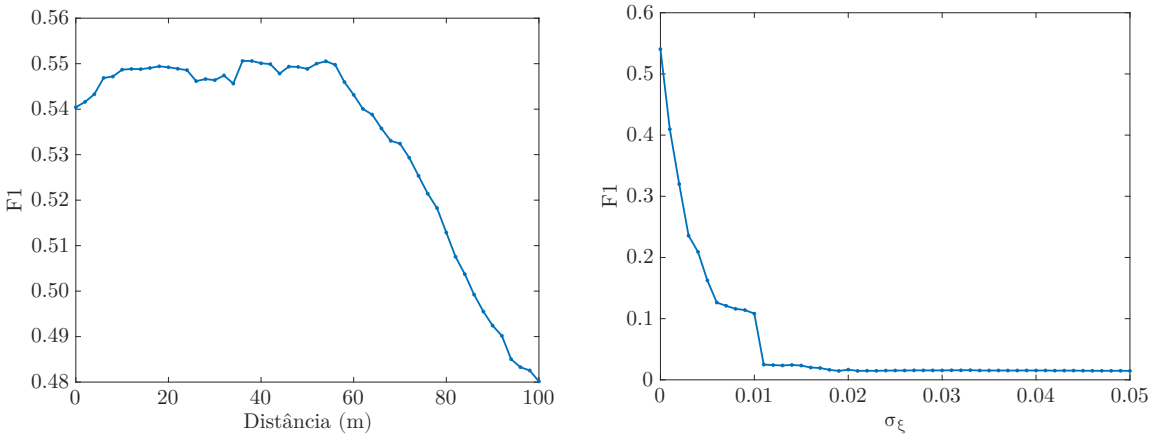
1.4 Problemas do monitoramento bioacústico automático através do ACR

Para que seja possível identificar a variação nas populações de anuros, o monitoramento ambiental deve ser distribuído geograficamente e mantido por longos períodos de tempo. O esforço humano requerido e os elevados custos acabam impactando de forma negativa nos resultados. Neste contexto, a relação custo-benefício das soluções tecnológicas que integram RSSF constituem uma alternativa às técnicas manuais. No entanto, a densidade e a posição de nós da rede; as atenuações sonoras; os ruídos ambientais; a acurácia das diferentes técnicas de classificação; a quantidade de espécies monitoradas; e a interação de diferentes espécies em cenários dinâmicos, são algumas das causas de erros das estimativas das populações.

As abordagens de monitoramento autônomas com sensores acústicos baseiam-se na utilização de técnicas de classificação para realizar o reconhecimento das espécies, tais como: *Support Vector machine* (SVM) ou *K-Nearest Neighbors* (*kNN*) dentre outras. Embora estas técnicas consigam resultados teóricos maiores a 95%, em alguns casos (Huang et al., 2009, Colonna et al., 2012), a aplicação em situações reais, quando são combinadas com redes de sensores, sofrem problemas que diminuem a taxa de acerto. Por exemplo: a atenuação do nível do sinal captado pelo microfone por efeito da distância ou o incremento no nível dos ruídos ambientais, são alguns destes problemas. A figura 1.3 ilustra a diminuição do F-Score (F1) no reconhecimento de 10 espécies de anuros em relação ao aumento da variância dos ruídos aleatórios aditivos (σ_ξ, ruído branco neste caso, figura 1.3(b)) e em relação à potência do sinal recebido no microfone (figura 1.3(a)). Aqui podemos observar que o aumento do nível de ruído prejudica o reconhecimento mais que as atenuações.

Nas técnicas de classificação, o sinal bioacústico proveniente do microfone é representado por um conjunto de características (Colonna et al., 2012), também chamadas *Acoustic Low-Level Descriptors* (LLDs). Tal conjunto de descritores constitui um mapeamento da série temporal em um espaço de dimensões menores, na qual diferentes eventos acústicos podem ser separados e reconhecidos com maior facilidade. Independente do método de classificação escolhido, detectar de forma precisa o começo e o fim do evento acústico possui uma relação direta com a acurácia geral do reconhecimento. Por exemplo, se o segmentador falha frequentemente, acusando falsos negativos, as sílabas serão perdidas aumentando o viés na estimativa de extinção das espécies.

A figura 1.4 apresenta três sílabas de diferentes espécies (segmentos a-b, c-d e e-f). Nesta figura, pode-se notar que os padrões dos sinais acústicos de cada sílaba



(a) (b)

Figura 1.3. Relação entre o F1 da classificação e: (a) a atenuação dos sinais causada pela distância até o nó sensor, e (b) a adição de ruído aleatório branco nas sílabas com diferentes valores de variância (σ_ξ), utilizando Análise Discriminante Quadrático (QDA) e validação cruzada por indivíduos.

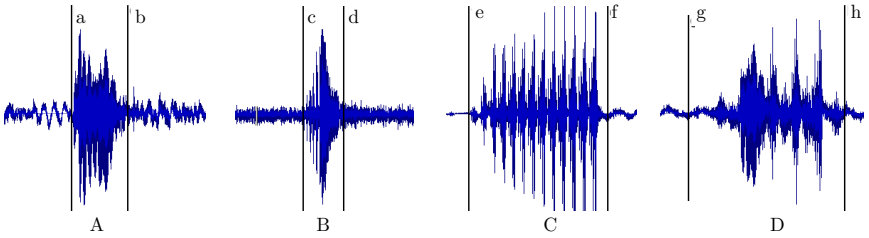


Figura 1.4. Três padrões temporais correspondestes a três espécies diferentes (A, B e C) e um padrão (D) resultado da combinação das vocalização dos segmentos A e C.

representam as espécies de forma única, possibilitando o reconhecimento. O quarto segmento da figura 1.4 (D), é uma combinação linear das vocalizações de duas espécies diferentes. As combinações de sílabas de diferentes espécies acontecem em cenários com um número maior de indivíduos (cenários mais confusos). A probabilidade de captar sinais combinados nos nós sensores aumenta proporcionalmente à biodiversidade do lugar. Estes "novos" padrões de onda não conhecidos pelo classificador aumentam a taxa de erro do reconhecimento, podendo levar a conclusões erradas, identificando espécies não existes (falsos positivos).

Além das atenuações, ruídos e combinações de sinais, existe um quarto fator que influencia no resultado. Este é a "capacidade de generalização do modelo". A generalização relaciona-se com a habilidade do método para detectar novos indivíduos (ou espécimens), os quais não foram incluídos na base de treino do classificador. Em

outras palavras, é a capacidade de detecção de novos indivíduos das espécies treinadas em cenários reais. A forma de avaliar a generalização do modelo, embora pareça trivial, é um erro frequente encontrado nos trabalhos relacionados à área (Colonna et al., 2016a).

Finalmente, a maioria dos ACRs não foram desenvolvidos especificamente para redes de sensores, possuindo elevados custos computacionais, incluindo memória e processamento, que diminuem a vida útil da bateria e desconsideram a capacidade de colaboração e troca de informação entre os nós sensores.

1.5 Por que utilizar RSSF?

A filtragem, a segmentação e a classificação podem ser insuficientes para reconhecer de forma precisa as espécies quando o cenário monitorado é complexo. Nestes cenários, combinações de vocalizações formam novos padrões, sendo necessário identificá-los. Este é um problema frequente em Redes de Sensores Acústicos Sem Fio (RSASF) (Cai et al., 2007, Bertrand, 2011). As RSASF são um subconjunto das RSSF tradicionais, nas quais o instrumento de sensoriamento principal é um microfone. Porém, em uma RSSF tradicional o microfone é simplesmente um sensor adicional dentro do conjunto de sensores ambientais, tais como: umidade, temperatura, pressão, etc.

A multiplicidade de sensores e a capacidade colaborativa entre eles é uma das maiores vantagens das RSSF, uma vez que, a troca de informações entre os nós permite que sejam aplicadas operações de fusão e agregação de dados entre os sensores membros de um *cluster* (ou comitê) (Akyildiz et al., 2002, Bal et al., 2009, Nakamura, 2007). Atualmente, esta nova classe de redes, é uma aposta ao futuro das técnicas de monitoramento ambiental bioacústico (Mainwaring et al., 2002, Acevedo and Villanueva-Rivera, 2006, Hu et al., 2009, Sueur et al., 2012, Wimmer et al., 2013).

Utilizar um ACR para monitoramento bioacústico de forma isolada desaproveita as vantagens das RSSF, por exemplo, a captação de sinais próximos correlacionados para aumentar a relação SNR. Colonna et al. (2014a) e Ribas et al. (2012) mostraram que a "opinião" dos sensores vizinhos pode ser utilizada para fortalecer a decisão do reconhecimento ou identificar cenários confusos. A figura 1.5 é um exemplo de cenário confuso no qual é utilizado um comitê com quatro sensores para identificar duas espécies.

Os experimentos realizados na Great Duck Island, por Mainwaring et al. (2002), e na Austrália, por Hu et al. (2009), respectivamente, são uma demonstração do alcance das técnicas de monitoramento ambiental que integram redes de sensores acústicos,

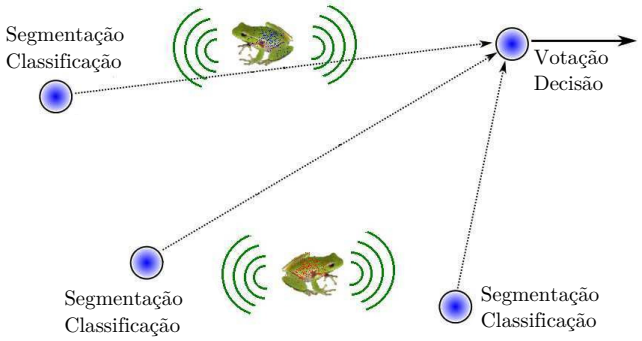


Figura 1.5. *Cluster* de sensores acústicos com votação e rejeição.

oferecendo vantagens únicas sobre as abordagens manuais.

1.6 Objetivos

Nosso objetivo principal é desenvolver um método para monitorar e reconhecer espécies de anuros, que possa ser utilizado junto com redes de sensores acústicas ou como ferramenta de análise e classificação bioacústica em bases de dados com gravações ambientais.

Para atingir este objetivo, os seguintes objetivos específicos foram propostos:

- Aumentar a robustez do método de reconhecimento de sinais bioacústicos eliminando ruídos ambientais da floresta minimizando as distorções causadas nos áudios, incluindo ruídos aleatórios branco e coloridos.

- Melhorar os métodos de análise bioacústico, incluindo: decomposição em funções oscilatórias com as frequências fundamentais dos cantos de cada espécie, filtragem dos ruídos ambientais e reconstrução dos sinais captados pelos sensores.

- Evitar classificar segmentos de áudio irrelevantes, tais como, ruídos ambientais de fundo, desenvolvendo uma técnica de segmentação automaticamente não supervisionada para encontrar as sílabas de cada espécie.

- Identificar os melhores LLDs para segmentar as sílabas de forma não supervisionada em condições adversas, como por exemplos contaminações por ruído braco e colorido.

- Melhorar a capacidade de generalização do modelo de reconhecimento bioacústico.

- Estender o método de reconhecimento de espécies para uma abordagem multirótulo (*multi-label*) de forma a identificar a família e o gênero dos indivíduos.
- Melhorar, ou evitar, classificações incorretas, abordando o problema de classificação colaborativa dos sensores.

1.7 Hipóteses de pesquisa e abordagem proposta

A primeira melhoria proposta é o acréscimo de uma etapa prévia de análise dos sinais bioacústicos, incluindo: decomposição em componentes principais (etapa de análise), filtragem de ruídos aleatórios para melhorar a qualidade do sinal (etapa de seleção dos componentes), e reconstrução dos sinais com os componentes mais relevantes (etapa de síntese). Nossa primeira hipótese é que a filtragem permite gerar um sinal mais limpo e, consequentemente, diminuir a probabilidade de erro na segmentação e classificação (figura 1.1). Para desenvolver nosso método de filtro escolhemos a técnica de decomposição em componentes oscilatórias *Singular Spectrum Analysis* (SSA) (Golyandina et al., 2001). Avaliamos diferentes metodologias para obter a entropia de cada componente e utilizamos estes valores como critério para escolher os componentes que farão parte da reconstrução. Além disso, SSA pode ser utilizada para filtragem não paramétrica e não supervisionada, podendo ser aplicada sem o conhecimento prévio dos sinais. Os fundamentos da análise e o método proposto para filtragem bioacústica com SSA são apresentados no Capítulo 6.

A segunda melhoria proposta é na etapa de segmentação das vocalizações. Primeiramente, realizamos um estudo comparativo entre diferentes LLDs aplicados aos problemas de segmentação não supervisionada em diferentes contextos. Nossa segunda hipótese, é que LLDs baseados em quantificadores de teoria da informação são essenciais para identificar segmentos com somente ruídos ambientais. Assim, simulamos diferentes tipos e níveis de ruídos (brancos e coloridos) e propomos uma nova técnica de segmentação. Esta nova técnica é incremental, fato que permite a detecção das sílabas em tempo real com complexidade computacional de processamento $\mathcal{O}(n)$ e custo de memória constante $\mathcal{O}(1)$, sendo n a quantidade amostras do áudio processado. Esta técnica também é não supervisionada e serve para evitar que o classificador receba segmentos de áudio sem informações relevantes (Colonna et al., 2015). Além disso, a segmentação automática ajuda a diminuir a quantidade de dados transmitidos no caso de ser utilizada em uma rede de sensores, ou simplifica o trabalho manual no caso de ser utilizada em um software de processamento. O método de segmentação é apresentado no Capítulo 4.

Para avaliar as diferentes interações das etapas de nossa abordagem na configuração em série, desenvolvemos uma metodologia que considera estas etapas como partes de um classificador multinível. Com a abordagem proposta, descrita na seção 4.6.2, conseguimos quantificar o impacto que a segmentação causa na taxa do reconhecimento das espécies. Nossa avaliação permite quantificar as perdas em termos de acurácia e taxas de falsos positivos e negativos (Colonna et al., 2015). Desta forma, podemos avaliar nossa terceira hipótese, onde se afirma que a segmentação impacta diretamente na taxa de classificação.

O quarto diferencial de nosso método é a forma de avaliar a capacidade de generalização do modelo de reconhecimento (ou classificador). Neste caso respondemos a seguinte questão: Qual seria a acurácia do reconhecimento num cenário real no qual aparecem novos indivíduos não existentes na base de treino? A resposta foi obtida aplicando validação cruzada por indivíduos (Colonna et al., 2016a). Este tipo especial de validação cruzada é própria dos sistemas bioacústicos e origina nossa quarta hipótese, onde se sugere que a generalização dos modelos de classificação e dos LLDs utilizados para reconhecer os diferentes indivíduos, depende do método de validação escolhido durante os experimentos. Este conceito é apresentado no Capítulo 5.

Abordar o problema de classificação de forma tradicional implica em utilizar um classificador plano (*flat classifier*). Geralmente os classificadores planos apresentam um desempenho aceitável, no entanto, quando o número de espécies aumenta consideravelmente, estes podem apresentar uma diminuição do desempenho causado pela complexidade da função de classificação. Nossa quinta hipótese, afirma que é possível utilizar a taxonomia filogenética para diminuir a complexidade do reconhecimento. Portanto, nos propomos uma abordagem de classificação hierárquica (multinível e multirótulo) para reduzir o espaço de possíveis soluções e simplificar as funções de classificação. Além dos ganhos no reconhecimento das espécies, com o método hierárquico é possível reconhecer a família, o gênero e a espécie de cada amostra, e com isto planejar diferentes estratégias de monitoramento dependendo da região estudada e das espécies envolvidas (Colonna et al., 2016b). Esta proposta é apresentada no Capítulo 5.

A sexta melhoria de nosso método é abordar o problema de classificação de forma colaborativa entre os nós da rede. Desta forma, cada sensor da rede classifica os diferentes eventos acústicos e envia as informações para o líder do *cluster*. Além desta configuração, outras são possíveis, por exemplo, cada sensor pode segmentar o áudio e transmitir as amostras originais do microfone, ou pode transmitir os valores dos descritores acústicos de cada sílaba. Estas possibilidades originam nossa sexta hipótese, que garante que a comunicação entre os sensores vizinhos ajuda no problema de classificação e possibilita reconhecer casos com maior incerteza. Nós propomos realizar a

classificação dentro de cada sensor, e transmitir somente o resultado da classificação, para que o nó líder escolha o melhor candidato, aplicando votação e reduzindo ainda mais a quantidade de dados transmitidos para economizar bateria. A classificação colaborativa é apresentada no Capítulo 5.

Do ponto de vista de aprendizagem de máquina, a combinação das opinião dos diferentes classificadores, que encontram-se nos sensores, é conhecido como *Ensemble Learning*. Nós utilizamos e avaliamos quatro técnicas de votação que não precisam ser supervisionadas: votação maioritária, votação maioritária ponderada, regra geométrica e regra aritmética (Colonna et al., 2014a). Finalmente, aplicamos entropia sobre o resultado das votações para eliminar casos de elevada incerteza. Esta colaboração entre os sensores ajuda a identificar casos confusos nos quais mais de uma espécie estão presentes e vocalizando ao mesmo tempo. A nossa proposta colaborativa inclui avaliar *ensembles* com diferentes classificadores criados com técnicas de *cluster* aplicadas nas RSSF.

1.8 Contribuições principais

As contribuições apresentadas nesta tese são:

- Uma avaliação e comparação de descritores acústicos de baixo nível, que utilizam teoria da informação para quantificar a entropia dos segmentos dos sinais bioacústicos, aplicados ao problema de segmentação não supervisionado em condições adversas de contaminação com ruído branco ou colorido. Nestas avaliações incluímos uma nova abordagem de análise de séries temporais chamada *Permutation Entropy*, e definimos um algoritmo para encontrar o limiar ótimo de segmentação.

- Um algoritmo de segmentação incremental com memória reduzida ($\mathcal{O}(1)$) para extrair as sílabas dos anuros em tempo real (em fluxo contínuo de áudio).

- Uma metodologia de avaliação multinível para quantificar o impacto que a segmentação causa na taxa final de reconhecimento, quando se utiliza um classificador plano.

- Um novo critério de escolha dos componentes principais das vocalização utilizando o espectro singular, aplicado ao problema de filtragem de ruídos ambientais da floresta utilizando diversos critérios de entropia.

- A construção de um banco de filtros com Resposta de Impulso Finito (FIR) otimizados para os sinais bioacústicos de cada espécie e que, quando combinados

com nosso critério de entropia da bases, consiga eliminar os componentes dos sinais com comportamento menos determinístico.

- Uma variante robusta do método SSA, utilizando um coeficiente de correlação robusto que permite extrair facilmente os componentes ambientais de baixa frequência (a tendência ou *trend*) e tolerar níveis maiores de ruído Gaussiano e não Gaussiano, incluindo ruído impulsivo.
- Uma metodologia aprimorada para avaliar o erro de generalização do modelo de classificação baseada na validação cruzada por indivíduos.
- Um método de classificação hierárquica para espécies de anuros capaz de decompor o problema em subproblemas, reconhecendo a família, o gênero e a espécie de cada amostra, e realizamos um estudo comparativo sobre as vantagens e desvantagens dos classificadores hierárquicos comparados com os métodos planos.
- Uma abordagem de monitoramento colaborativa utilizando RSSF incluindo votação e rejeição de casos confusos.

1.9 Organização desta tese

O conteúdo desta tese encontra-se organizado da seguinte maneira.

Os fundamentos e conceitos teóricos sobre os descritores acústicos e as medidas de informação utilizadas nas abordagens de segmentação e classificação encontram-se no Capítulo 2. As diferentes técnicas de aprendizagem de máquina para classificação são descritas no mesmo capítulo. Neste capítulo incluímos os conceitos de ruídos aleatórios aditivos e as diferentes técnicas de filtragem aplicadas, que posteriormente são utilizadas no Capítulo 6.

O estudo dos trabalhos relacionados encontram-se no Capítulo 3. Como a maioria dos trabalhos prévios focam em resolver cada etapa do ACR de forma isolada, deixando lacunas entre a integração dos elementos, separamos o estudo dos trabalhos relacionados de acordo com as etapas da abordagem proposta.

No Capítulo 4, abordamos o problema da segmentação bioacústica automática. Apresentamos um estudo comparativo das diferentes medidas de entropia aplicadas ao problema de segmentação não supervisionada. Ainda neste capítulo, desenvolvemos nossa abordagem de segmentação incremental, incluindo nossa proposta de avaliação multinível que relaciona as etapas de segmentação e classificação.

No Capítulo 5 apresentamos: (a) nossa proposta de avaliação para determinar a capacidade de generalização do modelo de reconhecimento bioacústico, (b) o método

hierárquico para classificação de espécies que permite reconhecer a família e o gênero das amostras, e, por último, (c) abordamos o problema de classificação colaborativa e rejeição de casos confusos entre os nós sensores da rede.

Nosso método de aprimoramento de sinais bioacústicos utilizando decomposição em componentes principais é desenvolvido no Capítulo 6. Dentro deste capítulo incluímos o novo critério proposto para a seleção das bases da reconstrução, que utiliza diferentes metodologias para obter a entropia dos componentes principais. Neste capítulo, também apresentamos a metodologia para encontrar os coeficientes do banco de filtros FIR otimizado para cada espécie. Diversos métodos de filtragem foram comparados quantificando os efeitos causados na segmentação dos sinais e na classificação final das espécies.

Finalmente, as conclusões e direções futuras encontram-se no Capítulo 7.

2

Fundamentos Teóricos

Neste capítulo apresentamos os conceitos necessários para analisar e interpretar as séries temporais bioacústicas (seção 2.1). Apresentamos também as técnicas de classificação utilizadas no reconhecimento das espécies e os princípios de funcionamento das RSSF.

Os fundamentos teóricos nos permitem entender como é realizado o mapeamento das vocalizações dos anuros em conjuntos de características chamados "descritores acústicos" (seção 2.2). Incluímos os cálculos e conceitos relacionados à entropia das permutações utilizada como descritor acústico de baixo nível. A teoria também possibilita compreender a relação entre os descritores e as diferentes funções de classificação utilizadas (seção 2.5).

Descrevemos o modelo de ruído aditivo dos sinais e a formação dos diferentes tipos de ruídos coloridos. Explicamos o funcionamento dos filtros de sinal baseados na subtração espectral média (SMS), na transformada *Wavelet* ($\mathcal{W}$) discreta e no *Singular Spectrum Analysis* (seções 2.3 e 2.4).

Por último, abordamos conceitos gerais sobre as Redes de Sensores que constituem o meio de aplicação de nosso trabalho (seção 2.7). Assim, a partir dos fundamentos a seguir é possível compreender o funcionamento básico da abordagem proposta.

2.1 Representação Digital das Vocalizações

A maioria das espécies de anuros produzem vocalizações intensas e agudas. O ar flui desde a laringe até o saco vocal que funciona como uma cavidade ressonante para reforçar o sinal (Gerhardt, 1975). O trato vocal modula um sinal portador de informação que viaja pelo ar até alcançar o microfone do sensor. A vocalização captada pelo microfone é um conjunto de valores ordenados no tempo representando as medições de pressão acústica que relacionam-se com a amplitude da onda. A medição do sinal analógico é chamado de série temporal discreta definida como $x[n] = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$. Em um hardware digital os valores dos índices temporais tornam-se valores de tempos discretos no intervalo $1 \leq i \leq n$ e, consequentemente, a série pode ser armazenada como um vetor de n dimensões.

O hardware de aquisição de sinal possui um módulo conversor analógico-digital (ADC). Este módulo discretiza a amplitude do sinal analógico em níveis, representando cada x_i pelo seu valor binário (ou de ponto flutuante) mais próximo, onde o índice temporal é representado pelo subíndice i. Essa conversão é chamada de quantização. A quantização possui uma perda de informação associada, denominada sinal-ruído de quantização. Uma análise mais detalhada sobre o impacto da quantização na taxa de reconhecimento das espécies pode ser encontrada em Colonna et al. (2012).

A representação da série temporal no formato de vetor possibilita a aplicação de diferentes técnicas de análise que serão apresentadas nas próximas seções. Também existem representações visuais das formas de onda com os valores das amplitudes e frequências (figura 2.1). O espectrograma é a variação da energia das frequências como função do tempo. Este, é calculado aplicando janelamento sucessivo a x, com ou sem sobreposição entre as janelas (*overlap*). Cada janela é denominada *frame*, e possui um tamanho pré-definido (N) e seu próprio rótulo temporal (ou *time-stamp*). Assim, os valores sucessivos contidos dentro do *frame* correspondem ao conjunto $\{x_{i-N/2}, \ldots, x_i, \ldots, x_{i+N/2}\}$.

A partir dos pontos que compõem um *frame*, é possível calcular diferentes características do sinal que os representem de forma única. Obviamente, existem características que separam melhor um sinal de outro. A nova representação do sinal pelo valor das características (ou descritores acústicos) constitui um mapeamento, ou redução de informação, e pode ser utilizado com o propósito segmentar, filtrar ou classificar o sinal.

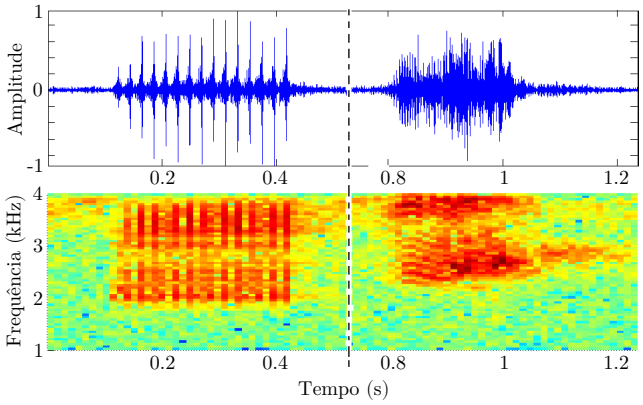


Figura 2.1. Forma de onda (superior) e espectrograma (inferior) de duas espécies diferentes.

2.2 Descritores Acústicos

Devido ao aumento de popularidade da aplicação de técnicas de aprendizagem de máquina, nos problemas de reconhecimento de fala, os termos "conjunto de características" ou "descritores acústicos" tornaram-se sinônimos. No entanto, os descritores acústicos podem ser classificados como de baixo nível (LLD) ou de alto nível (HLD) (Amatriain, 2004).

O descritores de baixo nível estão relacionados com as características físicas dos sinais. Amatriain (2004) classificou os LLD em três categorias de granularidade: os calculados ponto-a-ponto do sinal (ou instantâneos), *frame-a-frame* e segmento-a-segmento. Alguns descritores, como a energia do sinal (E) ou a taxa de cruzamento por zero (ZCR), podem ser calculados nos três níveis de granularidade. Entretanto, descritores como os Coeficientes Mel (MFCCs) ou a Entropia das Permutações (PE) somente podem ser calculados frame-a-frame ou segmento-a-segmento. Além da granularidade, os descritores podem ser agrupados pelo domínio de aplicação do sinal, seja este temporal ou espectral.

Os descritores de alto nível podem ter significados semânticos ou sintáticos, diferentemente dos LLD que somente carregam informações sobre as características físicas da série temporal. No contexto de processamento áudio, os HLD semânticos relacionam-se à interpretação ou significado de um trecho do áudio em um determinado contexto (exemplo: o som de um piano). Enquanto que os HLD sintáticos possuem informações concretas que o usuário final pode entender sem saber o significado total do áudio e também sem conhecer o processamento digital do sinal (exemplo: som do acorde "DO" executado em um violão). Um exemplo deste último caso é o timbre da voz, que pode ser obtido como combinação de vários LLD. Pela natureza de nossa

aplicação, na qual não existe interpretação semântica possível sobre o que um animal tenta dizer, focamos o estudo aplicando os LLD (Lasseck, 2014).

2.2.1 Energia e Taxa de Cruzamento por Zero

A energia de um sinal acústico quantifica as amplitudes acumuladas dentro do um *frame* (Vaca-Castaño and Rodriguez, 2010, Colonna et al., 2012, Jaafar and Ramli, 2013). Esta característica pertence ao domínio temporal e pode ser obtida como:

$$\mathrm{E} = \sum_{i=1}^{N} x_i^2, \tag{2.1}$$

onde x_i são os valores de amplitude que assume a série temporal x nos tempos discretos i. Acrescentado o fator multiplicativo $1/N$ ao somatório, Jaafar and Ramli (2013) se referem a esta característica como a energia média do *frame*.

O ZCR é outro descritor acústico obtido no domínio Temporal. Este valor, representa quantas vezes um sinal mudou de positivo para negativo e vice-versa (Huang et al., 2009). Desta forma, sinais de frequência maior tendem a possuir valores maiores de ZCR. O cálculo é realizado aplicando:

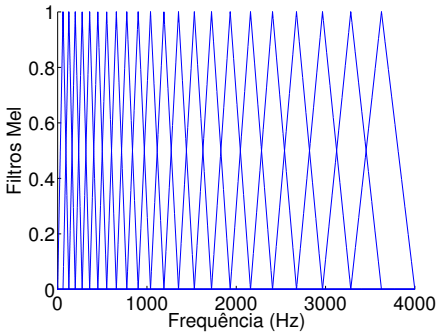
$$\mathrm{ZCR} = \frac{1}{2} \sum_{i=1}^{N-1} |\mathrm{sign}(x_i) - \mathrm{sign}(x_{i+1})|, \tag{2.2}$$

sendo a função *sign* é definida como:

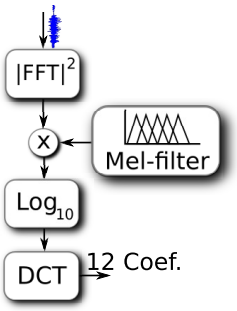
$$\mathrm{sign}(x_i) = \begin{cases} +1 & \text{se } x_i \geq 0 \\ -1 & \text{se } x_i < 0 \end{cases}. \tag{2.3}$$

Estes descritores contribuem com informações básicas sobre os sinais em questão, sendo pouco discriminativos para os problemas de reconhecimento das espécies (Colonna et al., 2012). Entretanto a complexidade computacional reduzida ($\mathcal{O}(N)$) e o consumo de memória constante ($\mathcal{O}(1)$), os torna atrativos para serem aplicados em problemas diferentes, mais abrangentes e com menos classes, como a segmentação automática do sinal (Jaafar and Ramli, 2013, Colonna et al., 2015).

O cálculo da Entropia das Permutações também pertence ao domínio temporal, mas por tratar-se de um descritor pouco explorado aplicado ao contexto bioacústico, é fornecida uma descrição mais detalhada na seção 2.2.5.



(a) Banco de filtros Mel.



(b) Bloco de extração dos coeficientes Mel.

Figura 2.2. Extração das características utilizando os MFCC.

2.2.2 Coeficientes Mel

Os sistemas de reconhecimento de fala humana (ASR), baseiam seu funcionamento no uso dos Coeficientes Mel (MFCCs) para a correta representação e classificação dos sinais de fala (Rabiner and Schafer, 2007). Os MFCCs foram originalmente desenvolvidos por Davis and Mermelstein (1980), sendo estes os descritores mais populares por causa de sua eficiência computacional, robustez aos ruídos e capacidade de capturar ressonâncias do trato vocal.

Estes coeficientes representam melhor os sinais da fala que os descritores que utilizam uma escala linear de frequência. A escala logarítmica Mel é consistente com a forma de percepção humana da voz, servindo para sinais periódicos e aperiódicos, reduzindo a quantidade de informação necessária para descrever o sinal sem uma perda de informação relevante (Cai et al., 2007). A ideia básica desta técnica é realizar uma análise espectral usando um banco de filtros triangulares espaçados logaritmicamente nas frequências, conforme figura 2.2(a).

O primeiro passo do cálculo dos MFCCs é obter a Densidade Espectral de Potência do sinal (PSD). Para calcular a PSD de forma digital deve-se utilizar a Transformada Discreta de Fourier:

$$X_f = \sum_{i=1}^{N} x_i e^{\frac{-i2\pi k(i-1)}{N}}, \quad k = \{0, 1, \ldots, N-1\}, \tag{2.4}$$

onde x_i são as amostras do áudio e X_f é o valor do componente de frequência no ponto $f = \frac{2\pi k}{N}$ do espectro. Computacionalmente, o algoritmo *Fast Fourier Transform* (FFT) realiza o cálculo numérico da transformada entregando como saída em uma série de números complexos. Finalmente a PSD é obtida como $|X_f|^2$.

No segundo bloco da figura 2.2(b) é aplicado um conjunto de filtros triangulares a $|X_f|^2$ distribuídos segundo a escala logarítmica Mel (Cowling and Sitte, 2003) definida como:

$$f_{mel} = 1127 \ \ln\left(1 + \frac{f}{700}\right). \tag{2.5}$$

A aplicação do banco de filtros sobre o espectro do sinal produz um valor M_r de saída para cada filtro triangular (figura 2.2(a)). Após a obtenção dos valores M_r, é aplicado o logaritmo e a transformada discreta do cosseno (DCT), para obter os R coeficientes finais (Rabiner and Schafer, 2007):

$$\mathrm{MFCC} = \frac{1}{R}\sum_{r=1}^{R} \log\left(\mathrm{M}_r\right) \cos\left(\frac{2\pi}{R}\left(r + \frac{1}{2}\right)c\right), \tag{2.6}$$

onde c é a quantidade de coeficientes desejados e R é a quantidade de filtros triangulares utilizados. Tipicamente $R \geq c$. A DCT é um método utilizado no processamento digital de sinais e de imagens pelo fato de proporcionar compressão de dados. Em outras palavras, o sinal original (neste caso a saída dos filtros) é representado em novos eixos perpendiculares. A função do cosseno é gerar novas bases ortogonais e atribuir um valor nestas bases para cada conjunto fornecido. Finalmente, é escolhido o sub-conjunto de coeficientes com maior valor dentre os R coeficientes para realizar a classificação.

Considerando que operações de adição e multiplicação possuem custo constante, é necessário, para se obter a saída dos filtros M_r, a realização de N multiplicações. Para a obtenção de cada coeficiente MFCC, são necessárias mais R multiplicações. Por fim, para obter o custo dos MFCC tem-se que somar o custo prévio da FFT, que no melhor caso, quando N é potência de 2, o custo é $\mathcal{O}(N \log N)$ (Cooley and Tukey, 1965). O que resulta em uma complexidade final assintótica dos MFCCs dominada pela FFT.

2.2.3 Entropia Espectral

A Entropia Espectral (H_f) foi utilizada por Shen et al. (1998) para segmentar sinais de fala e também por Sueur et al. (2008) para calcular o índice de atividade bioacústica em diferentes gravações. Esta característica é calculada frame-a-frame do sinal. Para o cálculo de H_f deve-se obter primeiro a transformada de Fourier do sinal $x \xrightarrow{\mathcal{F}} X_f$ e

normalizá-la conforme:

$$X_f = \frac{|X_f|}{\sum_{f=0}^{f_s/2} |X_f|}, \quad 0 \leq f \leq f_s/2, \tag{2.7}$$

onde f indica o índice das frequências e f_s a frequência máxima de amostragem. Posteriormente a entropia é obtida como:

$$\mathrm{H}_f = -\sum_{f=0}^{f_s/2} \frac{|X_f| \cdot \log_2 |X_f|}{\log_2(N)}, \tag{2.8}$$

na qual N é o tamanho do *frame* e também a quantidade de pontos no eixo de frequências f. Esta característica representa a concentração de energia nas bandas espectrais, resultando $\mathrm{H}_f \approx 0$ para espectros concentrados e $\mathrm{H}_f \approx 1$ para espectros dispersos. No caso de se utilizar a transformada rápida de Fourier (FFT) o espectro de frequências obtido X_f é simétrico respeito ao centro $N/2$, portanto a equação 2.8 pode ser reduzida a

$$\mathrm{H}_f = -\sum_{f=0}^{N/2} \frac{|X_f| \cdot \log_2 |X_f|}{\log_2 N/2}. \tag{2.9}$$

2.2.4 Entropia temporal

A Entropia das amplitudes (H_t), ou entropia temporal, é obtida a partir das oscilações da envoltória do sinal (Sueur et al., 2008). A envoltória do sinal, neste caso, é calculada utilizando a função analítica χ do sinal $x(t)$, definida como:

$$\chi(t) = x(t) + iX_{\mathcal{H}}(t), \tag{2.10}$$

onde $i^2 = -1$ e $X_{\mathcal{H}}$[1] é a transformada *Hilbert* ($\mathcal{H}$) de x definida pela convolução (Feldman, 1994):

$$X_{\mathcal{H}} = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau. \tag{2.11}$$

Esta transformação foi utilizada por Potamitis et al. (2014) com o nome de *Hilbert follower* para segmentar sinais acústicos identificando as sílabas. A partir da equação 2.10 a envoltória é obtida como $|\chi_i| = \sqrt{x_i^2 + X_{\mathcal{H}i}^2}$, $\forall\, i \in [1, N]$ (Benitez et al., 2001) e com

[1] A dependência temporal (t) será omitida das equações seguintes para melhorar a clareza do texto.

esta a função de densidade de probabilidades (PDF) das amplitudes:

$$A_i = \frac{|\chi_i|}{\sum_{i=1}^{N} |\chi_i|}, \tag{2.12}$$

na qual é aplicada a entropia de Shannon (equação 2.53, página 55):

$$\mathrm{H}_t = -\sum_{i=1}^{N} \frac{A_i \ \log_2 A_i}{\log_2 N}. \tag{2.13}$$

2.2.5 Entropia das Permutações

A Entropia das Permutação (PE) caracteriza a dinâmica de uma série temporal (Bandt and Pompe, 2002, Soriano et al., 2011). Esse quantificador, juntamente com a complexidade estatística de permutação, permite comparar ou distinguir comportamentos determinísticos, estocásticos ou caóticos em séries temporais (Rosso et al., 2007, 2010, Labate et al., 2013). A PE é um LLD do domínio temporal, ou seja, não é necessária nenhuma transformação do sinal. O cálculo considera o valor das amostras x_i e a ordem (ou desordem) destes valores em relação aos vizinhos próximos. Esta técnica possui dois parâmetros principais: o *time-embedding* (m), que determina o tamanho dos padrões que serão ordenados e comparados, e o *time-lag* (τ), que representa as unidades de tempo entre os valores que compõem os padrões ordinais.

O procedimento desenvolvido por Bandt and Pompe (2002) para obter a PE começa decompondo a série temporal em um conjunto de padrões ordinais (chamados também símbolos ou *motif*), criando uma nova representação simbólica $X_j^{m,\tau} = \{x_i, x_{i+\tau}, \ldots, x_{i+(m-1)\tau}\}$ com $J = N-(m-1)\tau$ padrões que pertencem a um dicionário $\Pi = \pi_1, \pi_2, \ldots, \pi_{m!}$ com $m!$ símbolos diferentes. Nesta nova representação é possível contar a frequência de ocorrência de cada símbolo (π_i) e dividí-la por J para obter as probabilidades de acordo com a equação:

$$p(\pi_i) = \frac{\|\{j : j \leq J, \ \mathrm{type}(X_j^{m,\tau}) = \pi_i\|}{J}, \tag{2.14}$$

na qual type$(\cdot)$ representa o mapeamento entre o espaço dos padrões ordinais e o espaço dos símbolos, e $\|\cdot\|$ é a cardinalidade do conjunto Π. Após obter a probabilidade de cada símbolo do dicionário é aplicado o cálculo de entropia de *Shannon*:

$$\mathrm{H}_s = -\sum_{i=1}^{\Pi} \frac{p(\pi_i)\ln p(\pi_i)}{\ln m!}, \tag{2.15}$$

para mensurar a incerteza ou previsibilidade da série temporal.

Para ilustrar o cálculo da PE vamos a utilizar um exemplo com $m = 3$ e $\tau = 1$. Seja e série:

$$x_i = \{x_1 = 1, x_2 = 3, x_3 = 2, x_4 = 1, x_5 = 3, x_6 = 1, x_7 = -1, x_8 = 1, x_9 = 3\},$$

os passos para obter o valor de PE são:

1. Criar um dicionário de símbolos com as possíveis permutações de três dígitos $\Pi = \{\pi_1 = (1,2,3), \pi_2 = (1,3,2), \pi_3 = (2,1,3), \pi_4 = (2,3,1), \pi_5 = (3,1,2), \pi_6 = (3,2,1)\}$;

2. Para cada instante de tempo (i) percorrer a série temporal formando sub-conjuntos de três elementos $(x_i, x_{i+1\tau}, x_{i+2\tau})$;

3. Ordenar cada sub-conjunto de forma ascendente, ex: para $i = 1$ o primeiro sub-conjunto de x seria $\{x_1 = 1, x_2 = 3, x_3 = 2\}$, que após a ordenação resulta $\{x_1 = 1, x_3 = 2, x_2 = 3\}$;

4. Obter os índices dos valores após a ordenação $(1, 3, 2)$ e encontrar dentro do dicionário Π, o símbolo que corresponde com esta ordem, que neste caso seria π_2;

5. Acumular um contador correspondente às ocorrências de cada π_i, para obter no final a frequência relativa de cada símbolo (ou histograma dos padrões ordinais da série);

6. Repetir os passos 1 a 5 até o final da série temporal (ou *frame*).

O resultado dos passos descritos são as frequências de ocorrência de cada padrão ordinal, e pode ser interpretado como o histograma de Π. Normalizando o histograma obtemos a função de distribuição de probabilidades dos π_i da série. A figura 2.3 resume os passos do procedimento para gerar um histograma de padrões ordinais de nosso exemplo.

Este procedimento permite verificar se a existência de um comportamento ordenado (caso o histograma esteja concentrado em algum padrão) ou completamente aleatório (caso o histograma seja uniforme) dos valores. Finalmente, é útil calcular o valor de entropia do histograma para quantificar o grau de desordem da série. Assim, sinais com tendências (*trend*) marcadas resultam em valores $\mathrm{H} \approx 0$.

Existem casos particulares de séries temporais nos quais alguns π_i nunca aparecem. Estes padrões foram batizados como "proibidos" por Amigó et al. (2008). A ausência de algum padrão pode ser devido a um comportamento específico da série,

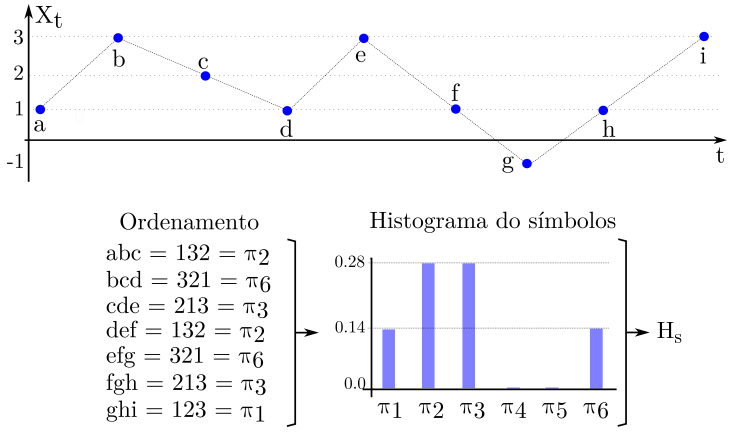


Figura 2.3. Exemplo de geração do histograma dos símbolos (ou padrões ordinais). Figura extraída do poster apresentado na Colonna et al. (2014b).

ou por não cumprir com a condição $m \ll N$ para garantir que existe uma probabilidade maior a zero de observar cada padrão. No outro extremo, encontra-se o caso particular em que a série é completamente aleatória, como por exemplo o ruído branco, no qual existe uma probabilidade aproximadamente uniforme de ocorrência para cada π_i (Rosso et al., 2010).

Voltando ao exemplo anterior e escolhendo um valor diferente para τ, por exemplo τ=2, o primeiro padrão formado seria x_1, x_3, x_5, e após a ordenação os índices resultariam iguais ao símbolo π_1, e não π_2. Em outras palavras, o efeito do parâmetro τ é permitir analisar a série em diferentes escalas temporais (Zunino et al., 2012).

O algoritmo completo para obter a PE possui dois parâmetros adicionais: o tamanho do *frame* e a quantidade de sobreposição entre *frames* vizinhos. O algoritmo 1 apresenta um pseudo-código para calcular H_s.

Algoritmo 1 - Pseudo-código PE

1: $\texttt{PE}(x, \Pi, m, \tau)$
2: $\texttt{para i} \leftarrow \texttt{1 até o comprimento de } x - \tau(m-1) \texttt{ faça}$
3: $\quad indices \leftarrow \texttt{sort}(x_{(i:\tau:i+\tau(m-1))})$
4: $\quad \pi_i \leftarrow \texttt{HashTable}(indices, \Pi)$
5: $\quad histograma(\pi_i) \leftarrow histograma(\pi_i) + 1$
6: $\texttt{fim para}$
7: $p(\pi) \leftarrow \frac{histograma}{\sum(histograma)}$
8: $\mathrm{H}_s \leftarrow \frac{-\sum(p(\pi) \quad \log(p(\pi)))}{\log(m!)}$

O laço "para" (linhas 2-6) do algoritmo 1 percorre a série $x(t)$ formando os padrões de tamanho m. A cada nova iteração os padrões são ordenados como foi explicado no exemplo anterior e os índices resultantes desta ordenação são utilizados como a chave

(*key*) de uma tabela *Hash* que retorna o π_i correspondente. A tabela *Hash* é criada antes da execução do algoritmo e após definir m, sendo o tamanho desta igual a $m!$. O mapeamento das chaves com seus padrões correspondentes possui custo $\mathcal{O}(1)$, portanto a complexidade depende somente do tamanho da série (N), resultando $\mathcal{O}(N)$. Um algoritmo diferente foi descrito por Parlitz et al. (2012) com complexidade $\mathcal{O}(mN)$.

Variações da PE

A metodologia original da PE teve duas extensões: a inclusão de pesos para separar melhor os padrões de diferentes amplitudes, denominada PE ponderada (WPE) e a derivação do quantificador de mínima entropia das permutações (PME) (Fadlallah et al., 2013, Zunino et al., 2015).

A figura 2.4(a) apresenta dois exemplos de padrões ordinais com $m = 3$ ($\pi_i = 1, 2, 3$). Fadlallah et al. (2013) notaram que a metodologia de cálculo da PE é independente das amplitudes do sinal e modificaram a equação 2.14 para incluir a variância do padrão ordinal em forma de peso w_j. Esta modificação permite identificar melhor mudanças abruptas dos sinas, como pode ser o caso dos áudios bioacústicos (figura 2.4(b)). A nova equação da WPE resulta:

$$p(\pi_i) = \frac{\|\{j : j \leq J, \text{type}(X_j^{m,\tau}) = w_j\pi_i\|}{J}, \tag{2.16}$$

na qual o peso w_j é:

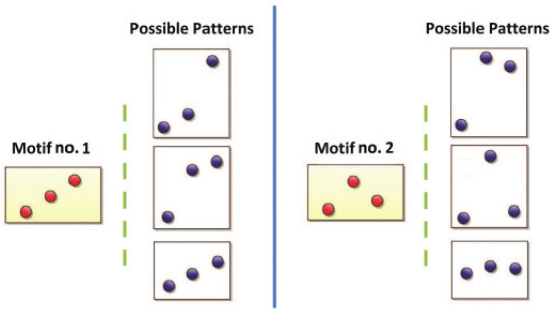
$$w_j = \frac{1}{m}\sum_{k=1}^{m}(x_{j+(k-1)\tau} - \bar{X}_j^{m,\tau})^2, \tag{2.17}$$

e a entropia $\mathrm{H_{WPE}}$ final obtêm-se aplicando a equação de Shannon 2.53.

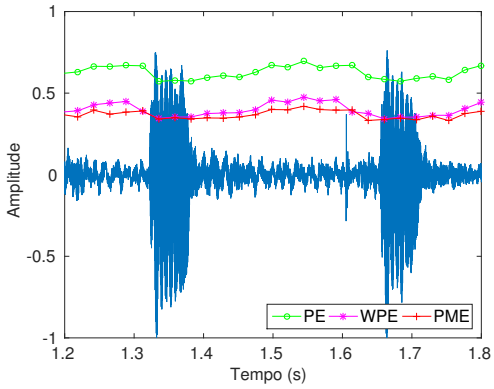
A PME foi apresentada como alternativa à PE para identificar a existência de correlações temporais ocultas nas séries temporais. Este novo quantificador conserva as propriedades da PE original: simplicidade de cálculo; baixa complexidade computacional; robustez os ruídos; e é invariante a transformações não lineares monótonas (Zunino et al., 2015). O cálculo é realizado aplicando a seguinte equação:

$$\mathrm{H_{PME}} = -\ln\Big(\max\big(p(\pi_i)\big)\Big). \tag{2.18}$$

Comparações de diferentes resultados obtidos por Zunino et al. (2015) mostraram que a PME possui melhor desempenho em situações nas quais o nível de ruído é maior, ou igual, à amplitude do sinal, confirmando a sua utilidade como uma alternativa, ou medida complementar, de correlações temporais.



(a) Figura extraída de Fadlallah et al. (2013).



(b) Vocalização da espécie Adenomera hylaedactyla.

Figura 2.4. (a) Exemplo de dois padrões ordinais iguais mas com diferente amplitude. (b) Exemplo de PE, WPE, e PME calculados a partir dos *frames* do sinal.

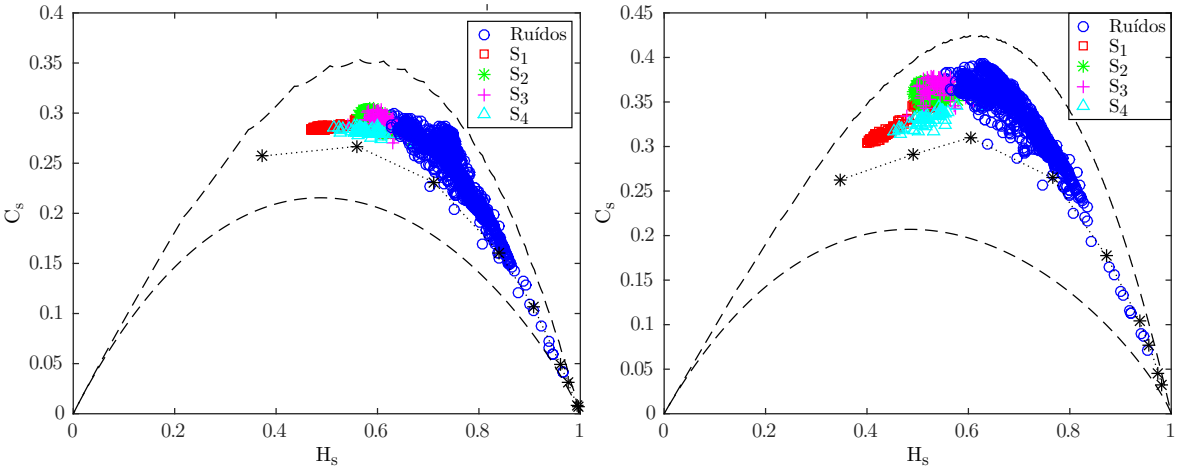
Em nossas abordagens de segmentação automática e filtragem, não supervisionadas, utilizamos estes quantificadores para separar *frames* e componentes principais (PCs) com padrões determinísticos de aqueles totalmente aleatórios (capítulo 4 e capítulo 6).

2.2.6 Medida de complexidade estatística

A partir da função de distribuição de probabilidades PDF, Rosso et al. (2007) definiram o conceito de Complexidade do sistema (*Statistical Complexity Measure*). Esta complexidade combina o conceito de "desordem", calculado pela entropia (H), com o conceito de "desequilíbrio" do sistema, obtido pela divergência. O desequilíbrio é a divergência entre dois histogramas, o histograma P da série temporal e um histograma de comparação (P_e) que representa o estado estacionário do sistema. O histograma de referência corresponde a uma distribuição uniforme, representando uma série aleatória na qual todas as ocorrências possuem o mesmo valor de probabilidades (e. g. um ruído aleatório). A complexidade estatística resulta definida como:

$$\mathrm{C}_s = \mathrm{Q}[P, P_e]\mathrm{H}_s[P], \tag{2.19}$$

na qual $\mathrm{H}_s[P]$ é a entropia de Shannon dos p_i e $\mathrm{Q}[P, P_e]$ é a divergência entre os histogramas P e P_e. Para o cálculo de divergência Q, Rosso et al. (2010) recomendam



(a) PE com parâmetros $m = 4$ e $\tau = 1$.

(b) PE com parâmetros $m = 5$ e $\tau = 1$.

Figura 2.5. Sílabas de quatro espécies diferentes de anuros, mais ruídos da floresta, representados no plano HxC.

utilizar a equação de Jensen-Shannon definida como:

$$\mathrm{Q}[P, P_e] = Q_0 \left\{ \mathrm{H}\left[\frac{P + P_e}{2}\right] - \frac{\mathrm{H}[P_e]}{2} - \frac{\mathrm{H}[P]}{2} \right\}, \tag{2.20}$$

na qual Q_0 é uma constante de normalização. Note-se que a metodologia de Bandt and Pompe (2002) permite calcular a função de distribuição de probabilidades (PDF) dos padrões ordinais ($P = \{p_{\pi_i}\}$) que representam as relações temporais da série, embora outras metodologias podem ser aplicadas para obter o histograma P, por exemplo: *Wavelet Entropy* (Rosso et al., 2001), *Symbolic Aggregate Approximation* (SAX) (Lin et al., 2003), ou *Local Binary Patterns* (LBP) (Esfahanian et al., 2013), entre outras.

A complexidade C_s pode ser representada visualmente no plano HxC. Neste plano é possível visualizar melhor a estrutura física da série temporal e comparar diferentes séries. A posição de cada ponto no plano possui uma interpretação física, assim pontos situados no canto inferior direito corresponde-se com uma série temporal de comportamento aleatório similar aos ruídos. A figura 2.5 apresenta um exemplo do plano de complexidade no qual foi utilizada a metodologia PE para obter os p_{π_i} do histograma de algumas séries de vocalizações de anuros.

2.3 Ruídos aleatórios e filtros de sinal

Diferentes fenômenos físicos naturais produzem diversos tipos de ruído que afetam a qualidade das medições. Neste caso, a gravação de som bioacústicos não é uma exceção. O modelo geral de ruído aditivo para sinais bioacústicos assume a forma $y = x+\xi$. Esta suposição considera que o sinal "limpo" x e o ruído ambiental ξ não são correlacionados. Esta suposição é razoável na maioria dos casos práticos nos quais o sinal e ruído são gerados por fontes independentes (Vaseghi, 2008).

No domínio espectral o sinal mais o ruído resulta

$$Y_f = X_f + \xi_f, \tag{2.21}$$

no qual Y_f, X_f e ξ_f podem ser variáveis complexas. A frequência f é obtida pela relação $\frac{2\pi k f_s}{N}$ na qual $k = 0, 1, 2, \ldots, N-1$, f_s é a frequência de amostragem em Hz e N o tamanho do *frame*.

A Densidade Espectral de Potência (PSD) dos sinais e dos ruídos ambientais é obtida a partir da magnitude de espectro como $|Y_f|^2 = |X_f|^2 + |\xi_f|^2$. A PSD pode ser calculada para o sinal completo ou para cada um dos seus *frames*. Com propósitos de simplificação da notação cada vez que seja referida uma variável no domínio espectral utilizaremos X_f no lugar de $|X_f|^2$.

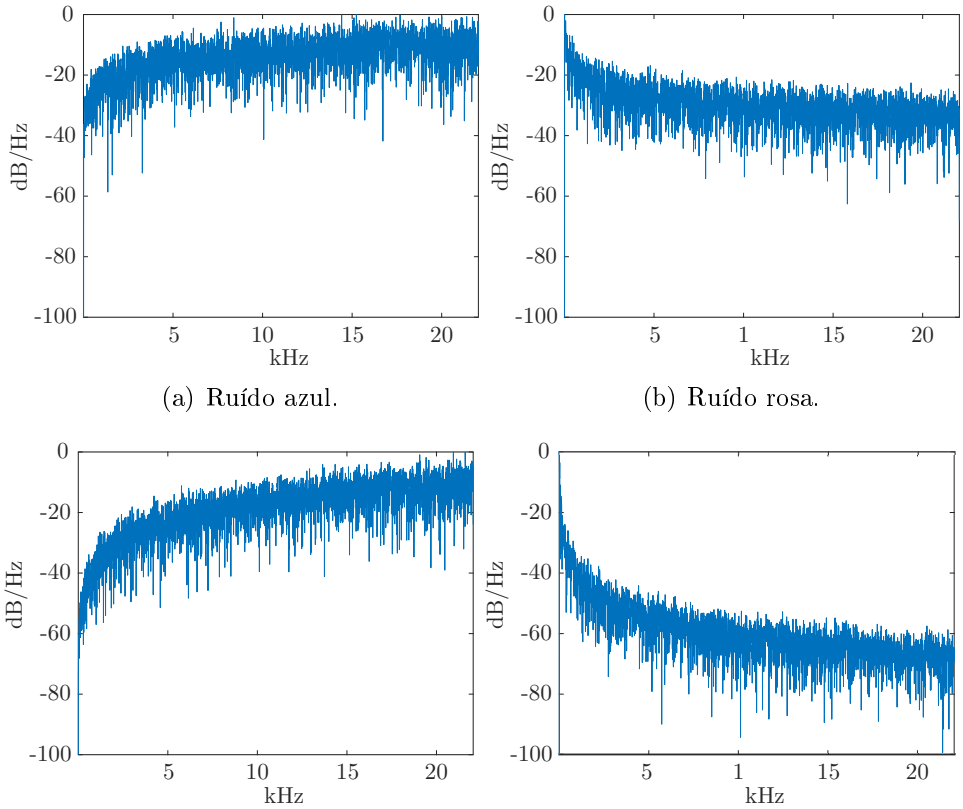
2.3.1 Diferentes tipos de ruídos

Muitos fenômenos físicos naturais produzem diversos tipos de ruído (Lowen and Teich, 1990, Vasseur and Yodzis, 2004), *e.g.*, ruído branco ou colorido. As gravações dos sinais acústicos não são uma exceção (Voss and Clarke, 1978). Neste caso, as séries temporais de ruído podem ser caracterizadas por uma variável aleatória com Densidade Espectral de Potência (PSD) que obedece a uma lei de potência da forma:

$$\xi_f^\alpha = \frac{\mathrm{L}}{|f|^\alpha}, \tag{2.22}$$

no qual L uma constante proporcional à variância do processo aleatório e o exponente α é um número real no intervalo $[-2, 2]$ (Kasdin, 1995, Plaszczynski, 2007, R2015a, 2015). Tipicamente este tipo de sinais é utilizado para teste de circuítos de áudio.

Este ruído chama-se colorido pela semelhança com o espectro de frequências da luz visível. Dependendo do valor de α será a forma da PSD do ruído (figura 2.6), assim este pode ser classificado segundo a cor (Vaseghi, 2008):



(a) Ruído azul.

(b) Ruído rosa.

(c) Ruído violeta.

(d) Ruído vermelho.

Figura 2.6. PSD dos diferentes tipos de ruídos coloridos de acordo com a equação 2.22.

- $\alpha = 0$ modela o ruído branco contendo quantidade igual de energia em todas as bandas de frequência;

- $\alpha = 1$ modela o ruído rosa, o qual possui o mesmo nível de pressão acústica em cada oitava, diminuindo a energia a medida que a frequência aumenta;

- $\alpha = 2$, modela o ruído vermelho no qual a energia das bandas de frequências diminuem conforme f aumenta, este é comum em gravações oceanográficas para descreve o ruído dos ambiente submarinos Rudnick and Davis (2003);

- $\alpha = -1$ modela o ruído azul, o qual contém mais energia a medida que as frequência aumentam (Ballón et al., 2011); e

- $\alpha = -2$ modela o ruído violeta, o qual aumenta ainda mais a energia em altas frequências comparado com o ruído azul.

2.3.2 Relação Sinal-Ruído

A Relação Sinal-Ruído (SNR) serve para quantificar a quantidade de ruído presente nas gravações em dB. Esta medida também é útil para quantificar a qualidade das gravações e pode ser obtida através de:

$$\mathrm{SNR} = 20 \log_{10} \frac{\sigma_x}{\sigma_\xi}, \tag{2.23}$$

onde σ_x e σ_ξ correspondem ao desvio padrão do sinal original e do ruído aditivo respectivamente. Nos experimentos pode-se variar esta razão para simular diferentes condições de ruído ambiental. O ponto crítico que demarca a separação entre baixa e alta contaminação dos sinais é dado pela relação $\sigma_x \approx \sigma_n$, a qual produz um $\mathrm{SNR} = 0\,\mathrm{dB}$.

2.3.3 Ruído impulsivo

O ruído impulsivo é a ocorrência esparsa de impulsos instantâneos com elevada energia e curta duração. Tipicamente, os impulsos são denotados por $\pm\delta_k$, com k representando a posição temporal do impulso. Se a amplitude do sinal x for normalizado no intervalo $[-1, 1]$, então $\delta(\cdot)$ pode assumir aleatoriamente os valores máximos ou mínimos ± 1. O tempo de ocorrência k entre cada impulso pode ser representado por uma variável aleatória com distribuição uniforme. A densidade deste ruído é definida pela razão entre a quantidade total de impulsos K que afetam o sinal e o comprimento N:

$$d_\delta = \frac{1}{N} \sum_{k}^{K} |\delta_k|. \tag{2.24}$$

Esta ruído é não correlacionado com o sinal acústico e representa uma condição extremamente adversa aparecendo devido a várias causas, por exemplo, descargas eléctricas que afetam os circuítos dos sensores, ou sons elevados causados pelas condições meteorológicas ou ambientais.

2.3.4 Filtros FIR

Uma tarefa relevante no desenvolvimento de sistemas para processamento de sinais digitais é o desenho de filtros com Resposta de Impulso Finito (FIR). Em bioacústica, este tipo de filtro é frequentemente usado para remover o ruído de alta frequência das gravações. Um filtro FIR é um filtro cuja resposta a qualquer sinal de entrada com comprimento finito é de duração finita. Este filtro possui propriedades úteis, tais como: não possui retroalimentação, a saída é estável e causal, a fase é linear

quando a sequência de coeficientes da resposta ao impulso é simétrica e possui uma implementação algorítmica simples (Oppenheim and Schafer, 2010).

A aplicação do filtro ao sinal é dada pela convolução discreta:

$$y[n] = h * x[n], \tag{2.25}$$

definida como:

$$y_i = \sum_{j=1}^{M} b_j \cdot x[i-j], \quad \forall i \in \mathbb{Z} : 1 \leq i \leq n, \tag{2.26}$$

onde x e y são os sinais de entrada e saída com comprimento máximo n, e $b_1, b_2, \ldots, b_M$ são os coeficientes que caracterizam a resposta ao impulso h do filtro com grau máximo M. Assim, a resposta em frequência do filtro pode ser interpretada como a função de transferência avaliada em $z = e^{j\omega}$. A figura 2.7 ilustra a forma de implementação direta do filtro mediante um diagrama de blocos. Os termos $x[i-j]$ são as amostras do sinal de entrada retrasadas pelos efeito do operador z^{-1} para poder aplicar multiplicação dos coeficientes.

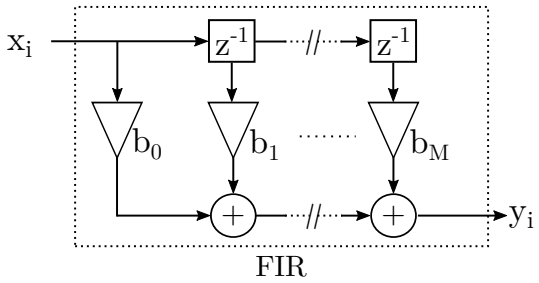


Figura 2.7. Exemplo de um diagrama em blocos de um filtro FIR discreto de ordem M^{th}. A parte superior é uma linha de $M-1$ atrasos, onde cada atraso é uma unidade temporal causada pelo operador temporal z^{-1} usando a notação da transformada $\mathcal{Z}$.

2.3.5 Eigenfilters

O método *eigenfilter* foi desenvolvido para projetar e desenhar filtros FIR digitais (Tkacenko et al., 2003). Este método baseia-se no cálculo dos autovetores de uma matriz complexa simétrica (i.e. uma matriz hermitiana) para obter os coeficientes do filtro. O cálculo de tais coeficientes é não trivial.

Nos trabalhos originais de Tomé et al. (2010, 2011) o cálculo foi simplificado utilizando SSA para obter os autovetores necessários. Nesses trabalhos, foi estabelecida uma correspondência entre SSA e um sistema Linear Invariante no Tempo (LTI), e

comprovou-se que as operações de decomposição e reconstrução do SSA, são equivalentes a uma única operação de convolução, o que simplifica a obtenção dos coeficientes dos *eigenfilters*. A partir dos autovetores, os autores chegaram a uma fórmula fechada para obter a resposta impulsiva de um filtro adaptativo FIR, ótimo para o sinal de entrada. Particularmente, no trabalho de Tomé et al. (2011), foi ilustrado um exemplo de aplicação desses filtros adaptativos FIR a um problema de aprimoramento de fala humana.

Com esta metodologia, a resposta em frequência de um *eigenfilter* depende do sinal de entrada. Isto é, a função FIR é ótima para o sinal a partir do qual obtiveram-se os autovetores, mas se o sinal de entrada muda a eficiência decai. Características adicionais dos *eigenfilter*, tais como: a diferença de atenuação entre a banda de frequências aceitas e atenuadas, a frequência central da banda passante principal, a largura de banda do filtro e outras características não foram explicitas. Portanto, uma descrição mais detalhada da resposta em frequência desses *eigenfilters* construídos utilizando sinais bioacústicos é mostrada na seção 6.7.4.

2.3.6 Subtração espectral

O filtro conhecido como *Spectral Subtraction* baseia-se na subtração dos valores do espectro de ruído ξ do espectro do sinal contaminado Y, obtido com a FFT de cada *frame*. Neste caso, os *frames* geralmente possuem uma duração dentre 20 ms até 30 ms, tamanho no qual considera-se o sinal estacionário. Uma vez que, a transformada de *Fourier* modela o sinal como uma combinação de funções bases sinusoidais ortogonais, cada uma destas pode ser processada individualmente ou em grupos para obter o sinal filtrado. Com esta formulação a amplitude espectral do ruído (ou "perfil de ruído") é obtida a partir do próprio sinal recebido no sensor.

A técnica de filtragem conhecida como subtração espectral média (*Spectral Mean Subtraction*) consiste em estimar a magnitude média da PSD do ruído $|\overline{\xi}_f|$ e subtraí-la da PSD do sinal observado $|Y_f|$. Assim, o sinal filtrado resulta em uma aproximação do sinal original dado por:

$$|\hat{Y}_f| = |X_f| + |\xi_f| - \alpha|\overline{\xi}_f| \approx |X_f|, \tag{2.27}$$

no qual o fator α controla a quantidade de ruído extraído (ou atenuação do filtro) (Va-

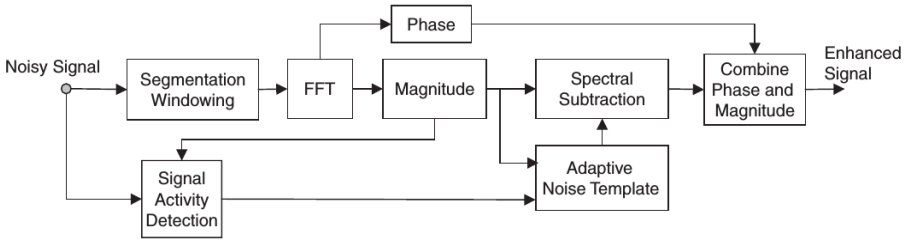


Figura 2.8. Técnica de subtração espectral representada por blocos de processamento. Figura extraída de Vaseghi (2008).

seghi, 2000) e $|\overline{\xi}_f|$ é obtida a partir dos T *frames* com somente ruídos:

$$|\overline{\xi}_f| = \frac{1}{T}\sum_{t=1}^{T}|\xi_{f_t}|, \tag{2.28}$$

onde t é o índice temporal ou identificador do *frame*.

Na abordagem clássica de subtração espectral os primeiros *frames* do sinal são utilizados para estimar $|\overline{\xi}_f|$ (Fukane and Sahare, 2011). Posteriormente, cada novo *frame* passa por um detector de atividade acústica. Assim, se o *frame* for considerado sinal é aplicada a equação 2.27, senão os valores deste são utilizados para atualizar o perfil de ruído (bloco *signal activity detector* da figura 2.8). Desta forma a regra de filtro aplicada a cada *frame* t resulta:

$$|\hat{Y}_{f_t}| = \begin{cases} |Y_{f_t}| - |\overline{\xi}_{f_t}| & \text{se o } \textit{frame } t \text{ for considerado ruído} \\ |Y_{f_t}| & \text{caso contrário.} \end{cases} \tag{2.29}$$

A fase do sinal deve ser preservada para não causar distorções na reconstrução. Portanto, o sinal filtrado é obtido aplicando a transforma inversa de *Fourier* como:

$$\hat{y} = \mathcal{F}^{-1}\{|\hat{Y}_f|e^{i\theta_{Y_f}}\}, \tag{2.30}$$

na qual é utilizada a mesma fase do sinal ruidoso original.

Este filtro possui duas desvantagens: (1) aparecem distorções isoladas nas altas frequências conhecidas como "ruído musical" e (2) a qualidade da filtragem depende do detector de atividade acústica (Vaseghi, 2000). No entanto, Cai et al. (2007) aplicaram este filtro ao problema de reconhecimento de aves eliminando o bloco detector e utilizando todos os *frames* do sinal para obter $|\overline{\xi}_f|$, relatando uma melhoria na taxa de reconhecimento. Por este motivo, este filtro é utilizado nas comparações realizadas no capítulo 6.

2.3.7 Transformada Wavelet

A transformada Wavelet ($\mathcal{W}$) é definida como a decomposição de um sinal em um conjunto de funções de bases ortogonais que podem ser dilatadas e trasladadas no tempo (Graps, 1995). Esta transformada, aplicada a um sinal contínuo (CWT), é definida como (Morettin, 1999):

$$\gamma(s,\tau) = \int x(t)\Psi^*_{s,\tau}(t)dt, \tag{2.31}$$

em que $x(t)$ é o sinal, $\Psi^*_{s,\tau}(t)$ são as funções bases nas quais o sinal é decomposto e * denota o conjugado complexo.

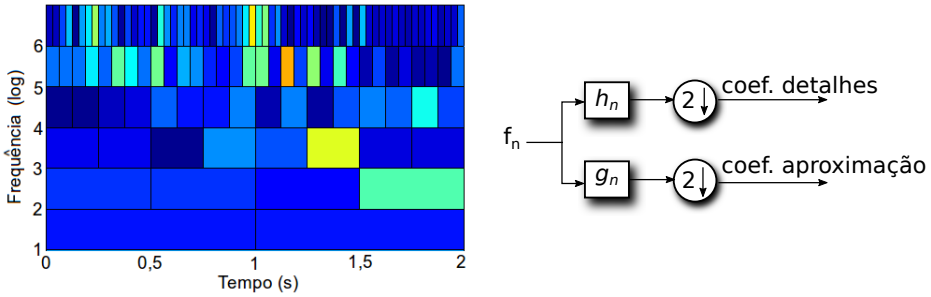
Os possíveis valores das variáveis s e τ, conhecidas como escalonamento e translação respectivamente, fornecem as bases ortogonais da família Wavelet. Formalmente, a função base (Ψ) ou protótipo, na qual é possível decompor o sinal $x(t)$, é conhecida como Wavelet "mãe" e resulta definida como:

$$\Psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-\tau}{s}\right), \tag{2.32}$$

esta equação pode ser interpretada como a aplicação de diversos filtros passa-banda sobre o sinal original.

A $\mathcal{W}$ oferece vantagens em relação a transformada de Fourier tradicional. A principal é possibilidade de variar o comprimento da janela de análise em função da escala, de forma a analisar o sinal com diferentes granularidades em frequências distintas (Johnson et al., 2007). Mudar o parâmetro s equivale a utilizar janelas menores para frequências maiores ou vice-versa, enquanto que o parâmetro τ especifica a localização da janela no tempo, assim a combinação destes dois parâmetros permite uma análise multirresolução tempo-frequência (figura 2.9(a)).

Os parâmetros s e τ da CWT podem ser discretizados sem perder a possibilidade de ter uma representação completa do sinal subjacente. Neste tipo de análise multirresolução discreta, cada escala (ou nível) do sinal é apresentado em diferentes graus de detalhes. Nos casos que a discretização dos parâmetros seja $s = 2^m$ e $\tau = n2^m$, pode-se obter uma decomposição eficiente da transformada através da aplicação de um par de filtros em quadratura (ou funções ortogonais). Assim, os coeficientes da decomposição em cada nível correspondem com a saída de um filtro passa-baixa e um passa-altas, caracterizados pela convolução com a função Wavelet mãe, e mais uma sub-amostragem por dois para causar as diferentes escalas temporais (figura 2.9(b)). Esta implementação da $\mathcal{W}$ utilizando filtros é conhecida como Transformada Wavelet



(a) (b)

Figura 2.9. (a) Escalograma Wavelet com as divisões da multirresolução tempo - frequência. (b) Obtenção dos coeficientes de detalhes e aproximação aplicando os filtros da DWT.

Discreta (DWT).

Matematicamente, a DWT permite representar o sinal discreto por dois conjuntos de coeficientes: os coeficientes de escala c_k (ou média) e os coeficientes de detalhes $d_{j,k}$ (ou diferença). Analiticamente isto resulta:

$$X(t) = \sum_{k=-\infty}^{\infty} c_k \phi(t-k) + \sum_{k=-\infty}^{\infty} \sum_{j=0}^{\infty} d_{j,k} \psi(2^j t - k), \tag{2.33}$$

as funções $\phi(t)$ e $\psi(t)$ são conhecidas como funções "pai" e "mãe" respectivamente. É possível obter uma sequência de coeficientes isolando h e g das seguintes equações:

$$\psi(t) = \sqrt{2} \sum_n h_n \phi(2t-n), \quad \forall n \in \mathbb{Z} \tag{2.34}$$

$$\phi(t) = \sqrt{2} \sum_n g_n \phi(2t-n), \quad \forall n \in \mathbb{Z}. \tag{2.35}$$

Tomando a transformada Z das duas equações anteriores, os valores h_n e g_n resultam coeficientes de filtros discretos. Desta forma, a DWT é a convolução do sinal com dois filtros, um filtro passa-baixas (g_n), para obter os coeficientes de escala, e um filtro passa-altas (h_n), para obter os coeficientes de detalhes. Em outras palavras, para realizar a DWT não é necessário escalonar ou trasladar a função Wavelet mãe, simplesmente pode-se filtrar sucessivamente o sinal de entrada. Assim, um procedimento recursivo pode ser utilizado para calcular os coeficientes em todos os níveis possíveis da decomposição (figura 2.10). Após a decomposição, a reconstrução do sinal original é obtida convolvendo os coeficientes de cada nível pelas funções inversas dos filtros g e h.

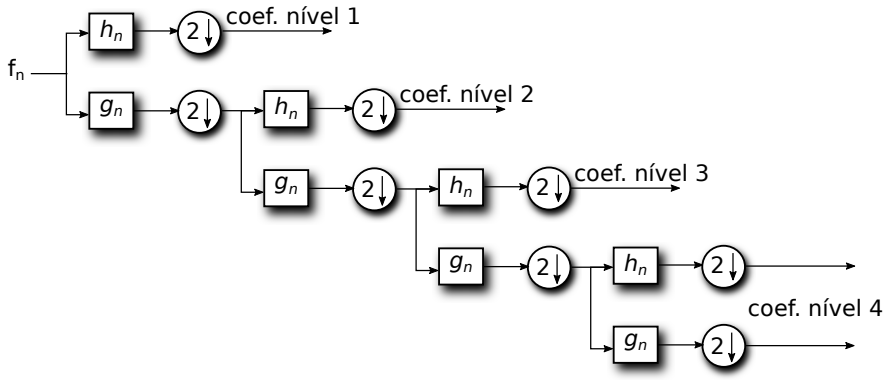


Figura 2.10. Obtenção dos coeficientes de detalhes e aproximação aplicando os filtros da DWT.

2.3.8 Filtro Wavelet

A DWT é uma ferramenta utilizada para aprimorar diferentes tipos de sinais pela eliminação dos ruídos aleatórios. Na filtragem as magnitudes dos coeficientes de detalhes e a aproximação, descritos na seção anterior, são reduzidos ou eliminados após comparação contra um limiar. Esta técnica é conhecida na literatura como Wavelet Denoising (Donoho, 1995, Joy et al., 2013).

Existem diferentes critérios para escolher o limiar de comparação e redução dos coeficientes menos significativos. Os métodos mais usuais para cortar os coeficientes são *hard* e *soft threshold*. No *hard thresholding* todos os coeficientes menores a T são igualados a zero e os maiores são mantidos sem modificação. O *soft thresholding* elimina os coeficientes menores que T e atenua linearmente os valores maiores. Finalmente, o *nonlinear thresholding* atenua os coeficientes realizando o mapeamento deste por uma função suave tipo exponencial, evitando mudanças ou cortes abruptos dos valores (Johnson et al., 2007). Existem também, diferentes formas de encontrar o limiar de corte T da figura 2.11, por exemplo: o limiar universal (equação 2.36) ou a minimização do risco empírico (SURE, equação 2.37). Sendo assim:

$$T = \sigma_\xi \sqrt{2\ \log N}, \tag{2.36}$$

no qual σ_ξ é uma estimativa da variância do ruído e N a quantidade de coeficientes, e

$$T = \underset{0 \leq T \leq \sigma_\xi \sqrt{2\ \log N}}{\arg\max} \{\sigma_\xi^2 N + \sum_{i=1}^{N} [\min(x_i, T^2) - 2\sigma^2 I(|x_i| \leq T)]\}, \tag{2.37}$$

em que x_i são os valores do sinal no domínio do tempo e a função indicadora I retorna 1, caso a condição seja satisfeita, ou 0 no caso contrário. Note-se que os limiares podem

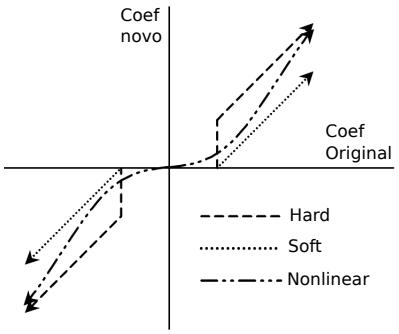


Figura 2.11. Funções *hard*, *soft* e *nonlinear thresholding*, figura adaptada de Johnson et al. (2007).

ser gerais, por exemplo o mesmo T em todos os níveis, ou pode ser ajustados para cada nível da decomposição T_m.

O desempenho de um filtro wavelet caracteriza-se não somente pela escolha correta de T_m, senão também pela escolha da função Ψ e o número de níveis da decomposição, para que possa ser eliminado o ruído mantendo as propriedades do sinal. Assim, uma vez escolhidas as funções g e h, a técnica resulta em um complexidade computacional $\mathcal{O}(n)$, sendo esta uma das principais vantagens desta abordagem. A figura 2.12 é um exemplo de aplicação do filtro wavelet (decomposição e reconstrução) usando a função *Haar* com dois níveis de detalhes. Nesta figura, encontram-se representados os limiares T_1 e T_2 de cada nível pelas linhas horizontais tracejadas. O critério utilizado para encontrar estes valores foi a minimização SURE e a filtragem obedeceu a atenuação dos coeficientes aplicando o *soft threshold*.

As técnicas de filtragem introduzidas aqui, *Spectral Mean Subtraction* e *Wavelet Denoising*, serão os pontos de comparação de nossa técnica de aprimoramento de sinais bioacústicos utilizando SSA, descrita no capítulo 6.

2.4 Sigular Spectrum Analysis

O *Sigular Spectrum Analysis* (SSA) é uma técnica para decompor uma série temporal em uma soma de componentes oscilatórios fisicamente interpretáveis. SSA pode ser aplicado com diferentes propósitos, por exemplo: 1) extração de tendências do sinal; 2) suavização dos dados; 3) extração de componentes de sazonalidade ou ciclos com diferentes períodos; 4) extração de periodicidades com diferentes amplitudes; e 5) detecção de pontos de mudança (Golyandina et al., 2001, Hassani, 2007). SSA é conhecido também como *Projective subspace* ou *Signal Subspace*, dependendo da área

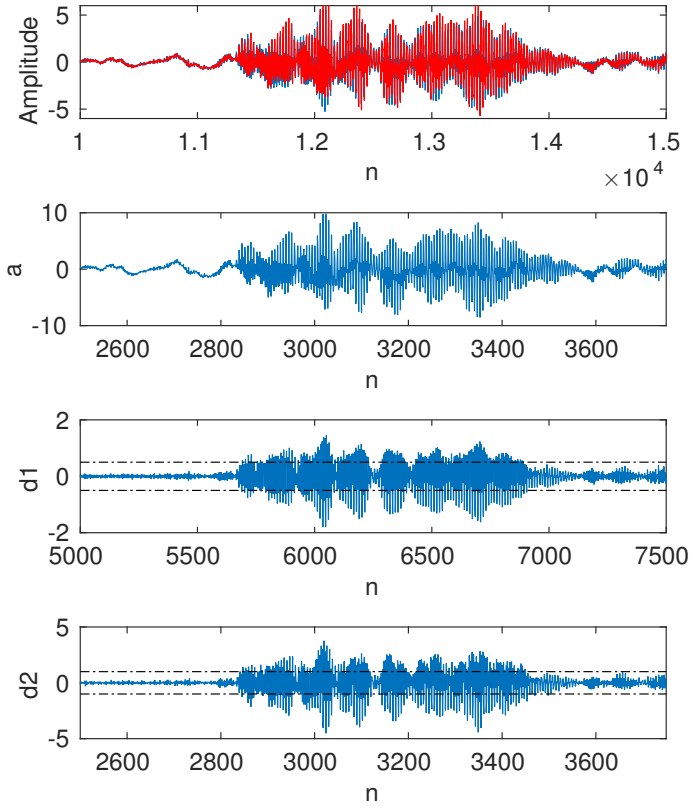


Figura 2.12. Sobreposição da sílaba da espécie Adenomera hylaedactyla antes e depois da filtragem, sinal azul e vermelho respectivamente. Decomposição aplicando *Haar* com *soft threshold* e SURE, na qual **a** representa os coeficientes de aproximação, **d1** os coeficientes de detalhes de nível um e **d2** do nível dois. Neste exemplo os limiares de corte são independentes em cada nível e identificados pelas linhas tracejadas.

de aplicação.

Esta técnica é não paramétrica e baseia-se na decomposição em componentes principais (PCs) utilizando os autovalores e autovetores, sendo possível realizar o processamento sem a necessidade de suposições prévias sobre o sinal ou sobre os ruídos (Golyandina et al., 2001). As aplicações e interesse por SSA têm aumentado nos últimos anos sendo aplicado em diferentes áreas, tais como: física, economia, meteorologia e saúde. SSA possui uma estreita relação com os métodos de filtro proposto por Ephraim and Van Trees (1995), Hansen and Jensen (2007), Hermus et al. (2007) que utilizam SVD. No entanto, a diferença principal com estes métodos é que SSA procura uma nova representação do sinal com diferentes componentes oscilatórios, enquanto que os filtros somente procuram reduzir os ruídos.

A aplicação do SSA consta de quatro passos básicos (Golyandina et al., 2001):

- Aplicar janelamento sucessivo no sinal para transformar este em uma sequência de vetores organizados em formato de uma matriz *Hankel* (ou de trajetória, conhecida também como *lagged vectors* ou *embedding matrix*);
- Obter o decomposição da matriz de trajetórias em bases ortogonais aplicando a decomposição em valores singulares (SVD);
- Selecionar os PCs para a reconstrução (este passo é comumente chamado de agrupamento); e
- Reconstruir a matriz de trajetórias utilizando somente os componentes escolhidos e calcular a média das anti-diagonais para obter a reconstrução da série univariada.

Através do SVD da matriz de trajetórias $X_{L\times K}$ é possível projetar o sinal original em um novo espaço vetorial ($x_N \rightarrow V_{N\times L}$) tornando mais evidentes as diferentes escalas temporais (ou estruturas) da série. Neste caso, N representa o comprimento do *frame*, devido a que SSA pode ser aplicado ao sinal completo ou a cada um do seus *frames*. L e K indicam o tamanho das matrizes. A matriz X é um mapeamento da série univariada na forma:

$$X_{L\times K} = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_K \\ x_2 & x_3 & x_4 & \cdots & x_{K+1} \\ x_3 & x_4 & x_5 & \cdots & x_{K+2} \\ \vdots & \vdots & & \ddots & \vdots \\ x_L & x_{L+1} & \cdots & \cdots & x_{K+L-1} \end{bmatrix}, \tag{2.38}$$

onde a quantidade de colunas é definida pela relação $K = N - L + 1$ com $K \geq 1$, e L componentes oscilatórias (ou autovalores). A característica principal de X é que os valores de x formam anti-diagonais. Aplicando a normalização $x = \frac{x-\bar{x}}{\sigma_x}$ ao sinal original e multiplicando $S = XX^T$ é possível obter as autocorrelações de x. A partir desta nova matriz de autocorrelações (S) é realizada a decomposição SVD.

O segundo passo do SSA é a decomposição em valores singulares aplicando SVD para obter as matrizes U e D, da forma:

$$S = U\Lambda^{\frac{1}{2}}D^T, \tag{2.39}$$

na qual, $U_{L\times L}$ são os autovetores esquerdos, $D_{L\times L}$ os autovetores direitos e $\Lambda_{L\times L}$ é uma matriz diagonal com os autovalores $\Lambda = \text{diag}(\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_L)$. As colunas da matriz

U formam as bases ortogonais do sub-espaço vetorial de dimensão L, e a matriz $V_{K\times L}$ é calculada projetando X da forma:

$$V = X^T U. \tag{2.40}$$

Esta duas matrizes possuem colunas ortonormais, sendo que V contém diferentes representações não correlacionadas de X (figura 2.14). A tripla $\sqrt{\lambda_i}U_i V_i$, definida por cada coluna i da decomposição SVD é chamada componente principal (PC).

O terceiro passo do SSA é o agrupamento ou seleção dos autovetores que irão participar da reconstrução. O critério tradicional de agrupamento é manter os autovalores de maior peso da matriz $\Lambda^{\frac{1}{2}}$, e eliminar os restantes antes da reconstrução do sinal. A contribuição dos primeiros λ_i, para o cálculo da norma de X, é maior do que os últimos λ_i. Portanto, os últimos autovalores representem o ruído do sinal. A figura 2.13 mostra os valores singulares normalizados segundo:

$$\lambda_i = \frac{\lambda_i}{\sum_{i=1}^{L} \lambda_i} \times 100, \tag{2.41}$$

de uma chamada da espécie Adenomera hylaedactyla. Os autovalores encontram-se em ordem decrescente[2] formando o espectro singular, que dá o nome ao método. Nesta figura, pode-se notar que o $\lambda_{i=5}$ separa o subespaço do sinal do subespaço que contem somente ruídos $(q-p)$. Diferentes métodos e critérios de seleção podem ser encontrados na literatura (Teixeira et al., 2005, Ghaderi et al., 2011, Romero et al., 2015). Parte de nossa proposta para filtrar sinais bioacústicos é desenvolver um critério próprio, aplicando os conceitos de entropia. Mais detalhes sobre esta metodologia encontram-se no capítulo 6.

O último passo de SSA é obter a reconstrução $\hat{X}$ e o sinal estimado $\hat{x}$. $\hat{X}$ deve ser calculado aplicando:

$$\hat{X} = UPV^T, \tag{2.42}$$

onde P é uma matriz diagonal com valores $p_{i,i} = 1$ na posição dos autovalores escolhidos no passo anterior e $p_{i,i} = 0$ nos lugares restantes. O P é um operador de seleção que serve para escolher as colunas de U e V que participam da reconstrução. A partir de $\hat{X}$ é necessário realizar a média das anti-diagonais para obter novamente a série

[2] As colunas das matrizes U e V devem ser intercaladas de acordo com a nova ordenação dos λ_i.

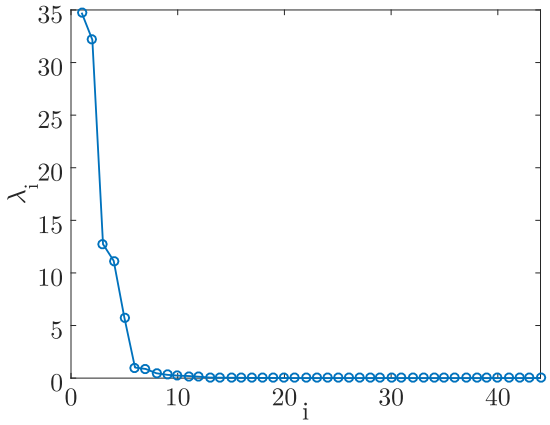


Figura 2.13. Espectro singular (*Singular Spectrum*). Autovalores contidos na diagonal da matriz Λ normalizados e ordenados.

univariada $\hat{x}$ aplicando as regras:

$$\hat{x} = \begin{cases} \frac{1}{k+1} \sum_{m=1}^{k+1} \hat{X}_{m,k-m+2} & \text{para: } 0 \leq k \leq L^* - 2 \\ \frac{1}{L^*} \sum_{m=1}^{L^*} \hat{X}_{m,k-m+2} & \text{para: } L^* - 1 \leq k \leq K^* - 1 \\ \frac{1}{N-k} \sum_{m=k-k^*+2}^{N-k^*+1} \hat{X}_{m,k-m+2} & \text{para: } K^* \leq k \leq N \end{cases}, \tag{2.43}$$

sendo $L^* = \min(L, K)$ e $K^* = \max(L, K)$. Nestas equações, os subscritos de y indicam o valor da matriz $\hat{X}$ posicionado na fila m e coluna $k - m + 2$.

A figura 2.14 mostra graficamente algumas colunas das matrizes U e V aplicadas à decomposição de uma vocalização da espécie Adenomera hylaedactyla utilizando $L = 22$. Neste caso, as colunas de U podem ser interpretadas como fontes geradoras de sinal, enquanto que as colunas de V são as projeções explicadas pela equação 2.40. As colunas de ambas matrizes estão ordenadas em forma decrescente segundo os valores de λ_i. Assim, que as projeções de X com maior variância (ou energia) ficam concentradas nas primeiras colunas, enquanto que os ruídos aleatórios se expandem em todas as colunas. Graficamente, notamos que aplicar a seleção P nas colunas de V pode ajuda a eliminar uma fração dos ruídos na reconstrução de $\hat{X}$, e.g., escolhendo as duas primeiras colunas de V para a reconstrução todo o ruído contido entre as colunas 3 e L seria eliminado.

Um exemplo completo de SSA (*Embedding*→*SVD*→*Grouping*→*Reconstruction*)

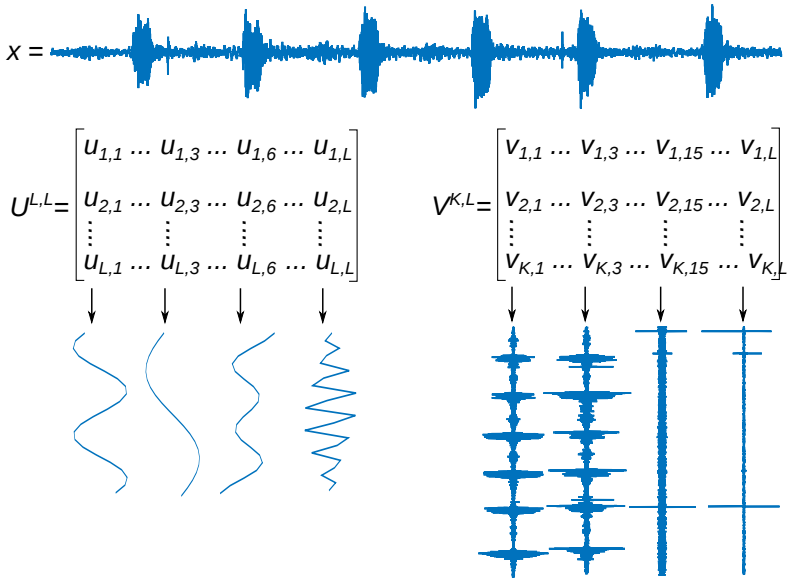


Figura 2.14. Exemplo de decomposição de uma vocalização da espécies Adenomera hylaedactyla. Neste exemplo, é possível observar de forma gráfica o conteúdo das matrizes U e V.

utilizando $L = 44$ e agrupamento[3] $\lambda_{1:4}$ é mostrada na figura 2.15. O sinal azul representa a gravação original da espécie Adenomera hylaedactyla e sobreposto em vermelho a reconstrução utilizando somente os primeiros quatro autovalores (2.15a). O sinal verde é o residual calculado como $r = x - \hat{x}$ (2.15b). Neste exemplo, observando os valores do residual não notamos perda ou distorção da amplitude das sílabas.

2.4.1 Revisão dos critérios de escolha dos componentes SSA

A matriz diagonal Λ obtida pela decomposição SVD é ordenada de maior para menor e, consequentemente, as colunas de U e V são intercaladas respeitando a mesma ordem. A regra de agrupamento mais tradicional é escolher os primeiros autovalores, segundo:

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_p}{\lambda_1 + \lambda_2 + \cdots + \lambda_m + \cdots + \lambda_L} \geq VAR, \tag{2.44}$$

na qual VAR geralmente é 0,95[4] (Teixeira et al., 2005). Com os índices dos autovalores escolhidos é formada a matriz P. Esta regra garante que pelo menos 95% da variância original do sinal é retida. Entretanto, o limiar de corte VAR pode ser alterado conforme

[3]Escolhemos utilizar o subscrito 1:4 para indicar o agrupamento correspondente aos autovalores λ_1, λ_2, λ_3 e λ_4.

[4]Aqui por simplicidade omitimos o radicando ($\sqrt{\lambda_i}$) e os valores foram normalizados para cumprir com $\sum_{i=1}^{l} \lambda_i = 1$.

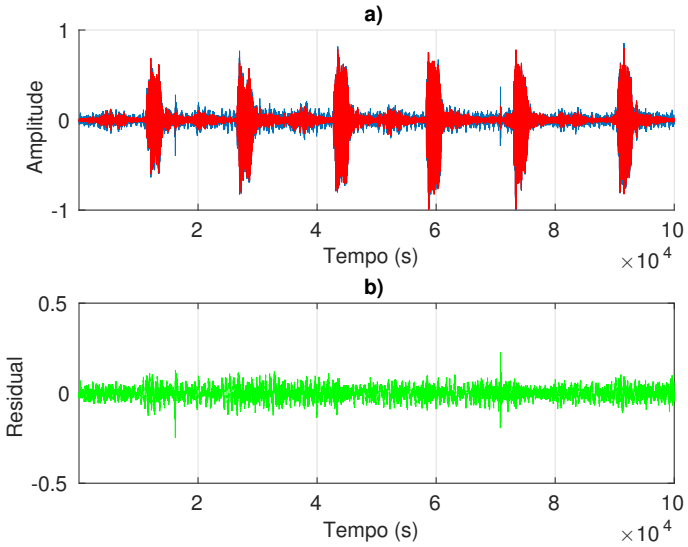
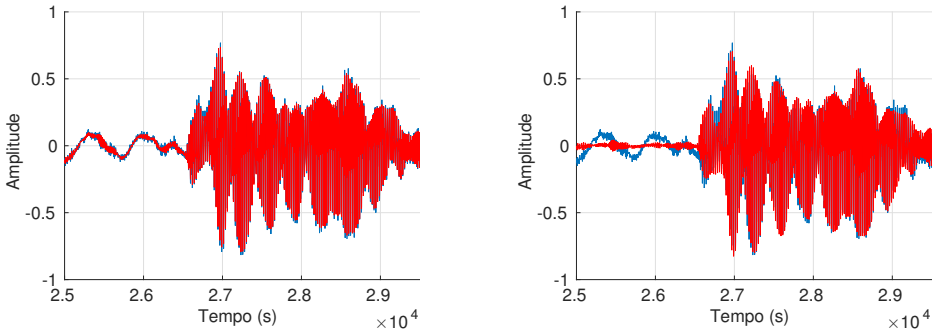


Figura 2.15. Vocalização original da espécie Adenomera hylaedactyla (azul) e reconstrução utilizando os primeiros quatro autovalores (vermelho) e o residual (verde).

a necessidade da aplicação. Aplicando a equação 2.44 aos autovalores da figura 2.13, o agrupamento resultante é $\lambda_{1:5}$ (figura 2.16).

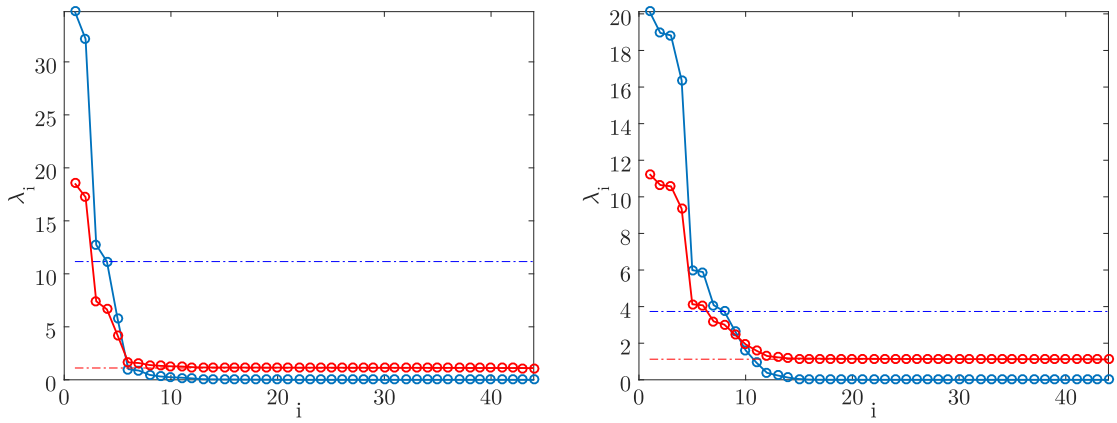


(a) (b)

Figura 2.16. Exemplo de reconstrução utilizando a segunda sílaba da vocalização da espécie Adenomera hylaedactyla. a) utilizando os 5 maiores autovalores ($\lambda_{1:5}$) conforme a regra 2.44 e b) utilizando os quatro maiores ($\lambda_{1:4}$).

Como mencionamos na seção anterior, uma característica da decomposição SVD é que a energia dos ruídos se distribui entre todas as bases da transformação. Consequentemente, um aumento na variância do ruído muda o peso dos autovalores e modifica o resultado da regra 2.44. Por exemplo, adicionando ruído aleatório branco na vocalização da figura 2.15 até alcançar uma relação SNR equivalente a -3 dB o agrupamento é $\lambda_{1:40}$. A figura 2.17 mostra a modificação do espectro singular devido ao aumento no

nível de ruído, no qual, os valores maiores diminuem e os menores aumentam conforme a SNR diminui. Assim, observamos com esta regra que aumentando o nível de ruído aumenta-se o agrupamento, fato que resulta em uma reconstrução menos filtrada.



(a) Adenomera hylaedactyla (b) Sinax ruber

Figura 2.17. Espectros singulares de duas espécies diferentes antes (azul) e depois de adicionar ruído aleatório branco (vermelho). As linhas tracejadas representam os limiares de agrupamento aplicando o critério da equação 2.44 antes e depois do aumento dos ruídos. Podemos notar que em ambas figuras o agrupamento aumentou após elevar o ruído.

Um critério diferente de escolha, utilizado nas abordagens clássicas de PCA, é selecionar os autovalores maiores ao limiar (King and Jackson, 1999):

$$\overline{\Lambda} = \frac{1}{L}\sum_{i=1}^{L}\lambda_i. \tag{2.45}$$

Esta regra mantêm o valor de $\overline{\Lambda}$ independente da variância do ruído. No exemplo anterior a aplicação desta regra conserva o agrupamento $\lambda_{1:5}$ até a diminuição da relação SNR alcançar o valor -6 dB. Consequentemente, esta regra de agrupamento é mais robusta.

Outros critérios, tais como: agrupamento com *k-means*, princípios baseados na Teoria da Informação ou estimação da Máxima Verosimilhança dos PCs, podem ser utilizados para criar o agrupamento ou separar os PCs (Wax and Kailath, 1985, King and Jackson, 1999, Teixeira et al., 2005, Romero et al., 2015).

2.4.2 Complexidade computacional do SSA

A aplicação do SSA requer de varias manipulações algébricas, tornando o método custoso quando comparado com os filtros descritos na seção 2.3.

O primeiro passo desta técnica é criar a matriz de trajetórias (equação 2.38) respeitando a condição $K = N-L+1$. Assim, um laço iterativo com uma janela deslizante de tamanho K demandaria somente L iterações. Neste caso, o custo temporal $\mathcal{O}(L)$ pode ser considerado desprezível. Entretanto, a memória necessária para armazenar a matriz $X^{L,K}$ é $\mathcal{O}(LK)$, que no pior caso, quando $L = {}^{N}/_{2}$, resulta $\mathcal{O}(N^2)$.

Antes de aplicar a decomposição em SVD é necessários obter a matriz de correlações definida como $S = XX^T$. O método mais simples para a multiplicação de matrizes precisa N^3 iterações (Cormen et al., 2009). Algumas abordagens mais eficientes, como a de *Strassen*, a multiplicação de matrizes quadradas possui custo menor ($\mathcal{O}(n^{2.807})$). Porém, quando as matrizes não são quadradas, como no caso $L < K$, a complexidade temporal resulta $\mathcal{O}(KL^2)$. A partir deste passo, uma liberação de memória é possível, porque somente deve-se armazenar $S^{L,L}$. Por exemplo, a série da figura 2.14 possui $N = 101122$ pontos, se neste caso for escolhido $L = \frac{N}{2}$, a matriz S teria tamanho 50561×50561, considerando uma representação com 32 bits seria necessária uma memória de tamanho 1.0226×10^{10} bytes.

A decomposição em PCs da matriz de correlações S é realizada com a implementação disponível no MATLAB, que aplica dois passos: (1) a fatoração QR e (2) a decomposição SVD aplicando um procedimento iterativo. A fatoração QR é utilizada para diagonalizar a matriz S respeitando a decomposição $S = QR$, na qual Q e R $\in \mathbb{R}^{L\times L}$. Q resulta ortogonal e R triangular superior. O custo desta fatoração devido a que S é quadrada resulta $\mathcal{O}(L^3)$ (Golub and Van Loan, 1996). O método iterativo para encontrar os autovalores finais precisa $\mathcal{O}(L)$ computações adicionais.

Finalmente, considerando o passo 2 completo do método SSA, no qual é obtida S e realizada a decomposição SVD, o custo resulta $\mathcal{O}(L + L^3 + KL^2)$ sendo dominado por $\mathcal{O}(KL^2)$. Porem, para valores $L \approx {}^{N}/_{2}$ o termo $\mathcal{O}(L^3)$ não deve ser simplificado, resultando em uma complexidade $\mathcal{O}(2L^3)$.

A complexidade de agrupamento, ou escolha dos PCs, depende da técnica escolhida. Se o critério for visual, observando os valores do espectro singular, pode se considerar que não existe custo computacional associado. Se aplicarmos o critério de reter o 95% da variância, o custo pode ser considerado constante $\mathcal{O}(1)$. No entanto, critérios mais complexos, como cálculo da entropia dos PCs, deve ser levado em consideração.

O passo final do SSA, a reconstrução, depende do produto UV^T. A obtenção da matriz $V.$ resulta da projeção $V = X^TU$. De maneira similar ao passo dois, que para obter V a complexidade é $\mathcal{O}(KL^2)$ e custo de memória é $\mathcal{O}(LK)$.

No último passo, para obter a reconstrução final $\hat{x}$ é necessário realizar o cálculo da média das anti - diagonais da matriz $Y = UV^T$. Conforme a equação 2.43 o custo

resulta em $\mathcal{O}(kN^2)$.

Finalmente, resumimos o custo computacional completo como $\mathcal{O}(KL^2 + KL^2 + kN^2)$ ou $\mathcal{O}(2L^3 + KL^2 + kN^2)$ no pior caso. A memória necessária para poder aplicar este método resulta: $\mathcal{O}(LK)$ para X, $\mathcal{O}(L)$ para S, $\mathcal{O}(L)$ para U, $\mathcal{O}(LK)$ para V e $\mathcal{O}(LK)$ para Y. Obviamente, algumas simplificações são possíveis, por exemplo, depois de obter U, a matriz S não é mais necessária, o mesmo ocorre depois dos produtos $V = X^T U$ ou $Y = UV^T$ em que as matrizes X e V não são mais necessárias e podem ser eliminadas da memória.

2.5 Técnicas de Classificação

As técnicas de aprendizagem de máquina são indispensáveis para o reconhecimento das espécies. Após o mapeamento dos sinais bioacústicos no novo espaço de características (utilizando os valores dos descritores), as técnicas de classificação são aplicadas para identificar padrões que não são evidentes nas amostras originais. Esta aplicação associa os conjuntos de LLD com padrões de exemplos, possibilitando desta forma, reconhecer anuros pelas similaridades entre as características acústicas.

Na literatura existem diversas técnicas de classificação tais como: Árvore de Decisão (C4.5), Redes Neurais Artificiais (ANN), Näive Bayes (NB), kNN, SVM entre outras. Cada uma destas técnicas possui vantagens, desvantagens e tipos de aplicação mais propícias. Em Kotsiantis (2007) é realizada uma revisão de técnicas de classificação usando métricas como acurácia, velocidade de treino e velocidade de classificação entre outras.

A escolhas dos classificadores k-NN, Näive Bayes, Análise Discriminante Quadrático e Árvore de Decisão, neste trabalho, foi baseada na revisão dos trabalhos relacionados e no intuito de experimentar funções de classificação diferentes, como as baseadas em: densidade, funções lineares, funções quadráticas e funções definidas por partes.

2.5.1 kNN

Em reconhecimento de padrões, kNN é o método usado para classificar um objeto (ou instância) desconhecido Y, baseando-se em um conjunto k de exemplos mais próximos, dentro de um espaço com n características e W classes. A decisão de classificar Y como pertencente à classe w_i depende unicamente dos valores das características e da classe a qual pertencem os k vizinhos mais próximos de Y (Cover and Hart, 1967).

Mais precisamente, o método atribui o novo objeto Y à classe W, à qual pertence a maioria de seus k vizinhos aplicando um cálculo de distância e realizando um ranking

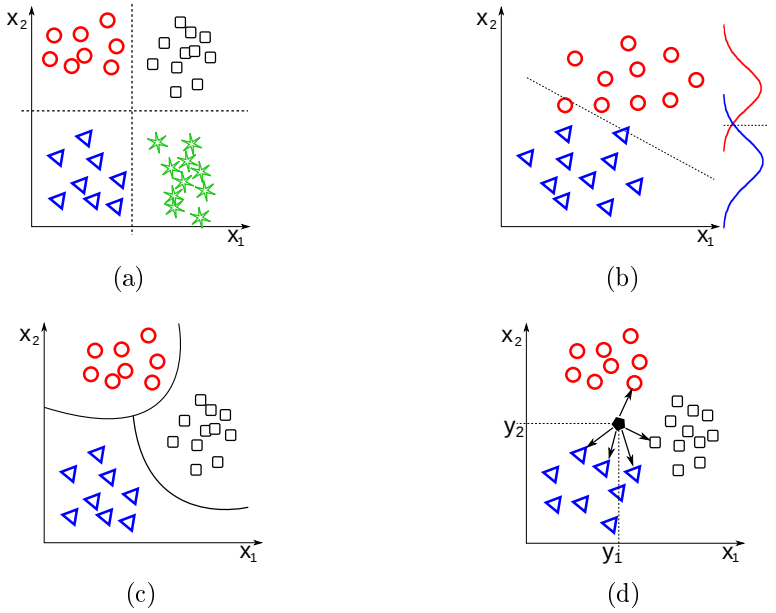


Figura 2.18. (a) Divisão do espaço de características utilizando árvore de decisão. (b) Separação linear aplicando Näive Bayes. (c) Funções de decisão criadas com QDA (figura extraída de Murphy (2012)). (d) Exemplo de decisão *5-NN* em um espaço vetorial de duas caraterísticas com três classes.

dos resultados. A figura 2.18(d) apresenta um exemplo com $k = 5$ e $W = 3$. A métrica de similaridade normalmente utilizada é a distância Euclidiana, que é calculada entre o vetor desconhecido e cada um dos vetores no espaço de exemplos como:

$$\mathrm{D} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}, \tag{2.46}$$

em que y_n é o valor numérico de cada característica desconhecida e x_n é o valor da mesma característica, mas da classe exemplo.

O funcionamento de kNN é descrito abaixo, de acordo com Theodoridis and Koutroumbas (2008):

1. Defina um valor para k, ou seja, a quantidade de vizinhos mais próximos (preferentemente ímpar para evitar empates);

2. Calcule a distância da nova amostra contra todos os exemplos da base;

3. Identifique os k vizinhos mais próximos (ranking), independentemente do rótulo das classes;

4. Conte o número de vizinhos mais próximos que pertencem a cada classe do problema;

5. Classifique a nova amostra atribuindo-lhe a classe mais frequente na vizinhança.

A particularidade do classificador é não precisar tempo de treinamento para gerar um modelo, mas consequentemente, o processo de classificação pode ser computacionalmente exaustivo e depende do tamanho do conjunto de exemplos. Por este motivo, a desvantagem de kNN é a complexidade computacional envolvida na obtenção dos k vizinhos mais próximos. Entretanto, uma vantagem é a independência quanto à distribuição de dados no espaço de características.

2.5.2 Classificador de Näive Bayes

O classificador de Näive Bayes (NB) estima as funções de densidade de probabilidade (PDF) $p(\mathbf{x}|w_i)$ necessárias para aplicar a regra de classificação:

$$w_i = \arg\max_{w_i} \prod_{j=1}^{l} p(x_j|w_i), \quad i = \{1, 2, \ldots, W\}, \tag{2.47}$$

baseado nos exemplos de treino, no qual existem W classes representadas por vetores de características com tamanho l (Theodoridis et al., 2010).

Para predizer a classe de uma amostra desconhecida a partir dos exemplos de treino, precisa-se descobrir, primeiro, a probabilidade de w_i dado $\mathbf{x} = [x_1, x_2, \ldots, x_l]$. Esta probabilidade pode ser obtida aplicando a regra do teorema de Bayes como

$$\mathrm{P}(w_i|\mathbf{x}) = \mathrm{P}(\mathbf{x}|w_i)\mathrm{P}(w_i)/\mathrm{P}(\mathbf{x}). \tag{2.48}$$

O ponto chave desta técnica de classificação é assumir que todas as características são independentes, e consequentemente a PDF resultante:

$$p(\mathbf{x}|w_i) = \prod_{j=1}^{l} p(x_j|w_i). \tag{2.49}$$

Apesar da suposição de independência entre características não ser sempre válida, a acurácia resulta em uma aproximação aceitável. Sem esta suposição, para obter boas estimativas das PDF, precisaria-se de um elevado número de exemplos, aumentando a demanda exponencialmente com a quantidade de características (l) (Theodoridis and Koutroumbas, 2008).

Por exemplo, dado um vetor desconhecido $\mathbf{x}$ e um problema bi-classe (w_1 e w_2)

a regra pratica de decisão é dada pela razão:

$$\frac{\mathrm{P}(w_1)\mathrm{P}(\mathbf{x}|w_1)}{\mathrm{P}(w_2)\mathrm{P}(\mathbf{x}|w_2)} > 1.$$

Pela suposição de independência entres as características, é possível evitar o cálculo de todas as combinações possíveis para os valores de $\mathbf{x}$ e obter as estimativas das PDF conforme a Equação 2.49, precisando-se de menos exemplos de treinamento. Assumindo que as PDF das classes sejam Gaussianas e aplicando o logaritmo, é possível transformar esta regra de decisão em um hiperplano linear da forma $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ e, consequentemente, concluir que o classificador de NB é um tipo de discriminante linear (Theodoridis and Koutroumbas, 2008).

2.5.3 Análise Discriminante Quadrático

Nos métodos de classificação mediante análise discriminante é assumido que as diferentes classes são geradas por uma mistura de diferentes distribuições normais multivariadas:

$$\mathrm{f}_i(x) = \frac{1}{\sqrt{(2\pi)^l|\Sigma_i|}} e^{\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}, \tag{2.50}$$

com dimensão l e matriz de covariância Σ_i para $i = 1, 2, \ldots, W$ classes (Theodoridis and Koutroumbas, 2008). Quando trata-se da análise discriminante linear (LDA), assumimos que as classes possuem diferentes valores médios (μ_i), mas com a mesma matriz de covariância para todas elas. No entanto, quando trata-se da análise quadrática (QDA), cada classe possui sua própria matriz de covariância (Murphy, 2012).

Partindo dos exemplos disponíveis na base de treino é possível inferir os parâmetros μ_i e Σ_i para cada classe de forma empírica. Assim, primeiro é calculada a média das amostras de cada classe, em seguida, está é subtraída dos valores e é calculada a matriz de covariância.

Para classificar um vetor de características desconhecido x segundo a regra de Bayes deve-se escolher a classe que maximiza a probabilidade a posteriori (equação 2.48), de forma a otimizar:

$$\hat{w} = \arg\max_i[\log(\mathrm{P}(w_i|x))] = \arg\max_i[\log(\mathrm{P}(x|w_i)\mathrm{P}(w_i))], \tag{2.51}$$

onde $\mathrm{P}(x|w_i)$ é dado pela equação 2.50 e $\mathrm{P}(w_i)$ é a probabilidade a priori da classe. O fator $\mathrm{P}(x)$ neste caso pode ser omitido pelo fato de ser uma constante. Aplicando algumas manipulações algébricas podemos escrever a função discriminante quadrática

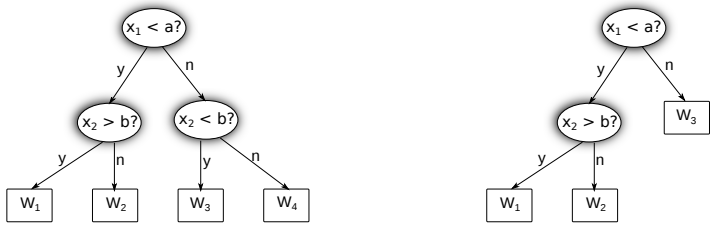


Figura 2.19. Diferentes exemplos de árvores de decisão balanceada (a) e desbalanceada (b).

δ como:

$$\delta_i(x) = -\frac{1}{2}\log(|\Sigma_i|) - \frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) + \log(\mathrm{P}(w_i)), \tag{2.52}$$

tornando a regra de classificação final QDA $\hat{w} = \arg\max_i \delta_i(x)$.

Observando a equação 2.52 podemos notar que no caso do LDA o termo $\log(|\Sigma_i|)$ é uma constante que não depende de i e portanto pode ser simplificado. Além desta observação, podemos destacar também que o termo $\sqrt{(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}$ é conhecido como distância de Mahalanobis (McLachlan, 2004) e, concluir que LDA e QDA são uma forma de calcular a distância de cada x até o centróide de cada classe.

2.5.4 Árvore de Decisão

A técnica mais popular de árvore de decisão é a C4.5 que foi desenvolvida por Quinlan (1993). Uma árvore de decisão é similar a um conjunto hierárquico de regras em que as classes são rejeitadas sequencialmente até alcançar uma classe final aceita (Theodoridis and Koutroumbas, 2008). Cada nó da árvore é uma regra que deve ser aplicada ao valor da característica correspondente, e dependendo do resultado, é escolhida da aresta que leva ao próximo nível da árvore. Desta forma, cada nó da árvore é um teste que compara o valor de uma característica contra uma constante (exemplo: `If` $x_i < b$ `then` w_i; ou para o caso de característica nominal: `If anuran==‘Leptodactylus’ then` w_i). O processo de classificação finaliza quando um nó folha, que indica a classe resultante, é alcançado (figura 2.19).

Existem algumas variações das árvores nas quais algumas características podem ser comparadas entre si, ou pode ser aplicada alguma função matemática de uma ou mais variáveis em cada nó (Witten and Frank, 2005).

Assumindo-se que as instâncias são independentes é aplicada a abordagem de divisão e conquista para criar a árvore. O critério normalmente utilizado para escolher

as características divididas primeiro é o ganho da informação (IG). Assim, o nó raiz particiona o espaço das classes pela característica mais discriminativa primeiro (ou seja com maior ganho de informação) e deixa por último as com menor ganho de informação (figura 2.18(a)). Consequentemente, esta técnica é capaz de criar uma função de classificação não linear. Este tipo de procedimento oferece vantagens quando um grande número de classes estão envolvidas. Um procedimento adicional conhecido como "poda" acontece após a criação completa da árvore, tendo por objetivo simplificar a estrutura permitindo que aconteçam alguns erros de classificação. Desta forma, evita-se o sobreajuste e o modelo generaliza melhor os conceitos aprendidos.

2.5.5 Decomposição de problemas multi-classe

Para resolver problemas de classificação multi-classe existem duas possibilidades. A primeira é utilizar um método de classificação que seja capaz de classificar todas as classes ao mesmo tempo, atribuindo um valor de probabilidade *a posteriori* a cada uma delas. A segunda possibilidade é decompor o problema original em sub-problemas binários. A decomposição binária possui duas vantagens: (a) permite utilizar classificadores naturalmente binários, tal como SVM, para resolver problemas multi-classes, e (b) simplificar a tomada de decisão mediante a simplificação das funções de classificação. Em alguns problemas a decomposição conseguiu aumentar a acurácia (Fürnkranz, 2001).

Em nossa aplicação a complexidade da função de classificação aumenta com o número de espécies monitoradas, sendo relevante avaliar a hipótese de simplificação mediante decomposição binária, para melhorar o resultado do reconhecimento das espécies de anuros. Esta hipótese é abordada no capítulo 5.

Exitem dois tipos de decomposição binária para problemas multi-classe: um-contra-um (*One-against-One* - 1A1) e um-contra-todos (*One-against-All* - 1AA). A decomposição é também conhecida como *Round Robin* (Fürnkranz, 2001). A figura 2.20 exemplifica esse conceito.

O procedimento um-contra-todos (1AA) começa separando todas as amostras (ou sílabas em nosso caso) em dois conjuntos: um conjunto com a classe alvo ("+1"), na qual é incluída somente uma espécie, e um segundo conjunto com todo as amostras das classes restantes ("−1"). Assim, o modelo $f(\cdot)$ é treinado e aplicado para estimar os rótulos do grupo de teste. Se o classificador decidir em favor da classe $+1$ então a espécie correspondente a esta classe ganha um voto, senão todas as espécies representadas pela classe -1 ganham o voto.

Na segunda rodada, este procedimento é repetido, mas as amostras que na rodada

anterior formaram a classe +1 agora são incluída dentro da classes −1, e uma classe diferente é escolhida para compor a classe +1. Novamente o modelo $f(\cdot)$ é treinado e avaliado, e os votos das classes são atribuídos seguindo a mesma regra anterior, isto é, um voto para a classe +1 se for escolhida, senão é atribuído um voto para cada uma das classes que compõem a classe −1. Este procedimento é repetido até que todas as classes sejam avaliadas como sendo +1.

Por último, a decisão final é obtida aplicando votação majoritária. Embora, a quantidade de iterações deste método de decomposição aumenta linearmente com o número de espécies, cada função de decisão avaliada tende a ser mais simples que no caso multi-classe (figura 2.20(b)).

O segundo procedimento de decomposição binária que propomos avaliar é o um-contra-um 1A1. Neste caso, o problema original é decomposto em problemas ainda menores que no caso 1AA. A estimativa do rótulo final de uma amostra desconhecida prossegue de forma semelhante realizando a votação majoritária, porém a forma que os votos são obtidos é diferente.

Primeiramente, são escolhidas duas classes qualquer do conjunto original, uma para compor a classe positiva e outra para a classe negativa. O modelo $f(\cdot)$ é treinado com somente as duas classes escolhidas e um voto positivo é atribuído à classe estimada. Na segunda iteração, a mesma classe que foi escolhida como sendo +1 é mantida e muda somente a classe do conjunto negativo. Novamente o modelo é treinado, avaliado e os votos correspondentes são atribuídos. Esse processo se repete até que a classe que foi escolhida como +1 seja compara um a um contra todas as outras classes do conjunto de dados. Após isso, outra classe do conjunto é escolhida para ser a classe positiva e todas classes restantes são avaliadas um a um contra esta.

Se considerarmos um problema com W classes, então o procedimento completo requer treinar e avaliar $W(W-1)/2$ modelos. O crescimento exponencial do número de modelos treinados e avaliados é a principal desvantagem do método 1A1. Porém, as funções de classificação são as mais simples possíveis (figura 2.20(a)). Em nosso domínio de aplicação, este método de decomposição simplifica o problema e permite que sejam embarcados classificadores binários nos nós sensores. Posteriormente, a comunicação entre os sensores, ou uma operação de fusão de decisões, pode levar à conclusão final.

2.5.6 Ganho da Informação

Além do custo computacional das características acústicas descritas na seção 2.2, é interessante quantificar o ganho de informação com a qual cada descritor acústico contribui para discriminar a natureza dos sinais. O ganho da informação (*Information*

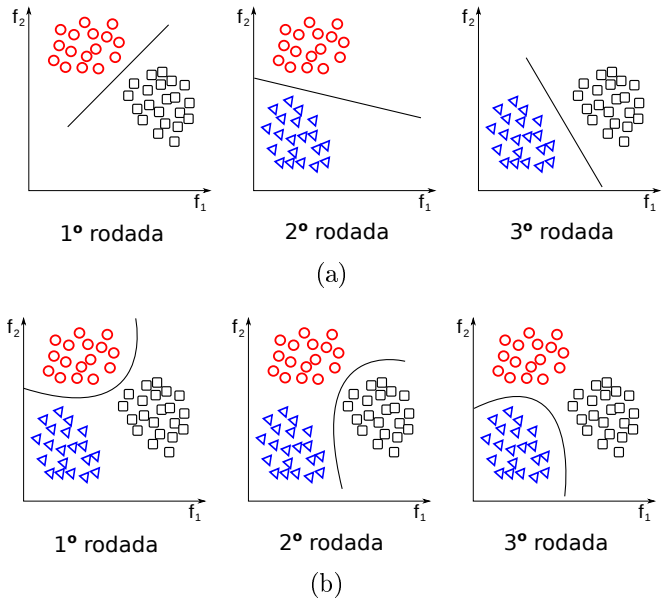


Figura 2.20. Figura adaptada de Fürnkranz (2001).

Gain - IG) possibilita realizar um ranqueamento dos LLDs baseado na quantidade de informação contida em cada um deles. Embora existam diversas métricas para construir um "ranking" das características, o IG é baseada em entropia, e portanto ajusta-se aos objetivos e métodos do Capítulo 4.

Shannon (1948) definiu a entropia como a quantidade real de informação contida em uma mensagem codificada dentro de um alfabeto com distribuição de probabilidades $P(c_i)$, onde c_i é um simbolo do alfabeto C. Em outras palavras, no contexto de classificação, a entropia é uma medida de incerteza para a tomada de decisões calculada como:

$$\mathrm{H}(C) = -\sum_i P(c_i)\log_2\big(P(c_i)\big). \tag{2.53}$$

A entropia condicional de uma classe dado um atributo é calculada da seguinte forma:

$$\mathrm{H}(C/Y) = -\sum_j \Big(P(\pi_j)\sum_i P(c_i/\pi_j)\log_2\big(P(c_i/\pi_j)\big)\Big), \tag{2.54}$$

onde $P(c_j)$ é a probabilidade a priori para todos os valores de C e $P(c_i/y_j)$ é a probabilidade de C dados os valores dos atributos Y. O ganho da informação (IG) representa a quantidade entropia reduzida por cada atributo, ou em outras palavras, a quantidade de informação que os atributos Y fornecem para determinar a classe C (Witten and

Frank, 2005). O IG é calculado como:

$$\mathrm{IG} = \mathrm{H}(C) - \mathrm{H}(C/Y). \tag{2.55}$$

Intuitivamente, o IG pode ser utilizado para avaliar e ranquear os atributos medindo a redução da incerteza em relação à classe e considerando a entropia como uma medida de "impureza". Desta forma, o IG mede a redução de impurezas causada por cada característica em uma coleção de amostras para identificar os LLDs que dividem perfeitamente o conjunto de sinais nas classes "sinal" ou "ruído". Assim, os LLDs altamente discriminativos devem fornecer informações máximas, enquanto que os menos discriminativos não fornecem informações úteis.

2.6 Métricas de Avaliação

As métricas para avaliar a classificação ou a segmentação utilizadas nos experimentos realizados nesta tese são apresentadas a seguir. Estas métricas encontram-se separadas por: métricas exclusivas de segmentação acústica, métricas exclusivas para avaliar o resultado da classificação, e métricas que podem ser utilizadas em ambos os casos.

2.6.1 Métricas de Classificação

As diferentes métricas para avaliar o acerto da classificação podem ser utilizadas nos casos supervisionados e não supervisionados, como o reconhecimento das espécies ou a segmentação automática do áudio. A métrica mais tradicional é a taxa de *erro* do classificador que contabiliza quantas vezes, do total de exemplos classificados, a classe resultante é diferente da real (ou esperada). A acurácia é definida a partir dos verdadeiros positivos (tp), falsos positivos (fn), falsos negativos (fn) e verdadeiros negativos (tn), da forma (Witten and Frank, 2005):

$$\mathrm{Acc} = \frac{tp + tn}{tp + tn + fp + fn}, \tag{2.56}$$

onde tp é o número de vezes que a espécie foi corretamente identificada, fn é a quantidade de vezes que uma sílaba de outra espécie é identificada como sendo da espécie alvo, fn o número de sílabas da espécie alvo que não foram perdidas e tn o número de sílabas de outras espécies que não foram reconhecidas com da espécie alvo. Com estes valores é possível montar uma tabela de contingência (ou matriz de confusão),

em que as filas representam as classes reais e as colunas as classes estimadas para cada exemplo classificado.

A partir da matriz de confusão é possível obter a Precisão (Prec) e a Revocação (Rec) do sistema. A Precisão indica a quantidade de sílabas classificadas corretamente do total de sílabas recuperadas e a Revocação quantifica a relação entre o total de sílabas corretas que o sistema deve recuperar do total recuperado. Para comparar o desempenho de dois ou mais sistemas é útil combinar a Precisão e a Revocação em uma única métrica. O F-Score ou F1 é obtido como a média harmônica entre a Precisão e a Revocação. As equações respectivas são:

$$\text{Prec} = \frac{tp}{tp + fp}, \qquad (2.57) \qquad \text{Rec} = \frac{tp}{tp + fn}, \text{e} \qquad (2.58)$$

$$\text{F1} = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}. \qquad (2.59)$$

Em problemas de classificação binários é interessante avaliar o grau de correlação entre as classes estimadas pela técnica de classificação e o rotulo verdadeiro. Para isto pode-se utilizar o coeficiente de correlação de *Matthews* (MCC) (Powers, 2007):

$$\text{MCC} = \frac{(tp \cdot tn) - (fp \cdot fn)}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}. \qquad (2.60)$$

Utilizando a tabela de contingência podem ser definidas as taxas, ou proporções de: verdadeiros positivos ou sensibilidade (TPR), de verdadeiros negativos (TNR), de falsos negativos ou taxa de perda (FNR) e de falsos positivos (FPR), descritos pelas equações:

$$\text{TPR} = \frac{tp}{tp + fp}, \qquad (2.61) \qquad \text{TNR} = \frac{tn}{tn + fp}, \qquad (2.62)$$

$$\text{FNR} = \frac{fn}{fn + tp}, \text{ e} \qquad (2.63) \qquad \text{FPR} = \frac{fp}{fp + tn}. \qquad (2.64)$$

Estas taxas são utilizadas para avaliar a relação custo-benefício da toma de decisões do classificador ou criar as curvas ROC (*Receiver Operating Characteristic*).

2.6.2 Macro-métricas

As métricas tradicionais descritas na seção anterior podem ser diretamente calculados a partir dos valores das tabelas de confusão do classificador. Por exemplo, a (micro) acurácia é a soma da diagonal principal da matriz de confusão dividida pelo total de exemplos classificados. A principal desvantagem das micro-métricas é que as classes mais numerosas, as quais possuem a maior proporção de exemplos para teste e treino, ponderam muito os resultados. Em outra palavras, se o classificador errar nas classes menos numerosas os resultados ainda podem parecer elevados. Este problema é geralmente apresentado em situações nas quais o número de exemplos nas bases de dados para cada classe não é balanceado (Colonna et al., 2016a).

As Macro-métricas foram definidas para contornar os possíveis viés causados nas Micro-métricas pelo desbalanceamento no número de exemplos das bases de dados (Sokolova and Lapalme, 2009). Neste caso, para cada classe individual C_i são obtidos seus respectivos tp_i, fn_i, tn_i e fp_i, a partir dos quais é possível calcular a Prec_i e a Rec_i de cada classe. Dado que os valores de precisão a revocação são obtidos por classe, os valores Macro- são obtidos como a média aritmética destes valores individuais. Finalmente, a Macro-Fscore é calculada pela equação 2.59 utilizando as médias de Prec_i e Rec_i. Desta forma, todas as classes são tratadas de forma equitativa.

2.6.3 Taxa de Eventos Acústicos Errados

O *Acoustic Event Error Rate* (AEER) foi projetado para avaliar algoritmos de segmentação em problemas de detecção de contexto, no qual os eventos acústicos podem pertencer a mais de uma classe. Esta métrica foi definida por Giannoulis et al. (2013) como:

$$\text{AEER} = \frac{D + I + S}{E}, \tag{2.65}$$

onde E é o número total de eventos em cada áudio, D é o número de eventos perdidos, I os eventos adicionados extras e S o número de substituições calculadas como $S = \min(D, I)$ [5]. Esta métrica considera que um evento é corretamente segmentado se começa e termina entre $\pm 100\,\text{ms}$ do limites reais do evento e se possuir pelo menos 50% do tempo total deste. Além disso, eventos duplicados são considerados falsos alarmes. Assim, quanto menor o AEER, melhor é a segmentação. Em nosso contexto de segmentação bioacústica, as palavras "evento" e "sílaba" das vocalizações são consideradas sinônimos. Para cada fluxo de áudio (ou gravação) podemos obter um valor de

[5] A implementação do AEER de Giannoulis et al. (2013) pode ser encontrada em `http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/`.

AEER. A partir de um conjunto de gravações definimos o Macro-AEER como a média aritmética de todos os AEER individuais.

2.6.4 Curvas ROC

A curva ROC é um gráfico que resume o desempenho de um classificador binário sobre todas as combinações de limiares de decisão possíveis. Esta curva é gerada traçando a relação entre a taxa verdadeiros positivos (TPR) e a taxa de falsos positivos (FPR), quando variá-se o limiar de decisão entre $[0, 1]$. Para cada comparação uma classe é atribuída e comparada à classe correta. Assim, a probabilidade de detecção torna-se função da probabilidade dos falsos alarmes e nos ajuda a selecionar o melhor sistema, independentemente da distribuição de classes (Fawcett, 2006, Slaby, 2007).

A área sob a curva (AUC) é uma medida de desempenho. Esta métrica é adequada para comparar diferentes sistemas de decisão por um único valor escalar. A propriedade mais importante da AUC é que este representa a probabilidade de sortear um par de instâncias (positivas e negativas) e classificá-las corretamente. No capítulo 4 utilizamos estas curvas e seus valores de AUC para avaliar as melhores características para o problema de segmentação de sinais bioacústicos.

2.7 Redes de Sensores Sem Fio

Uma Redes de Sensores Sem Fio (*RSSF*) é um tipo especial de rede *ad-hoc* composta por um grande número de dispositivos, distribuídos geograficamente, com capacidade de coleta e transmissão de dados, chamados nós sensores (Akyildiz et al., 2002). Esses sensores são capazes de monitorar o ambiente, coletar dados, realizar o processamento localmente e disseminar os dados coletados. O maior desafio das RSSF é detectar e avaliar eventos de interesse e reagir ao contexto.

Devido ao grande número de nós distribuídos, o custo por unidade deve ser baixo, levando a restrições de processamento, distância de transmissão e tempo de vida das baterias. Embora existam desvantagens, o baixo custo dos dispositivos também permite que quantidades maiores destes sejam utilizados para aumentar a confiabilidade da rede. Por exemplo, no caso de falhas nas transmissões dos dados, estes podem seguir rotas alternativas entre os nós da rede, ou no caso de detecção de eventos geograficamente distribuídos, podem-se eliminar falsos positivos pela votação dos nós vizinhos. Independente da aplicação, a redundância é uma vantagem deste tipo de redes.

As informações coletadas pelos sensores descrevem as condições físicas do ambiente, fornecendo aos usuários dados precisamente localizados no tempo e no espaço

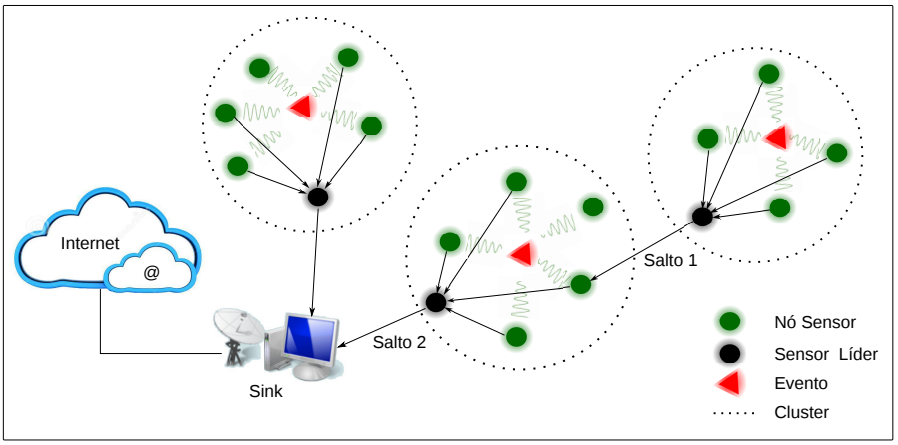


Figura 2.21. Topologia de uma RSSF.

(Zhao and Guibas, 2004). Cada nó da RSSF é equipado com um conjunto dispositivos sensores tais como: acelerômetros, sensor de temperatura, umidade, luminosidade, sensores acústicos como microfones, detectores de movimento ou câmeras fotográficas, infravermelho (IR), sísmicos, ou sensores magnéticos. Onde, cada nó se comunica através da interface sem fio com seu vizinhos mais próximos dentro do rádio de comunicação.

Ao contrário de um sistema centralizado, o processamento dos dados em uma RSSF pode ser realizado de forma distribuída pelos nós da rede. Conforme os dados se propagam pela RSSF, até chegar ao nó *sink*, duas operações podem ser aplicadas: agregação de dados, onde as informações podem ser complementadas com valores de varáveis correlacionadas obtidas pelos nós vizinhos, ou fusão de dados, na qual as informações podem ser misturadas para, por exemplo, diminuir ruídos aleatórios (Nakamura et al., 2007b). Uma topologia comum deste tipo de rede inclui: um ou vários nós sorvedouros (*sink*), os sensores líderes, e os nós finais que compõem os *clusters* (figura 2.21). Os *clusters* de sensores podem ser combinados em série ou em paralelo para estender o alcance da rede e poder se aprofundar no ambiente físico de forma não intrusiva.

2.7.1 Aplicações de Redes de Sensores

As RSSF são projetadas para executar tarefas de alto nível, tais como: detecção e seguimento de alvos, ou classificação e reconhecimento eventos. As aplicações destas redes são inúmeras e se caracterizam por serem ubíquas, pervasivas e menos intrusivas do que a presença humana. Os modos de implantação, os requisitos de energia, e inclusive a topologia da rede podem variar dependendo do objetivo: ambiental, comercial ou militar. Os exemplos de aplicações incluem:

Operação	Custo (nAh)
Transmitir um pacote	20,000
Receber um pacote	8,000
Ler o ADC	0,011
Ler EEPROM	1,111
Escrever/Apagar EEPROM	83,333

Tabela 2.1. Energia requerida por diferentes operação do sensor MICA Weather Board. Tabela adaptada de Mainwaring et al. (2002).

- Monitoramento ambiental, qualidade do habitat ou rastreamento de animais (Wang et al., 2003, Hu et al., 2009);
- Detecção de falhas em máquinas industriais (Tiwari et al., 2007);
- Aplicações militares, posição dos alvos, detecção de intrusos (Durisic et al., 2012);
- Cuidado da saúde (*Health Care*) (Chen et al., 2006b, Magaña-Espinoza et al., 2014); e
- Infra-estrutura e segurança (Balageas et al., 2010, Mansour et al., 2013).

2.7.2 Redução de informação e consumo de energia

Em uma RSSF cada pacote de dados transmitido implica em um custo de energia alto quando comparado com outras tarefas. Um dos maiores desafios é maximizar a quantidade de informação transmitida em cada pacote. Nas aplicações ambientais a quantidade de informação coletada depende da necessidade, ou seja, depende do fenômeno físico que se deseja mensurar e do tipo de sensor utilizado. A tabela 2.1, extraída do trabalho realizado por Mainwaring et al. (2002), mostra o custo de energia da cada operação.

Em nossa aplicação, um sensor de áudio com uma taxa de amostragem de 8 kHz geraria 8000 amostras por seg, considerando que cada amostra é representada por um byte, seriam gerados 8000 $^{\text{bytes}}/_{\text{seg}}$, uma taxa de informação elevada para um nó sensor. Um pacote de um sensor MICA possui uma carga útil (*payload*) de 25 bytes, o qual requer 320 pacotes para transmitir as amostras correspondentes de cada segundo de áudio. Esta operação consumiria 6.4 mAh, implicando que, em um segundo, seria consumida a mesma quantidade de energia que, segundo Mainwaring et al. (2002), consumiam em um dia inteiro. Além disso, existe uma limitação de processamento, sendo improvável para um nó sensor transmitir 320 pacotes por segundo.

Representar cada vocalização com um número reduzido de LLD permite diminuir a quantidade de informação necessária a ser transmitida, diminuindo também a quantidade de pacotes necessários, aumentando a vida útil da rede.

2.8 Comentários Finais

Neste Capítulo apresentamos como calcular os descritores acústicos a partir das sílabas dos anuros. Este mapeamento constitui um forma de redução de informação, que além da economia na representação, são o conjunto de características que o classificador precisa como entrada para reconhecer as espécies. Adicionamos explicações detalhadas sobre o metologia de cálculo da Entropia das Permutações e suas variações pelo motivo que serão aplicadas como descritores para identificar diferentes condições de ruídos.

Descrevemos diferentes tipos de ruídos aleatórios coloridos e o funcionamento de duas técnicas de filtragem para aprimorar a qualidade dos sinais e melhorar a resposta de nosso ACR.

Também apresentamos os conceitos de funcionamento das diferentes técnicas de classificação e as métricas que serão utilizadas nos capítulos subseqüentes. E como o nossa abordagem de monitoramento bioacústico será embarcada nos nós sensores, resumimos as características principais das RSSF, consumo de bateria e as aplicações encontradas.

3

Trabalhos Relacionados

Este capítulo apresenta as soluções atuais de monitoramento ambiental bioacústico. Os trabalhos revisados abordam a descrição, comparação e aplicação das diferentes abordagens utilizadas para reconhecimento de espécies animais. O capítulo está organizado em quatro seções principais, abordando os assuntos relacionados à filtragem, segmentação, extração de características e classificação. Esta organização mantém a coerência com as etapas do ACR apresentado na figura 1.1 (página 5).

A flexibilidade do ACR possibilita que cada estratégia de reconhecimento combine diferentes técnicas de classificação, diferentes LLD e diferentes formas de segmentação e filtragem. Consequentemente, o objetivo deste capítulo é apresentar uma descrição organizada em categorias que facilitem a combinação e escolha das componentes para desenhar e planejar uma abordagem com propósitos específicos e ou gerais.

3.1 Introdução

A organização geral dos conteúdos deste capítulo é ilustrada na figura 3.1. Com esta organização pretendemos esclarecer a flexibilidade das abordagens de reconhecimento bioacústico. Por exemplo, ao se planejar um novo sistema poderíamos combinar: (a) um filtro Wavelet, (b) uma segmentação baseada nos valores de energia do sinal, (c) uma representação mediante LLD no domínio da frequência e, finalmente, (d) aplicar a cada sílaba alguma técnica de aprendizagem de máquina no nó *sink*, o que resultaria em um sistema centralizado.

Ao se desenhar um novo sistema nem todas as etapas são obrigatórias, i.e., a filtragem e a utilização de RSSF são opcionais. Com as evidências apresentadas na seção 1.3 do capítulo 1, mostramos que os ruídos ambientais causam uma diminuição

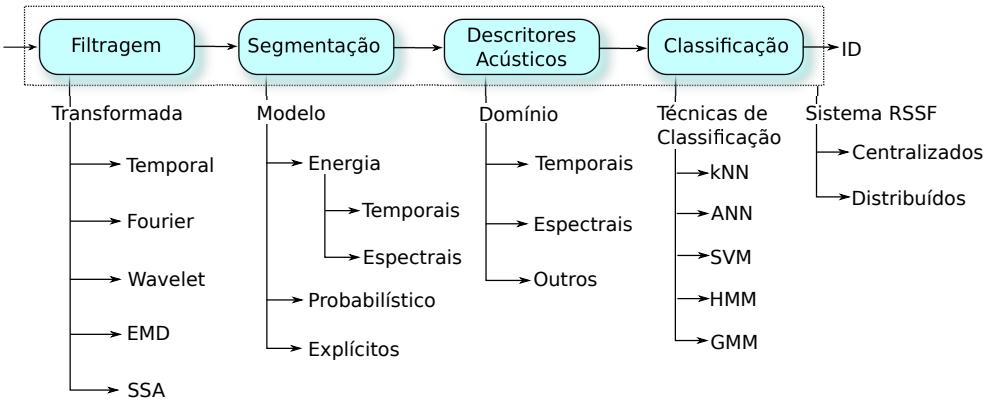


Figura 3.1. Classificação e organização das partes que compõem o ACR de reconhecimento de anuros.

na taxa de reconhecimento das espécies. Entretanto, a utilização de filtros de sinais, e o impacto destes, é um assunto menos frequente nas abordagens de reconhecimento bioacústico, porém, complexo e necessário.

Na seção 3.2 apresentamos os tipos de filtro de ruído mais utilizados no contexto bioacústico, separados segundo o domínio de aplicação, caracterizado pela transformada prévia aplicada aos sinais. As abordagens incluem: (1) o domínio espectral utilizando Fourier; (2) o domínio da transformada Wavelet; (3) o domínio das funções intrínsecas (IMF) utilizando a transformada EMD; e (4) o domínio das componentes principais utilizando SSA.

Na seção 3.3 apresentamos as diferentes abordagens de segmentação bioacústica. A segmentação serve para selecionar os trechos das vocalizações com informações relevantes para a classificação. Este é um dos componentes principais do ACR e impacta diretamente nos resultados finais. Organizamos a segmentação segundo três modelos fundamentais, baseados em: (a) energia (ou características físicas dos sinais), (b) em probabilidades ou (c) funções explícitas (ou classificadores supervisionados). Os modelos de energia foram separados segundo o domínio dos LLD em: temporais ou espectrais. Além disso, identificamos as propostas supervisionadas e não supervisionadas.

Os diferentes conjuntos de descritores acústicos utilizados para a classificação também foram organizados segundo o domínio no qual são obtidos, ressaltando a complexidade adicional e as vantagens de realizar transformações dos áudios antes da classificação (seção 3.4). As diferentes técnicas de aprendizagem de máquina geralmente aplicadas no contexto bioacústico estão resumidas na seção 3.5.1.

No que diz respeito as aplicações práticas, utilizando Redes de Sensores Sem Fio (RSSF), nas seções 3.5.2 e 3.5.3 separamos as abordagens como centralizadas ou

colaborativas. Nos casos colaborativos, identificamos três possibilidades de fusão de dados frequentemente utilizadas entre os sensores: fusão de áudio, fusão de descritores acústicos e fusão de decisões. E por último, na seção 3.6 apresentamos as conclusões.

3.2 Filtragem de sinais bioacústicos

Os diferentes tipos de ruídos provenientes da floresta degradam a qualidade dos sinais causando variações nos padrões a serem classificados. Assim, poderíamos definir ruído, desde o ponto de vista do classificador, como qualquer sinal com comportamento estocástico que não pertence ao conjunto de treinamento. Os diferentes ruídos podem ser aleatórios Gaussianos brancos ou coloridos, como é o caso dos sons ambientais, ou aqueles produzidos artificialmente. No estudo de Bardeli et al. (2010) foram relevados três tipos de ruídos que prejudicam a classificação: (1) os *biogenic sound* que são causados por outras espécies animais e possuem uma forte correlação com o horário do dia, (2) os *anthropogenic sounds* que são artificias causados principalmente por motores como carros ou barcos, e (3) os *ambiental sounds* que são ambientais como a chuva ou o vento.

Os ruídos são sinais com função de densidade espectral de potência (PSD) que respeita uma lei exponencial proporcional a $1/f^{\alpha}$. Os ruídos aleatórios brancos possuem a característica de ocuparem a totalidade do espectro de frequências ($\alpha = 0$), enquanto que os ruídos coloridos possuem correlações que se manifestam com maior intensidade em determinadas bandas de frequências dependendo do valor de α (Rosso et al., 2007, Lau et al., 1998). Os sons da chuva e do vento, por exemplo, podem ser considerados ruído coloridos.

3.2.1 Filtros temporais

Na teoria clássica de processamento de sinais existem diversas técnicas de filtragem a partir da definição das frequências de interesse, tais como filtros passa-faixa ou passa-baixas (Oppenheim and Schafer, 2010). Nestes casos as funções de transferência possuem parâmetros que permitem determinar o valor de atenuação nas diferentes bandas de frequências do sinal. Embora os requisitos para definir o filtro sejam definidos em termos das frequências de corte, a função de transferência h é um polinômio com j coeficientes, resultando em uma equação temporal da forma $x = h_1 x_{i-1} + h_2 x_{i-2} + \cdots + h_j x_{i-j}$.

No ACR desenvolvido por Xie et al. (2015b) (figura 3.11) a etapa de redução de ruídos foi baseada na proposta de Towsey et al. (2014b). Neste trabalho os autores aplicaram, primeiro, um filtro passa-baixos com frequência superior de corte igual a

8.82 kHz e posteriormente um filtro de média móvel ($w = 3$) para suavizar o sinal. Com este procedimento Xie et al. (2015b) concluíram que a filtragem teve um impacto positivo no cálculo dos descritores acústicos, que consequentemente, melhorou a acurácia no reconhecimento das espécies.

Um filtro passa-banda define uma faixa de frequência de interesse, filtrando as demais. Determinar as frequências de corte *a priori*, em nossa aplicação não é trivial, devido as diferentes espécies que possuem bandas de frequências sobrepostas, dificultando a separação. Além disso, estes filtros não são dinâmicos, i.e. não se adaptam as diferentes condições da floresta.

O filtro de Wiener é uma abordagem que atualiza a resposta do filtro dinamicamente (Ahlén and Sternad, 1991). Esta técnica ajusta os coeficientes do polinômio h para minimizar o MSE entre o sinal de entrada y e o sinal estimado $\hat{x}$. Aplicando-se um procedimento iterativo é possível obter os valores de h que maximizam a relação SNR. Este filtro, cuja função de transferência é:

$$H_f = \frac{X_f}{X_f - \xi_f}, \tag{3.1}$$

é considerado ótimo porque ajusta H_f para minimizar a diferença entre os espectros complexos $\mathrm{E}[(\hat{X}_f - X_f)^2]$, sendo X_f o espectro do sinal. No entanto, este filtro não é perfeito, porque modifica tanto a fase quanto a amplitude das frequências (Loizou, 2013). Esta técnica foi utilizada no trabalho de Ren et al. (2008) como método de comparação aplicado a duas espécies de macacos e de uma baleia.

Para aplicar o procedimento iterativo e encontrar H_f é utilizando Y_f (o sinal ruidoso) e o espectro limpo X_f. Entretanto, para encontrar os coeficientes h não é necessário realizar explicitamente a transformada de Fourier, basta simplesmente aplicar um procedimento para diminuir o MSE. Desta forma, devem-se utilizar sinais de referência livres de ruídos. Porém, isto não é possível, pois todas as espécies de nossa base foram gravadas diretamente na floresta em condições reais. Além disso, este filtro causa algumas distorções relacionadas com a fase dos sinais recuperados (Chen et al., 2006a).

O requisito de precisar de um sinal de referência livre de ruído pode ser evitado utilizando uma versão da mesma vocalização atrasada algumas unidades de tempo. Este é o conceito da técnica conhecida como ALE (*Adaptive Line Enhancer*), que foi aplicada como método de comparação no trabalho de Gur and Niezrecki (2011) sobre vocalizações de peixe-boi. A ALE baseia-se no conceito de predição linear, no qual um sinal quase periódico pode ser previsto utilizando combinações lineares de suas últimas amostras, enquanto que um sinal aleatório não. Isto ocorre porque o ruído aleatório

não tem nenhum grau de autocorrelação. No entanto, os sons ambientais no fundo das gravações não são totalmente aleatórios, fato que diminui a eficácia deste método.

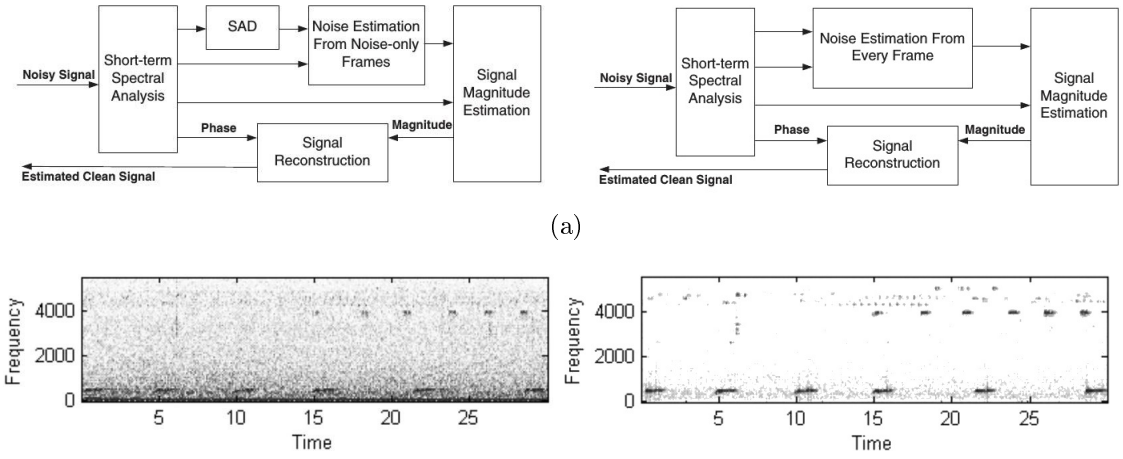
3.2.2 Filtros baseados na transformada de Fourier

O objetivo do método de substração espectral média (SMS), aplicado por Cai et al. (2007) aos cantos de pássaros, é estimar o nível de ruído ambiental em cada banda de frequência como o valor médio de todos os *frames* do sinal de entrada $\overline{\xi}_f$ (seção 2.3.6, página 34).

Na abordagem clássica de SMS é incluído um detector de atividade acústica (VAD) para separar os *frames* que possuem somente ruído dos que contém sinal. Assim, o espectro de frequências do ruído é baseado na informação dos *frames* que não possuem sinal, para posteriormente realizar $\hat{X}_f = Y_f - \overline{\xi}_f$. A operação de substração é aplicada no domínio da frequência e, mediante a transformada inversa de Fourier ($IFFT$), é obtido o sinal livre de ruídos no domínio temporal. Esta abordagem precisa de uma etapa de calibração prévia e um mecanismo de adaptação contínuo, que permita estimar o espectro do ruído a partir do novos *frames* de áudio. A seleção dos *frames* para estimar $\overline{\xi}_f$ deve ser cuidadosa, para não perder informação das frequências que sejam características da espécie monitorada e prejudicar a taxa de acerto do classificador.

No caso de Cai et al. (2007) o sinal completo, incluindo os *frames* com sílabas, foram utilizados para estimar o perfil do filtro. Desta forma, os autores evitaram utilizar um detector de atividade acústica (VAD) para separar os *frames*, mostrando que não existe diferença significativa (figura 3.2(a)). Isto ocorre porque as sílabas das vocalizações bioacústicas são esparsas e os segmentos do sinal sem atividade acústica são mas extensos e frequentes, fazendo com que a média de cada frequência se aproxime do piso de ruído real. Os espectrogramas resultantes encontram-se na figura 3.2(b). Os autores também concluíram que a etapa de filtragem melhorou o resultado do reconhecimento.

As abordagens de classificação baseadas na utilização do espectrograma como imagem, que aplicam máscaras binárias sobre os *pixels*, podem ser consideradas filtros no domínio de Fourier. Por exemplo, no trabalho de Neal et al. (2011) foi utilizada uma máscara binária para reduzir ruídos. Nos trabalhos de Potamitis (2014) e Ventura et al. (2015) os filtros morfológicos, além de reconhecer formas dos agrupamentos dos *pixels*, também filtram ruídos. Em outras palavras, estas máscaras cumprem o papel de filtrar e melhorar o sinal antes de realizar a operação seguinte, melhorando o resultado do reconhecimento das espécies.



(a)

(b)

Figura 3.2. Figuras extraídas de Cai et al. (2007). (a) Sistemas de filtro de sinal com e sem VAD. (b) Espectrogramas do sinal original e filtrado.

3.2.3 Filtros baseados na transformada Wavelet

Existem técnicas de filtragem em diferentes domínios de transformações além de Fourier. Este é o caso do *Soft* e *Hard Threshold* aplicando no domínio Wavelet. Exemplos deste tipo de filtro encontram-se nos trabalhos de Gur and Niezrecki (2007) e Gur and Niezrecki (2011), que aplicaram estes filtros para melhorar a qualidade das gravações dos peixes-boi.

Neste tipo de filtro não é necessário estimar *a priori* nenhum parâmetro dos sinais. O procedimento consiste em aplicar a transformada para obter os coeficientes de detalhes e aproximação, atenuar o valor destes coeficientes aplicando uma regra baseada em limiar de corte e aplicar a transformação inversa para recuperar o sinal filtrado (seção 2.3.8, página 38).

Desta forma encontra-se uma relação custo-benefício entre quantidade de coeficientes retidos, pela aplicação do limiar, e a distorção da reconstrução. Assim, limiares baixos filtram menos, enquanto que, limiares altos deixam resultados com maior distorção. No trabalho de Gur and Niezrecki (2007) os autores adotaram um critério para encontrar o limiar ótimo baseado nos valores da autocorrelação dos coeficientes $\mathcal{W}$. Entretanto no trabalho de Gur and Niezrecki (2011) foi adotado o critério de minimização SURE com um limiar diferente para cada nível da transformada $\mathcal{W}$. A figura 3.3 apresenta comparativa entre o filtro ALE e *Wavelet Package Transform* extraída de Gur and Niezrecki (2011).

Finalmente, Ren et al. (2008) também aplicaram uma variante da transformada

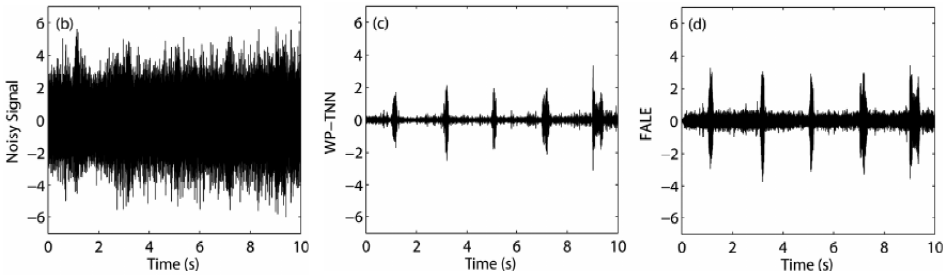
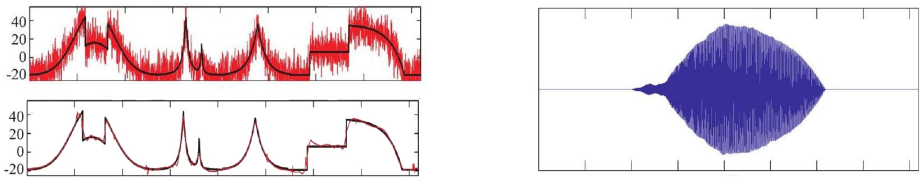


Figura 3.3. Filtragem de uma vocalização de peixe-boi com ruído branco aplicando (c) WPT e (d) ALE extraída Gur and Niezrecki (2011).

$\mathcal{W}$ e compararam esta com o filtro de Wiener e de *Ephraim-Malah*. Neste caso foram utilizadas as vocalizações de duas espécies de macacos e uma de baleia, concluindo que a melhor alternativa é a transformada *Wavelet*.

3.2.4 Filtros baseados no EMD

O método EMD decompõe o sinal original em um conjunto componentes conhecidas como funções intrínsecas (IMF) (Rilling et al., 2003). Com um procedimento similar ao *Soft* e *Hard Threshold* da transformada Wavelet, Kopsinis and McLaughlin (2009) aplicaram um limiar aos coeficientes de cada IMF antes de realizar a reconstrução. Um exemplo resultante deste procedimento é apresentado na figura 3.4(a). A figura 3.4(b) é um exemplo de vocalização de morcego filtrada a partir deste método, utilizada por Kopsinis and McLaughlin (2009). Embora o procedimento de obtenção dos coeficientes EMD seja totalmente empírico, os resultados gerados mostraram-se interessantes, principalmente pela vantagem de ser uma transformação não linear.



(a) (b)

Figura 3.4. Figuras extraídas de Kopsinis and McLaughlin (2009). (a) Sinal definida por partes com ruído aleatório Gaussiano e sua versão filtrada utilizando EMD. (b) Sinal filtrado correspondente a uma vocalização de um morcego

3.2.5 Filtros baseados em SSA

O *Singular Spectrum Analysis* é uma técnica similar com a decomposição em componentes principais apropriado para decompor sinais temporais de uma dimensão (seção 2.4, página 39). As funções obtidas pela decomposição SSA são ordenadas de acordo com o grau de contribuição segunda a variância de cada componente. Isto significa que as componentes de maior variância carregam informação sobre as frequências com maior energia, e portanto, representam melhor o sinal original.

Com este método a filtragem acontece durante a seleção dos autovetores que formam as bases da reconstrução. Comumente são escolhidos os autovetores que corresponde com os maiores autovalores. Desta forma, uma fração da variância dos ruídos contidos nas componentes principais representadas pelos menores autovalores é cortada da reconstrução (Allen and Smith, 1997). Recentemente, Tomé et al. (2010) mostraram que SSA possui uma equivalência com um banco de filtros em paralelo, os quais podem ser obtidos a partir dos autovetores da matriz de correlações do sinal de entrada. No artigo de Tomé et al. (2011) é relacionada a teoria de sistema invariantes linear (LTI) mostrando a equivalência com os filtros de resposta impulsiva finita (FIR). Além disso, mostraram as implicâncias de aplicar esta transformada no contexto de sinais biomédicos.

3.2.6 Considerações Sobre Filtragem

A tabela 3.1 resume as propriedades principais das técnicas de filtragem segundo a transformada de sinal utilizada. A tabela evidencia as propriedades de cada técnica, para simplificar a escolha ao se desenhar um novo sistema de reconhecimento bioacústico. Por exemplo, ao se utilizar a subtração espectral, fica claro que não pode ser realizada de forma incremental, fato que requer uma capacidade de processamento maior em comparação a transformada Wavelet, para aplicação de processamento de áudio.

No caso específico da subtração espectral existem três considerações adicionais. A primeira é a necessidade de selecionar corretamente os *frames* que serão utilizados para gerar o perfil de ruído, a segunda é que com este método aparecem artefatos nas altas frequências que podem prejudicar a classificação e, terceiro, no caso que desejar realizar a reconstrução do sinal, deve-se manter a fase do sinal original (Vaseghi, 2000, Verteletskaya and Simak, 2011).

O filtro baseado na transformada Wavelet é o mais apropriado para hardware com baixos recursos. A capacidade incremental do *Lifting Scheme* permite realizar a transformação em tempo real em RSSF (Rein and Reisslein, 2011). No entanto no

Tranformada	Bases	Linear	Representação	Método	Procedimento	Complexidade
Fourier	*a priori*	Sim	Freq	Substração Espectral	Bloco	$\mathcal{O}(n \log n)$
Wavelet	*a priori*	Sim	Tempo Freq.	Soft-Hard Threshold	Incremetal (Real Time)	$\mathcal{O}(n)$
EMD	Adaptável	Não	Tempo Freq.	Soft-Hard Threshold	Bloco (Pseudo Real Time)	$\mathcal{O}(n \log n)$
SSA	Adaptável	Sim	Tempo Freq.	LTI (FIR)	Bloco	$\mathcal{O}(2KL^2 + Kn^2)$

Tabela 3.1. Características das diferentes abordagens de filtrado no contexto bioacústico classificadas segundo a transformação do sinal utilizada. Neste tabela n representa o comprimento do sinal, L e K o tamanho da matriz de trajetórias da decomposição SSA.

capítulo 6 mostramos que certos tipos de ruídos coloridos, com energia concentrada nas baixas frequências, afetam a qualidade da filtragem.

O filtro baseado em EMD necessita de fundamentação teórica, pois é simplesmente um procedimento empírico, fato pelo qual, provavelmente, tem sido menos explorado na literatura deste contexto. A principal vantagem desta técnica é a transformação não linear do sinal nas componentes IMF. No entanto, os requisitos de memória dificultam a implementação em um sensor de baixo custo.

Tanto EMD quanto SSA possuem a vantagem de criar as bases da transformada em função do sinal de entrada. Diferente de Fourier ou Wavelet que utilizam um conjunto de bases fixas, SSA otimiza a transformação tornando esta mais compacta. Outras vantagens e propriedades de SSA são exploradas no capítulo 6. Além disso, apresentamos exemplos que mostram a eficácia desta técnica para decompor as vocalizações dos anuros e a capacidade de filtragem de ruídos diferentes na natureza, sejam estes aleatórios brancos, coloridos, ou ambientais. Apesar das vantagens, este método trabalha com matrizes grandes, tornando-se custoso para redes de sensores sem fio.

3.3 Segmentação

O problema de segmentação de áudio não é exclusivo das aplicações ambientais. A segmentação é objeto de estudo em abordagens de reconhecimento de voz (Bhandari et al., 2014, Kaur and Kaur, 2013, Ramírez et al., 2004, Rybach et al., 2009), música (Foote, 2000) e *streaming* (Cettolo et al., 2005, Giannakopoulos et al., 2008). Em relação a estes trabalhos existe uma classificação de três categorias utilizadas por Cettolo et al. (2005). Assim, podemos classificar os modelos de segmentação bioacústica em: (1) modelos baseados em energia; (2) modelos baseados em probabilidades; e (3) modelos explícitos.

3.3.1 Modelos de segmentação baseados em energia

Nestes modelos de segmentação são utilizadas as mudanças nos valores dos descritores acústicos com o objetivo de encontrar o início e o fim das sílabas. Os descritores utilizados geralmente relacionam-se com os valores de energia do sinal e podem ser obtidas no domínio temporal ou espectral. Neste caso, é monitorada a mudança destes valores em relação ao tempo para encontrar e separar os padrões. Estes são os modelos mais simples e frequentemente utilizados devido ao baixo custo computacional e a abrangência para segmentar as vocalizações das diferentes espécies.

3.3.1.1 Segmentação com descritores no domínio temporal

A técnica de extração de sílabas mais simples utiliza um limiar de amplitude, uma janela com tamanho fixo pré-definido e um procedimento iterativo temporal. A amplitude do sinal relaciona-se com a energia deste através da equação 2.1 (página 20).

O protocolo de segmentação encontrado nos trabalhos de Huang et al. (2009), Cheng et al. (2010), Colonna et al. (2012) consiste em três passos básicos: (1) percorrer o áudio até encontrar o valor máximo de amplitude; (2) a partir deste valor selecionar α amostras a direita e a esquerda; e (3) extrair a sílaba e apagar estes valores. Este procedimento deve ser repetido até que o máximo valor de amplitude seja menor que β (figura 3.5). Embora esta técnica seja simples, podemos destacar três possíveis causas de erros: o primeiro, é a sensibilidade a ruídos impulsivos elevados maiores que β; o segundo problema, é determinar o comprimento da sílaba (2α) *a priori* e mantê-lo constante para todas as espécie; e o terceiro problema, é a necessidade de armazenar o áudio completo antes de percorrê-lo, o que torna uma abordagem com custo alto de memória para um nó sensor.

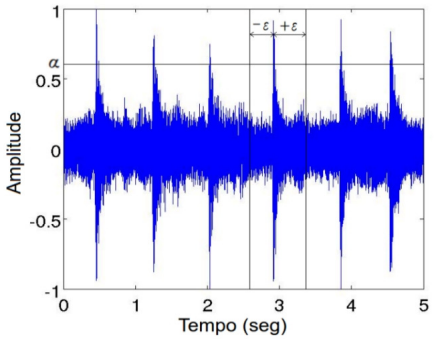


Figura 3.5. Vocalização da espécie *Adenomera andreae* indicando o limiar de amplitude β e temporal α. Figura extraída de Colonna et al. (2012).

Fagerlund (2007) também utilizaram um limiar de amplitude para a segmentação. Entretanto, a diferença de Huang et al. (2009), é aplicado um procedimento iterativo que ajusta automaticamente o valor ótimo de corte β (T dB). Neste caso, antes de começar a segmentação é calculada a energia da envoltoria do sinal (normalizada 0 dB). O valor inicial do limiar é estimado como a metade do nível de ruído inicial, que por sua vez, é definido como o mínimo global da energia. O algoritmo consta de três passos: 1) encontrar as sílabas maiores que o limiar T_{dB}, 2) atualizar o valor de amplitude dos ruídos (ξ_{dB}) com as amostras que sobraram, e 3) atualizar o valor do limiar com $T_{dB} = {}^{\xi_{dB}}/_{2}$. Este procedimento é repetido até a convergência de T_{dB}. Finalmente, Fagerlund (2007) aplicaram uma técnica de suavizado que agrupa segmentos muito próximos para evitar micro-segmentações.

O algoritmo de segmentação proposto por Harma and Somervuo (2004) utiliza a base dos procedimentos com os valores da energia do sinal e um limiar T_{dB}. Entretanto, para encontrar o valor ótimo os autores utilizaram uma regra diferente. O método iterativo de quatro passos é:

1. encontrar os valores máximos (M_{dB}) de energia da envoltória do sinal;
2. separar as regiões entre os pontos M_{dB} e o limiar T_{dB};
3. estimar ξ_{dB} usando os valores que sobram entre as duas regiões máximas;
4. atualizar o valor do limiar da forma $T_{dB} = {}^{(M_{dB}-\xi_{dB})}/_{2}$ e voltar ao segundo passo até não encontrar variações significativas no valor de ξ_{dB}.

Ao finalizar, o limiar T_{dB} indica os valores do começo e final de cada sílaba. Neste caso a regra que atualiza os valores de T_{dB} é mais robusta separando melhor as médias dos grupos sinal e ruído.

Um procedimento iterativo similar ao de Fagerlund (2007) foi utilizado por Somervuo et al. (2006). Neste trabalho, para corrigir os micro cortes, e formar uma sílaba maior, foram agrupados segmentos com menos de 15 ms entre eles. Notamos que a fragmentação das sílabas é um problema mais frequente do que a detecção de falsos positivos, ao se utilizar características temporais dos sinais.

Uma estratégia diferente para melhorar a segmentação das sílabas é combinar mais de um descritor acústico (equação 3.2). Este tipo de abordagem foi aplicada por Rahman and Bhuiyan (2012), Jaafar and Ramli (2013), Jaafar et al. (2014) para sinais bioacústicos, e também por Hemakumar and Punitha (2014) para sinais de fala. A abordagem utiliza os valores de energia do sinal, para encontrar as variações das amplitudes, e a taxa de cruzamento por zero, para aproximar a frequência do sinal.

Com esta regra as comparações feitas são em relação aos *frames* anteriores, ou aos *frames* considerados ruídos. Assim, uma sílaba é detectada quando ambas características superam os limiares:

$$T_h = \begin{cases} 1 & \text{se } E_n \geq T_E \text{ e } Z_n \geq T_Z \\ 0 & \text{outro caso} \end{cases} \tag{3.2}$$

na qual T_h é o limiar composto que indica presença ou ausência de sílaba (1 ou 0), $E_n \geq T_E$ é a condição do limiar (T_E) de energia e $Z_n \geq T_Z$ é a condição de limiar para o ZCR no tempo n. O ZCR e E são características frequentemente utilizadas em processamento de sinal, por serem computacionalmente simples comparadas com as transformações dos sinais e fornecer informações complementares. A condição com dois limiares torna o método menos sensível a falsos positivos, porém, é necessário identificar as melhores características para combinar e encontrar os respetivos limiares.

Potamitis et al. (2014) utilizaram a transformada de *Hilbert* ($\mathcal{H}$) para segmentar o canto de aves. Embora nesta abordagem exista uma transformação do sinal, esta representa uma nova série temporal não relacionada com o espectro de frequências. Assim, o *Hilbert follower* é utilizado para recuperar a envoltoria do sinal e extrair o sinal modulante das sílabas.

O resultado da transformada $\mathcal{H}$ é um sinal analítico com partes reais e imaginarias da forma $\mathcal{H}(x) = x + jx$, na qual o sinal modulante é obtido aplicando $y = (|\mathcal{H}(x)|)^{1/2}$. A partir de y é possível realizar a segmentação aplicando um limiar de amplitude detectando o começo e o final de cada sílaba. Embora Potamitis et al. (2014) tenha apresentado resultados levemente melhores aos trabalhos anteriores, a obtenção da envoltória requer cálculos adicionais e uma memória capaz de armazenar a série temporal completa.

No trabalho de Garcia et al. (2014), ao invés de utilizar uma transformação matemática de bases diferentes, é utilizado o próprio sinal para obter a autocorrelação. Neste caso, a autocorrelação é uma alternativa à FFT para encontrar a frequência fundamental (f_0) da espécie (Shimamura and Kobayashi, 2001). Este método não depende do conteúdo harmônico do sinal, sendo mais apropriado para caracterizar vocalizações de animais com frequências concentrados em uma única banda, constituindo uma diferença fundamental dos sistemas de reconhecimento de voz.

3.3.1.2 Segmentação com descritores espectrais

As abordagens de segmentação que utilizam características espectrais podem ser separadas em dois grupos: as que utilizam o espectro de frequências, para extrair carac-

terísticas dos sinais, e as que a partir do espectro geram o espectrograma e aplicam técnicas de processamento digital de imagens.

Rickwood and Taylor (2008) utilizaram a energia das diferentes bandas de frequências em sua proposta para segmentar o canto das baleias. Com este modelo é assumido que as vocalizações contenham maior concentração de energia em determinadas bandas espectrais, diferentemente dos ruídos aleatórios que contenham energia uniforme em todas as bandas. Assim, é aplicada a FFT para gerar o espectro de frequências, a partir do qual é realizado o histograma da densidade espectral de potência de cada banda. Com os histogramas é estimado o limiar da segmentação, que é utilizado para comparar cada *frame*. Após a segmentação é aplicado um método de suavização que considera a classe dos *frames* vizinhos para diminuir os falsos positivos e negativos.

Aplicando a FFT é possível gerar o espectrograma do sinal em uma representação que relaciona o tempo com a frequência, formando uma matriz $\mathrm{S}(f,t)$. Nos trabalhos de Harma (2003) e Lee et al. (2006) é utilizada esta matriz como imagem, na qual a segmentação é realizada de forma iterativa:

1. procurar o par f,t com o valor máximo de $|\mathrm{S}(f,t)|$ expressado em dB ($\mathrm{A} = 20\log_{10}|\mathrm{S}(f,t)|$); e

2. a partir deste ponto procuram-se os valores de $t_1 < t < t_2$ que satisfazem a condição $\mathrm{A}(t_{1,2}) < \mathrm{A} - \beta$ para determinar o começo da sílaba (t_1) e o final (t_2);

3. extrair a sílaba e apagar os valores respectivos $\mathrm{S}(t_{1\leq 2}, f) = 0$; e

4. repetir estes passos até não encontrar valores superiores a $\mathrm{A} > -20dB$.

A condição β (dB) é um critério de finalização que deve ser pré-definido.

Xie et al. (2015a) complementaram este procedimento adicionando um filtro Gaussiano, em formato de máscara com $5x5$ *pixels*, para suavizar o sinal antes de aplicar a segmentação de Harma (2003) e Lee et al. (2006). Desta forma, o sinal original torna-se menos ruidoso, melhorando o resultado do procedimento descrito. Note-se, que aplicar a suavização antes da segmentação é a principal diferença dentre os trabalhos já descritos, indicando que a segmentação pode ser melhorada se o sinal de entrada for mais apropriado.

No espectrograma a cor dos *pixels* representa a energia em cada ponto $\mathrm{S}(f,t)$. O tratamento descrito por Aide et al. (2013) explora esta representação e agrupando os pixeis segundo os valores dos vizinhos formando Regiões de Interesse (ROIs). Posteriormente, a segmentação ocorre eliminando os pixeis com valores inferiores aos 10%

da média total. Assim, os pixeis resultantes são armazenados em uma matriz esparsa que será usada para a classificação. Este procedimento requer que seja aplicada uma técnica de agrupamento binária não supervisionada, que resulta computacional mais complexa comparada com os procedimentos iterativos descritos. No entanto, se abrem as possibilidades de aplicar algoritmos para reconhecimento de imagens mais robustos.

Potamitis (2014) mantiveram a ideia de aplicar sobre o espectrograma diversas técnicas de processamento digital de imagens. O primeiro processo consiste de operadores morfológicos que permitam derivar funções de máscaras para encontrar as ROIs. Estas máscaras otimizadas para as espécies de interesse permitem eliminar a maioria dos ruídos e padrões não necessários. Posteriormente a imagem do espectrograma é binarizada. No final do processo ficam mais evidentes as faixas de frequências de cada vocalização. A figura 3.6 apresenta o espectrograma resultante desta técnica, embora o resultado seja muito interessante o custo de aplicação tornasse inviável em um nó sensor.

Oliveira et al. (2015) estenderam a aplicação de filtros morfológicos aplicados ao espectrograma para definir o "Detector de Atividade Acústica" da figura 3.7. Este sistema é otimizado para encontrar e separar as vocalizações de uma espécie de pássaro (*Lapwing Vanellus chilensis*), que foi utilizada como prova de conceito pelos autores.

Na proposta de Evangelista et al. (2014) os *frames* do espectrograma são convertidos em duas sequências, uma com a energia das frequências e outra com o espectral centroide. Neste caso, para cada nova sequência é gerado um histograma ao qual aplica-se um filtro da mediana. O resultado gera dois grupos como máximo local X_1 e X_2. A partir destes valores é aplicada uma regra composta de dois limiares similar aos trabalhos de segmentação com características temporais de Rahman and Bhuiyan (2012), Jaafar and Ramli (2013), Jaafar et al. (2014). Esta abordagem não utiliza o espectrograma como imagem, sem embargo, é necessário manter os áudios completos para poder gerar os histogramas.

A aplicação de técnicas de processamento digital de imagens para segmentar espectrogramas de aves é longamente utilizada. Recentemente o trabalho de Lasseck (2014), citado no concurso de reconhecimento bioacústico de Joly et al. (2014), aplicou:

1. segmentação binária da imagem;
2. operadores de contração e dilatação (*closing and dilatation*);
3. o filtro da mediana; e
4. remoção de pequenos objetos,

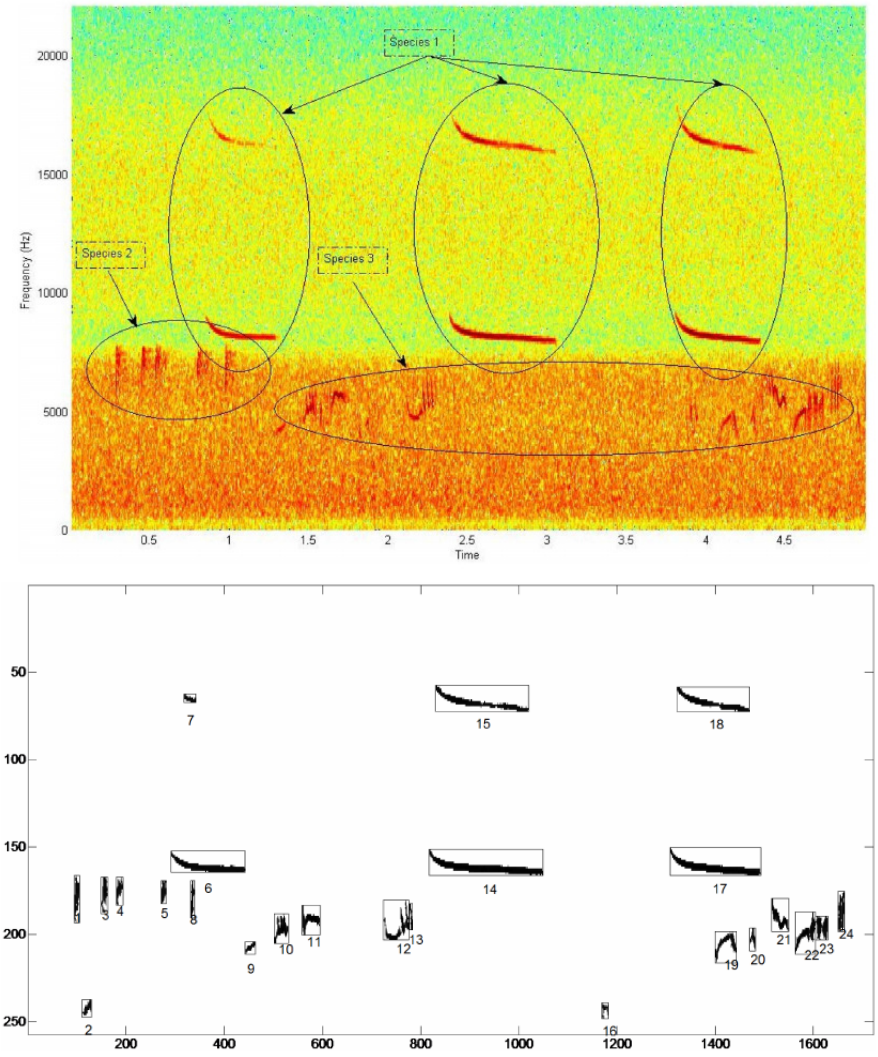


Figura 3.6. (a) Espectrograma original e (b) espectrograma binarizado. Figuras extraídas de Potamitis (2014).

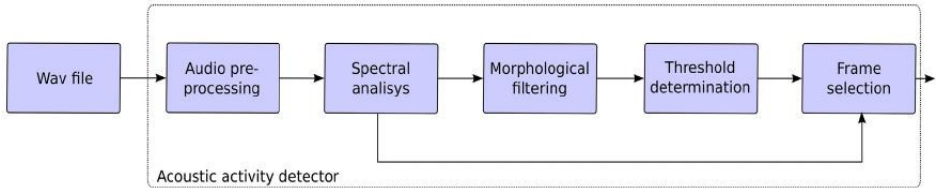


Figura 3.7. Sistema de detecção de atividade acústica aplicada à espécie de pássaro *Lapwing Vanellus chilensis*. Figura extraída de Oliveira et al. (2015).

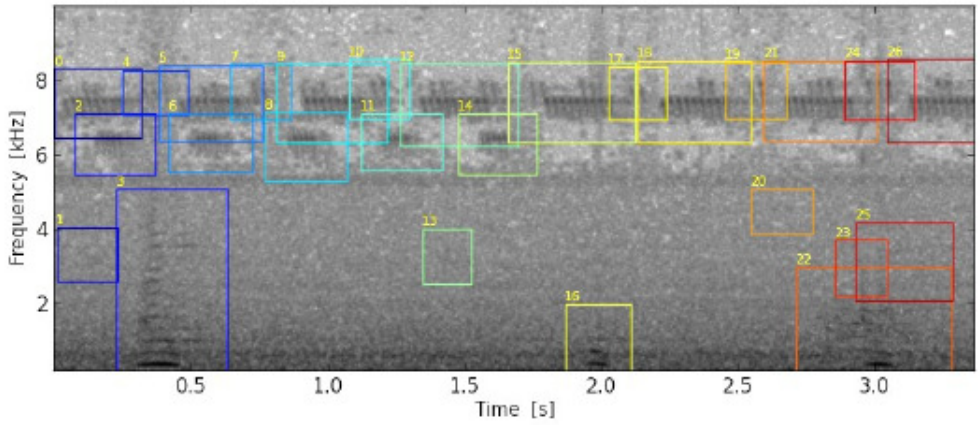


Figura 3.8. Agrupamentos encontrados no espectrograma pela técnica descrita por Lasseck (2014).

obtendo o agrupamento da figura 3.8.

Os quatro passos descritos resumem o procedimento aplicado na maioria das abordagens de segmentação que utilizam o espectrograma. Podemos destacar que estas técnicas baseiam-se, principalmente, na aplicação de um método para identificar os *pixel* de maior intensidade e a aplicar uma máscara (ou filtro) para diminuir as micro-segmentações. Embora os resultados sejam interessantes, utilizar processamento de imagens aumenta consideravelmente os requisitos de memória necessária no nó sensor. No entanto, se desconsideramos as restrições de hardware, novas técnicas de processamento digital de imagens podem ser utilizadas, por exemplo: as *Deep Convolutional Neural Networks* (Krizhevsky et al., 2012, Espi et al., 2015) ou outras técnicas derivadas da teoria *Deep Learning* (Lee et al., 2009, Deng et al., 2010). Sendo esta, novas oportunidades dentro do monitoramento ambiental bioacústico.

3.3.2 Modelos baseados em probabilidades

Esta categoria inclui os trabalhos que utilizam modelos probabilísticos para descrever as séries temporais. Com estes cada classe acústica, por exemplo, música, fala ou ruído, é representada por uma função de densidade de probabilidades (*Probability Density Function* - PDF), a partir da qual é aplicado um critério de comparação que permite encontrar as fronteiras da segmentação. O *Bayesian Information Criterion* (BIC) é um destes critérios (Cheng and Wang, 2003, Cettolo et al., 2005, Heinicke et al., 2015). Outras possibilidades incluem: comparar as diferentes regiões segmentadas utilizando os valores de entropia de cada PDF, ou aplicar um cálculo de similaridade, como a divergência de *Kullback-Leibler* (KL) ou Maximum Mean Discrepancy (Shen et al.,

1998, Fagerlund and Laine, 2014, Sinn et al., 2013).

Utilizar a PDF possibilita a aplicação de diferentes quantificadores de informação, como a entropia de Shannon. No trabalho de Shen et al. (1998) a PDF é obtida a partir do histograma normalizado dos valores da FFT. Com essa PDF os autores calcularam a entropia da vocalização, e chamaram este quantificador de entropia espectral (H_f). Desta forma, a segmentação é realizada calculando H_f em cada *frame* e aplicando um limiar sobre os valores obtidos, para detectar o ponto final dos segmentos de áudio correspondentes às sílabas.

Lakshminarayanan et al. (2009) aplicaram o mesmo procedimento para obter a PDF espectral e, no lugar de calcular a entropia de Shannon, utilizaram a divergência de KL como uma PDF de referência uniforme. Desta forma, a PDF uniforme é considerada o ponto de equilíbrio do sistema, caracterizando os *frames* com comportamento aleatório. Após obter a KL de cada *frame*, os mínimos locais são utilizados para determinar as fronteiras entre os segmentos da vocalização. As regiões dentro dos limites são tratados como segmentos e é calculada a energia de cada um destes. Em seguida, é aplicado um procedimento para determinar o limiar de separação, baseado nos valores de energia, e apenas os elementos superiores a este limiar são tratados como sílabas.

A PDF de um segmento de áudio pode ser obtida como a distribuição de probabilidade dos descritores acústicos de cada *frame* dentro do próprio segmento. Esta abordagem foi utilizada por Heinicke et al. (2015) para segmentar vocalizações de primatas. Neste caso, as fronteiras entre os segmentos foi descrita pelo valor mínimo do *Generalized Likelihood Ratio* (GLR), e posteriormente, os segmentos foram agrupados de acordo com os parâmetros da PDF utilizando o critério BIC. Isto significa que para saber se dois *frames* consecutivos pertencem ao mesmo segmento duas hipóteses H_0 e H_1 devem ser comparadas. Neste caso $H_0 : (x_1, x_2, \cdots, x_i, x_{i+1}, \cdots, x_N) \sim N(\mu_x, \Sigma_x)$ são os valores do sinal correspondentes aos dois *frames* consecutivos, com parâmetros descritos pela média (μ_x) e a covariância (Σ_x), e $H_1 : (x_1, x_2, \cdots, x_i) \sim N(\mu_1, \Sigma)$ e $(x_{i+1}, \cdots, x_N) \sim N(\mu_2, \Sigma_2)$ é a hipótese que representa os dois *frames* separadamente. A regra de decisão aplicada é:

$$\mathrm{BIC} = \frac{N_x}{2} \log |\Sigma_x| - \frac{N_1}{2} \log |\Sigma_1| - \frac{N_2}{2} \log |\Sigma_2| + \lambda P, \tag{3.3}$$

onde N_x, N_1 e N_2 indicam o comprimento do sinal e dois *frames* respectivamente, e P relaciona-se com os graus de de liberdade ou a dimensão acústica (mais detalhes desta regra encontram-se no trabalho de Delacourt and Wellekens (2000)). O termo λ controla a penalidade e serve como parâmetro de sensibilidade para ajustar a regra de decisão que rejeita H_0 quando $\mathrm{BIC} < 0$.

Os áudios são um tipo particular de séries temporais e a principal dificuldade destes métodos resulta de encontrar a PDF correspondente. Para isto existem diferentes metodologias, como exemplo: ajustar os dados utilizando uma distribuição gaussiana ou converter primeiro a série em uma representação simbólica como SAX ou PE e posteriormente utilizar os histogramas dos símbolos (Bandt and Pompe, 2002, Lin et al., 2003). No caso da PE podemos destacar o trabalho de Sinn et al. (2013), no qual foi aplicada a PE para segmentar sinais elétricos EEG (eletroencefalografia). Nesta trabalho o critério utilizado foi *Maximum Mean Discrepancy*, que diferente do BIC este considera a discrepância entre três *frames* consecutivos, tornando o critério mais robusto.

A ideia de aplicar diferentes metodologias para obter as PDF's que representam a série resulta interessante e promissor para entender a natureza ou estrutura dos sons no cenários de floresta. Entretanto, este tipo de abordagens são menos frequentes nas aplicação bioacústicas. Com estas evidências propomos estudar, no capítulo 4, a aplicabilidade de diferentes metodologias de padrões ordinais para segmentar as vocalizações dos anuros.

3.3.3 Modelos explícitos

Nestes modelos uma técnica de classificação é treinada de forma supervisionada para reconhecer cada *frame* do sinal. Os *frames* devem ser representados por um conjunto de descritores e um especialista deve rotular manualmente cada um destes como "sinal" ou "ruído". Desta forma, diferentes tipos de classificadores podem ser utilizados para identificar os *frames* que correspondem as fronteiras de separação entre as sílabas.

Chu and Blumstein (2011) apontaram que uma vantagem deste tipo de abordagens é a possibilidade de realizar a segmentação e a classificação em um único passo. O modelo de segmentação/classificação de Chu and Blumstein (2011) é baseado em HMM e pode ser aplicado em um *stream* de áudio continuo sem a pré-segmentação. O sistema foi otimizado para detectar uma espécie de ave resultando em uma precisão superior a 70%. Entretanto, os autores relataram algumas dificuldades, como o fato de ter que aplicar um agrupamento primeiro para descobrir a possível quantidade de sílabas diferentes da mesma espécie, antes de criar os modelos HMM. Isto leva a um dificuldade adicional, devem-se treinar modelos HMM diferentes para uma mesma espécie e um modelo HMM específico para detectar segmentos que somente possuem ruídos ambientais.

Nesta categoria devemos incluir os trabalhos de Aide et al. (2013), Potamitis (2014), que empregam tratamento digital de imagens baseados nas figuras do espectro-

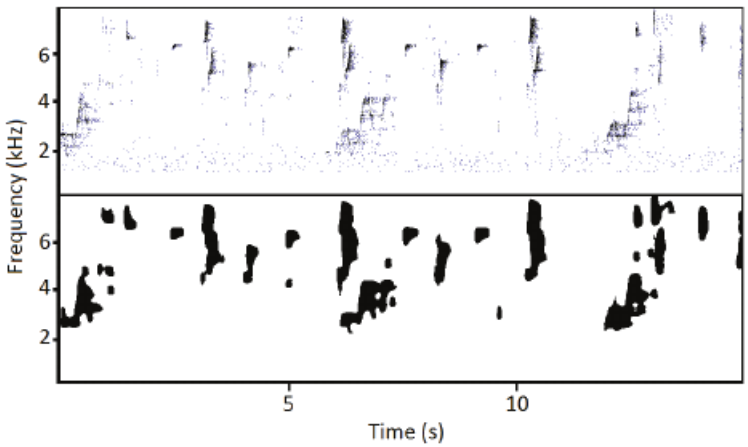


Figura 3.9. Espectrograma antes e depois da classificação binária. Figura extraída de Neal et al. (2011).

grama para segmentar e reconhecer as espécies. Nestes sistemas um especialista deve reconhecer as formas do agrupamento que os *pixels* possuem antes de criar os filtros morfológicos. No enteando, apesar de ser uma abordagem computacionalmente custosa para um nó sensor de baixos recursos, a segmentação e classificação podem ter lugar em um único passo.

Uma alternativa aos filtros morfológicos, mas também utilizando o espectrograma, é o trabalho de Neal et al. (2011). Neste caso os autores aplicaram a técnica de aprendizagem supervisionada *Random Forest* para binarizar a imagem do espectrograma e destacar as vocalizações de diferentes espécies de aves. Neste caso os descritores acústicos utilizados foram a energia dos *pixels* do espectrograma, que relacionam-se diretamente com a cor da imagem. Após a classificação uma máscara de suavização gaussiana é aplicada deixando como resultado a figura 3.9.

Finalmente, podemos mencionar que pelo fato de existir uma etapa de treinamento para gerar os modelos de classificação, estes conseguem melhores resultados nos domínios específicos. Porém, são menos abrangentes e geralmente devem ser supervisionados. As abordagens supervisionas requerem conhecer todas os possíveis padrões de sinal *a priori* dificilmente possível de se aplicar em um cenário desconhecido. No entanto, estas soluções são mais adequadas aaos sistemas cujo objetivo é rastrear uma, ou umas poucas espécies de interesse.

3.3.4 Avaliações das técnicas de segmentação

Diversos autores destacaram a importância da segmentação dentro dos sinais nas abordagens de monitoramento ambiental bioacústico. Por exemplo, Evangelista et al. (2014)

destacaram a importância de utilizar segmentação em sílabas de forma manual e automática comparada contra a utilização do áudio completo. Eles mostraram que a segmentação manual produz um ganho de 23% na taxa de reconhecimento e a segmentação automática 7%. Isto revela que a segmentação automática escolhe os trechos mais relevante dos áudios, melhora e facilita a classificação. No entanto, não existe consenso na forma de avaliar os ganhos.

Diferentes metodologias para avaliar o impacto da segmentação podem ser utilizadas. Desde o ponto de vista de detecção de eventos acústicos é necessário quantificar quantas sílabas foram perdidas, ou incorretamente recuperadas. Somervuo et al. (2006) compararam a eficiência da segmentação automática e manual utilizando 1068 sílabas de referência. Nos resultados reportados perderam-se 157 sílabas, 26 foram segmentadas com micro cortes e 40 foram incorretas. Assim, a acurácia do sistema foi de aproximadamente 90%, mas com alta variabilidade de precisão ao segmentar diferentes espécies. Podemos notar que além de diferentes resultados, nas diferentes espécies, o método de validação depende do critério do especialista ao realizar a segmentação manual.

Do ponto de vista do monitoramento ambiental, é necessário quantificar a informação relevante, para o reconhecimento da espécie, que o evento detectado e segmentado possui. Neste sentido, Rahman and Bhuiyan (2012), Jaafar and Ramli (2013), Jaafar et al. (2014) avaliaram a segmentação automática diretamente com o resultado da classificação. Desta forma, foi mostrado que existe uma correlação entre o resultado do classificador e a eficácia da segmentação. Entretanto, avaliar desta forma não permite identificar se os ganhos e as perdas foram devidas à capacidade do classificador ou do segmentador.

A importância de quantificar o impacto da segmentação no reconhecimento bioacústico de anuros foi ressaltada por Chen et al. (2012). Neste trabalho os autores desenvolveram uma técnica de classificação com dois níveis, aplicando *Multi-Stage Average Spectrum* (MSAS). No primeiro nível as sílabas das vocalizações são agrupadas e separadas segundo a duração temporal, para serem posteriormente classificadas no segundo nível. Este trabalho mostra a relação entre agrupar e não agrupar as sílabas de diferentes comprimentos, mostrando que existe um ganho na taxa de classificação. A abordagem constitui mais um indicador de que a segmentação é uma peça chave do sistema de monitoramento.

3.3.5 Considerações sobre a segmentação

Considerando os três modelos de segmentação notamos que cada um deles possui vantagens e desvantagens. Os modelos mais simples que utilizam energia possuem baixo custo computacional para o processamento dos dados e são mais abrangentes. No entanto, são também mais suscetíveis a falsos positivos ou negativos causando micro cortes ou recuperando sílabas inexistentes.

Os modelos probabilísticos nas abordagens bioacústicas são menos frequentes. Isto é devido à dificuldade em obter a PDF das vocalizações considerando as variações nas condições acústicas do ambiente. Entretanto, no capítulo 4 mostramos a aplicação de uma nova abordagem baseada na comparação de padrões ordinais para obter tais PDF's.

Finalmente, os modelos de classificação supervisionados precisam ser treinados por um especialista, demandando conhecimento e trabalho manual. Neste caso, inclusive um classificador "bem treinado" teria dificuldades em reconhecer padrões novos em situações reais, ou não seria capaz de identificar variações dentro da própria classe, se não foi treinado para tal propósito. Uma observação importante é que um classificador consegue reconhecer um segmento do sinal com ruído, considerando que o ruído não possui padrão, se e somente se os descritores acústicos utilizados representam este ruído.

A partir desta observação e destacando que a maior dificuldade dos métodos de segmentação, é lidar com as diferentes condições de ruídos ambientais. Propomos uma mudança de paradigma, na qual identificamos segmentos dos sinais como ruídos utilizando a entropia como descritor acústico. A vantagem principal de mudar a abordagem consiste em eliminar a necessidade de conhecer todos os padrões de sinais bioacústicos possíveis e obter independência do nível de ruído no ambiente. As comparações dos diferentes descritores e nossa abordagem de segmentação encontram-se no capítulo 4.

Em nossa proposta de segmentação incluímos também uma nova metodologia para avaliar a iteração entre o segmentador e o classificador. Desta forma, desenvolvemos uma método de avaliação seguindo a teoria de confiabilidade e modelando nosso ACR como um sistema multinível (página 120).

3.4 Descritores bioacústicos para classificação

Como foi mencionado na introdução, a capacidade de reconhecimento das abordagens bioacústicas depende principalmente da qualidade dos descritores acústicos, i.e., deseja-se elevada correlação intra-classe e baixa correlação inter-classe. Em outras palavras,

os descritores de uma mesma espécie devem ter a mínima variância possível, enquanto que é maximizada a separação entre as diferentes espécies. Por este motivo, encontrar o conjunto ótimo de descritores é a parte mais crítica dos sistemas de reconhecimento bioacústico e na literatura encontramos frequentemente trabalhos que focam em identificar a combinação de descritores e classificadores que melhor resolve o problema de monitoramento (Clemins, 2005, Colonna et al., 2012, Gunasekaran and Revathy, 2010, Yuan and Ramli, 2013). Nesta seção o objetivo não é dar uma descrição matemática destes descritores, porque isto é abordado na seção 2.2 dos fundamentos, mas mostrar a relevância e o impacto que os diferentes descritores possuem na classificação bioacústica.

Ao se utilizarem técnicas de aprendizagem de máquina (*Machine Learning*) é necessário representar o sinal acústico como um conjunto de características para que possa ser comparado. Assim um mapeamento ou transformação leva a serie temporal a um espaço vetorial de dimensões diferentes. Esta nova representação é conhecida com vetor de características (ou descritores acústicos). A função de classificação que separa as classes é criada usando exemplos destes vetores como base de treinamento.

Geralmente um segmento do sinal, ou *frame*, é representado por mais de um descritor formando grupos de coeficientes (*feature vector*). Em alguns casos estes coeficientes podem ser obtidos unicamente no domínio temporal, como é o caso dos *Linear Prediction Coding* (LPC), o *Zero Crossing Rate* (ZCR) ou o *Pitch*. Em outros casos podem ser obtidos somente no domínio espectral, como os MFCC, o *Spectral Centroide*, a frequência fundamental, a largura de banda, etc. Também existem abordagens que utilizam outros domínios como a transformada Wavelet, da qual pode se obter a energia ou a entropia dos coeficientes (Yen and Fu, 2002, Colonna et al., 2012). Por último, encontramos abordagens híbridas que utilizam escritores temporais e espectrais (Huang et al., 2009, Vaca-Castaño and Rodriguez, 2010).

3.4.1 LLD temporais

Nesta categoria os mais utilizados são os coeficientes LPC. A análise LPC produz um conjunto de coeficientes que modela o trato vocal como um filtro. Estes coeficientes são usados por si só ou como base para o cálculo características *cepstrais* ou outros bancos de filtros (Rabiner and Schafer, 2007). Nos trabalhos de Clemins (2005) e Yuan and Ramli (2013) foram comparados os LPC contra os MFCC concluindo-se que os LPC modelam melhor as características físicas do trato vogal mas tem uma acurácia de classificação menor que os MFCC.

Outras características físicas dos sinais como: a energia ou potência, a taxa de

cruzamento por zero, o pitch a entropia temporal H_t também foram propostas como descritores acústicos (Colonna et al., 2012). Neste caso a energia relaciona-se com a amplitude do sinal e serve para separar fontes sonoras distantes dos microfone. Tanto o pitch quanto o ZCR servem par caracterizar a frequência fundamental do sinal, no entanto o ZCR é mais sensível aos ruídos ambientais sendo menos apropriado para o cenário da floresta (Colonna et al., 2014a). A entropia temporal, ou outras medidas de entropia como as descritas por Han et al. (2015) ou Dayou et al. (2011), são úteis para identificar características estruturais dos sinais, como o grau de correlação. No entanto, não são úteis para diferenciar padrões com a mesma entropia.

3.4.2 LLD espectrais

Para obter os descritores espectrais é necessário realizar a transformada de Fourier primeiro. Um exemplo é o trabalho pioneiro de Taylor et al. (1996) no qual foram utilizados os valores do espectro par treinar e reconhecer as diferentes espécies. Nesta categoria devem ser incluídos também todos os trabalhos que utilizam o espectrograma e aplicam processamento digital de imagens (Oliveira et al., 2015, Xie et al., 2015c). O conceito principal aqui é aprender a distribuição da energia do sinal em cada banda de frequências utilizando os valores dos coeficientes de Fourier (Ganchev et al., 2015).

Uma vantagem de utilizar uma transformação do sinal é que a energia fica representada de forma compacta, utilizando poucos coeficientes. Consequentemente, utilizar todos os coeficientes do espectro para compor o vetor de características é uma abordagem pouco recomendada, porque eleva a complexidade das funções de classificação e diminui a acurácia total. As abordagens mais robustas aplicam um conjunto de filtros sobre os valores do espectro para conseguir uma representação mais compacta. As propostas que utilizam filtros morfológicos para agrupar *pixels* vizinhos são um exemplo disto. Outra possibilidade é utilizar os MFCC.

Os MFCC se destacam dentro do conjunto de LLD espectrais por serem robustos aos ruídos e pouco sensíveis as diferentes frequências de amostragem do hardware (Colonna et al., 2012). Além de serem coeficientes decorrelacionados, é possível obter a diferença destes entre *frames* sucessivos (ΔMFCC), para modelar as variações temporais-espectrais (Jaafar et al., 2014).

Estudos mais recentes em reconhecimento de fala utilizam a transformada Wavelet como uma alternativa à transformada de Fourier, para conseguir ter uma multiresolução em tempo e frequência e diminuir a complexidade computacional. Uma comparação entre Wavelet e os MFCC foi realizada por Modic et al. (2003), os quis destacaram a abrangência da transformada Wavelet sobre todo o espectro de frequên-

cias. No trabalho de Adam et al. (2013) foi combinada a escala de frequências Mel dos MFCC com a transformada Wavelet conseguindo melhorar os resultados.

Além dos tradicionais MFCC, com escala Mel, existem variações destes úteis no contexto do monitoramento bioacústico. Por exemplo, Zhou et al. (2011) comparam os MFCC com os LFCC que utilizam uma escala de frequências lineares aplicados ao reconhecimento de voz. Ganchev et al. (2015) aplicou os LFCC para o reconhecimento da espécie *Vanellus chilensis lampronotus*. Devido que estas espécies possui um espectro com maior largura de banda, ocupando as altas frequências, os LFCC tiveram melhor desempenho. Howard and López (2009) estenderam os MFCC invertendo a escala de frequências Mel (aMFCC). No entanto, ainda existe uma carência de avaliações utilizando os LFCC e os aMFCC no contexto bioacústico.

3.4.3 Considerações sobre os LLD

Finalmente, no que diz respeito aos conjuntos híbridos de descritores podemos destacar duas abordagens: (a) a adotada por Vaca-Castaño and Rodriguez (2010), na qual é utilizado o maior número possível de descritores e posteriormente é aplicada um redução com PCA, ou (b) realizar uma seleção de dos LLD mais discriminantes, aplicando critérios como o ganho da informação (Colonna et al., 2012). Podemos destacar que a combinação entre LLD temporais e espectrais torna o sistema mais robustos, entretanto os MFCC são considerados o estado da arte nestas abordagens.

Atualmente novas técnicas baseadas em *autoencoder* surgiram. Com os *autoencoders* é possível prender o conjunto de coeficientes automaticamente, de forma não supervisionada e não linear (Deng et al., 2010). Se a entrada do *autoencoder* é uma série temporal, então este irá aprender um mapeamento temporal, porém se a entrada são os valores do espectrograma este aprende um mapeamento espectral (Deng and Yu, 2014). Embora esta seja uma técnica promissora requer um volume de dados de treinamento consideravelmente maior nem sempre disponível.

3.5 Classificação Bioacústica

Após serem obtidos os coeficientes LLD das sílabas uma técnica de aprendizagem de máquina identifica os padrões e separa as espécies. Dentre técnicas mais utilizadas encontram-se SVM, kNN, ANN e HMM, as quais foram aplicadas a diferentes espécies animais além de anfíbios, tais como: aves, morcegos, lobos, baleias e elefantes, dentre outras (Skowronski and Harris, 2006, Root-Gutteridge et al., 2014, Rickwood and Taylor, 2008, Clemins et al., 2005, Weninger and Schuller, 2011).

3.5.1 Técnicas de classificação

Os trabalhos de Yuan and Ramli (2013), Han et al. (2011), Dayou et al. (2011) e Vaca-Castaño and Rodriguez (2010) utilizaram kNN para reconhecer espécies separando tanto descritores temporais quanto espectrais. Huang et al. (2009) e Colonna et al. (2012) comparam SVM contra kNN mostrando que kNN obteve melhores resultados. No entanto, Colonna et al. (2012) observaram que a complexidade kNN aumenta proporcionalmente ao tamanho da base de treino, tornando-se custoso para um sensor de baixos recursos. Por outro lado, os parâmetros de SVM devem ser ajustados de forma a minimizar a quantidade de vetores de suporte necessários, para diminuir a complexidade computacional de classificação e que possa ser embarcado no sensor.

Yen and Fu (2002) e Cai et al. (2007) mostraram uma alternativa utilizando Redes Neuronais (ANN). Entretanto, a aplicação de ANN não foi comparada com outra técnica de classificação. A capacidade de reconhecimento e a complexidade da ANN depende da quantidade de neurônios utilizados, principalmente na camada oculta. A camada de saída aumenta proporcionalmente ao número de diferentes espécies que deseja distinguir e o processo completo de reconhecimento é um produto matricial que pode ser tornar demorado para um sensor com memória limitada.

Oliveira et al. (2015) comparou seu próprio método de classificação, baseado em filtros morfológicos aplicados ao espectrograma, contra a abordagem paramétrica proposta por Harma (2003) e a abordagem de reconhecimento de fala, que utiliza modelos de mistura gaussiana (GMM) proposta por Sahidullah and Saha (2012). Os modelos ocultos de Markov (HMM) utilizados por Potamitis et al. (2014) e os modelos GMM de Xie et al. (2015b) são modelos longamente utilizados em reconhecimento de fala. No entanto são técnicas sensíveis aos ruídos que afetam aos LDD, o que resulta em modelos menos acurados para sistemas bioacústicos que utilizam dados reais da floresta. Entretanto, HMM e GMM são modelos probabilísticos mais simples de codificar que SVM ou ANN, ganhando em simplicidade e diminuindo a complexidade.

3.5.2 Sistemas de classificação centralizada

Os sistemas de monitoramento centralizados são similares aos sistemas de recuperação de áudio (figura 3.10) (Dong et al., 2015). Isto significa que a partir de uma sílaba (ou vocalização) é realizada uma consulta em um banco de dados com informações sobre as diferentes espécies. Em algumas abordagens a ação de consulta ao banco de dados é substituída por uma técnica de aprendizagem de máquina (figura 3.11). Assim, um modelo de classificação é treinado *off-line* com as informações das espécies contidas no banco, e quando se deseja classificar novas sílabas não é necessário fazer

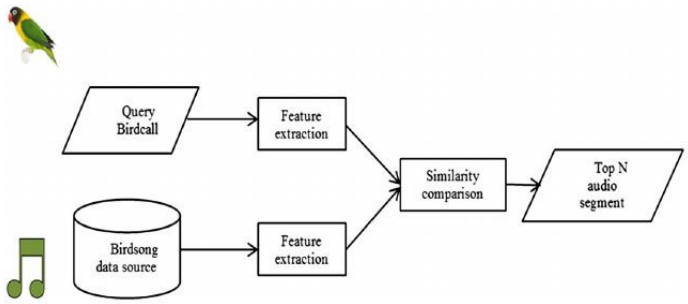


Figura 3.10. Sistema de recuperação de áudio. Figura extraída de Dong et al. (2015).

uma consulta explícita, bastando simplesmente aplicar o modelo treinado e obter a resposta da espécie.

A estratégia de utilizar um modelo de classificação possibilita que este seja embarcado na rede RSSF, nos próprios nós ou no *sink*. Esta é a diferença entre colaborativo ou centralizado. Um exemplo de sistema centralizado é o ARBIMON desenvolvido por (Aide et al., 2013), utilizado para o reconhecimento de pássaros, insetos, mamíferos e anfíbios.

O ARBIMON é uma plataforma web[1] híbrida que permite processar e classificar áudios manualmente escolhidos, ou receber os dados automaticamente dos sensores da RSSF e realizar a classificação no nó *sink*. Esta ferramenta permite experimentar e comparar diferentes métodos de classificação, mas utiliza principalmente as ROIs do espectrograma, em uma abordagem similar a Potamitis (2014). Desta forma, são extraídas as frequências máximas e mínimas, a duração do canto, a intensidade e a largura de banda. Como foi ressaltado nas seções anteriores, utilizar o espectrograma com imagem possui elevados custos de memória e processamento, motivo pelo qual o sistema não é embarcado. Assim, os nó sensores distribuídos devem enviar o áudio completo até o *sink*, transformando-se em uma abordagem centralizada cliente-servidor, consumindo maior energia dos nós.

Potamitis et al. (2014) e Xie et al. (2015b) realizaram a coleta dos áudios com Unidades Autônomas de Gravação (ARUs), mas a classificação final acontece de forma centralizada em um computador. Xie et al. (2015b) definiram uma abordagem centralizada para o reconhecimento de anuros, no qual é utilizado GMM para selecionar e classificar cada *frame* do espectrograma. Esta abordagem possui duas particularidades: a inclusão de uma etapa de filtragem antes de obter os valores das características que são utilizadas para segmentar e a correlação da detecção (ou atividade acústica)

[1] `arbimon.net`

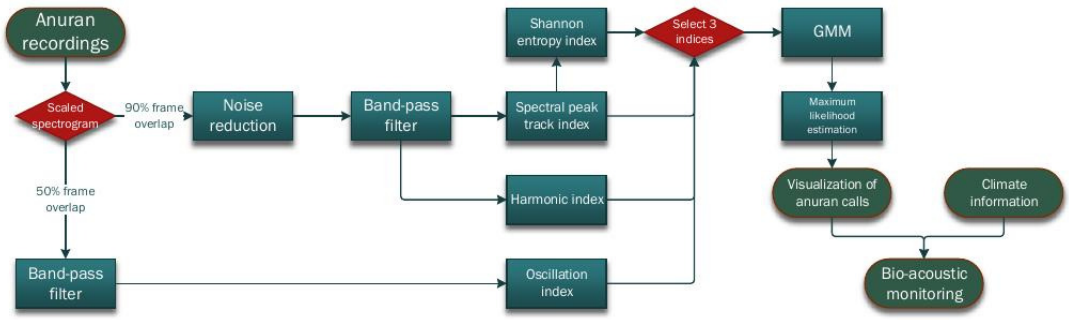


Figura 3.11. ACR para reconhecimento de anuros proposto por Xie et al. (2015b).

das espécies com as variáveis ambientais temperatura, umidade e velocidade do vento (figura 3.11).

3.5.3 Sistemas de classificação colaborativa

Nos sistemas colaborativos a classificação é realizada nos próprios sensores da rede espalhados pela área monitorada. Desta forma existem diferentes possibilidades de combinação, isto é, cada sensor pode realizar a classificação e transmitir a resposta para o *sink* ou pode enviar os LLD e deixar que a classificação seja realizada pelo sensor líder do *cluster*. Um exemplo disto é o trabalho de Wang et al. (2003), no qual foram utilizados os nós finais da RSSF para coletar os áudios e enviá-los para o sensor líder. Este último aplica uma correlação entre os espectrogramas para reconhecer a espécie e transmite o resultado final.

A classificação e a troca de informação entre os sensores pode ser interpretada como uma fusão de dados. Na proposta de classificação colaborativa de Ribas et al. (2012) foram definidas três formas de fusão: dos áudios, dos LLD e do resultado das classificações. Assim, neste trabalho é aplicada uma técnica de agrupamento para formar os *clusters* de sensores com sinais acústicas similares, posteriormente cada sensor classifica a espécie e finalmente o nó líder realiza uma votação maioritária para escolher a espécie mais provável. Esta abordagem focada no agrupamento dinâmico dos sensores e para a classificação utiliza os MFCC junto com a distância de *Mahalanobis*.

A metodologia de classificação de Ribas et al. (2012) é interessante devido a dois fatores fundamentais: 1) não é necessário transmitir o áudio completo senão simplesmente o resultado da classificação, reduzindo a quantidade de informação transmitida, e 2) a votação permite aumentar a certeza da classificação dos membros do *cluster*. Baseados nestas vantagens, Colonna et al. (2014a) propuseram uma abordagem de classi-

ficação colaborativa, experimentando diferentes técnicas de votação entre os sensores e adicionando um critério de rejeição para descartar os casos confusos (capítulo 5.3). Recentemente a proposta de Colonna et al. (2014a) foi adotada por Silva and Ruiz (2015) para avaliar o impacto dos parâmetros da rede: energia consumida, quantidade de pacotes transmitidos e recebidos. Concluindo-se que embarcar o modelo de classificação no nó sensor economiza energia e estende a vida útil da rede.

3.5.4 Considerações sobre a classificação

Diferentes estudos comparativos sobre as abordagens para reconhecimento de espécies animais podem ser encontrados na literatura. Alguns destes comparam diferentes descritores ou diferentes classificadores, enquanto que outros compram ambos (Jaafar et al., 2014, Armitage and Ober, 2010, Huang et al., 2009, Weninger and Schuller, 2011, Acevedo et al., 2009). A tabela 3.2 resume a combinação entre técnicas de classificação, descritores acústicos e resultados dentre os trabalhos mais relevantes encontrados.

Podemos notar que os métodos de monitoramento perseguem dois objetivos: ser abrangentes tentando classificar o maior número de espécies possíveis, como no caso de Cai et al. (2007), Huang et al. (2009), Vaca-Castaño and Rodriguez (2010), Jaafar et al. (2014), ou ser específicos focando em uma espécie de interesse, como Hu et al. (2009), Oliveira et al. (2015). O sistemas mais especializados de Harma (2003) e Somervuo et al. (2006) utilizam abordagens paramétricas, modelando o canto de cada espécie como uma função específica. Os sistemas específicos são úteis para realizar rastreamento ou localização, além disso, por serem projetados para cumprir com tarefas específicas podem ser otimizados. As abordagens colaborativas mostraram ter ganhos nas taxas de reconhecimento provavelmente pela coleta e utilização de informações correlacionadas entre os sensores.

3.6 Comentários finais

Ao longo deste capítulo identificamos diversas propostas de classificação bioacústica. Dentre as características mais usuais destacamos os MFCC, que realizam um mapeamento completo do espectro de frequências das diferentes espécies. As técnicas de classificação mais utilizadas são GMM, SVM e kNN que, embora sejam conceitualmente diferentes, quando combinadas com os MFCC conseguem separar as diferentes espécies com alta acurácia. No entanto, não foi definida nenhuma abordagem para esta tarefa. Assim, cada trabalho combina LLD e classificadores diferentes tornando o estudo comparativo dificultoso.

Autor	Animal	Características	Classificador	Resultados	RSSF
Colonna et al. (2012)	9 anuros	MFCC	kNN	97%	Sim
		Espectral Centroide Largura de Banda Wavelet	SVM		
Taylor et al. (1996)	Bufo marinus	Espectrograma	C4.5	60%	Sim
Yen and Fu (2002)	4 anuros	Wavelet	ANN	71%	Não
		Fisher's			
Cai et al. (2007)	14 pássaros	MFCCs	ANN	81-86%	Sim
Huang et al. (2009)	5 anuros	Espectral Centroide	kNN	83-100%	Não
		Largura de banda ZCR	SVM	82-100%	
Vaca-Castaño & Rodriguez [2010]	10 pássaros 20 anuros	MFCCs PCA	kNN	86% 91%	Sim
Han et al. (2011)	9 anuros	Espectral Centroide	kNN	83-100%	Não
		Entropia de Shannon Entropia de Rényi			
Jaafar et al. (2014)	28 anuros	MFCC	SVM	97-98%	Não
		ΔMFCC	SRC LMkNN-FDW		
Yuan and Ramli (2013)	8 anuros	MFCC	kNN	93-98%	Não
		LPC			
Potamitis et al. (2014)	3 pássaros	MFCC	HMM	84-88%	Não
Evangelista et al. [2014]	75 pássaros	MARSYAS framework	SVM	60%	Não
Dayou et al. (2011)	9 anuros	Entropia de Shannon	kNN	90%	Não
		Entropia de Rényi Entropia de Tsallis			
Xie et al. (2015b)	4 anuros	Índice de Oscilação	GMM	63-83% (Prec)	Não
		Índice harmônico		76-96% (Rec)	
		Pico espectral Entropia de Shannon			

Tabela 3.2. Resumos dos trabalhos de reconhecimento de diferentes espécies animais.

A dificuldade para definir um método padrão (ou *baseline*) é devido, também, à não existência de uma base de dados pública com vocalizações de anuros que permita avaliar os métodos em igualdade de condições. Entretanto, encontramos na literatura uma extensa variedade de trabalhos de classificação de animais, nos quais o interesse principal é apenas o método de classificação, dispensando a possibilidade de aplicação sobre uma RSSF.

Identificamos aqui que os trabalhos de classificação colaborativa possuem vantagens que melhoram os resultados. Estes geralmente consideram os parâmetros da RSSF tais como: topologia, custos e viabilidade de implementação, modelando a classificação como uma técnica de *ensemble learning*. A técnica de classificação demarca como será implementado o monitoramento na RSSF. Em outras palavras, esta define qual o processamento que cada nó deverá executar, quais características devem ser extraídas e transmitidas e como será realizada a classificação.

As tarefas secundarias como filtragem e segmentação não estão sempre presentes e detalhadas na literatura embora sejam peças fundamentais dos sistema para alcançar resultados satisfatórios. Por exemplo, a segmentação é um fator comum na maioria dos métodos de classificação bioacústicos. Nestes, a lógica de detectar e isolar o evento antes da classificação está presente como um pré-processamento que evita erros na etapa de classificação (Evangelista et al., 2014). Embora existam diversas abordagens de segmentação, a escolha dependerá muito das espécies envolvidas e dos requerimentos da aplicação (Garcia et al., 2014). Na maioria dos trabalhos se atribui menos importância aos métodos de segmentação e as avaliações correspondentes.

O desconhecimento *a priori* de todos os possíveis sinais presentes na floresta adiciona uma dificuldade extra na tarefa de segmentação, motivo pelo qual os métodos baseados no conteúdo espectral das frequências são interessantes, assemelhando-se a técnicas de filtragem. Embora sejam mais robustos, há uma perda de abrangência no número de sinais diferentes que podem ser segmentadas sem se ter que modificar o conjunto de frequências de interesse. Em outras palavras são menos abrangentes, mas podem ser muito úteis em situações onde pretenda-se realizar a identificação de uma ou poucas espécies.

Enquanto que a detecção das sílabas usando características temporais parece menos precisa que os métodos espectrais, as características temporais são mais econômicas em termos de memoria e processamento. Isto as torna mais atrativas para sensores com limitações de hardware, cumprindo um papel fundamental na implementação. No contexto das RSSF devem-se considerar as estratégias não supervisionadas. Por estes motivos decidimos utilizar somente características temporais dos sinais e técnicas de detecção de mudanças de padrões baseadas em limiares que não precisem ser treinadas

e que sejam adaptativos e dinâmicos (capítulo 4). Avaliar o qualidade da segmentação com o resultado do classificador pode mascarar erros das técnicas de segmentação. Por este motivo é interessante mostrar a relação entre acurácia de segmentação e acurácia de classificação.

A capacidade de detectar as sílabas dos cantos depende do nível dos ruídos aleatórios e ambientais. A maioria das abordagens de filtragem baseadas em *Spectral Subtraction* utilizam um detector de atividade acústica para separa os *frames* e estimar o nível de ruído. A estimação correta é importante para não causar distorções excessivas nos áudios e prejudicar a classificação. Desta forma entendemos que existe uma relação entre manter as informações principais, correspondentes às bandas de frequências de todas as espécies monitoradas e melhorar os valores dos descritores (MFCCs) eliminando ruídos aleatórios e ambientais.

Do ponto de vista qualitativo a filtragem dos ruídos melhora a qualidade das gravações, aumentando possivelmente, o resultado final do reconhecimento. Embora os resultados encontrados sejam interessantes, nós procuramos definir um método mais flexível e adaptativo às mudanças dinâmica do ambiente.

Finalmente, entendemos que existem uma falta de trabalhos que combinem e avaliem o sistema completo em situações reais considerando as restrições impostas pelas RSSF.

4

Segmentação de sinais bioacústicos

Neste capítulo apresentamos um estudo comparativo entre diferentes descritores acústicos de baixo nível baseados principalmente em quantificadores de informação, utilizados nas estratégias de segmentação bioacústica não supervisionadas para fluxos contínuos de áudio.

A seguir, avaliamos e comparamos diversas medidas de entropia incluindo as recentemente desenvolvidas *Permutation Entropy* (PE), *Weighted Permutation Entropy* (WPE), *Permutation Min-Entropy* (PME), e as comparamos com a Energia do sinal (E), a Taxa de Cruzamento Zero (ZCR) e a Entropia Espectral (H_f). Além disso, apresentamos um algoritmo para estimar o limiar de segmentação ótimo utilizado pelas técnicas de segmentação. Realizamos três avaliações diferentes aplicando métricas *frame-to-frame*, *point-to-point* e *event-to-event*. Geramos diferentes tipos de ruídos (branco e coloridos) e avaliamos combinações de tais descritores com o intuito de melhorar o resultado da segmentação. Realizamos um ranqueamento destes descritores baseado no ganho da informação (*Information Gain* - IG) e mostramos que em um cenário com severas condições de ruído, os descritores baseados em entropia são robustos, alcançando 97% de desempenho e mantendo um baixo custo computacional. Concluímos que não há LLD que seja adequado para todos os cenários possíveis, devendo-se adotar diferentes LLDs dependendo das condições de ruído esperadas. Posteriormente, criamos um método de segmentação incremental vantajoso para sensores com recursos limitados. Este método incremental possui complexidade de memória reduzida $\mathcal{O}(1)$ comparada com as janelas deslizantes que custam $\mathcal{O}(n)$. Além do mais, esta adaptação permite obter a resposta em tempo real. Finalmente, desenvolvemos um conjunto de equações para mensurar o impacto que a segmentação causa no resultado final do reconhecimento.

4.1 Introdução

Uma tarefa fundamental na maioria dos sistemas de reconhecimento de espécies é a segmentação das vocalizações em unidades menores. A segmentação é o ato de "cortar" o sinal em regiões homogêneas em termos dos descritores acústicos. Desta forma, o objetivo pode ser definido como: identificar segmentos das vocalizações, utilizando as propriedades de sinal, para que possam ser classificados em função do seu conteúdo.

Existem diferentes nomes para as unidades menores que compõem um sinal bioacústico dependendo da hierarquia e da duração. Por exemplo, o canto dos pássaros é composto por diferentes vocalizações, dentro das quais existem frases formadas por sílabas que foram geradas por elementos (Fagerlund, 2007). As vocalizações dos anuros são menos complexas, em outras palavras a linguagem dos anuros é mais reduzida. Por este motivo, a maioria dos sistemas de reconhecimento de anuros utiliza como unidade menor as sílabas (figura 4.1) (Huang et al., 2009).

A identificação e extração das sílabas é o primeiro passo para o correto reconhecimento das espécies. Swiston and Mennill (2009) realizaram uma comparação entre a segmentação manual e automática examinando: o tempo requerido para o análise, a precisão e compreensão das gravações. Lopes et al. (2011) compararam os benefícios da segmentação contra a utilização do áudio completo, e Jaafar and Ramli (2013) mostraram o impacto da segmentação na taxa de classificação. No entanto, poucos estudos focaram em identificar quais são os melhores descritores acústicos para esta tarefa.

Técnicas de segmentação automática de sinais foram extensivamente estudadas no reconhecimento de fala humana. No entanto, estas técnicas não são totalmente adequadas às vocalizações bioacústicas, porque possuem características acústicas diferentes (Rickwood and Taylor, 2008). Além disso, para melhorar a classificação das espécies, é importante selecionar os segmentos mais representativos de cada vocalização, uma vez que estes contêm longos períodos de ruído ambiente (Evangelista et al., 2014). Adicionalmente, a maioria das abordagens de segmentação automática envolvem procedimentos não sequenciais que consomem grandes quantidades de memória. Estas abordagens não são adequadas para cenários nos quais os dados devem ser transmitidos ou processados por sistemas com recursos limitados, como os nós das RSSF (Nakamura et al., 2014).

A segmentação lida com a relação custo-benefício entre tempo, ou esforço humano, e a qualidade na recuperação das sílabas. Quando trata-se de uma técnica automática e não supervisionada, embarcada em um nó sensor, esta relação se transforma em custo de processamento, ou memória, versus a taxa de erro dos eventos acústicos, ou acurácia no reconhecimento. Portanto, é fundamental responder:

- Quais são os melhores descritores acústicos para segmentar as vocalizações de maneira não supervisionada;
- Qual é o melhor limiar de decisão;
- Como reduzir a complexidade do processamento ou a quantidade de memória necessária; e
- Qual é o impacto que a segmentação têm no resultado final do reconhecimento.

Nas RSSF, a aquisição de som é realizada de forma não intrusiva pelos nós de sensor, fato que nos permite monitorar o ambiente por longos períodos de tempo. Em regiões afastadas, de difícil acesso, substituir as baterias dos sensores frequentemente torna-se operativamente inviável. Portanto, é necessário desenvolver um método eficiente de segmentação para minimizar a quantidade de dados processados, transmitidos ou armazenados pelos nós da rede. Assim, manter o método de segmentação o mais econômico possível, desde o ponto de vista da complexidade computacional, mediante a utilização de um conjunto reduzido de LLDs, é nosso desafio principal.

Em situações reais, tais como as florestas tropicais, os cenários podem ser complexos apresentando uma elevada riqueza acústica, como resultado da interação de várias espécies no mesmo local (Depraetere et al., 2012). Portanto, é impossível conhecer *a priori* todos os padrões de sinal existentes. Assim, propomos uma mudança no paradigma de segmentação: no lugar de identificar diferentes padrões dos sinais, tentaremos identificamos apenas segmentos com características de ruído. Assim, os segmentos restantes poderão ser considerados sílabas. Isto é possível porque a segmentação é equivalente a um classificador binário não supervisionado, no qual separamos características acústicas que pertencem a segmentos das classes "sinal" ou "ruído". Após a segmentação, o classificador será o responsável pelo reconhecimento da espécie.

Finalmente, escolhemos uma abordagem não supervisionada para evitar rotular sinais desconhecidos, mediante a intervenção de um especialista humano, e também evitamos atualizar os modelos de classificação embarcados, o que resulta em maior consumo de energia devido à comunicação *full-duplex*. Além disso, decidimos focar apenas no uso de LLDs temporais e espectrais para lidar com as restrições de hardware dos sensores.

4.2 Definição do Problema

A segmentação considera uma vocalização, emitida por um anuro, como uma composição de várias sílabas repetidas ao longo do tempo. A classificação acontece utilizando

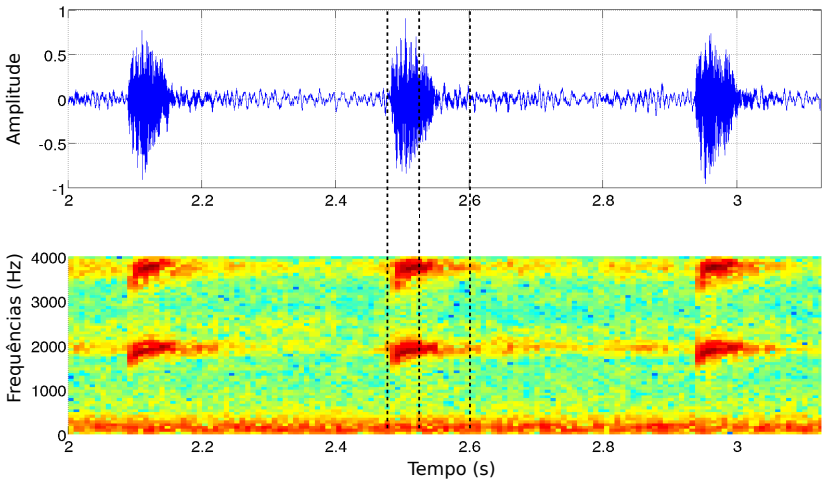


Figura 4.1. Vocalização com três silabas da espécie Adenomera hylaedactyla. Figura adaptada de Colonna et al. (2015). Na parte de cima da figura temos a série temporal e abaixo dela seu espectrograma.

as sílabas como unidades bioacústicas básicas para identificar a espécie. A figura 4.1 mostra uma vocalização típica da espécie Adenomera hylaedactyla com três sílabas. O começo, meio e fim das sílabas são delimitadas por linhas verticais representando dois tipos diferentes de mudanças que caracterizam esta: (1) uma mudança abrupta no nível de sinal, i.e., o começo da sílaba indicado pela primeira linha pontilhada vertical; e (2) uma variação gradual crescente ou descendente, i.e., um aumento suave, ou atenuação, da amplitude do sinal, como pode ser visto entre a segunda e terceira linhas verticais.

O terceiro padrão observado nas vocalizações é a semelhança e repetição das sílabas ao longo do tempo. No espectrograma da figura 4.1 pode-se observar também que a vocalização desta espécie possui duas bandas espectrais, e em cada uma destas a largura de banda é maior na primeira metade da sílaba.

As características das sílabas podem mudar entre as diferentes espécies, e.g., começar com uma mudança gradual e terminar com um corte abrupto, ao contrário de nosso exemplo. Assim, podemos definir o problema de segmentação como: detectar o início e o fim de cada sílaba. Desta forma, podemos considerar uma vocalização como um fluxo contínuo de áudio (*stream*) no qual busca-se extrair todas as sílabas. Pode-se notar que os intervalos de tempo entre as sílabas, o som ambiental ou ruído, não possui energia nas bandas de frequências que caracterizam a vocalização, portanto não acrescenta informação útil para identificar a espécie. Os segmentos de ruído devem ser descartados, porque estes podem confundir o classificador; aumentar os custos de transmissão; e reduzir o tempo de vida da rede. Por esta razão, em situações reais, é conveniente detectar alterações nos sinais monitorados para decidir quando iniciar e

terminar o processamento dos dados.

Formalmente, o sinal bioacústico captado pelo sensor $x = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ é uma série temporal em que os valores representam os níveis de pressão acústica (ou amplitude) dentro do intervalo temporal $1 \leq i \leq n$, no qual N representa o comprimento máximo do sinal. Um *frame* (ou janela temporal) $\mathbf{x} = \{x_i, x_{i+1}, \ldots, x_{i+N}\}$ é definido como um subconjunto de valores consecutivos com tamanho N. Desta forma, o sinal pode ser representado por um conjunto de *frames* utilizando janelas deslizantes de tamanho N. Assim, o desafio principal é classificar estes *frames* de acordo com as classes "sinal" ou "ruído" ("1" ou "0"). Para resolver esse problema, nós representamos os *frames* $\mathbf{x}$ através de um conjunto de LLDs. Por exemplo, utilizando o valor correspondente de entropia do *frame* $\mathbf{x}$, podemos aplicar a regra de decisão binária:

$$\text{classe}(\mathbf{x}) = \begin{cases} 1 & \text{se } \mathrm{H}(\mathbf{x}) \leq T_H \\ 0 & \text{se } \mathrm{H}(\mathbf{x}) > T_H \end{cases}, \tag{4.1}$$

onde T_H é o limiar de decisão ótimo para o valor de entropia correspondente ao *frame* $\mathbf{x}$. Com esta regra, podemos atribuir a classe "sinal" aos *frames* com valores baixos de entropia e definir uma sílaba por um conjunto sucessivo de *frames* identificados como "sinal". Já que a entropia pode ser interpretada como uma medida de "impureza", quanto maior é seu valor, maior é a probabilidade do sinal subjacente ser um ruído aleatório. Regras de decisão similares podem ser construídas para outros LLDs.

A utilização desta regra requer encontrar o valor ótimo T_H. Consequentemente, este é um desafio secundário que surge desta abordagem. O ótimo T_H é uma relação custo-benefício entre a sensibilidade ao ruído e a precisão nas fonteiras das sílabas extraídas, ou em outras palavras, a relação entre sílabas recuperadas e perdidas. Para encontrar o melhor limiar, apresentamos uma técnica de agrupamento descrita no Algoritmo 2. Os resultados dos experimentos da Seção 4.5.1 suportam a hipótese de que esse algoritmo produz uma divisão ótima dos *frames*. Assim, a aplicação desta regra de decisão nos permitirá comparar e identificar individualmente os melhores descritores acústicos para o problema de segmentação das vocalizações, de forma não supervisionada no contexto das RSSF.

4.3 Base de dados

Gerar uma base de dados das vocalizações com anotações manuais da segmentação é uma etapa crucial. As anotações são necessárias para avaliar o desempenho das diferentes abordagens de segmentação automática comparadas contra a capacidade

do especialista humano (*Ground Truth* - GT). Para os anuros, não encontramos uma base de dados pública de vocalizações com anotações correspondentes à segmentação, provavelmente porque fazer estas é uma tarefa que requer elevado esforço e tempo de um especialista humano.

Para conduzir nossos experimentos montamos nossa própria base de áudios. Para isto, utilizamos espécies gravadas no campus da Universidade Federal do Amazonas, gravações cedidas por biólogos especialistas e gravações comerciais disponíveis em Haddad (2005), Márquez et al. (2002), Marty (1999). No total usamos vinte e três gravações com quatorze espécies e rotulamos manualmente 6324 segmentos totais (3.155 sílabas) em duas classes: "sinal" ou "ruído ambiente". Para cada sílaba segmentada identificamos também as espécies correspondentes[1]. A lista completa das espécies utilizadas nos experimentos deste capítulo encontram-se na primeira coluna da tabela 4.1.

[1]Nossa base de dados com as respectivas anotação encontra-se disponível em `https://goo.gl/aZRhPJ`

Tabela 4.1. Espécies utilizadas nos experimentos de segmentação automática. A primeira coluna é o nome científico da cada espécie, a segunda coluna indica a quantidade de sílabas encontradas pela inspeção manual (GT). As colunas restantes apresentam a quantidade de sílabas recuperadas utilizando cada LLD com a metodologia descrita nas próximas seções.

		LLDs					
Espécies	**GT**	E	PE	WPE	PME	ZCR	H_f
Adenomera h.	58	57	83	91	94	155	158
Hyla m.	39	51	93	89	97	217	12
Adenomera a.	50	50	193	164	194	168	156
Ameerega t.	86	92	105	99	104	128	0
Osteocephalus o.	26	33	310	248	323	324	582
Rhinella g.	2	3	2	3	2	66	0
Scinax r.	57	27	17	34	20	58	0
Hypsiboas c.	1548	1403	2533	1941	2971	4021	2233
Brachycephalus e.	1184	116	132	115	131	2	0
Aplastodiscus albof.	28	28	147	133	151	125	0
Aplastodiscus albos.	8	7	155	181	215	158	3
Aplastodiscus p.	13	13	13	13	24	110	0
Dendropsophus a.	49	46	256	194	213	182	2
Dendropsophus e.	7	7	75	81	123	1	0
Total	3155	1933	4114	3386	4662	5715	3146

4.4 Métricas para avaliar a segmentação

A segmentação considera-se correta quando é possível identificar com precisão o começo e o final de cada sílaba. Assim, para avaliarmos a precisão da segmentação automática entre os diferentes LLDs e métodos, nós comparamos cada resultado com a segmentação manual (GT), quantificando três tipos de erros: a taxa de eventos acústicos errados (*event-to-event*); a diferença entre *frames* com tamanho fixo (*frame-to-frame*; e a diferença ponto-a-ponto (*point-to-point*) entre a segmentação automática e GT. Além destas comparações, na abordagem de segmentação incremental, apresentada no final deste capítulo, utilizamos a distância entre as fronteiras das sílabas manualmente identificadas e as estimadas automaticamente.

Para contar a taxa de eventos acústicos errados (do inglês *Acoustic Event Error Rate* - AEER) comparamos cada segmento da GT com a segmentação automática e aplicamos o cálculo do AEER descrito na seção 2.6.3 pela equação 2.65. Esta métrica considera o número de sílabas perdidas, sílabas adicionas extras e sílabas substituídas, pela quantidade total de sílabas. Portanto, quanto menor é o AEER melhor é o resultado da classificação. Como nossa abordagem de segmentação é um problema binário, as palavras "evento" e "sílaba" podem ser consideradas sinônimos.

Aplicando uma janela deslizante de tamanho N, podemos representar o sinal por um conjunto de *frames*. Desta forma, os eventos acústicos, sejam estes sílabas ou ruídos, são formados por frames sucessivos do mesmo rótulo. Portanto, podemos quantificar a qualidade da segmentação utilizando uma métrica que avalie se o rótulo de cada *frame* é o correto, ou seja uma avaliação *frame-a-frame*. Para isto, utilizamos as curvas ROC e o valor da área sob a curva AUC (seção 2.6.4). Assim, quanto maior é o AUC, melhor é a classificação dos *frames*, e portanto, dos eventos detectados.

Uma vez que um rotulo é atribuído a um *frame*, indiretamente todos os pontos do sinal que compõem o *frame* recebem o mesmo rotulo. Portanto, é possível comparar cada ponto do sinal manualmente segmentado contra cada ponto do sinal automaticamente segmentado, e determinar as proporções de tp, fn, tn e fn do segmentador. A comparação ponto-a-ponto permite quantificar a qualidade da segmentação em termos de precisão e revocação considerando as distâncias entre as fronteiras dos eventos acústicos segmentados manualmente e automaticamente. Posteriormente, podemos obter as métricas tradicionais de aprendizagem de máquina, tais como: a taxa de verdadeiros positivos (TPR), a taxa de falsos negativos (FNR), a precisão (Prec), a revocação (Rec) e o F-score (F1), descritas na seção 2.6.1. Estas métricas são baseadas na tabela de contingência, ou na matriz de confusão do classificador, e servem tanto para avaliar a qualidade da segmentação, quanto a acurácia na classificação das espécies como

apresentado no capítulo 5. As interpretações das taxas TPR e FNR indicam perdas mínimas dos pontos do sinal quando os valores de TPR aumentam e os valores de FNR diminuem.

4.5 Comparação de descritores acústicos aplicados à segmentação automática não supervisionada

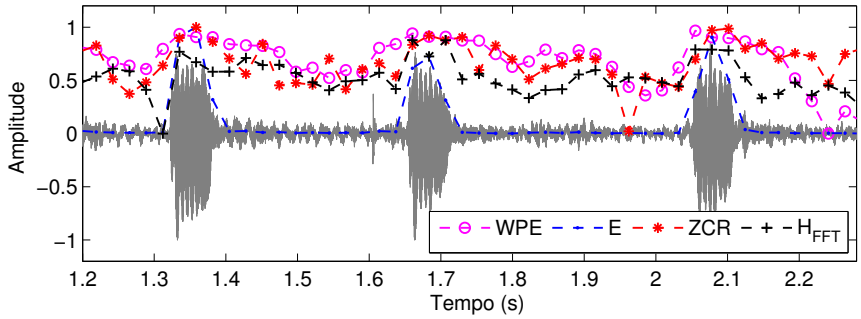
Em nossos experimentos, primeiro comparamos os descritores acústicos individualmente adicionando diferentes tipos e níveis de ruídos, para identificar os melhores LLDs no contexto das vocalizações com ruído ambiental. Avaliamos o desempenho da segmentação através das curvas ROC e a área sob a curva (seção 4.5.2). Posteriormente combinamos diferentes características para provarmos a hipótese de que a combinação de LLDs melhora o resultado da segmentação, quantificando o número de sílabas recuperadas de cada espécie, a taxa de eventos acústicos (AEER) e F1 do segmentador (seção 2.6.3).

4.5.1 Metodologia experimental

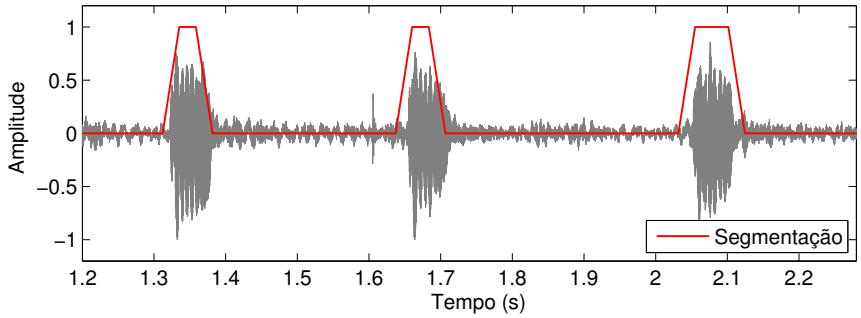
Com base na revisão bibliográfica decidimos comparar seis descritores acústicos do domínio temporal e um do domínio espectral. Estes são: a energia do sinal (E), o *Zero Crossing Rate* (ZCR), a entropia espectral (H_f), a *Permutation Entropy* (PE), a *Weighted Permutation Entropy* (WPE) e *Permutation Min-Entropy* (PME); equações 2.1, 2.2, 2.8, 2.14, 2.16, 2.18, respectivamente.

Para poder avaliar a qualidade da segmentação gerada por cada descritor, primeiro fragmentamos cada áudio formando um conjunto de *frames* (D). O tamanho escolhido para os *frames* foi 23,21 ms, ou seja 1024 pontos, o que nos permite respeitar a condição da PE ($m! \leq N$), com parâmetros $m = 4$ e $\tau = 1$. Decidimos não aplicar sobreposição entre *frames* sucessivos para evitar a contagem de pontos repetidos e dificultar as avaliações ponto-a-ponto. Posteriormente, cada *frame* com seu respectivo índice temporal (*timestamp*) recebeu um valor de LLD. Como resultado, obtemos uma nova série temporal para cada LLDs da forma $D_{id} = \{S(LLD_l, t_0), S(LLD_l, t_1), \ldots, S(LLD_l, t_n)\}$, onde *id* representa o identificador da vocalização, o índice temporal do *frame* e l o descritor escolhido.

A figura 4.2(a) ilustra uma chamada da espécie Adenomera hylaedactyla e sua representação através de diferentes LLDs. O resultado deste mapeamento é um novo conjunto de séries temporais no espaço vetorial das características acústicas. Antes de



(a) Representação mediante LLDs.



(b) Segmentação aplicando E.

Figura 4.2. Vocalização da espécie Adenomera hylaedactyla e sua representação mediante séries temporais de LLDs.

prosseguir com nossa análise, normalizamos as séries dos LLDs no intervalo $[0, 1]$. Para normalizar a energia E e a taxa de cruzamento por zero ZCR aplicamos:

$$\hat{\mathrm{D}}_{id} = \frac{\mathrm{D}_{id} - \min(\mathrm{D}_{id})}{\max(\mathrm{D}_{id})} - \min(\mathrm{D}_{id}), \tag{4.2}$$

e para normalizar os LLDs que utilizam entropia (PE, WPE, PME e H_f) utilizamos:

$$\hat{\mathrm{D}}_{id} = 1 - \left(\frac{\mathrm{D}_{id} - \min(\mathrm{D}_{id})}{\max(\mathrm{D}_{id})} - \min(\mathrm{D}_{id}) \right). \tag{4.3}$$

A intenção desta normalização é deixar os *frames* com sinal mais próximos de "1", caso contrário, mais próximos de "0". Além disso, esta normalização permite utilizar os valores LLDs como se fosse a pontuação dada por um classificador binário.

Os modelos de segmentação não supervisionados baseados em descritores precisam de um limiar de decisão T_f. Para encontrar o limiar ótimo propomos aplicar o algoritmo 2, que se baseia em um procedimento de binarização de imagens (Sezgin and Sankur, 2004). Este procedimento divide iterativamente o conjunto de *frames* ($\hat{\mathrm{D}}_{id}$) em dois subgrupos enquanto tenta-se maximizar a distância entre as médias de cada

grupo. Em outras palavras, a separação ótima entre as classes acontece quando a distância entre as médias encontra-se balanceada. Além de equiparar as distribuições de probabilidade do dois subgrupos, o limiar encontra-se em uma situação de equilíbrio, na qual pequenas perturbações não causam mudanças relevantes no seu valor T_f. Diferente de outras técnicas de clustering, *e.g.* k-means, este método tenta evitar a criação de clusters finos com poucas amostras, o que poderia causar a perda de sílabas com curta duração.

Algoritmo 2 Procedimento para encontrar o limiar ótimo Tf.

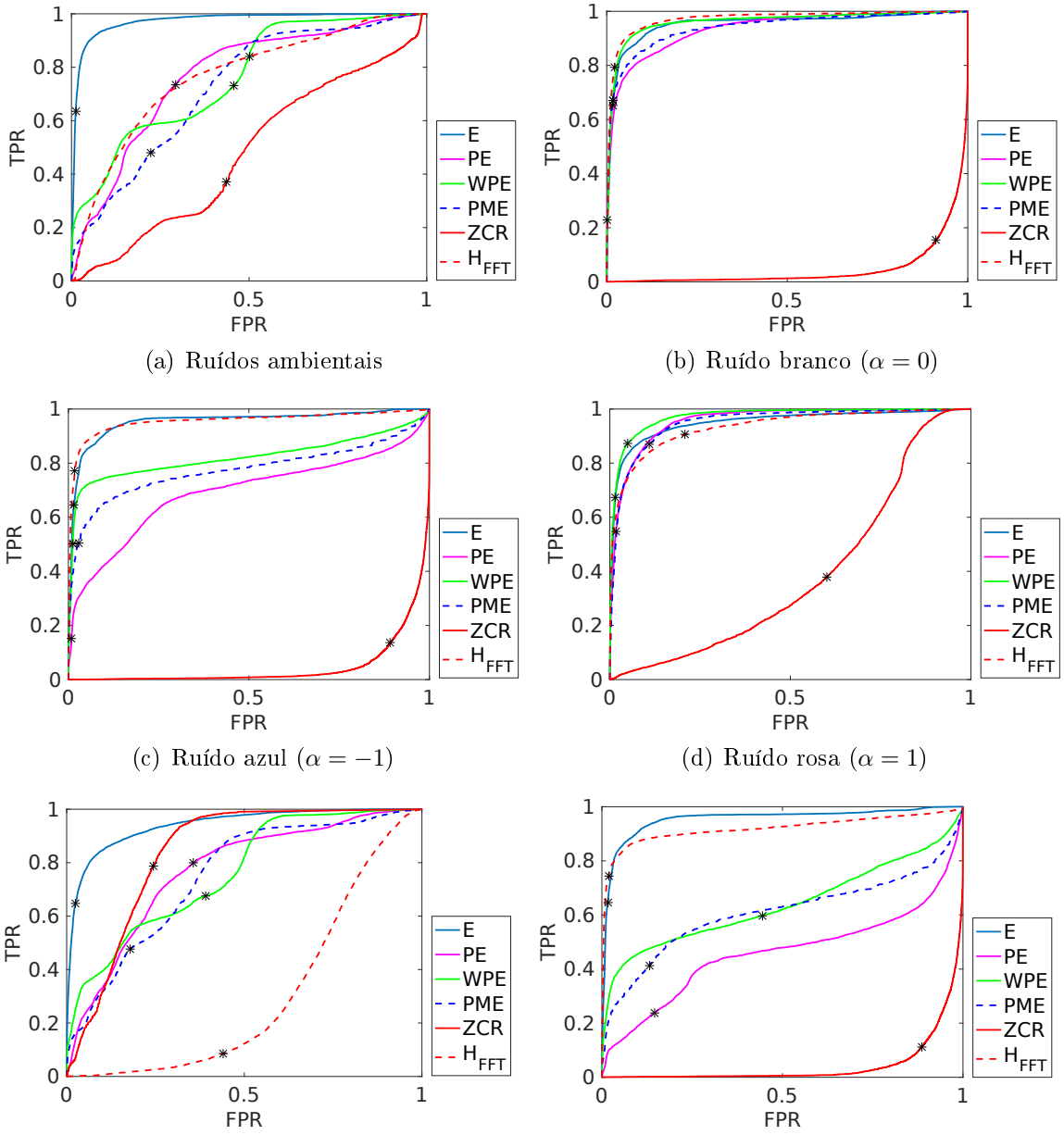
1: $Tf = \bar{x}$;
2: $T_i = 0$;
3: **enquanto** $|Tf - T_i|$ **faça**
4: $\quad m_1 = \text{média}(x \geq Tf)$;
5: $\quad m_2 = \text{média}(x \leq Tf)$;
6: $\quad T_i = Tf$;
7: $\quad Tf = \frac{m_1+m_2}{2}$;
8: **fim enquanto**

Uma vez encontrado o valor T_f, uma regra de comparação simples é aplicada para decidir se o *frame* correspondente é sinal ($\hat{\text{D}}_{id}(f,t) \geq T_f$) ou ruído de fundo ($\hat{\text{D}}_{id}(f,t) < T_f$). A escolha do T_f lida com uma relação custo-benefício entre perda de sílabas (por culpa de escolher um limiar elevado) e tolerância aos falsos positivos (por escolher limiares baixos). A figura 4.2(b) ilustra um exemplo de segmentação utilizando a energia dos *frames* e o limiar encontrado pelo algoritmo 2. Um exemplo da relação entre a taxa de falsos positivos e a taxa verdadeiros positivos dada por esta regra é ilustrada pelos pontos em negrito da figura 4.3(a).

4.5.2 Análise *frame-a-frame* utilizando curvas ROC

A análise ROC permite quantificar o desempenho do segmentador, em termos das taxas TPR e FPR. A metodologia para criar a curva ROC implica que devem ser testados todos os valores possíveis de limiares ($0 \leq T_f \leq 1$). A interpretação de cada curva, neste caso, representa capacidade do segmentador para selecionar *frames* da classe "sinal" utilizando cada descritor individualmente.

Para poder gerar tais curvas realizamos um experimento no qual atribuímos a cada frame a classe +1 se pelo menos 30% de seus valores pertencem a uma sílaba segmentada manualmente, ou seja, 30% dos pontos do frame são GT. Desta forma, podemos relacionar o rótulo da classe aos valores LLDs. A partir destes valores, construímos as curvas e calculamos o valor da área sob estas (AUC) para comparar o



(a) Ruídos ambientais

(b) Ruído branco ($\alpha = 0$)

(c) Ruído azul ($\alpha = -1$)

(d) Ruído rosa ($\alpha = 1$)

(e) Ruído vermelho ($\alpha = 2$)

(f) Ruído violeta ($\alpha = -2$)

Figura 4.3. Curvas ROC de segmentação para cada descritor acústico com diferentes tipos de ruídos.

desempenho final da segmentação de cada LLDs. As figuras 4.3 apresentam as curvas ROC para todas as espécies listadas na tabela 4.1 com os ruídos ambientais das gravações, mais a adição de cinco tipos diferentes de ruídos.

Em um cenário real podemos encontrar diferentes tipos de ruídos. Portanto, para simular tais cenários, contaminamos artificialmente o conjunto de dados original com ruídos branco, azul, rosa, vermelho e violeta (seção 2.3.1), utilizando o mesmo

valor de variância que os sinais originais $\sigma_x \approx \sigma_n$. De acordo com a equação 2.23, esta contaminação resulta em SNR = 0 dB (página 32). A figura 4.3 mostra as curvas obtidas para cada LLDs em todos os cenários simulados. Os pontos em negritos ($*$) representam a relação FPR-TPR dada pelo limiar T_f encontrado com o algoritmo 2. Aqui podemos notar que o limiar T_f causa uma FPR-TPR que pode ser considerada ótima (baixos falsos positivos e elevados verdadeiros positivos), concluindo-se que o algoritmo 2 produz um limiar de decisão ótimo. O número de sílabas recuperadas aplicando este limiar na condição BN é resumido na tabela 4.1, entre a terceira e oitava coluna.

Observamos na figura 4.3(a) que a energia do sinal E produz uma curva "bem comportada", isto é, maximiza o AUC nos cenários com BN e $\alpha = \{-1, 2, -2\}$, o que significa que esta curva alcança rapidamente valores elevados de TPR enquanto que os FPR continuam baixos. Este comportamento deve-se às características das vocalizações dos anuros que crescem rapidamente em amplitude, separando-se do nível de ruído ambiente. O conjunto de características baseadas em entropia das permutações (PE, WPE, PME) tiveram um comportamento similar a E nos casos $\alpha = \{0, 1\}$. A H_f resultou "bem comportada" nos casos $\alpha = 0, -1, 1, 2$. Finalmente a característica com pior desempenho foi o ZCR, por ser sensível às mudanças de frequência dos sinais, tanto dos ruídos quanto das sílabas. Uma possível inversão da regra que utiliza o ZCR poderia melhorar os resultados, no entanto, inverter a regra de decisão é impraticável para uma aplicação real devido ao fato de que não é possível conhecer o tipo de contaminação a *priori* ou mudar a regra conforme a mudança do cenário acústico.

Para avaliar e comparar o desempenho de todas os descritores de maneira quantitativa, calculamos a área das curvas AUC. Os resultados são apresentados na tabela 4.2 são os cálculos das áreas das curvas presentes na figura 4.3. Nesta tabela, podemos confirmar que E possui bom desempenho para discriminar a amplitude do sinal, sendo pouco afetado pelo tipo de ruído adicionado. Dentre as medidas de entropia das permutações, podemos destacar que WPE foi melhor, pois esta considera valores de amplitude do sinal. E finalmente percebemos que quando os ruídos afetam as altas frequências ou todas as frequências de sinal em igual medida, o H_f apresenta um resultado ótimo.

Na figura 4.4 apresentamos a variação do AUC em função da variância dos ruídos σ_n para todos os casos simulados, no intervalo $-35\,\text{dB} \leq \text{SNR} \leq 70\,\text{dB}$. Estas figuras representam o desempenho do sistema nos seus dois extremos, em uma situação ideal quase sem ruídos aleatórios (70 dB ou melhor caso) e em uma situação altamente desfavorável (-35 dB ou pior caso). Como podemos observar, para uma $\text{SNR} \leq -25\,\text{dB}$, a decisão de segmentação é quase aleatória com qualquer LLDs, exceto no caso de contaminação com ruído vermelho o qual possui alta concentração de energia nas bandas

Tabela 4.2. Desempenho de cada descritor quantificado através do AUC para diferentes condições de ruído. BN ruídos ambientais sem ruído artificial. Colunas 4-7 ruídos coloridos aditivos segundo a equação 2.22. Em todos os casos os ruídos adicionados possuem a mesma variância do sinal original.

	BN	branco	azul	rosa	vermelho	violeta
E	**0,97**	0,95	**0,95**	0,95	**0,93**	**0,95**
PE	0,76	0,93	0,69	0,95	0,77	0,46
WPE	0,76	0,96	0,81	**0,97**	0,77	0,62
PME	0,72	0,94	0,77	0,94	0,73	0,61
ZCR	0,47	0,04	0,04	0,38	0,84	0,04
H_f	0,76	**0,97**	**0,95**	0,93	0,32	0,91

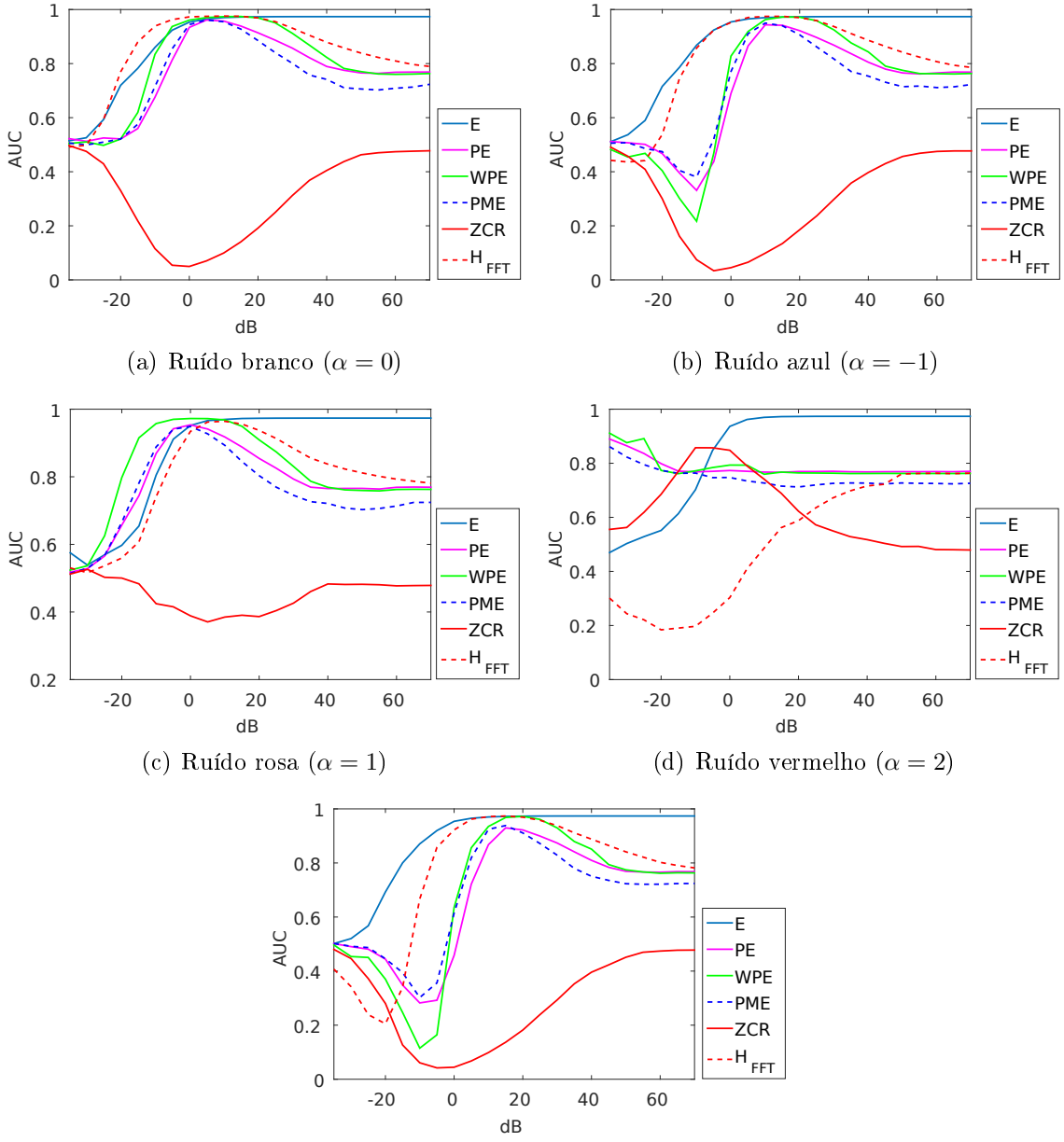
de frequências mais baixas.

Para valores de SNR elevados, a segmentação baseada em E apresenta um resultado ótimo em todos os casos. Os LLDs baseados em PE mostraram um comportamento diferente, quando a SNR diminui, o resultado da segmentação melhora, porque as correlações fracas são quebradas pela adição de ruído aleatório de alta variância. No caso particular do ruído vermelho, os descritores PE, WPE e PME resultaram melhor do que E para valores extremamente desfavoráveis (SNR $\leq -10\,\mathrm{dB}$). Isto deve-se a duas razões fundamentais, primeiro as autocorrelações internas das sílabas foram pouco afetadas por este tipo de ruído, e segundo o ruído ambiental caraterístico das gravações dentro da floresta contém elevada energia nas bandas de baixas frequências[2] e portanto o fator aleatório introduzido ajudou a romper os padrões dos sinais que geram as correlações de baixas frequências. Entre as características de entropia, o WPE capta melhor as diferenças de amplitude nas sílabas.

Um comportamento contrário em condições de ruído vermelho verifica-se para o descritor H_f, o qual basei-se no espectrograma para obter a entropia. Neste caso, a concentração de energia nas baixas frequências diminui a entropia espectral. Entretanto, H_f obteve um desempenho ótimo, inclusive melhor que E, em uma ampla variação da SNR ($-10\,\mathrm{dB} \leq \mathrm{SNR} \leq 25\,\mathrm{dB}$), em condições com ruído branco. Em outras palavras, como o ruído branco espalha energia uniformemente em todas as bandas de frequências, o histograma da H_f é equalizado, destacando-se somente os padrões correspondentes às sílabas. No caso em que a SNR supera os 25 dB, o desempenho começa a cair, indicando que a energia espectral do ruído alcançou o nível de energia espectral das sílabas.

Como regra geral, destacamos que os valores máximos de AUC são atingidos entre $-5\,\mathrm{dB} \leq \mathrm{SNR} \leq 10\,\mathrm{dB}$ com a maioria dos LLDs exceto para o ZCR, inclusive no caso

[2]Esta observação pode ser constata na figura 4.1, na qual existe elevada energia nas bandas $0\,\mathrm{Hz} \leq f_s \leq 300\,\mathrm{Hz}$.



(a) Ruído branco ($\alpha = 0$)

(b) Ruído azul ($\alpha = -1$)

(c) Ruído rosa ($\alpha = 1$)

(d) Ruído vermelho ($\alpha = 2$)

(e) Ruído violeta ($\alpha = -2$)

Figura 4.4. Variação do AUC em relação à SNR. Estas curvas ilustram o desempenho da segmentação entre níveis extremos de SNR. Os valores de AUC em 0 dB são consistentes com os valores apresentados na tabela 4.2.

mais favorável com ruído vermelho, pois este não conseguiu o mesmo desempenho que os demais LLDs.

Por último, a figura 4.5 ilustra a variação da AUC em termos de densidade do ruído impulsivo, adicionado aos sinais de acordo à equação 2.24 (página 32). Neste caso, a densidade de picos $d_\delta\,\%$ representa percentagem de pontos do sinal alterados por $\pm\delta$. Aqui observamos que o desempenho de todos os LLDs diminui rapidamente

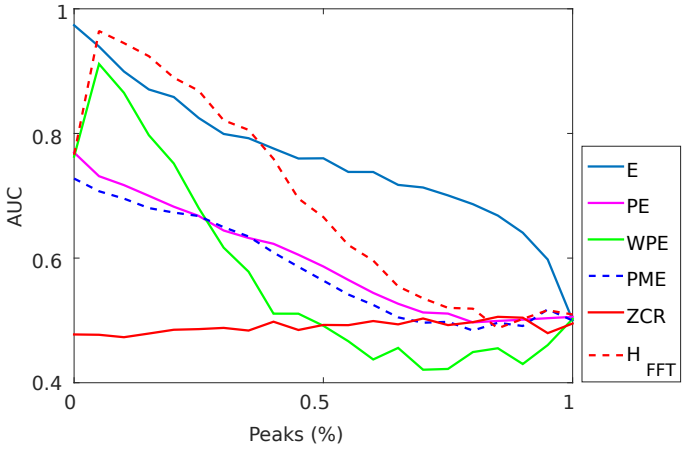


Figura 4.5. Variação da AUC em relação à porcentagem de ruído impulsivo. Os valores referentes a 0% são consistentes com os da coluna BN na tabela 4.2.

quando a quantidade de ruído impulsivo aumenta. Sem embargo, existem as exceções WPE e H_f, que após o ponto inicial a segmentação melhora até alcançar seu máximo em aproximadamente $d_\delta = 0,1\,\%$. Neste caso, a H_f supera inclusive a E mantendo-se acima até o ponto $d_\delta = 0,4\,\%$. A razão é que a adição de um componente estocástico completamente descorrelacionado com o sinal, ajuda a quebrar as autocorrelações fracas. Observamos que o WPE apresenta maior queda de AUC após seu ponto máximo, isto ocorre devido ao aumento na frequência relativa de alguns padrões (π_j) ponderados pela alta variâncias do ruído impulsivo.

4.5.3 Análise dos eventos acústicos aplicando AEER

O sistema bioacústico para reconhecimento de espécies baseia-se na capacidade do classificador em distinguir cada sílaba individualmente. Portanto, cada sílaba pode ser considerada como um evento acústico isolado. Do ponto de vista da detecção de eventos acústicos, precisamos quantificar quantos eventos são perdidos ou incorretamente recuperados.

Na tabela 4.1 foi apresentado o número de eventos identificados pelo especialista (segunda coluna) em comparação à quantidade de eventos recuperados aplicando a regra automática descrita. As linhas da tabela foram separadas por espécies para mostrar claramente a dificuldade em segmentar os padrões das vocalizações das diferentes espécies. Em alguns casos, tais como Adenomera a. ou Osteocephalus o., a técnica encontrou mais sílabas que as existentes, o que indica ocorrência de micro cortes na segmentação (*e.g.*, uma sílaba dividida em dois). No caso oposto, como na espécie Brachycephalus e., o número total de sílabas recuperadas foi inferior ao GT, o que indica

que os descritores foram suficientemente sensíveis, e se perderam eventos do fluxo de áudio.

Dentre todas as colunas da tabela 4.1, notamos que E produz valores próximos ao GT, entretanto é necessário saber se os eventos recuperados foram realmente as sílabas esperadas. Uma métrica útil para quantificar tais erros é o *Acoustic Event Error Rate* - AEER (equação 2.65, página 58). Esta métrica é frequentemente utilizada nas abordagens de detecção de contexto através do áudio, e foi aplicada ao problema de segmentação bioacústica no trabalho de Colonna et al. (2015). O AEER considera que um evento é corretamente detectado se as fronteiras deste encontram-se próximas às do evento real, com uma tolerância de ± 50 ms, e possui duração igual ou maior a 50% do tempo total da sílaba comparada. Além disso, eventos duplicados são considerados falsos alarmes. Assim, no melhor dos casos o AEER tende a zero.

Os valores de AEER apresentados na tabela 4.3 mostram que a menor taxa de eventos errados foi obtida com E e o segundo melhor valor com H_f, na condição BN. A última linha desta tabela corresponde à média dos AEER, também conhecida como Macro-AEER. Embora o teste estatístico indique empate ente E e H_f, comparando a coluna H_f desta tabela com a respectiva coluna da tabela 4.1 identificamos que o desempenho não foi aceitável, pois muitas sílabas foram perdidas ou inclusive, nenhuma sílaba foi extraída em algumas espécies. Além disso, comparado as duas tabelas, deduzimos que a utilização dos PE, WPE, PME e ZCR produz micro segmentações dos sinais, situação que pode ser desfavorável para o classificador subsequente.

4.5.4 Análise ponto-a-ponto

Para medir a precisão dos limites estimados, comparamos o GT com a segmentação automática contando os erros ponto-a-ponto. Assim, cada ponto do sinal segmentado é comparado com cada ponto da segmentação GT aplicando as métricas Prec, Rec e F1 baseadas na matriz de confusão (seção 2.6.1, página 56). Estas métricas são úteis para comparar os pontos recuperados das sílabas que são relevantes e a fração de pontos relevantes que são recuperados. Quanto maior o valor destas métricas melhor é a segmentação automática.

A tabela 4.4 apresenta os valores Prec, a Rec e o F1 de cada LLD na condição BN. Destacamos em negrito os resultados que mostraram ganhos estatisticamente significativos usando o melhor valor de cada linha para a aplicação do teste, considerando um nível de confiança de 95%. Como podemos notar, E teve os melhores valores de Prec e F1. Isto significa que, segmentar o sinal com somente a energia é possível identificar corretamente os pontos pertencentes às sílabas. Por outro lado, as características

Tabela 4.3. AEER utilizado para quantificar a qualidade na recuperação das sílabas (*evento-a-evento*). A última linha apresenta a Macro-AEER. O *t-test* com significância $p \leq 0,05$ foi aplicado para comparar E aos restantes LLDs. Os valores considerados empate encontram-se em negrito.

	LLDs					
Espécies	E	PE	WPE	PME	ZCR	H_f
Adenomera h.	0,06	2,08	2,67	2,74	3,32	3,64
Hyla m.	2,18	5,21	4,84	5,15	14,46	1,11
Adenomera a.	0,23	5,44	4,92	5,60	5,13	3,96
Ameerega t.	0,88	0,93	0,97	1,02	2,86	1,02
Osteocephalus o.	2,52	13,66	11,32	14,09	14,24	23,86
Rhinella g.	1,60	0,00	1,00	0,60	28,60	1,40
Scinax r.	1,35	1,78	2,06	2,05	3,46	1,04
Hypsiboas c.	0,12	3,09	3,11	3,44	3,43	1,74
Brachycephalus e.	0,72	0,74	0,97	0,86	1,00	1,00
Aplastodiscus albof.	0,00	6,70	6,26	6,94	6,24	1,03
Aplastodiscus albos.	1,70	18,70	22,29	26,29	20,64	1,82
Aplastodiscus p.	1,55	0,00	0,00	1,03	10,18	1,07
Dendropsophus a.	0,45	6,12	4,47	5,37	5,17	1,10
Dendropsophus e.	0,00	12,06	12,86	18,46	1,40	1,13
Macro-AEER	**0,95**	5,46	5,55	6,69	8,58	**3,21**

Tabela 4.4. Precisão, Revocação e F-Score para a avaliação ponto-a-ponto das fronteiras dos eventos acústicos na condição BN. Os números em negrito foram considerados empate com significância estatística $p \leq 0,05$ comparando todos os LLDs com o melhor valor cada linha.

	E	PE	WPE	PME	ZCR	H_{FFT}
Pre	**0,87**	0,39	0,36	0,34	0,12	0,15
Rec	0,53	**0,93**	**0,96**	**0,95**	0,44	0,23
F1	**0,61**	**0,48**	0,46	0,44	0,16	0,13

baseadas em entropia (PE, WPE e PME) tiveram melhor Rec, o que significa que estes LLDs identificaram melhor os pontos que não pertencem às sílabas.

Nas figuras 4.6(a) e 4.6(b) ilustramos as variações das taxas de falsos negativos (FNR) e falsos positivos (FPR) para diferentes valores de ruído branco aditivo. Com estas métricas, quanto menor é o seu valor, melhor é o resultado da segmentação. Na figura 4.6(c) apresentamos a acurácia (Acc), que ao contrario das taxas anteriores, quanto maior é seu valor, melhor é a segmentação. Estas métricas foram obtidas ponto-a-ponto usando uma vocalização da espécie Aplastodiscus perviridis como exemplo, a qual foi escolhida por ter baixo nível de ruído ambiental. Neste exemplo, o ruído branco adicionado foi variando no intervalo $-35\,\text{dB} \leq \text{SNR} \leq 70\,\text{dB}$.

Na figura 4.6(a) observamos que a partir do valor $\text{SNR} \geq -10\,\text{dB}$ a E produz uma FNR constante, enquanto que as medidas de entropia geram menos falsos negativos.

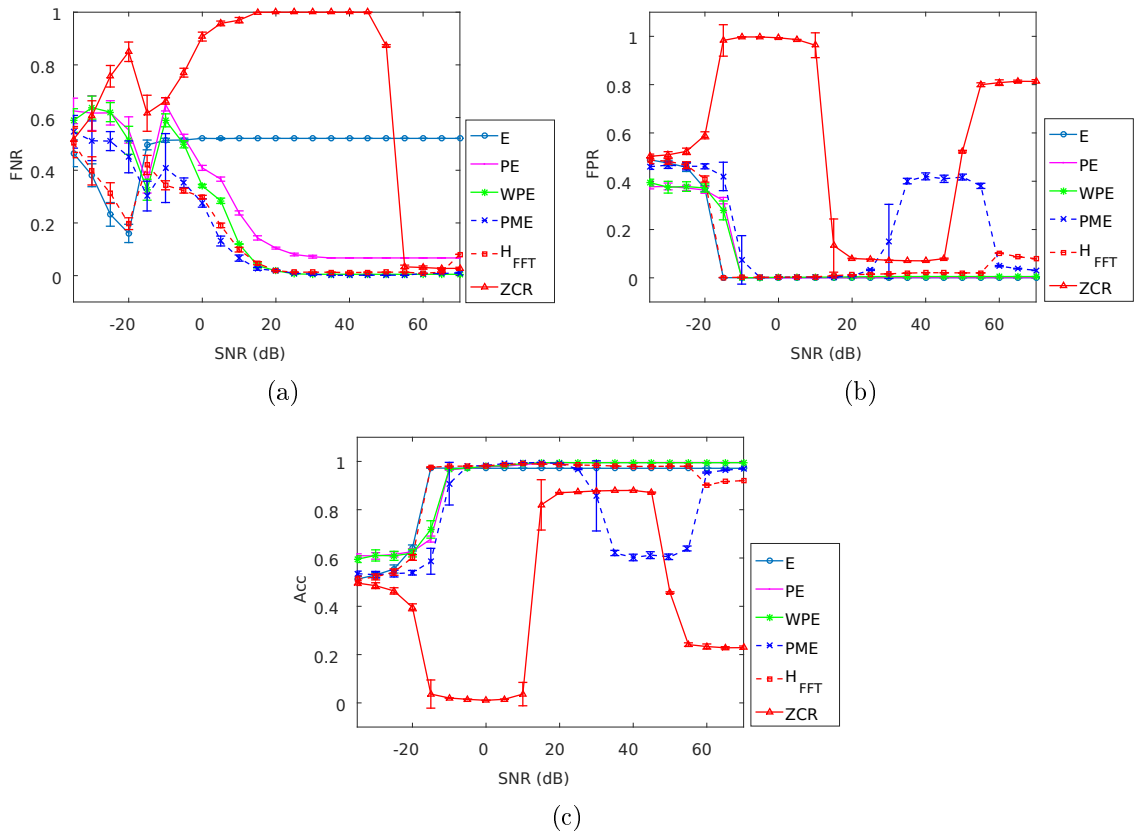


Figura 4.6. Exemplo do desempenho de segmentação utilizando uma vocalização da espécie Aplastodiscus p. quando se adiciona ruído branco variável, quantificado através de métricas ponto-a-ponto.

Em outras palavras, os descritores baseados em entropia perdem menos pontos das sílabas quando o ruído adicionado é branco e possui baixa variância. No que diz respeito ao ZCR, notamos que este melhora a partir SNR $\geq 55\,\text{dB}$, o que indica que adicionar artificialmente uma variável completamente descorrelacionada ajuda na segmentação.

Conclusões similares para as medidas de entropia podem ser obtidas inspecionando a figura 4.6(b), com exceção do PME, o qual aumenta sua taxa de falsos positivos no intervalo $10\,\text{dB} \leq \text{SNR} \leq 60\,\text{dB}$. Novamente, o ZCR obteve um desempenho inferior comparado aos restantes LLDs, exceto no intervalo $15\,\text{dB} \leq \text{SNR} \leq 45\,\text{dB}$ no qual melhorou a FPR levemente. Comparando a FNR do ZCR no mesmo intervalo podemos deduzir que este LLD perdeu aproximadamente todos os pontos do sinal, incluído sílabas e não sílabas. Além disso, notamos que todos os LLDs tiveram um comportamento aleatório para valores inferiores $\text{SNR} < -20\,\text{dB}$, que se correspondem a situações de ruído muito desfavorável, onde os sinais são praticamente imperceptíveis. Finalmente, concluímos que o LLD com melhor desempenho neste exemplo foi H_f. Pois este foi favorecido pelo incremento no piso de ruído do espectro, separando

melhor os valores de entropia.

4.5.5 Ranqueamento e combinação de LLDs

Como cada caraterística acústica não esta diretamente relacionada com as demais, exceto pelo subgrupo baseado na entropia das permutações (PE, WPE e PME), podemos supor que a combinação de LLDs melhoraria o resultado da segmentação. Para abordar esta hipótese, duas questões fundamentais devem ser consideradas: (1) como devem-se combinar os LLDs evitando cair em um problema combinatório com complexidade exponencial, e (2) como reduzir a combinação para um vetor de uma dimensão e aplicar o algoritmo 2. Dadas estas premissas, decidimos primeiro classificar os LLDs de acordo com o Ganho da Informação (IG) de cada um, para gerar um ranqueamento e evitar realizar combinações desnecessárias. Com base no ranqueamento, geramos combinações sequenciais começando pelo LLD com maior IG e incluindo sequencialmente os restante LLD de acordo a sua posição no *ranking*. Posteriormente, para reduzir a dimensão da combinação e poder aplicar o algoritmo 2, aplicamos a Análise de Componentes Principais (PCA).

O ganho de informação (IG) avalia os atributos medindo a redução da incerteza em relação às classes, considerando a entropia como medida de "impureza" (Witten and Frank, 2005). Em outras palavras, IG quantifica a redução na impureza da classificação causada por cada caraterística. Assim, os LLDs que segmentam perfeitamente o conjunto de sílabas deveriam maximizar o IG, enquanto que os LLDs que não fornecem informações relevantes para segmentar os sinais deveriam minimizar o IG. O ranqueamento dos LLDs utilizados e seus IGs, para as condições com ruído de fundo (BN) e com ruído branco a 0 dB, são mostrados na tabela 4.5. Comparando as colunas desta tabela, percebemos que um aumento na SNR leva a uma diminuição no IG e causa uma nova ordem dos LLDs, na qual H_f lidera o ranqueamento. Independentemente disso, os valores de AUC na tabela 4.6 apresentam melhorias para algumas combinações de LLDs quando o ruído branco é adicionado. Este fato confirma mais uma vez que a presença de ruído branco melhora o desempenho dos LLDs.

Após o ranqueamento, os LLDs foram combinados sequencialmente e reduzidos via PCA. As figuras 4.7(a) e 4.7(b) mostram as curvas ROC de tais combinações aplicando a metodologia de segmentação descrita anteriormente. Os valores AUC de cada curva são apresentados na tabela 4.6. Entre todas as curvas, a energia do sinal é melhor na condição BN, enquanto que a H_f é melhor na presença de ruído branco. As combinações que incluem PE, WPE e PME obtiveram desempenho semelhante entre elas. Além disso, quando adicionamos ZCR ao conjunto de combinações na condição

Tabela 4.5. Ranqueamento IG em condições normais de ruído ambiental (BN) e com ruído branco (0 dB. A ordem do ranqueamento foi alterada, mostrando o efeito de decorrelação causado pela adição de uma variável aleatória.

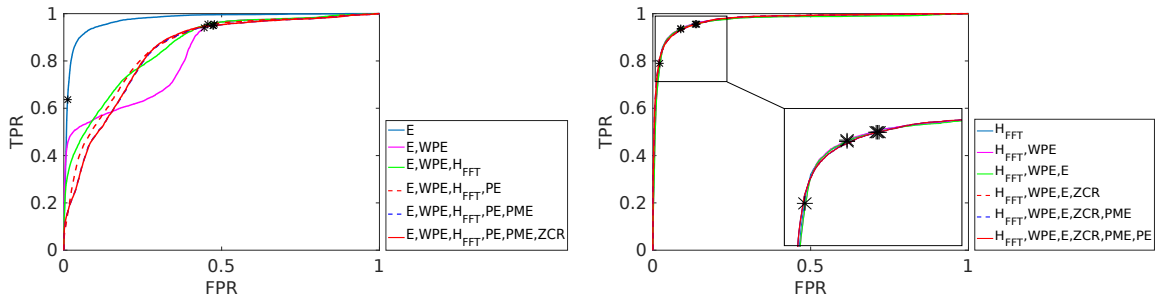
	BN		Ruído Branco	
Ranking	LLD	IG	LLD	IG
1^{st}	E	0,4811	H_{FFT}	0,3163
2^{nd}	PE	0,4809	WPE	0,2738
3^{rd}	WPE	0,4807	E	0,2727
4^{th}	H_{FFT}	0,4807	ZCR	0,2643
5^{th}	PME	0,2282	PME	0,2623
6^{th}	ZCR	0,1480	PE	0,1243

Tabela 4.6. Valores de AUC correspondentes às curvas das figuras 4.7(a) e 4.7(b), gerados pelas combinações sequenciais dos LLDs de acordo ao ranqueamento apresentado na tabela 4.5.

BN		White Noise	
LLDs	AUC	LLDs	AUC
E	0,973	H_{FFT}	0,972
E,WPE	0,830	H_{FFT},WPE	0,971
E,WPE,H_{FFT}	0,862	H_{FFT},WPE,E	0,969
E,WPE,H_{FFT}	0,854	H_{FFT},WPE,E	0,974
PE		ZCR	
E,WPE,H_{FFT}	0,841	H_{FFT},WPE,E	0,974
PE,PME		ZCR,PME	
E,WPE,H_{FFT}	0,842	H_{FFT},WPE,E	0,975
PE,PME,ZCR		ZCR,PME,PE	

BN, o desempenho reduz ainda mais. Concluímos que a combinação de características no primeiro cenário (figura 4.7(a)) não melhora o desempenho de segmentação como esperávamos. Contudo, no cenário com ruído branco (figura 4.7(b)), todas as combinações obtiveram um desempenho ótimo. Adicionalmente, considerando as variações nos valores AUC da tabela 4.6, notamos que a combinação de LLDs baseada no coeficiente IG pode ser enganosa.

Por último, a figura 4.8 mostra um exemplo visual de segmentação utilizando três características separadamente e a sua combinação e redução via PCA. Estas características (E, WPE e H_f) são essencialmente não relacionadas porque foram obtidas através de diferentes procedimentos. Esta figura também ilustra uma mistura de sílabas pertencentes a três espécies diferentes, isto é, uma concatenação de três vocalizações de diferentes espécies em apenas um áudio. As linhas pontilhadas verticais marcam o ponto inicial e final de cada vocalização. Além do ruído de fundo já presente nas vocalizações, adicionamos ruído branco a 20 dB. Conseguimos assim, verificar visualmente as fronteiras das sílabas encontradas utilizando cada LLD. Neste caso, E se mostrou



(a) **BN**. (b) Ruído branco.

Figura 4.7. Curvas ROC correspondentes à redução de cada combinação de LLDs apresentadas na tabela 4.6.

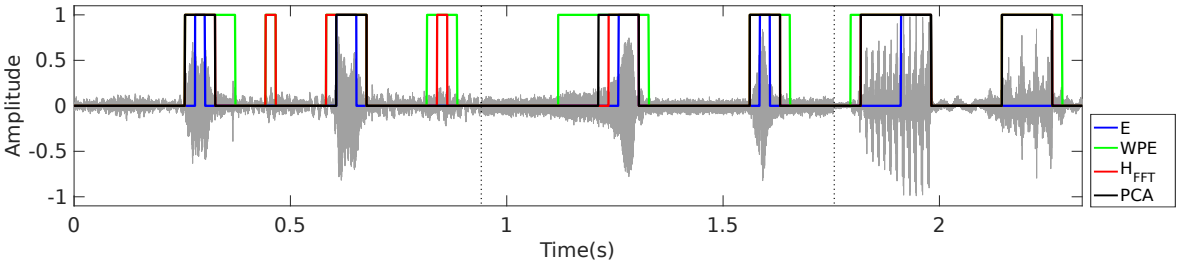


Figura 4.8. Exemplo de limites de segmentação encontrados em um sinal misturado com três especies diferentes Adenomera h., Hyla m. e Scinax r., contaminado com ruído branco a 20 dB. Comparação visual de segmentação utilizando E, WPE, H_f e a redução aplicando PCA.

demasiado estrito, porque corta partes das sílabas ou perde algumas destas (*e.g.* a última sílaba). Em contraste, os limites de WPE e H_f foram menos rigorosos incorrendo em dois falsos positivos. A redução PCA conseguiu um desempenho equilibrado, nem muito rigoroso nem muito tolerante. Entretanto, a combinação de LLDs e a redução PCA aumentaram a quantidade de operações, e portanto, o custo computacional do método de segmentação.

4.5.6 Conclusões sobre a comparação de LLDs aplicados à segmentação

Nesta seção, apresentamos uma série de avaliações que abrangem a comparação de diferentes descritores acústicos de baixo nível usados para segmentar automaticamente chamadas dos anuros. As avaliações apresentadas incluíram métricas *frame-a-frame*, evento-a-evento e ponto-a-ponto. Avaliamos também o ganho de informação de cada descritor e a qualidade da segmentação para diferentes combinações. Simulamos diver-

sos cenários, incluindo ruído brancos e coloridos com diferentes níveis de variância, e apresentamos um algoritmo para encontrar o limiar de segmentação ótimo. Finalmente, mostramos que, a ideia de combinar a simplicidade dos modelos de energia (baseados em limiares de decisão) com a robustez dos modelos probabilísticos, de forma não supervisionada, é ótima para resolver o problema da segmentação bioacústica com poucos recursos computacionais, o que possibilitaria a utilização em sensores de baixo custo.

Através dos experimentos mostramos que a inclusão da metodologia baseada na entropia das permutações pode melhorar a segmentação, mas a melhoria depende do tipo de ruído que afeta os sinais. Isto é, a PE e suas variantes apresentam bom desempenho quando o ruído das gravações é do tipo rosa ou vermelho, nos quais a baixas frequências são predominantes, mas experimentam uma diminuição no desempenho da segmentação quando o tipo de ruído é azul ou violeta, nos quais predominam as altas frequências. Além disso, não aconselhamos utilizar a WPE em cenários propensos a ruídos impulsivos.

Em nosso experimentos, concluímos que a adição de ruído branco melhora o resultado da segmentação dos LLDs baseados em entropia, porque "rompem" as correlações fracas do ruído ambiental. Este efeito, causado pela adição de uma variável aleatória, também foi relatado por Bandt and Pompe (2002) e Veisi et al. (2007), sendo uma consequência da propriedade de transformação invariante de PE. Entretanto, verificamos que esta propriedade é mantida quando trata-se da análise de sinais bioacústicos. No caso da entropia espectral, a adição de ruído branco melhora a segmentação, porque este ruído espalha energia uniformemente entre todas as frequências, mascarando os pequenos picos do espectrograma. Assim, o espectro torna-se mais concentrado nas frequências fundamentais diminuindo a entropia.

No que diz respeito à complexidade, considerando um tamanho de *frame* n, o cálculo dos LLDs E e ZCR possuem complexidade computacional linear $\Theta(n)$; a complexidade da H_f depende da transformada de Fourier (FFT), a qual tem uma complexidade $\Omega(n \log_2(n))$ no melhor caso e $\mathcal{O}(n^2)$ no pior caso; e os LLDs derivados da metodologia PE possuem complexidades $\mathcal{O}(mn)$, onde m é o tamanho dos padrões ordinais.

Na próxima seção apresentamos a transformação incremental de dois LLDs aplicados ao problema de segmentação em tempo real com baixos recursos em sensores acústicos.

4.6 Metodologia de segmentação incremental

Várias abordagens de segmentação automática utilizam procedimentos não-sequenciais que consomem quantidades maiores de memória e processamento (seção 3.3). Por uma razão prática esses tipos de abordagens não são adequados para cenários com fluxo contínuo de dados, como no caso de *stream* de áudio. Além disso, quando trata-se de monitoramento ambiental com sensores bioacústicos, grandes quantidades de dados devem ser processadas em tempo real por dispositivos com limitações de recursos (Nakamura et al., 2014). Como a segmentação é o primeiro passo do procedimento de reconhecimento, deseja-se consumir o mínimo possível de recursos.

Nesta seção, introduzimos nossa proposta de segmentação incremental com complexidade computacional constante e consumo mínimo de memória. Nossa técnica fornece a resposta em tempo real (ponto-a-ponto do sinal). Em contraste ao uso de janelas deslizantes, a estratégia incremental armazena apenas estatísticas simples das séries temporais. Assim, conseguimos uma redução de memória igual a $1/N$ (sendo N o tamanho da janela). A complexidade também é reduzida de $\mathcal{O}(N \times (n/(N-m)))$ para $\mathcal{O}(n)$ (onde m é a sobreposição entre as janelas sucessivas).

Aplicar janelas deslizantes para detectar o início e fim de cada sílaba, lida com um *trade-off* entre processar dados desnecessariamente ou aumentar a taxa de falsos positivos. Portanto, se o tamanho da janela for muito maior do que a sílaba, dados irrelevantes serão utilizados para calcular os descritores, fato que pode causar a perda do sinal. Por outro lado, um tamanho muito reduzido de janelas pode causar que ruídos impulsivos sejam detectados como sílabas de curta duração, confundindo o classificador. Consequentemente, observamos que os parâmetros N e m afetam a precisão do segmentador, sendo difícil definir um par de valores que operam corretamente para todas as espécies descritas.

De acordo com Jaber (2013) técnicas incrementais devem satisfazer os seguintes requisitos: (a) adaptabilidade aos tipos de mudanças, sejam estas abruptas ou graduais; (b) transferência de conhecimentos lembrando as decisões passadas; (c) redução do tempo de reação; e (d) baixa taxa de erro, evitando falsos positivos e negativos. Com base nestes requisitos, estamos interessados em técnicas de segmentação que sejam sensíveis às mudanças, robustas aos ruídos e eficientes em relação à memória necessária, o tempo de processamento e o consumo de energia. Para conseguir isto, refizemos a abordagem de janelas para lembrar o histórico das séries temporais por meio de descritores que podem ser calculados de forma incremental.

4.6.1 Proposta incremental

O desafio aqui é detectar o início e o fim de uma sílaba sem aplicar janelas deslizantes, simplesmente identificando mudanças dos sinais. Para isto, começamos rescrevendo o cálculo de energia como:

$$\mathrm{E} = \frac{1}{N}\sum_{i=1}^{N} x_i^2. \tag{4.4}$$

Acrescentar o fator $\frac{1}{N}$ nesta equação equivale a calcular a média dos valores x_i^2 dentro do *frame*. Desta forma, podemos reescrever o cálculo da energia com pesos $w_1, w_2, \ldots w_n \geq 0$ semelhante à média ponderada exponencial (Finch, 2009):

$$\mathrm{E}_n = \frac{1}{\mathrm{W}_n}\sum_{i=1}^{n} w_i x_i^2, \tag{4.5}$$

com pesos $\mathrm{W}_n = \sum_{i=1}^{n} w_i$ e $\alpha = \frac{w_n}{\mathrm{W}_n}$, e índice temporal n. Assim:

$$\begin{aligned}
\mathrm{E}_n &= \frac{1}{\mathrm{W}_n}(w_n x_n^2 + \sum_{i=1}^{n-1} w_i x_i^2) \\
\mathrm{E}_n &= \frac{1}{\mathrm{W}_n}(w_n x_n^2 + \mathrm{W}_{n-1}\mathrm{E}_{n-1}) \\
\mathrm{E}_n &= \frac{1}{\mathrm{W}_n}(w_n x_n^2 + (\mathrm{W}_n - w_n)\mathrm{E}_{n-1}) \\
\mathrm{E}_n &= \frac{1}{\mathrm{W}_n}(\mathrm{W}_n\mathrm{E}_{n-1} + w_n(x_n^2 - \mathrm{E}_{n-1})) \\
\mathrm{E}_n &= \mathrm{E}_{n-1} + \frac{w_n}{\mathrm{W}_n}(x_n^2 - \mathrm{E}_{n-1})
\end{aligned}$$

e, finalmente,

$$\mathrm{E}_n = \mathrm{E}_{n-1} + \alpha(x_n^2 - \mathrm{E}_{n-1}). \tag{4.6}$$

A equação 4.6 permite calcular o valor de E_n sem utilizar uma janela deslizante. Nesta nova formulação, o parâmetro α controla o *trade-off* entre importância da amostra atual e o peso das amostras passadas, provendo um valor de E a cada novo dado de entrada. Esta equação adapta-se a mudanças graduais dos sinais, tais como o aumento progressivo no nível de ruído, sem perder mudanças abruptas de amplitude. O custo de memória é $\mathcal{O}(1)$ e o custo de processamento é $\mathcal{O}(n)$. Esta transformação incremental suporta transferência de conhecimento, lembrando das decisões passadas quando novas amostras são processadas. Além disso, possui somente um parâmetro para ajustar e

satisfazer as restrições de memória e processamento dos sensores de baixo custo. A partir de agora nos referimos a esta técnica como E-I.

A variável α está diretamente relacionada com a sensibilidade do segmentador. Dependendo do valor da sensibilidade os tempos de transição entre sinais com diferentes amplitudes causam flutuações nos valores de E. A variável de decisão neste método é binária ("1" sinal ou "0" ruído). Assim, para reduzir a probabilidade de falsas decisões, ou micro-segmentações, adicionamos um filtro da moda[3]. Aqui identificamos esta nova versão do segmentador incremental como E-IMF. Com esta modificação, a amostra atual é identificada como parte de uma sílaba se a maioria das últimas amostras também a conformam. O algoritmo 3 apresenta o E-IMF e a figura 4.10 ilustra o resultado de sua aplicação.

Algoritmo 3 Segmentador E-IMF

```
    função SEGMENTADOR(x_n, α, E_(n−1), S)
        E_n = E_(n−1) + α(x_n^2 − E_(n−1))
3:      se E_n ≥ T_E e mc < S então
            mc = mc + 1
        senão se mc > 0 então
6:          mc = mc − 1
        fim se
        se mc ≥ S/2 então
9:          ENVIAR(x_n)
        fim se
    fim função
```

Neste algoritmo, E_n é atualizada com cada nova amostra x_n (linha 2). Se o limiar T_E é ultrapassado, a memória do filtro (mc) é incrementada até o valor máximo S (linha 4). Isto significa que a variável de decisão é atualizada sempre que um possível valor de sinal é detectado. Neste caso, o filtro foi implementado com um único contador, para manter baixo os requerimentos de memória no sensor. No caso de novas amostras não serem detectadas como sinal, a memória do filtro é decrementada até o valor mínimo (linha 6). Finalmente, uma amostra é considerada parte de uma sílaba se $S/2$ amostras anteriores também foram. Assim, a resposta final enviada ao classificador, acontece depois da condição da linha 8.

[3]O filtro da moda considera o valor mais frequente observado, dentro de um intervalo, como o valor correto.

4.6.2 Técnica para avaliar a cascata Segmentação-Classificação

Como apontado anteriormente, as tarefas de segmentação e classificação podem ser tratadas como partes integrais do sistema de reconhecimento. Esta configuração remete ao funcionamento de um classificador multinível com a primeira camada não supervisionada, sendo a saída do segmentador (o primeiro classificador) a entrada para o reconhecedor de espécies que toma a decisão final.

Para ilustrar o funcionamento completo do sistema, supomos que uma sílaba é corretamente identificada pelo segmentador (tp_s do segmentador) e passada para o classificador. Este último pode reconhecê-la como: "é a espécie alvo" (tp_c), "não é a espécie alvo" (tn_c) ou cometer um dos dois tipos de erros possíveis (fp_c e fn_c). Assim, a taxa final de verdadeiros positivos neste exemplo, é obtida aplicando a equação 4.7. Da mesma forma, todas as combinações possíveis entre entrada e saída do sistema são consideradas nas equações:

$$tp_f = tp_s \cdot tp_c, \tag{4.7}$$

$$tn_f = tp_s \cdot tn_c + tn_s - tp_s \cdot tn_c \cdot tn_s, \tag{4.8}$$

$$fn_f = tp_s \cdot fn_c + fn_s - tp_s \cdot fn_c \cdot fn_s, \tag{4.9}$$

$$fp_f = tp_s \cdot fp_c + fp_s - tp_s \cdot fp_c \cdot fp_s, \tag{4.10}$$

em que os índices s, c e f indicam segmentador, classificador e resultado final, respectivamente.

Este sistema multinível possui algumas características especiais (figura 4.9). Quando o segmentador produz um tn_s ou um fn_s nada é passado para o classificador, evitando erros de classificação por culpa dos segmentos correspondentes aos ruídos ambientais (equação 4.8). A equação 4.9 quantifica a perda total do sistema, e não somente a perda do classificador (diferente da avaliação realizada por Jaafar et al. (2014)). Esta configuração em série do sistema também leva a uma simplificação do número de combinações possíveis entre segmentador e classificador.

No exemplo anterior, é claro que o resultado da classificação final depende da resposta do segmentador. No entanto, a resposta do segmentador é representada ponto-a-ponto do sinal, sendo que o classificador identifica segmentos completos. Consequentemente, existe uma consideração prática de implementação, na qual os valores

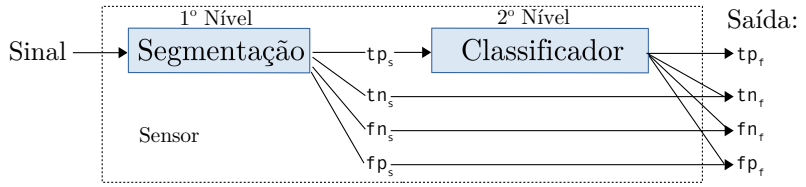


Figura 4.9. Interação entre segmentador e classificador. Apenas segmentos reconhecidos são enviados para o classificador espécies.

Tabela 4.7. Média dos resultados do E-I, E-IMF e E-WIN.

Método	TPR	FNR	MCC	Prec	Rec	F1	AEER
E-I	0,73	0,26	0,61	0,70	0,73	0,65	1,65
E-IMF	0,72	0,27	0,60	0,69	0,72	0,64	1,19
E-WIN, N=128	0,47	0,52	0,59	0,93	0,47	0,59	19,96
E-WIN, N=256	0,49	0,50	0,60	0,93	0,49	0,60	11,63
E-WIN, N=512	0,50	0,49	0,60	0,92	0,50	0,61	7,11
E-WIN, N=1024	0,51	0,48	0,60	0,90	0,51	0,60	5,78
E-WIN, N=2048	0,52	0,47	0,58	0,86	0,52	0,59	1,28
E-WIN, N=4096	0,52	0,47	0,53	0,73	0,52	0,55	1,13

tp, *tn*, *fn* e *fn*, sejam estes do segmentador ou do classificador, devem ser calculados como porcentagem do total de pontos. A partir das equações 4.7, 4.8, 4.9 e 4.10, é possível calcular as métricas tradicionais: precisão, revocação e F1 do sistema completo, sendo esta, uma contribuição nossa que permite avaliar quanto é o incremento, ou decremento, da taxa de reconhecimento das espécies causado pela ação do segmentador.

4.6.3 Metodologia experimental

Para avaliar o método de segmentação incremental, comparamos este com o resultado da segmentação manual e com a segmentação utilizando diferentes tamanhos de janelas deslisantes. Quantificando três tipos de erros: ponto-a-ponto, a distância entre as fronteiras reais e as estimadas (WinPR), e a taxa de erro dos eventos acústicos (AEER). Avaliamos também a acurácia de classificação das sílabas segmentadas e o impacto que a segmentação causa no sistema completo, utilizando a metodologia descrita na seção 4.6.2.

No método de janelas deslizantes (E-WIN) quanto maior é o tamanho de N, maior é a memória usada para se lembrar do passado. No método incremental, quanto menor for α, maior é o peso dos valores antigos. Em nossos experimentos, utilizamos os parâmetros $\alpha = 0.001$, $T_E = 0.02$, $S = 40$ e janelas de tamanhos N= $\{128, 256, 512, 1024, 2048\}$ com 50% de sobreposição (*overlap*). A tabela 4.7 apresenta os resultados comparando os métodos E-I, E-IMF e E-WIN.

Observando os valores AEER nesta tabela inferimos que aumentar o tamanho da

Tabela 4.8. Impacto da segmentação incremental no sistema de reconhecimento e comparado com janelas deslizantes de diferentes tamanhos. Requisitos aproximado de memória para cada técnica em Byte [B]

Métricas	GT	W_{128}	W_{256}	W_{512}	W_{1024}	W_{2048}	W_{4096}	E-I	E-IMF
Prec	0,99	0,90	0,90	0,89	0,89	**0,96**	0,94	0,92	0,91
Rec	0,98	0,46	0,47	0,48	0,48	0,52	0,50	**0,75**	0,73
F1	0,99	0,58	0,59	0,61	0,60	0,64	0,62	**0,81**	0,80
Memória	-	128 B	256 B	512 B	1024 B	2048 B	4096 B	2 B	3 B

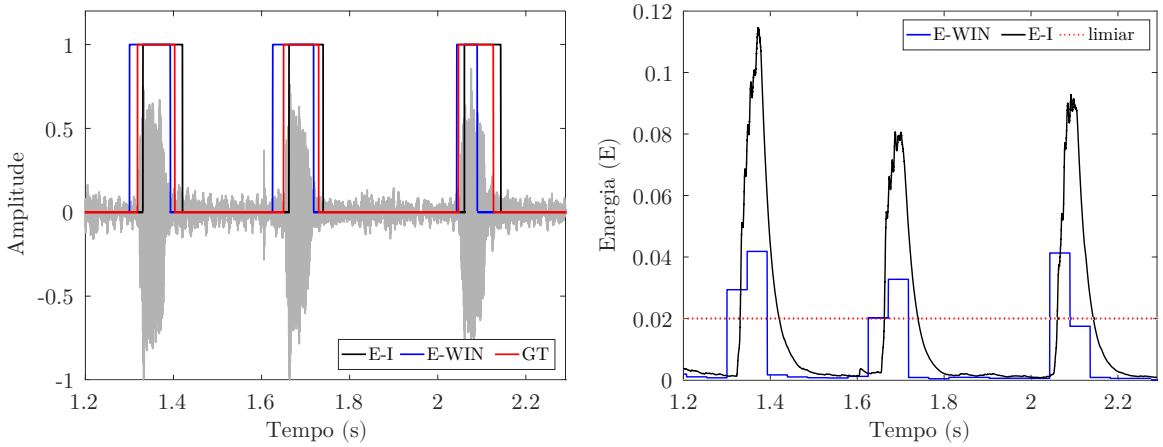
janela diminui as micro-segmentações[4], mas com custo de memória maior. Tamanhos de janela menores melhoram a precisão perdendo menos pontos das sílabas, mas diminuem a revocação e aumentam o AEER. Utilizar o filtro da moda para evitar os micro cortes na abordagem incremental melhorou o valor de AEER, no entanto, o atraso na detecção do começo e final das sílabas aumentou, causando uma diminuição de Prec, Rec e F1, e aumentando FNR, indicando assim, uma perda maior de valores das sílabas. Além disso, o método incremental possui taxas menores de erro, perdendo aproximadamente 25% dos pontos anotados manualmente, enquanto que utilizar janelas causa uma perda aproximada de 50%.

O TPR e FNR são indicadores de perda de sinal (ponto-a-ponto). De modo geral, quanto maior são Rec e TPR e menor FNR, melhor é o *trade-off* entre o número de pontos perdidos e transmissão de amostras desnecessárias. Os valores de FNR nesta tabela são consistentes com a revocação, indicando um aumento com o tamanho da janela e conseguindo capturar mais amostras no começo e no final das sílabas. A revocação das janelas menores significa que a duração da sílaba é muito maior do que a janela. Finalmente, o valor MCC do E-I indica uma correlação maior com a segmentação manual que os métodos restantes.

A tabela 4.8 mostra os resultados finais das três métricas Prec, Rec, e F1 do sistema completo (aplicando as equações da seção 4.6.2). Para cada técnica de segmentação, obtivemos valor final utilizando a média de cada espécie (*macro-average*). A última linha é a quantidade aproximada de memória necessária de cada método.

Nesta tabela a coluna GT representa o desempenho do classificador kNN para identificar os segmentos definidos pelo especialista, ou seja, sem *fn* nem *fn*. A diferença entre as colunas W_{2048} e W_{4096} mostra que aumentar o tamanho da janela também causa perda de pontos, i.e. a métrica Rec diminui quando a janela aumenta de 2048 para 4096. Os melhores valores de cada linha são destacados em negrito. Assim, podemos observar que a técnica incremental superou o F1 das janelas. Com estas comparações conseguimos identificar também que a melhor janela é W_{2048}.

[4]Consideramos "micro" cortes das sílabas quando estas são segmentadas em vários fragmentos menores.



(a) Segmentação manual (GT), incremental (E-I) e utilizando janelas deslizante (E-WIN$_{2048}$).

(b) Valores da energia do sinal calculado em cada janela (E-WIN) e de forma incremental (E-I), mais o limiar de decisão.

Figura 4.10. Exemplo de segmentação da espécie Adenomera hylaedactyla.

A figura 4.10(a) exemplifica o resultado da segmentação manual (GT), utilizando janelas de 2048 pontos, e da segmentação incremental da vocalização da espécie Adenomera hylaedactyla. Observamos que é necessário um tempo maior para detectar o final de cada sílaba utilizando a segmentação incremental. No entanto, comparado com GT os limites são mais próximos do real do que utilizando janelas. Em outras palavras, utilizar janelas causa um deslocamento dos limites de detecção para à esquerda, enquanto que a abordagem incremental causa um deslocamento a direita. A figura 4.10(b) mostra os valores da energia do sinal para as janelas e para nossa abordagem incremental.

A abordagem de cálculo incremental apresentada tem o objetivo de diminuir a utilização de recursos dos sensores. Mostramos aqui que o método incremental é eficaz na detecção de segmentos de áudio de interesse, quando comparado com as janelas deslizantes. Assim, ao segmentar o sinal antes da transmissão, o método reduz a quantidade de dados enviados pelos sensores, e como a transmissão é a tarefa com custo de energia mais significativo, essa redução dos dados aumenta a vida útil da rede. Além disso, o cálculo incremental é adequado para um contexto de *big data*, fornecendo uma resposta em tempo real (Gama and Gaber, 2007).

4.6.4 Avaliação do sistema completo

Para quantificar o impacto que a segmentação automática causa na taxa de reconhecimento das espécies, aplicamos a metodologia de avaliação descrita na seção 4.6.2. Nesta prova utilizamos as sílabas resultantes da segmentação manual para criar a base

Tabela 4.9. Resultados do sistema multinível (segmentação mais classificação). Segunda coluna (GT) segmentação manual, terceira coluna (BN) segmentação automática com ruídos ambientais, e quarta até decima coluna (dB) segmentação automática adicionando ruído branco com diferentes níveis de variância.

			Ruído branco						
	GT	BN	10 dB	7 dB	5 dB	3 dB	0 dB	-1,7 dB	-3 dB
Prec	0,98	0,96	0,72	0,57	0,31	0,32	0,17	0,10	0,10
Rec	0,96	0,55	0,31	0,22	0,18	0,11	0,09	0,08	0,08
F1	0,96	0,70	0,43	0,30	0,23	0,17	0,12	0,09	0,09

de treino do classificador que irá reconhecer a saída do segmentador. Representamos cada sílaba por um conjunto de doze coeficientes Mel obtidos a partir de um banco com 24 filtros (seção 2.2.2, página 21). O classificador utilizado foi o kNN com $k = 1$. Escolhemos esta configuração com base no estudo de Colonna et al. (2012).

Nesta configuração de sistema multinível, os *frames* contíguos, reconhecidos pelo segmentador como sinal, são agrupados em um segmento completo representando a sílaba. Este segmento passa pela extração de características (12 MFCCs) e entra no classificador para ser reconhecido. A tabela 4.9 apresenta os resultados do reconhecimento utilizando como única característica de segmentação a energia do sinal (E). Aplicamos as equações 4.7, 4.8, 4.9 e 4.10 para calcular a precisão, a revocação e F-Score do sistema multinível.

Da tabela 4.9, observamos que embora a segmentação GT seja realizada por um especialista, o classificador que encontra-se no seguinte passo não é capaz de reconhecer 100% das sílabas. Comparando a segunda e terceira coluna desta tabela observamos que a segmentação automática causa a perda de algumas sílabas se compararmos com a segmentação manual. Isto reflete-se na diminuição da revocação, e consequentemente, do F1. Vemos também, que a precisão sofreu pouco impacto, concluindo que as sílabas segmentadas foram bem reconhecidas pelo classificador. Em outras palavras, o segmentador escolheu as melhores sílabas do áudio, sem recortar estas, de forma que prejudique o reconhecimento.

Entre a quarta e décima coluna encontram-se os resultados de adicionar ruído branco aos áudios com diferentes níveis de variância antes de entrar no sistema. Assim, comprovamos que o aumento dos ruídos afeta mais o classificador do que o segmentador. Isto é devido as variações que os ruídos causam nos espectros de frequências, que consequentemente, alteram os valores dos MFCCss. Entre a nona e a décima coluna, o F1 não variou, indicando que o classificador chegou em um ponto de saturação a partir do qual não é mais capaz de reconhecer as sílabas, razão pela qual uma etapa de filtragem é justificada.

4.6.5 Conclusões sobre a segmentação incremental

A estratégia de segmentação incremental apresentada nesta seção atingiu os requerimentos de desempenho, em termos de consumo de memória e tempo computacional, e ao mesmo tempo manteve a qualidade da classificação das sílabas. No entanto, precisamos explorar futuramente, em mais detalhes, a relação entre a qualidade do áudio (considerando a frequência de amostragem) e os parâmetros α e N. A transformação do método de segmentação em incremental eliminou a possibilidade de utilizar o procedimento descrito no algoritmo 2 para encontrar os limiares ótimos. A adaptação deste algoritmo ao caso incremental constitui também uma extensão futura.

Diferentes valores do parâmetro α podem causar mais, ou menos, atraso na detecção do início e fim da sílaba. Consequentemente, o classificador deve ser capaz de reconhecer a espécie utilizando somente uma fração das sílabas. Como foi mostrado na figura 4.10(a), utilizar janelas deslizantes pode causar o mesmo problema. Variar a quantidade de pontos do sinal utilizados para extrair os coeficientes Mel, altera a escala destes coeficientes e dificultam o casamento dos padrões pelo classificador. Uma solução utilizada aqui foi normalizar os coeficientes no intervalo $[-1, 1]$. Além deste problema, a precisão na detecção do ponto final da sílaba pode causar que segmentos maiores do que a memória do sensor consegue processar, sendo este um problema geral das abordagens de segmentação.

4.7 Considerações finais sobre a segmentação

Neste capítulo, apresentamos o problema de detectar mudanças nos sons bioacústicos que identifiquem o começo e o fim de cada sílaba dentro de uma vocalização de anuros. O propósito fundamental da segmentação é separar trechos dos áudios que contenham somente ruídos ambientais, para diminuir a quantidade de dados: transmitidos, armazenados, processados e/ou classificados pelos sensores. Propomos resolver este problema utilizando descritores acústicos que permitam identificar cada *frame* do sinal de forma não supervisionada. Assim, a segmentação foi realizada verificando o tempo específico no qual muda a distribuição dos valores de cada descritor.

Na primeira parte deste capítulo, realizamos um estudo comparativo entre diferentes descritores acústicos do domínio temporal e espectral. Comparamos estes individualmente e combinados em pares. Além disso, utilizamos a entropia das permutações e suas variações como descritor para identificar segmentos dos sinais com comportamentos aleatórios similares aos ruídos. Avaliamos a tolerância da técnica de segmentação aos diferentes tipos e níveis de ruídos, sejam estes branco ou coloridos.

Incluímos também um procedimento automático para encontrar os limiares de corte ótimos para separar os *frames*.

Para avaliar nossos experimentos, montamos uma base de dados com vocalizações de quinze espécies e realizamos as anotações manualmente de 3155 sílabas. O fato de não encontrar uma base pública para este propósito dificulta o estudo comparativo com outras abordagens. No entanto, nós incluímos a combinação de descritores mais frequentes encontrados na literatura (E e ZCR). Finalmente, com as espécies presentes em nossa base de dados, mostramos que o melhor descritor para realizar a segmentação é a energia do sinal, considerando a relação entre complexidade de cálculo e erro de segmentação.

Como apontado anteriormente, as tarefas de segmentação e classificação foram tratados como um classificador multinível com camadas não supervisionadas. Nesta configuração, a saída do segmentador (o primeiro classificador) é a entrada para o reconhecedor de espécies (o segundo classificador), que toma a decisão final. Mostramos através de nossos resultados que a segmentação possui impacto direto na taxa de reconhecimento. Assim, utilizar simplesmente a taxa de acerto do classificador para avaliar o ACR completo pode gerar um resultado enganoso. Para isso, desenvolvemos uma forma de avaliação com quatro equações que quantificam o resultado final, considerando todas as possibilidades de interação entre segmentador e classificador.

Nossa última contribuição é uma adaptação do método de segmentação utilizando janelas deslizantes de forma incremental. Comparamos esta nova abordagem com a clássica utilizando janelas deslizantes de diferentes tamanhos. Com a metodologia aplica, conseguimos identificar o melhor tamanho de janela, que tem relação com a memória mínima que o hardware precisa para processar o áudio.

Os resultados deste capítulo reforçam as evidências que as atenuações e os ruídos são as causas principais de erro no reconhecimento das espécies, como foi mencionado na introdução (figura 1.3). Por este motivo, consideramos necessário desenvolver uma técnica de filtro que seja capaz de lidar com sinais bioacústicos e cenários dinâmicos, tais como a floresta amazônica. O filtro de ruídos é o motivo do Capítulo 6. As atenuações são consequência da distância entre o indivíduo e o microfone do sensor. Por isso, no Capítulo 5.3 propomos uma abordagem colaborativa, que utiliza a opinião dos sensores próximos para detectar o evento acústico e melhorar a taxa de reconhecimento.

5

Método de Classificação

Neste capítulo apresentamos a etapa de classificação do nosso ACR de monitoramento bioacústico. Após a segmentação dos sinais em sílabas, estas devem ser utilizadas para identificar as espécies presentes. Para alcançar este objetivo treinamos, comparamos e avaliamos diferentes técnicas de aprendizagem de máquina. Como parte de nossa proposta, apresentamos primeiro, um método de validação cruzada por indivíduos para evitar o viés na classificação, que se produz ao utilizar uma abordagem baseada em sílabas, e nos permite interpretar a capacidade de generalização do modelo. A seguir, investigamos uma nova abordagem de classificação *multi-output* (*multi-class* e *multi-label*) que permite identificar a taxonomia da espécies (família e gênero), e utilizar estas informações para ganhar conhecimento sobre a estrutura das amostras no espaço de características acústicas. Para isso, criamos e comparamos duas estruturas hierárquicas de classificadores do tipo *top-down*, respeitando a taxonomia biológica filogenética, utilizando: 1) um classificador por nó pai e 2) um classificador por nível. Finalmente, apresentamos e avaliamos uma abordagem apropriada para fusão de classificações combinando sensores acústicos que pertencem ao mesmo *cluster* de uma rede RSSF.

5.1 Introdução ao método de validação proposto

Atualmente, o método mais utilizado por especialistas para monitorar populações de anuros é a aplicação de *surveys* acústicos (Marques et al., 2013, Silva, 2010). A aplicação de tais *surveys* é uma tarefa manual que requer um especialista treinado para reconhecer as espécies presentes no lugar monitorado pelas vocalizações ouvidas durante um período de tempo. Neste tipo abordagem, é considerada a presença ou ausência da espécie como uma variável binária, a qual é utilizada posteriormente para realizar as estatísticas populacionais, sem levar em consideração a quantidade de indivíduos ou o tipo e a finalidade de cada vocalização.

Para viabilizar a aplicação de *surveys* acústicos muitos recursos humanos e econômicos são necessários, bem como conhecimento especializado, sendo difícil a aplicação em áreas tropicais remotas tais como a floresta amazônica durante longos períodos de tempo. Portanto, nosso objetivo é desenvolver um sistema de Reconhecimento Automático de Chamadas (ACR) para monitorar populações de anuros detectando de forma binária a "presença" ou "ausência" deles, de forma menos invasiva por meio de sensores acústicos.

A detecção da presença de uma determinada espécie dentro da área de alcance do nó sensor é dada pela identificação dos eventos acústicos classificados em um intervalo de tempo. Em nosso caso, os eventos são as sílabas das vocalizações. Em contrapartida, a ausência da espécie é a "não detecção" de suas sílabas. Assim, uma elevada taxa de falsos negativos no reconhecimento pode levar as conclusões erradas. Portanto, faz-se necessário desenvolver uma técnica de avaliação que permita quantificar o erro de generalização do método para situações reais.

No capítulo 3 foram apresentamos trabalhos relacionados com a segmentação e o pré-processamento, trabalhos de análise e seleção de descritores acústicos para a classificação e, finalmente, trabalhos que comparam diferentes algoritmos ML de reconhecimento. O conjunto de todos esses trabalhos é um exemplo da adaptabilidade do *framework* aos diferentes objetivos. A maioria dos sistemas apresentados baseiam-se em abordagens de reconhecimento de sílabas utilizando validação cruzada (k-CV) para estimar o desempenho do classificador e as capacidades de generalização do modelo. O procedimento k-CV tradicional divide o conjunto total em k subconjuntos separados para treino e teste, mas durante a separação é ignorado o fato de que sílabas (ou amostras) do mesmo indivíduo podem estar presentes em mais de um dos k conjuntos.

No decorrer de nossos experimentos percebemos que, quando sílabas de um espécimen em particular encontram-se ao mesmo tempo em mais de um dos k subconjuntos, aparece um viés nos resultados. Quando isso acontece, a precisão do classificador é

superestimada. Nossa nova proposta de validação cruzada incorpora o rótulo dos indivíduos com informação adicional para ser utilizado durante a divisão k-CV, evitando misturar sílabas do mesmo espécimen nos conjuntos de treino e teste, e assim, reduzir o viés nos resultados.

5.1.1 O problema da validação cruzada tradicional nos sistemas bioacústicos

Uma vez que os áudios foram segmentados e as silabas extraídas, procedimentos padrões de validação cruzada podem ser aplicados ao conjunto de amostras para estimar o erro do modelo de classificação. Entretanto, quando sílabas do mesmo indivíduo estão presentes, tanto no conjunto de treino quanto no conjunto de teste, é produzida uma sobrestimação da acurácia do reconhecimento. Este problema é agravado ainda mais no caso extremo quando se utiliza *Leave-one-Out* CV, no qual somente uma sílaba é utilizada para teste, deixando as restantes amostras do conjunto para treinamento. Por exemplo, os trabalhos apresentados por Huang et al. (2009) e Colonna et al. (2012) são exemplos deste problema, no qual podem ser observados resultados próximos ao 100% de reconhecimento para algumas espécies.

Nós chamamos à sobrestimação no reconhecimento bioacústico de "problema de generalização", originado pela seguinte questão de pesquisa: Dada uma função de classificação $f(\cdot)$, treinada a partir de um conjunto de sílabas pertencentes a k indivíduos, seria possível reconhecer as sílabas de m novos indivíduos não presentes no conjunto de treinamento original? Isto é, uma vez que o modelo for embarcado no nó sensor e posicionado na floresta, queremos saber se será possível identificar novos indivíduos da mesma espécie.

O problema da similaridade entre as sílabas pertencentes ao mesmo indivíduo é ilustrado na figura 5.1. Nesta figura pode-se observar um sinal de áudio (domínio temporal) com três sílabas de um único espécimen da espécie Adenomera hylaedactyla. Visualmente estas sílabas parecem ser diferentes, mas no domínio dos descritores acústicos suas diferenças não são realmente perceptíveis. Esta observação é confirmada pelos valores dos MFCCss apresentados na tabela 5.1. A última linha desta tabela resume a média e o desvio padrão (std) de cada coluna. Aqui, a baixa variância dos MFCCs é um indicativo da elevada similaridade entre as sílabas.

Se neste exemplo aplicarmos um classificador baseado na distância euclidiana entre os vetores, por exemplo kNN, e separarmos a primeira sílaba para teste e as duas sílabas restantes para treinamento, conseguiríamos um similaridade igual a 0,055 entre a primeira e segunda sílaba, e de 0,049 entre a primeira e a terceira sílaba. Consequen-

Tabela 5.1. MFCCs extraídos das três sílabas da figura 5.1.

	MFCCs									
sílaba$_1$	0,00	0,14	0,42	0,69	0,81	0,90	0,93	0,97	1,00	0,91
sílaba$_2$	0,00	0,13	0,42	0,72	0,84	0,93	0,94	0,96	1,00	0,92
sílaba$_3$	0,00	0,14	0,44	0,72	0,83	0,91	0,92	0,95	1,00	0,92
Média	0,00	0,14	0,42	0,71	0,83	0,91	0,93	0,96	1,00	0,92
std	0,000	0,005	0,010	0,018	0,017	0,014	0,012	0,008	0,000	0,003

temente, esta situação é susceptível de resultar em uma alta taxa de reconhecimento. No entanto, somente conseguiríamos identificar o próprio indivíduo. Tecnicamente, esta situação não é considerada *overfiting*, mas sim como uma situação que apresenta viés.

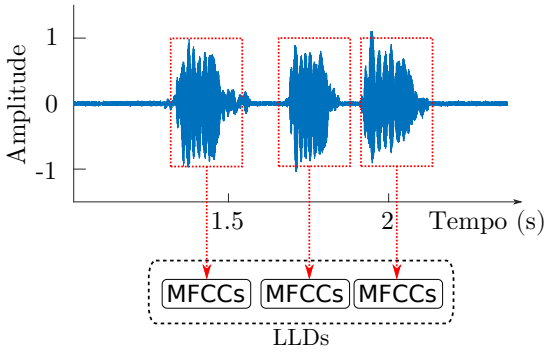


Figura 5.1. Trecho de uma gravação com três sílabas da espécie Adenomera hylaedactyla.

Este exemplo ilustrativo ajuda a entender por que alguns trabalhos relacionados alcançam quase 100% de precisão. Portanto, podemos levantar a hipótese seguinte: aprender os parâmetros da função de classificação e testá-la em sílabas provenientes do mesmo indivíduo é um equívoco metodológico que aumenta artificialmente a acurácia do modelo. Como foi ilustrado neste exemplo, o modelo teria uma pontuação quase perfeita repetindo os rótulos das amostras, mas deixaria de prever sílabas de novos espécimens, devido à baixa generalização da função de classificação treinada.

O problema da CV tradicional é que parte dos exemplos disponíveis são separados para testes e outra parte para treinamento, mas no contexto bioacústico devemos evitar gerar uma divisão aleatória (ou estratifica) contendo sílabas do mesmo indivíduo nos dois subconjuntos (teste e treino). Portanto, o procedimento clássico k-CV não é adequado para este contexto de aplicação.

5.1.2 Validação cruzada por indivíduos

Uma prática padronizada para estimar o error esperado dos modelos de classificação, em uma situação real, é treinar e testar utilizando CV. Com k-CV o conjunto de dados original é dividido em k subconjuntos disjuntos e para cada um desses o erro condicional (e_k) é estimado treinando o modelo $f(\cdot)$ com $k-1$ subconjuntos. Este procedimento é repetido k vezes, Assim, o erro de generalização esperado pode ser obtido como a média de e_k. Como mencionado anteriormente, podemos esperar que este procedimento aproxime o erro real do sistema, mas quando a informação dos indivíduos é omitida cairemos numa situação em que a divisão poderia deixar sílabas do mesmo espécimen nos conjuntos de treino e teste.

Para resolver este problema, propomos uma avaliação mais justa considerando a identificação dos espécimens durante a divisão k-CV, deixando no mesmo subconjunto todas as sílabas que pertencem ao mesmo indivíduo, evitando desta forma, misturá-las nos subconjuntos. Em seguida, propomos utilizar ***Leave-one-Out* por indivíduos** (ou gravações em nosso casso específico), para medir o desempenho dos algoritmos de classificação, ou seja, fazendo k igual ao número de espécimens diferentes. Assim, um indivíduo é separado para testes e os restantes para treino. Estas divisões são repetidas até que cada indivíduo é avaliado no subconjunto de teste. Desta forma, para cada iteração as classificações das sílaba são armazenadas. Após a conclusão do LOOCV, as precisão Micro- e Macro-métricas são calculadas utilizando a matriz de confusão.

Devido a que estamos lidando com um problema supervisionado e precisamos identificar cada indivíduo durante a avaliação, agora cada sílaba deve estar associada a dois rótulos, um para identificar o espécimen id_k e outro para identificar a espécie (s_i). Um exemplo do novo conjunto de dados é:

$$\text{Base de dados} = \begin{bmatrix} \mathbf{c}_1 = [c_1, c_2, \ldots, c_l], & s_1, & \text{id}_1 \\ \mathbf{c}_2 = [c_1, c_2, \ldots, c_l], & s_1, & \text{id}_1 \\ \vdots & \vdots & \vdots \\ \mathbf{c}_d = [c_1, c_2, \ldots, c_l], & s_i, & \text{id}_k \end{bmatrix}$$

Note-se que neste exemplo as duas primeiras sílabas pertencem ao mesmo indivíduo, isto é devido que em cada gravação diversas sílabas podem ser encontradas por cada indivíduo. Finalmente, assumimos que o erro de generalização aplicando CV por indivíduos é mais realista, pelo fato de se treinar com um espécimen para prever outro diferente. A tabela 5.2 apresenta a quantidade de sílabas totais e individuais em relação aos espécimens da nossa base de dados.

Esta nova forma de validação cruzada apresenta duas particularidades principais.

Primeiro, é necessário ter representada cada espécie no conjunto de treinamento com pelo menos dois indivíduos. Em segundo lugar, o número de instâncias de treino e teste em cada subconjunto k pode estar desbalanceado.

5.1.3 O problema do desbalanceamento das classes

Cada espécie de anuros possui sua própria taxa de sílabas (diferente quantidade de sílabas por unidade de tempo) em suas chamadas. Esta é uma característica específica das vocalizações de cada espécie, fato que causa que um número desigual de amostras seja recuperado de cada indivíduo (Colonna et al., 2015). Neste contexto, um modelo de classificação que sempre decide em favor da espécie mais numerosa poderia ter uma alta precisão, inclusive perdendo todas as sílabas das outras classes. Para evitar este viés nas métricas, sugerimos utilizar Macro-métricas em vez das tradicionais Micro-métricas (seção 2.6.2, página 58). Isso significa que o valor final das métricas é obtido calculado a média da métricas de cada espécie individualmente (Colonna et al., 2015, 2016a, Sokolova and Lapalme, 2009). As Macros-métricas foram introduzidas na seção 2.6.1. Embora, este problema seja secundário na maioria das aplicações, nos sistemas bioacústicos deve ser considerado para realizar uma avaliação justa dos métodos.

5.1.4 Simplificação dos modelos multiclasses

Nos sistemas de reconhecimento bioacústicos o número de classes dos modelos de classificação é igual ao número de espécies que se desejam reconhecer. Portanto, a complexidade das funções de classificação aumenta proporcionalmente ao número de espécies. No entanto, os princípios de aprendizagem de máquina sugerem que sempre devem-se preferir funções de classificação simples, uma vez que estas generalizam melhor o conhecimento aprendido. Por outro lado, existem métodos de classificação que são naturalmente binários (e.g., SVM) e portanto, devem ser adaptados a contextos multiclasse mediante a adoção de alguma estratégia de decomposição do problema. Assim, para proporcionar uma comparação justa entre os métodos de classificação adotados e reduzir a complexidade das funções, nos experimentos apresentados a seguir utilizamos e comparamos duas estratégias de decomposição binárias: um-contra-um (*One-against-One* - 1A1) e um-contra-todos (*One-against-All* - 1AA) (Fürnkranz, 2001), como apresentadas na seção 2.5.5 (página 53).

Particularmente, no domínio de aplicação das RSSF, resolver o problema de reconhecimento mediante um conjunto de classificadores binários apresenta duas vantagens:

(1) cada sensor pode estar especializado no reconhecimento de uma única espécie, sujeito desta forma a uma probabilidade menor de erro, e (2) a decisão final do comitê de sensores pode ser realizada no nó líder mediante uma operação de fusão de dados. Desta forma, a classificação resulta numa colaboração entre os diferentes sensores. Uma abordagem de classificação colaborativa é apresentada na seção 5.3.4.

No caso de avaliação utilizando LOOCV por indivíduos, como foi introduzido na seção anterior, o procedimento 1AA começa separando todas as sílabas do primeiro espécimen no conjunto de teste e agrupando as sílabas restantes da mesma espécie no conjunto de treinamento da classe alvo ("+1"). Todas as sílabas das espécies restantes, que não pertencem à espécie alvo, são agrupadas na classe de treinamento negativa ("-1"). Assim, o problema de reconhecer a espécie alvo resulta num problema de decisão binária. Em seguida, o modelo $f(\cdot)$ é treinado e aplicado para estimar os rótulos do grupo de teste.

Na segunda iteração este procedimento é repetido, mas separando todas as sílabas do segundo espécimen para teste e reincorporando as sílabas que foram separadas na iteração anterior ao conjunto de treino. A validação é repetida até que todos os espécimes no conjunto de dados sejam avaliados. O resultado final é obtido aplicando votação majoritária. Com o 1AA o número de problemas binários que devem ser avaliados é igual ao número de classes m, e ao se combinar com k-CV por indivíduos resulta em $r = m \times k$ iterações, onde k representa o total de indivíduos (Colonna et al., 2016a). No entanto, a complexidade do modelo avaliado $f(\cdot)$ a cada iteração é menor do que numa configuração multi-classe tradicional (figura 5.2).

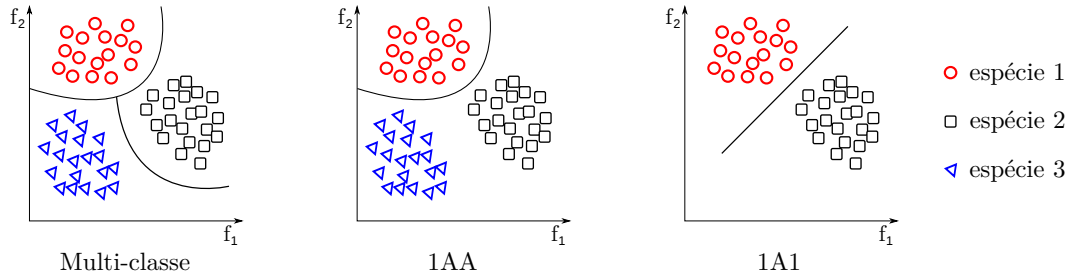


Figura 5.2. Exemplos de decomposição de problemas multi-classes.

O segundo procedimento de decomposição binária aplicado é o 1A1. Este procedimento separa o problema original em problemas ainda menores do que o 1AA. A estimativa do rótulo procede de forma semelhante ao 1AA, mas com a principal diferença de que existe agora uma classe negativa para cada espécie sem ser a espécie alvo. Em outras palavras, a classe negativa no caso anterior 1AA é decomposta em

$m-1$ classes negativas menores, uma para cada espécie (figura 5.2). Depois disso, o resultado de cada sub-problema é combinado usando a regra de votação majoritária.

Tipicamente, com o 1A1 tem-se $^{m(m-1)}/_{2}$ iterações, e ao se combinar com k-CV por indivíduos o número total de iterações torna-se $^{m(m-1)}/_{2} \times k$ (Colonna et al., 2016a). No entanto, esta decomposição reduz a complexidade de cada sub-problema, em comparação com a abordagem multi-classe e a 1AA (Fürnkranz, 2001). Neste caso, cada nó sensor da rede poderia incluir um modelo binário para diferenciar unicamente duas espécies de interesse, e novamente, as votações teriam lugar no nó líder do *cluster*. Entretanto, o número de sensores necessários para monitorar uma área cresce em função da quantidade espécies monitoradas de acordo com $^{m(m-1)}/_{2}$.

5.1.5 Metodologia experimental e resultados

A base de dados utilizadas nestes experimentos de classificação (tabela 5.2) inclui dez espécies diferentes com um total de 60 indivíduos, e após a segmentação obtiveram-se 7195 sílabas (instâncias de treino e teste). Essas amostras foram colhidas *in situ* em condições reais de ruído ambiental[1]. Algumas destas espécies foram gravadas no campus da Universidade Federal do Amazonas*, outras são gravações realizadas na região da Mata Atlântica†, e a última espécie foi obtida em Córdoba‡, Argentina. Estas gravações foram armazenadas em formato *.wav* sem compressão para evitar perdas de qualidade, com frequência de amostragem de 44,1 kHz e 32 bits de resolução, o que nos permite analisar sinais de até 22,05 kHz. A partir de cada sílaba foram extraídos 24 MFCCss utilizando 44 filtros triangulares, introduzidos na seção 2.2.2. Estes vetores de coeficientes foram normalizados para amplitude máxima $-1 \leq c_l \leq 1$.

Para realizar a extração das sílabas baseamos nossa abordagem de segmentação no método por *frames* apresentado no capítulo 4. Assim, aplicamos a entropia espectral H_f em *frames* com 0,0464 s de duração e 66% de sobreposição para obter uma ótima resolução tempo-frequência. Foi adicionada uma variável aleatória Gaussiana com variância igual ao 10% da variância do sinal original para quebrar as correlações fracas sem causar distorções graves no sinal como foi mostrado na seção 5.1.5. A figura 5.3 apresenta um exemplo do resultado desta segmentação. A escolha de se utilizar a H_f foi baseada no fato que para obtermos esta e também os MFCCss para representar as sílabas, deve-se realizar a Transformada de Fourier, e portanto, um único cálculo da FFT seria compartilhado pelas etapas de segmentação e extração dos LLDs, diminuindo consequentemente o número de operações realizadas no nó sensor.

[1] Uma cópia da nossa base de gravações encontra-se pública na internet, no link `https://goo.gl/aZRhPJ`.

Tabela 5.2. Base de dados. Os índices s e k representam o número de sílabas e o número de indivíduos (espécimens) respectivamente.

Família	Gênero	Espécies	s	k
Leptodactylidae	Leptodactylus	Leptodactylus fuscus*	4	270
	Adenomera	Adenomera andreae*	8	672
		Adenomera hylaedactyla†	11	3478
Hylidae	Dendropsophus	Hyla minuta†	11	310
	Scinax	Scinax ruber†	5	148
	Osteocephalus	Osteocephalus oophagus*	3	114
	Hypsiboas	Hypsiboas cinerascens*	4	472
		Hypsiboas cordobae‡	4	1121
Bufonidae	Rhinella	Rhinella granulosa*	5	68
Dendrobatidae	Ameerega	Ameerega trivittata†	5	542

*Amazonas, †Mata Atlântica, ‡Córdoba

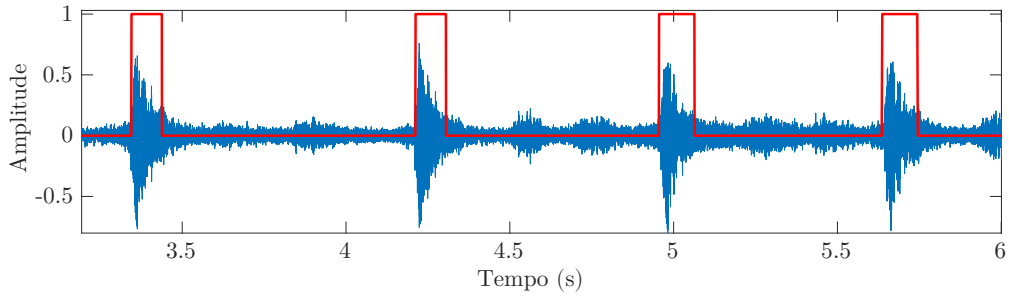


Figura 5.3. Exemplo de segmentação e extração de sílabas aplicando H_f e o critério dado pelo algoritmo 2.

Uma vez que a base de dados foi representada por seus LLDs, nós comparamos o desempenho de cinco métodos de classificação utilizando as decomposições 1AA e 1A1. Utilizamos: kNN com $k = 3$, SVM com kernel RBF (RBF-SVM), SVM com kernel polinomial de grau 3 (Poly-SVM), Árvore de Decisão (*Decision Tree* - DT) e QDA[2]. Os parâmetros do RBF-SVM foram automaticamente configurados pelos software de cálculo numérico MATLAB. Os resultados da precisão e revocação separados por espécies encontram-se nas tabelas 5.3 e 5.4. Nas duas últimas linhas destas tabelas são resumidas as Macro-métricas. O teste estatístico t-test foi aplicado para comparar os resultados entre as colunas da precisão e entre as colunas da revocação dos diferentes métodos. Os melhores valores de precisão e revocação foram escolhidos para as comparações, assim, os valores em negrito foram determinados como empate com nível de confiança $p = 0,05$.

Comparando os resultados das tabelas 5.3 e 5.4 observamos que a abordagem

[2]Uma comparação mais extensiva entre estes classificadores e seus parâmetros foram publicadas em (Colonna et al., 2016a).

Tabela 5.3. Decomposição um-contra-todos (1AA).

	kNN		RBF-SVM		Poly-SVM		DT		QDA	
Species	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
Adenomera a.	0,74	0,37	0,82	0,24	0,61	0,35	0,40	0,36	0,42	0,37
Adenomera h.	0,99	0,98	0,99	0,96	0,99	0,99	0,97	0,94	0,98	0,99
Ameerega t.	0,85	0,61	0,95	0,34	0,74	0,57	0,62	0,40	0,82	0,22
Hyla m.	0,66	0,73	0,87	0,43	0,70	0,78	0,25	0,47	0,48	0,80
Hypsiboas cin.	0,86	0,85	0,67	0,59	0,72	0,89	0,81	0,72	0,64	0,91
Hypsiboas cor.	0,90	0,96	0,38	0,89	0,94	0,96	0,80	0,89	0,92	0,91
Leptodactylus f.	0,53	0,78	0,95	0,40	0,57	0,80	0,43	0,39	0,41	0,47
Osteocephalus o.	0,37	0,46	0,97	0,28	0,70	0,60	0,02	0,03	0,09	0,21
Rhinella g.	0,16	0,83	1,00	0,77	0,25	0,83	0,32	0,60	1,00	0,16
Scinax r.	0,84	0,61	0,91	0,52	0,91	0,84	0,32	0,19	0,98	0,65
Médias	**0,69**	**0,72**	**0,85**	0,54	**0,71**	**0,76**	0,49	0,50	0,67	0,57
Macro-F1	0,70		0,66		0,74		0,50		0,62	
Micro-F1	0,85		0,74		0,86		0,75		0,80	

Tabela 5.4. Decomposição um-contra-um (1A1).

	kNN		RBF-SVM		Poly-SVM		DT		QDA	
Species	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
Adenomera a.	0,73	0,37	0,87	0,20	0,47	0,35	0,47	0,33	0,50	0,34
Adenomera h.	0,99	0,98	1,00	0,74	0,99	0,99	0,97	0,98	0,99	0,99
Ameerega t.	0,83	0,61	1,00	0,19	0,85	0,44	0,64	0,30	0,88	0,26
Hyla m.	0,67	0,73	0,51	0,49	0,70	0,80	0,36	0,56	0,47	0,93
Hypsiboas cin.	0,84	0,87	0,12	0,75	0,91	0,89	0,61	0,76	0,59	0,87
Hypsiboas cor.	0,90	0,96	0,86	0,82	0,89	0,96	0,84	0,94	0,92	0,92
Leptodactylus f.	0,43	0,78	1,00	0,12	0,75	0,77	0,74	0,63	0,43	0,42
Osteocephalus o.	0,29	0,43	1,00	0,19	0,47	0,37	0,10	0,14	0,08	0,23
Rhinella g.	0,27	0,83	0,98	0,83	0,13	0,82	0,33	0,76	1,00	0,16
Scinax r.	0,84	0,62	1,00	0,17	0,97	0,68	0,85	0,49	1,00	0,64
Médias	**0,68**	**0,72**	**0,83**	0,45	**0,71**	**0,71**	**0,59**	**0,59**	**0,69**	**0,58**
Macro-F1	0,70		0,48		0,71		0,59		0,56	
Micro-F1	0,85		0,61		0,84		0,79		0,80	

1AA foi superior à 1A1 no F1 em três métodos de classificação (QDA, RBF-SVM e Poly-SVM) e produz um empate (kNN). Percebe-se também que entre as colunas da tabela 5.3 existiram menos situações de empate do que na tabela 5.4 identificadas pelo teste estatístico, o que sugere que a abordagem 1A1 beneficiou os métodos DT e QDA. No que diz respeito aos resultados individuais para cada espécie, observamos que as espécies Rhinella g. e Osteocephalus o. foram as mais difíceis de identificar pela maioria dos métodos, portanto deveriam ser evitadas durante a aplicação de *surveys* acústicos automáticos para monitoramento em determinadas áreas. Além disso, podemos notar que existe uma variabilidade marcada nos valores de precisão e revocação para cada espécie entre os diferentes métodos, por exemplo, a espécie Hypsiboas cor. obteve uma precisão de 90% com kNN, 38% com RBF-SVM e 94% com Poly-SVM, o que indica que futuramente poderia se aplicar uma técnica de *ensemble learning* para melhorar estes resultados.

A tabla 5.5 apresenta a matriz de confusão para o caso kNN com 1AA. Inspecionando esta matriz descobrimos que o maior número de confusões aconteceram entre

as espécies Adenomera andreae e Leptodactylus fuscus, Osteocephalus oophagus, Rhinella granulosa; e entre Ameerega trivittata e Rhinella granulosa. Estas confusões causaram uma redução considerável na precisão do reconhecimento das espécies Osteocephalus oophagus e Rhinella granulosa, e na revocação da espécie Adenomera andreae.

Tabela 5.5. Matriz de confusão utilizando kNN com 1AA e validação cruzada por indivíduos. Rótulos: (a) Adenomera andreae, (b) Adenomera hylaedactyla, (c) Ameerega trivittata, (d) Hyla minuta, (e) Hypsiboas cinerascens, (f) Hypsiboas cordobae, (g) Leptodactylus fuscus, (h) Osteocephalus oophagus, (i) Rhinella granulosa, and (j) Scinax ruber. Prec_i e Rec_i representam a precisão e a revocação respectivamente.

	a	b	c	d	e	f	g	h	i	j	Rec_i
a	**249**	0	35	26	12	2	139	59	150	0	0,37
b	0	**3436**	0	37	0	4	0	0	0	1	0,98
c	39	0	**333**	14	0	29	0	0	126	1	0,61
d	34	14	19	**229**	0	0	0	0	1	13	0,73
e	7	0	0	0	**405**	29	20	11	0	0	0,85
f	1	6	0	2	13	**1077**	11	10	0	1	0,96
g	3	4	1	8	8	23	**211**	2	10	0	0,78
h	1	0	0	0	29	20	9	**53**	2	0	0,46
i	1	1	0	0	1	1	6	0	**57**	1	0,83
j	0	1	1	26	1	11	0	8	9	**91**	0,61
Prec_i	0,74	0,99	0,85	0,66	0,86	0,90	0,53	0,37	0,16	0,84	

Na seção 5.1.3 foi introduzido o problema do desbalanceamento das classes. Com a finalidade de exemplificar e quantificar este problema, na última linha das tabelas 5.3 e 5.4 apresentamos a Micro-Fscore, a qual considera a matriz de confusão total sem distinção das classes. Podemos observar uma diferença considerável entre os valores Micro-F1 e Macro-F1. Isto indica que, neste tipo de sistemas bioacústicos, escolher métricas mais "otimistas" pode levar a conclusões erradas sobre a acurácia final do sistema, e portanto causar erros nas estimativas populacionais das espécies.

Com o propósito de destacar o viés introduzido nas abordagens de validação cruzada tradicionais, nas quais os indivíduos não são identificados, isto é, a validação é realizada por sílabas, nós executamos um kNN e obtivemos uma precisão de 97%, uma revocação de 96% e um F1 de 96%. Estes resultados são equivalentes às abordagens descritas por vários autores (Huang et al., 2009, Han et al., 2011, Colonna et al., 2012, Dayou et al., 2011, Jaafar et al., 2014, Vaca-Castaño and Rodriguez, 2010, Yuan and Ramli, 2013). Pelas diferenças apresentadas nestes resultados e comparados às tabelas anteriores, utilizando k-CV por indivíduos, confirmamos nossa hipótese de generalização dos modelos.

5.1.6 Conclusões sobre a validação cruzada proposta e a simplificação dos problemas multi-classe

Nesta seção apresentamos um procedimento alternativo de validação cruzada e dois métodos para decomposição e simplificação de problemas multiclasse. Para poder realizar estas comparações introduzimos um rótulo adicional às instâncias de nossa base, de forma a poder identificar as sílabas de cada indivíduo. Através dos resultados experimentais mostramos que a validação cruzada proposta reduz a acurácia quando comparada contra os métodos encontrados na literatura, nos quais não se realiza distinção entre as sílabas dos indivíduos. Desta forma foi possível quantificar de forma mais precisa o erro esperado dos métodos de reconhecimento das espécies, sem sobreestimar a acurácia dos mesmos. Apresentamos também a diferença entre os resultados utilizando Micro- e Macro-métricas e concluímos que, pelas caraterísticas das aplicações bioacústicas, as Macro-métricas produzem um resultado mais justo devido ao desbalanceamento inerente das classes.

Finalmente comparamos duas abordagens para decompor problemas multiclasse em um "pool" de classificadores binários. Uma comparação dos ganhos entre as colunas da tabela 5.3 (1AA) e as colunas da tabela 5.4 (1A1) são apresentados na tabela 5.6. Nesta tabela, os resultados positivos indicam que a abordagem 1AA teve ganhos sobre a abordagem 1A1, por exemplo, a precisão do *RBF-SVM* com 1AA resultou positivo comparado com a precisão do *RBF-SVM* e 1A1, indicando que o 1AA foi melhor. Os valores da Macro-F1 foram comparados ponto-a-ponto entre cada um dos métodos avaliados e apresentados na última linha desta tabela. Em todos os casos aplicamos o t-teste e concluímos que não existe diferença estatística significativa entre os resultados, portanto, não existiu diferença entre decompor o problema aplicando 1AA ou 1A1. Embora os ganhos apresentados por 1AA na maioria dos casos sejam frações inferiores a 10%, podemos notar que a decomposição 1A1 favoreceu os métodos DT e QDA. Sem embargo, a Macro-F1 do 1AA foi superior em três dos cinco métodos de classificação avaliados.

Além dos ganhos positivos alcançados pela decomposição 1AA, o menor custo computacional é sua segunda vantagem, lembrando que o número de problemas binários resolvidos 1AA é igual ao número total de classes. Por este motivo, o 1AA requer uma quantidade menor de iterações e consequentemente, um menor custo computacional. Além disso, no caso em que se deseje embarcar um classificador binário dentro de cada nó sensores da rede, com 1AA o número de sensores seria menor e portanto mais econômico.

Tabela 5.6. Ganhos do 1AA sobre o 1A1

	kNN	RBF-SVM	Poly-SVM	DT	QDA
Macro-Prec.	+0,01	+0,02	0,00	−0,10	+0,02
Macro-Rec.	0,00	+0,09	+0,05	−0,09	+0,01
Macro-F1	0,00	+0,18	+0,03	−0,09	+0,06

5.2 Classificação bioacústica hierárquica

Como foi apontado no capítulo 1, a objetivo geral de nossa abordagem consiste em tratar o desafio de monitorar anuros como uma tarefa de reconhecimento de espécies usando seus cantos, combinando técnicas de processamento digital de sinais e ML. Neste contexto existem dois possibilidades: (1) utilizar técnicas de classificação multi-classe planas, como tradicionalmente é feito, ou (2) construir um método especializado que aproveite caraterísticas do nosso domínio de aplicação, tal como a taxonomia das espécies.

Nos métodos de reconhecimento bioacústico tradicionais, geralmente um classificador "plano" multi-classe é treinado para identificar diversas espécies animais, onde o número de classes é igual ao número de espécies que se deseja reconhecer (figura 5.4(a)). Resultados de aplicação de classificadores multi-classe planos foram apresentados na seção 5.1.5. Neste tipo de métodos, a complexidade da função de classificação aumenta proporcionalmente à quantidade de espécies que se deseja monitorar. Para evitar esse problema, propomos uma abordagem "hierárquica" que decompõe o problema em três níveis taxonômicos utilizando a família, o gênero e a espécie.

A taxonomia filogenética visa organizar as espécies animais em categorias hierárquicas (figura 5.5). O conhecimento da taxonomia dos anuros é uma característica específica do nosso domínio de aplicação, a qual pode ser aproveitada para construir um método com melhor desempenho e que permita investigar as relações acústicas entre as amostras das diferentes espécies com maior detalhe. Utilizando esta organização pré-definida desenvolvemos um sistema de classificação hierárquico combinando classificadores planos em forma de árvore hierárquica do tipo *top-down*. Para consegui-lo, primeiro tivemos que transformar o problema original com um único rótulo em um problema *multi-output* (i.e. multi-rótulo e multi-classe). Assim, nossa hipótese é que, utilizando a taxonomia filogenética, é possível: (1) discriminar melhor as classes, e (2) reduzir a complexidade dos modelos envolvidos na classificação.

Nossa proposta inclui comparar dois métodos de classificação hierárquica: (1) utilizando um classificador por nó pai e (2) um classificador por nível, contra uma abordagem plana (figura 5.4). Estas abordagens são frequentemente chamadas de *Local Classifier per Parent Node* (LCPN) e *Local Classifier Per Level* (LCPL), respectiva-

mente. Aplicando LCPN ganhamos em simplicidade, sendo possível decompor o espaço de caraterísticas (ou LLDs) do problema original em sub-problemas com um número menor de classes. Com o segundo método, o LCPL, podemos combinar classificadores com uma determinada organização hierárquica, e melhorar os resultados aproveitando a diversidade dos classificadores. Além disso, em ambas estratégias podemos inspecionar as matrizes de confusão em cada nível hierárquico para obter informações adicionais sobre a relação entre as amostras e suas classes.

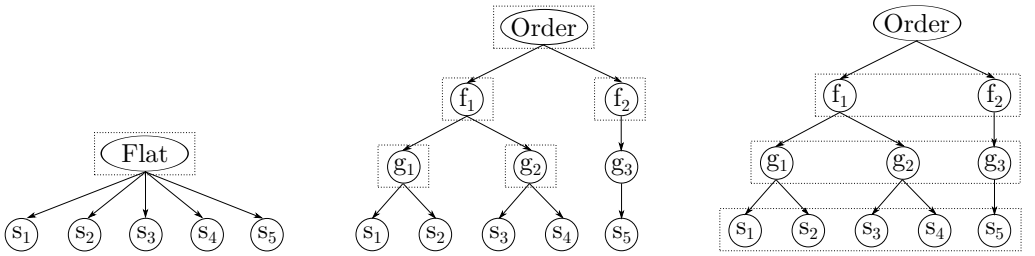
5.2.1 Fundamentos dos métodos hierárquicos

Métodos hierárquicos são freqüentemente utilizados para resolver problemas multi-rótulo, nos quais as classes possuem uma estrutura taxonômica inerente, isto é, uma instância que pertence a uma subclasse, naturalmente pertence a sua classe de nível superior. Estes métodos ajudam a simplificar problemas complexos com várias classes, transformando-os em uma abordagem multi-rótulo que considera a relação hierárquica entre tais rótulos. Por exemplo, com a configuração LCPN (figura 5.6(a)), cada vez que descemos a um nível na hierarquia o número de soluções possíveis é reduzido simplificando a função de decisão.

Existem dois modelos para descrever as relações hierárquicas entre as classes nos problemas multi-rótulo: (a) árvores, e (b) grafos acíclicos direcionados (*Direct Acyclic Graphs* - DAG). Uma estrutura de árvore conecta um conjunto de nós de folhas a um único nó pai formando várias subárvores não interconectadas no mesmo nível. Um DAG é uma estrutura mais flexível, a qual permite que nós folhas tenham mais de um nó pai (Freitas and Carvalho, 2007). Em nossa aplicação descartamos as estruturas DAGs, devido às restrições taxonômicas do nosso problema, as quais ditam que cada espécie pode pertencer unicamente a um gênero e a uma família.

A figura 5.4 ilustra um classificador plano e duas abordagens hierárquicas frequentemente encontradas na literatura. Nas figuras 5.4(b) e 5.4(c), um conjunto de classificadores planos é empregado para construir duas árvores hierárquicas multi-rótulo (Silla Jr and Freitas, 2011). Estas árvores podem ser desbalanceadas dependendo da estrutura taxonômica do problema. Os classificadores dentro de cada nó são treinados separadamente antes de montar a estrutura final. Durante a fase de reconhecimento, a estratégia *top-down* é adotada para determinar a classe de uma nova amostra. Esta estratégia começa a partir dos nós superiores executando as classificações correspondentes e desce até um nó folha no último nível hierárquico. Assim, a decisão final resulta em uma relação única entre o conjunto de rótulos previstos e o caminho da árvore seguido.

Estes métodos são bem adequados para o contexto de reconhecimento de espécies de anuros, onde os rótulos possuem uma taxonomia inerente. Sem embargo, uma desvantagem importante associada a este tipo abordagens é a propagação de erros provenientes dos níveis superiores.



(a) Classificador plano. (b) Um classificador por nó pai. (c) Um classificador por nível.

Figura 5.4. Classificador multi-classe plano (a). Diferentes estratégias para criar um método multi-rótulo hierárquico combinando classificadores planos. Os nível superior classifica os rótulos das famílias (**f**), o nível do central identifica o gênero (**g**) e o nível inferior reconhece as espécies (**s**). As figuras 5.4(b) e 5.4(c) são exemplos das abordagens LCPN e LCPL respectivamente.

5.2.2 Motivação para utilizar uma abordagem hierárquica

Anuran é o nome de uma Ordem de animais dentro da Classe de Anfíbios que inclui sapos e rãs. Segundo estudos recentes existem mais de 6600 espécies de anuros diferentes no mundo, classificadas em 56 famílias e vários gêneros (Frost, 2016). Precisamente, a diversidade de anuros nas áreas tropicais da América do Sul é a maior do mundo, concentrando aproximadamente 70% da biodiversidade global de anfíbios (IUCN, 2016). Neste contexto, desenvolver um classificador plano que consiga reconhecer a maioria das espécies existentes, inclusive para monitorar áreas pequenas, não é uma tarefa trivial. O modelo deveria ser treinado com o número de classes igual ao número de espécies que pretende-se reconhecer, o que causaria que a complexidade da função de decisão aumente conforme aumenta a quantidade de espécies monitoradas.

Como mencionamos na seção anterior, a abordagem hierárquica LCPN pode aliviar estes problemas e reduzir a complexidade das funções de classificação decompondo o espaço de caraterísticas acústicas. Assim, propomos usar a taxonomia de *Linnaeus*, atualmente conhecida como taxonomia filogenética, para construir uma árvore com três níveis hierárquicos que consiga reconhecer para cada amostra os rótulos referentes à família, o gênero e a espécie (figura 5.5).

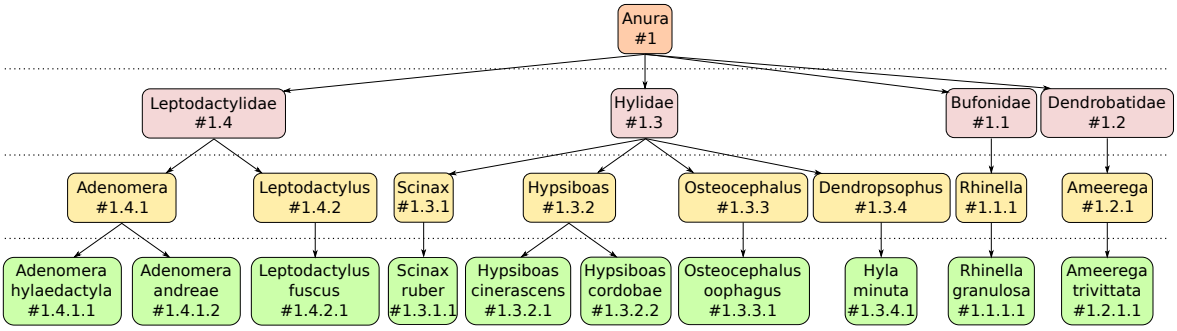


Figura 5.5. Taxonomia das espécies dentro da nossa base de dados. Do nível superior ao inferior: famílias, gêneros e espécies. O marcador # representa o identificador de cada nó da nossa hierárquica.

Em contraste, a abordagem LCPL não reduz o espaço de caraterísticas, nem a complexidade da função de classificação no último nível da árvore, a qual se corresponde com um classificador plano. No entanto, a LCPL permite tratar o problema com diferentes níveis de granularidade, isto é, combinando modelos com maior ou menor complexidade de forma semelhante a um *ensemble learning* (figura5.6(b)). Outra vantagem do LCPL, em relação ao LCPN, reside nos casos em que as ramas da hierarquia possuem apenas um nó de folha, como no caso da família Bofunidae representada na figura 5.5. Nesses casos especiais, a abordagem LCPL fornece uma probabilidade posterior para esses nós folha. Embora existam vantagens, as aplicações desta abordagem têm sido limitadas na literatura sendo utilizada a maioria das vezes como método de comparação (ou *baseline*) (Silla and Kaestner, 2013, Silla Jr and Freitas, 2011).

5.2.3 Descrição da abordagem hierárquica proposta

A taxonomia filogenética visa organizar as espécies animais em categorias hierárquicas. Usando essa organização predefinida para anuros, podemos construir nosso sistema de classificação hierárquico adicionando dois rótulos extras ao nosso conjunto de dados original:

$$\text{Base de dados} = \begin{bmatrix} \mathbf{c}_1 = [c_1, c_2, \ldots, c_l], & s_1, & g_1, & f_1, & \text{id}_1 \\ \mathbf{c}_2 = [c_1, c_2, \ldots, c_l], & s_1, & g_1, & f_1, & \text{id}_1 \\ \vdots & \vdots & \vdots & & \\ \mathbf{c}_d = [c_1, c_2, \ldots, c_l], & s_i, & g_j, & f_m, & \text{id}_k \end{bmatrix}$$

com estes novos rótulos (g_j e f_m), transformamos nosso problema multi-classe com um único rótulo em um problema multi-rótulo e multi-classe ao mesmo tempo (MM). Este MM é uma generalização dos problemas tradicionais multi-rótulo, onde as classes são

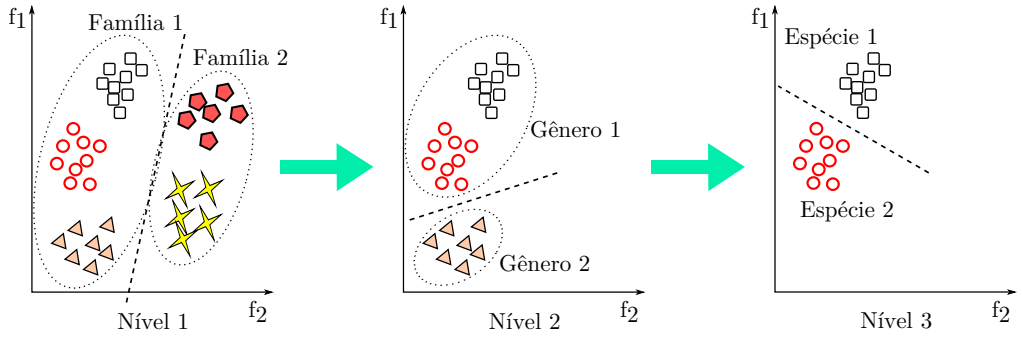
binárias em cada coluna. Os MM também são chamados de *multi-output* porque a saída é composta por uma tupla de rótulos (Borchani et al., 2016), que em no nosso caso é uma tripla da forma $[s, g, f]$. A base de dados introduzida na tabela 5.2, apresenta os rótulos das famílias e os gêneros correspondentes a cada uma das espécies utilizadas em nossos experimentos.

Em nossa aplicação a transformação MM é viável porque há uma relação inequívoca entre os nomes das espécies e seu gênero e família. Isto é, um subconjunto $S^0 = \{s_1, \ldots, s_p\}$ de espécies pertence a um único gênero ($S \subseteq g_m$), enquanto que um subconjunto de gêneros $G^0 = \{g_1, \ldots, g_p\}$ por sua vez pertencem a uma família em particular ($G \subseteq f_m$) tal que $f_m \subseteq F^0$. Portanto, qualquer $s_i \subseteq G^0 \subseteq F^0$ sem ambigüidade. Assim, se um classificador plano consegue predizer corretamente uma determinada espécie, por conseguinte estaria também predizendo as classes gênero e família. Dado este conceito, podemos aplicar engenharia reversa, e desenvolver uma abordagem hierárquica *top-down* conforme ilustrado nas figuras 5.4(b) e 5.4(c). Assim, a nossa árvore hierárquica é construída respeitando as ligações entre gêneros e famílias ilustrada na figura 5.5.

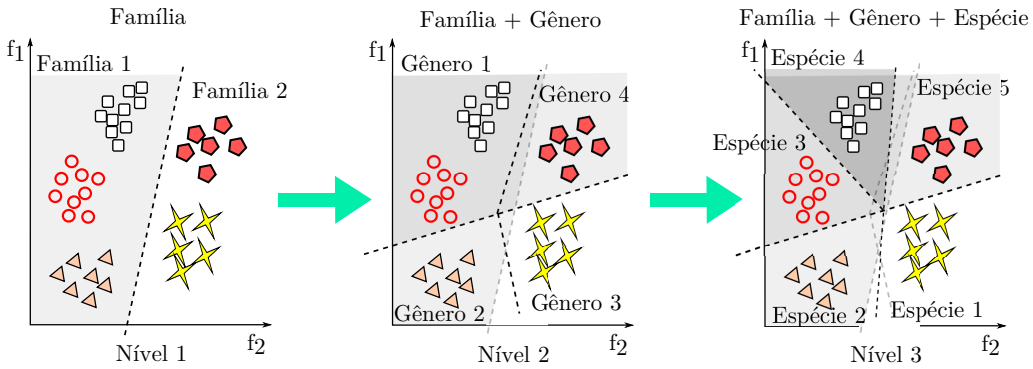
Uma vantagem notável da abordagem hierárquica é que não precisamos realizar as classificações em todos os níveis para algumas bifurcações da hierarquia. Esta é a principal vantagem do LCPN quando combinada com a taxonomia de *Linnaeus*. Por exemplo, observamos na figura 5.5 que se o primeiro classificador atribuir o rótulo Bufonidae a uma nova amostra no nível superior, não é necessário continuar realizando as classificações dos níveis restantes, porque não há mais divisões nesse ramo da hierarquia. Portanto, os rótulos do gênero Rhinella e da espécie Rhinella granulosa são atribuídos automaticamente.

A figura 5.6(a) ilustra como procede a simplificação do problema usando a decomposição LCPN com dois atributos do ponto de vista do espaço dos LLDs. Como podemos notar, no início, todas as amostras pertencem a duas famílias. Após a classificação do primeiro nível o problema é reduzido e simplificado pela decomposição do espaço de possíveis soluções, no qual apenas permanecem as amostras da primeira família. Este processo é repetido até que o último nível de classificação seja atingido identificando o rótulo final da espécie. Assim, a classe de um nó de folha é usada para estimar o rótulo de novas amostras.

A figura 5.6(b) ilustra o procedimento LCPL e como a combinação de funções de decisão ocorre. Assim, a decisão final corresponde à combinação das funções de decisão sobrepostas dos três níveis (cinza escuro). Uma vez que, cada nível da hierarquia é treinado com um número diferente de classes, mas com o mesmo número de amostras, podemos considerar que são visões diferentes do mesmo problema com granularidades



(a) Etapas de decomposição e simplificação do problema executando o LCPN hierárquico.



(b) Sobreposição das funções de decisão dos níveis executando LCPL. A decisão final corresponde à sobreposição de três decisões individuais.

Figura 5.6. Comparação entre as abordagens LCPN e LCPL desde a perspectiva de espaço de caraterísticas.

de decisão distintas. Observe que os limites de tais funções não são necessariamente iguais em todos os níveis. As definições restantes e configurações experimentais da nossa abordagem hierárquica são detalhadas a seguir.

5.2.4 Metodologia experimental e resultados

O conjunto de dados utilizados nestes experimentos foi introduzido na tabela 5.2[3], o qual foi obtido aplicando o procedimento de segmentação e extração de caraterísticas da seção 5.1.5. Os algoritmos ML embarcados nos nós da hierarquia são: kNN com 3 vizinhos, dois SVMs, um com kernel RBF (RBF-SVM) e outro com kernel polinomial (Poly-SVM) de grau três, e uma árvore de decisão (DT). Os parâmetros do kernel RBF-SVM foram configurados automaticamente usando a implementação padrão do software Matlab. O mesmo é válido para a árvore de decisão. Além disso, em todos classificadores optamos pela decomposição multi-classe 1AA apresentada ante-

[3]Disponível em `https://goo.gl/10iCRp`.

riormente, pelo fato de apresentar melhor desempenho e realizar menos comparações binárias.

Como estamos lidando com um problema supervisionado, e queremos quantificar as capacidades de generalização do sistema hierárquico, aplicamos o procedimento de validação cruzada por indivíduos desenvolvido na seção 5.1.2. Esta validação ajuda-nos a estimar o erro esperado em uma situação real. Além disso, utilizamos as Macro-métricas descritas na seção 2.6.2 para diminuir o impacto do desbalanceamento das classes. Ressaltamos que uma vantagem importante da abordagem hierárquica, comparada com o classificador plano, é a possibilidade de inspecionar as matrizes de confusão para cada nível da hierarquia. As tabelas 5.7, 5.8 e 5.9 ilustram tais matrizes para os rótulos das famílias, os gêneros e as espécies, utilizando kNN com 1AA em cada nó da hierarquia LCPL.

Tabela 5.7. Matriz de confusão dos rótulos das famílias com kNN e LCPL. Prec_i e Rec_i representam a precisão e a revocação respectivamente.

	Bufonidae	Dendrobatidae	Hylidae	Leptodactylidae	Rec_i
Bufonidae	**57**	0	3	8	0,83
Dendrobatidae	126	**333**	44	39	0,61
Hylidae	12	21	**2030**	102	0,93
Leptodactylidae	158	38	182	**4042**	0,91
Prec_i	0,16	0,84	0,89	0,96	

Tabela 5.8. Matriz de confusão dos rótulos de gênero com kNN e LCPL. Legenda: (a) Adenomera, (b) Ameerega, (c) Dendropsophus, (d) Hypsiboas, (e) Leptodactylus, (f) Osteocephalus, (g) Rhinella, e (h) Scinax. Prec_i e Rec_i representam a precisão e a revocação respectivamente.

	a	b	c	d	e	f	g	h	Rec_i
a	**3685**	37	64	19	138	57	148	2	0,88
b	39	**333**	13	29	0	1	126	1	0,61
c	48	20	**228**	0	0	0	1	13	0,73
d	17	0	2	**1537**	26	10	0	1	0,96
e	7	1	8	30	**212**	2	10	0	0,78
f	1	0	0	50	9	**52**	2	0	0,45
g	2	0	0	2	6	0	**57**	1	0,83
h	1	1	26	12	0	8	9	**91**	0,61
Prec_i	0,96	0,84	0,66	0,91	0,54	0,40	0,16	0,83	

A partir destas matrizes de confusão é possível obter as Macro-métricas da precisão, revocação e F1 para cada nível da hierarquia, permitindo realizar análises adicionais que não seriam impossíveis com um classificador plano. A tabela 5.10 apresenta o resultado destas métricas para os três níveis.

Tabela 5.9. Matriz de confusão dos rótulos das espécies com kNN e LCPL. Legenda: (a) Adenomera andreae, (b) Adenomera hylaedactyla, (c) Ameerega trivittata, (d) Hyla minuta, (e) Hypsiboas cinerascens, (f) Hypsiboas cordobae, (g) Leptodactylus fuscus, (h) Osteocephalus oophagus, (i) Rhinella granulosa, e (j) Scinax ruber. Prec_i e Rec_i representam a precisão e a revocação respectivamente.

	a	b	c	d	e	f	g	h	i	j	Rec_i
a	**249**	0	37	27	13	2	138	57	148	1	0,37
b	0	**3436**	0	37	0	4	0	0	0	1	0,98
c	39	0	**333**	13	0	29	0	1	126	1	0,61
d	34	14	20	**228**	0	0	0	0	1	13	0,73
e	11	0	0	0	**412**	31	16	2	0	0	0,87
f	0	6	0	2	13	**1081**	10	8	0	1	0,96
g	3	4	1	8	8	22	**212**	2	10	0	0,78
h	1	0	0	0	30	20	9	**52**	2	0	0,45
i	1	1	0	0	1	1	6	0	**57**	1	0,83
j	0	1	1	26	1	11	0	8	9	**91**	0,61
Prec_i	0,73	0,99	0,84	0,66	0,86	0,90	0,54	0,40	0,16	0,83	

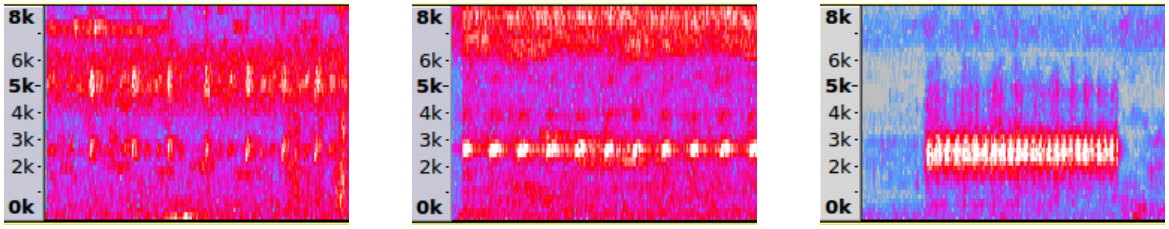
No que diz respeito às classificações, podemos notar que a baixa precisão da família Bufonidae foi causada pelas elevadas confusões entre as espécies Adenomera a. e Ameerega t. com a Rhinella g., mas sem intervenção da espécie Adenomera h. a qual pertence à mesma família que a Adenomera a.. Isto sugere dois fatos importantes. Primeiro, é possível monitorar com um elevado grau de precisão um cenário real contendo as espécies Rhinella g. e Adenomera h. utilizando somente o classificador treinado com os rótulos das famílias. Segundo, existem semelhanças acústicas importantes contidas no espaço dos LLDs entre as famílias Bufonidae, Leptodactylidae e Dendrobatidae, e consequentemente entre os gêneros Adenomera, Ameerega e Rhinella de nossa base de dados.

Na figura 5.7 são comparados três espectrogramas que ilustram tais semelhanças acústicas. Além disso, comparando as tabelas 5.5 e 5.9 observamos que a abordagem hierárquica LCPL atingiu um maior número de verdadeiros positivos nas espécies Hypsiboas cin., Hypsiboas cor. e Leptodactylus f. comparado ao classificador plano.

Tabela 5.10. Resultado da classificação hierárquica por nível com kNN e LCPL.

	Família	Gênero	Espécie
Macro-Precisão	0,71	0,66	0,69
Macro-Revocação	0,82	0,73	0,71
Macro-F1	0,76	0,69	0,69

Adicionalmente, analisando a matriz de confusão no primeiro nível e comparando-a com o nível de gênero, notamos que Rhinnela h. perdeu aproximadamente 50% de suas amostras no segundo nível contra a classe Hypsiboas. Também percebemos a



(a) Adenomera a. (b) Ameerega t. (c) Rhinella g.

Figura 5.7. Comparação visual entre espectrogramas de três espécies. A cor branca representa frequências com maior concentração de energia. Nestas figuras podemos notar que as bandas de frequências centralizadas em 2.5 kHz são comuns para as três espécies. A espécie Adenomera a. também contem uma alta concentração de energia nas bandas frequências próximas a 5 kHz.

dificuldade em reconhecer a Rhinella g. na presença do Hypsiboas cin. e Hypsiboas cor.. No entanto, o caso oposto não é igualmente verdadeiro. Isto sugere que as amostras do gênero Rhinella encontram-se provavelmente sobrepostas com as amostras de Hypsiboas no espaço dos descritores acústicos. Conclusões semelhantes podem ser obtidas para outros gêneros e espécies. Por exemplo, várias amostras de Scinax foram confundidas com Dendropsophus e Hypsiboas, e também dentro do gênero Hypsiboas, a Hypsiboas cor. foi principalmente confundida com espécie Scinax r..

Uma vantagem prática desta abordagem hierárquica reside no fato que, para um cenário real contendo as espécies Rhinella g., Hila m., Scinax r., Osteocephalus o., Hypsiboas cin., e Hypsiboas cor., o método pode obter uma alta taxa de reconhecimento, e inclusive maior ao classificador plano, porque essas espécies pertencem a famílias diferentes. Uma outra alternativa é treinar o método hierárquico até o segundo nível para separar, com alta precisão, o conjunto de espécies {Adenomera a. e Adenomera h.} do conjunto {Hypsiboas cin. e Hypsiboas cor.} ou do Scinax r., diminuindo assim a carga computacional.

Os resultados comparando as duas abordagens com todos os métodos de classificação mencionados, são apresentados nas tabelas 5.11 e 5.12. Novamente, o teste estatístico t-test foi aplicado às colunas da precisão e revocação. Aqui, podemos notar que a melhor precisão foi obtida pelo RBF-SVM do método LCPN. No entanto, a melhor Macro-F1 foi obtida usando a combinação Poly-SVM e LCPL.

Embora estes métodos permitam investigar a relação taxonômica entre as amostras existem duas desvantagens principais desta abordagem. A primeira é a propagação dos erros desde os níveis superiores até o nível inferior. A segunda, é o custo computacional de se treinar três classificadores no lugar de um, com a abordagem LCPL, ou treinar cinco modelos de classificação, com a abordagem LCPN. A tabela 5.13 mostras

Tabela 5.11. LCPN. Os valores em negrito foram detectados como empate pelo teste estatístico.

	kNN		RBF-SVM		Poly-SVM		DT	
Species	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Adenomera a.	0,73	0,37	0,80	0,22	0,60	0,34	0,72	0,50
Adenomera h.	0,99	0,98	0,99	0,94	0,99	0,98	0,98	0,97
Ameerega t.	0,84	0,61	0,97	0,29	0,70	0,58	0,71	0,43
Hyla m.	0,67	0,73	0,83	0,43	0,70	0,84	0,32	0,64
Hypsiboas cin.	0,84	0,87	0,16	0,72	0,89	0,84	0,67	0,60
Hypsiboas cor.	0,89	0,96	0,87	0,82	0,86	0,96	0,88	0,91
Leptodactylus f.	0,54	0,78	0,98	0,30	0,71	0,72	0,35	0,20
Osteocephalus o.	0,42	0,43	0,97	0,33	0,15	0,35	0,21	0,13
Rhinella g.	0,16	0,83	0,96	0,82	0,27	0,79	0,11	0,72
Scinax r.	0,84	0,61	0,88	0,47	0,94	0,70	0,72	0,45
Macro-	**0,69**	**0,72**	**0,84**	**0,54**	**0,68**	**0,71**	0,57	**0,56**
Macro-F1	0,70		0,65		0,70		0,56	

Tabela 5.12. LCPL. Os valores em negrito foram detectados como empate pelo teste estatístico.

	kNN		RBF SVM		Poly SVM		Tree	
Species	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Adenomera a.	0,73	0,37	0,87	0,22	0,60	0,35	0,54	0,44
Adenomera h.	0,99	0,98	0,99	0,95	0,99	0,99	0,98	0,93
Ameerega t.	0,84	0,61	0,97	0,30	0,74	0,56	0,71	0,43
Hyla m.	0,66	0,73	0,83	0,43	0,65	0,83	0,31	0,50
Hypsiboas cin.	0,86	0,87	0,36	0,56	0,72	0,85	0,58	0,70
Hypsiboas cor.	0,90	0,96	0,43	0,93	0,90	0,96	0,87	0,84
Leptodactylus f.	0,54	0,78	0,98	0,31	0,70	0,74	0,25	0,20
Osteocephalus o.	0,40	0,45	0,97	0,28	0,48	0,55	0,07	0,08
Rhinella g.	0,16	0,83	0,96	0,82	0,22	0,79	0,11	0,72
Scinax r.	0,83	0,61	0,94	0,45	0,92	0,79	0,34	0,17
Macro-	**0,69**	**0,72**	**0,83**	**0,52**	**0,69**	**0,74**	0,47	0,50
Macro-F1	0,70		0,64		0,72		0,49	

os ganhos (ou as perdas) da Macro-precisão, Macro-revocação, e a Macro-F1, causados pela propagação dos erros das abordagens hierárquicas em relação ao método de classificação plano 1AA, ambos aplicados ao mesmo conjunto de dados. Os resultados positivos indicam que a bordagem hierárquica da coluna correspondente superou o mesmo tipo de classificador plano. Os resultados negativos indicam o caso oposto. Como podemos notar, a abordagem LCPN foi a única que conseguiu superar sua correspondente plana utilizando árvore de decisão.

Tabela 5.13. Ganhos das abordagens hierárquicas comparadas contra as abordagens planas. Valores positivos indicam que a a bordagem hierárquica da coluna correspondente superou o mesmo tipo de classificador plano. Os resultados negativos indicam o oposto.

	Prec		Rec		Fscore	
	LCPN	LCPL	LCPN	LCPL	LCPN	LCPL
kNN	0,00	0,00	0,00	0,00	0,00	0,00
RBF-SVM	-0,01	-0,02	0,00	-0,02	-0,01	-0,02
Poly-SVM	-0,03	-0,02	-0,05	-0,04	-0,04	-0,02
DT	+0,08	-0,02	+0,06	0,00	+0,06	-0,01

5.2.5 Conclusões sobre os métodos de reconhecimento hierárquicos

Na metodologia apresentada nesta seção, primeiro convertemos o problema multi-classe original em um problema *multi-output*, o que nos permitiu investigar a correspondência entre a taxonomia das espécies e as amostras, no espaço dos descritores bioacústicos das vocalizações. Concluindo-se que, a combinação entre taxonomia filogenética, os MFCCss e a proximidade obtida através do classificador kNN, é útil para identificar semelhanças bioacústicas entre as espécies do ponto de vista do sistema de reconhecimento. No que diz respeito ao método LCPN, sua principal vantagem algorítmica é podar os dados de treinamento para simplificar o espaço de decisões. Já a principal vantagem do método LCPL, é a combinação (ou agregação) das funções de decisão de para aumentar os acertos ou diminuir a incerteza.

Especificamente as espécies Adenomera h., Ameerega t., Hypsiboas cin. e Scinax r. foram claramente reconhecíveis na presença de outras espécies usando um LCPL com kNN e, portanto, são bons candidatos para um programa de monitoramento bioacústico automatizado. Gostaríamos de enfatizar que essas espécies pertencem a diferentes famílias e gêneros, confirmando que nossa estratégia hierárquica é útil neste contexto de aplicação. Além disso, um fato adicional interessante descoberto é que a espécie Hypsiboas cor., que pertence a outro país, e habita em uma área no tropical, foi fácil de reconhecer.

Adicionalmente podemos comparar o desempenho do classificador plano contra o hierárquico utilizando as mesmas configurações (tabelas 5.13). Infelizmente, nestas comparações obtivemos desempenhos semelhantes ou inferiores, sendo impossível reivindicar qual método foi superior. No entanto, com o nossa abordagem hierárquica obtivemos várias informações complementares relacionadas à taxonomia. Além disso, percebemos que os erros dos classificadores ficaram mais dependentes do tipo de valida-

ção cruzada utilizada do que ao próprio método de classificação. Finalmente, destacamos como principal desvantagem da maioria das abordagens hierárquicas a propagação dos erros.

5.3 Classificação colaborativa

A capacidade de monitoramento ambiental de um sensor acústico na floresta é limitada pela sensibilidade do microfone (figura 1.3(a)). Consequentemente, para poder cobrir uma área maior é necessário distribuir um conjunto de nós formando uma rede. A seguir apresentamos uma abordagem de monitoramento bioacústico colaborativo que aproveita as capacidades de troca de informação entre os nós vizinhos da rede. Partimos do suposto que um classificador plano é embarcado em cada nó do *cluster* e que existe sobreposição entre as áreas de cobertura dos sensores próximos. Assim, nosso objetivo é identificar como aproveitar a natureza colaborativa da rede para melhorar o reconhecimento dos anuros.

Nossa modelagem de classificação colaborativa aplica diferentes técnicas de *ensemble learning*. Para isto, combinamos diferentes classificadores e diferentes técnicas de votação, sempre considerando as restrições de processamento dos nós. Nos experimentos seguintes simulamos cenários com ruídos aleatórios, atenuações e confusões (cenários com mais de uma espécie). Assim, investigamos como descartar cenários confusos utilizando as percepções dos diferentes sensores dentro de um mesmo *cluster*.

5.3.1 Fundamentos do monitoramento colaborativo

Os sistemas de monitoramento de anfíbios modernos utilizam métodos de classificação automática e Redes de Sensores Sem Fio (RSSF), para estimar mudanças nas populações, com o objetivo de determinar suas causas. As RSSF constituem uma alternativa aos métodos manuais para capturar os sinais bioacústicos de forma não intrusiva, permitindo manter o monitoramento no longo prazo com menor esforço humano e menores custos operacionais. Estas redes são compostas por sensores de baixo custo, fato que permite distribuir uma grande quantidade de nós sobre as áreas desejadas (Akyildiz et al., 2002). A capacidade de comunicação sem fio permite a colaboração entre os sensores através da troca de informações (Nakamura et al., 2007b). No entanto, o hardware de baixo custo impõe restrições, tais como: menor capacidade de processamento e baterias com vida útil reduzida (Xing et al., 2005).

Do ponto de vista das redes de sensores acústicos, cada vocalização emitida por um anuro, pode ser considerada um evento acústico que deve ser detectado e processado

pelos sensores próximos. O grupo de nós que detectam o evento acústico formam um comitê de sensores. O comitê de sensores é um conceito mais amplo do que cluster, sendo este um agrupamento na camada de aplicação da rede, enquanto que o cluster de sensores é definido pelo algoritmo de roteamento da rede (Nakamura et al., 2009). Desta forma, o comitê pode incluir sensores próximos de diferentes clusters ou inclusive ser um subconjunto de nós dentro do próprio cluster.

Em nossos experimentos consideramos que cada sensor é capaz de processar o áudio do microfone aplicando a abordagem apresentada na figura 1.1. A cada novo evento acústico a saída do sensor será um vetor com as probabilidades de ocorrência das espécies que foram utilizadas durante o treino do classificador. Assim, disseminando L sensores na área desejada, cada um deles terá sua própria opinião sobre o evento detectado. A partir deste cenário, nossa hipótese é: *combinando as diferentes opiniões de cada sensor é possível decidir qual é a espécie presente com grau maior de certeza.* Baseado nesta hipótese, podemos considerar que a área monitorada por cada nó pode ter sobreposição com os nós vizinhos mais próximos, conseguindo capturar sinais correlacionados e, consequentemente a colaboração entre eles deve melhorar o resultado final.

No cenário colaborativo, o conjunto de sensores é considerado equivalente a um *ensemble learning*. Diferentemente do *ensemble* tradicional, aqui a decisão final da classificação é realizada por um nó intermediário, que encontra-se entre os sensores acústicos que detectam o evento e o destino final (*sink*). Este nó é chamado de líder do comitê (ou simplesmente *sensor líder*) e concentra e combina as classificações dos sensores acústicos (figura 5.8). Diferentes técnicas de *ensemble* podem ser utilizadas para combinar a saída dos sensores, neste trabalho experimentamos quatro regras diferentes: a votação majoritária, a votação majoritária ponderada, a regra geométrica e a regra aritmética (Theodoridis and Koutroumbas, 2008).

A combinação das respostas dos sensores do comitê foi referenciada por Ribas et al. (2012) como "fusão de decisões" (*decision fusion*). Esta forma de fusão permite reduzir a quantidade de dados capturados pelos microfones, e representá-los por um byte identificando a espécie. A fusão, neste caso, lida com uma relação custo-benefício entre aumentar o processamento em cada nó, para ter uma classificação eficiente, e diminuir a quantidade de dados transmitidos pela rede até o *sink*. Além disso, realizar a classificação no sensor permite armazenar os resultados localmente, economizando memória quando comparado com o áudio completo.

Visto como um conjunto de classificadores, cada sensor da rede desempenha um papel fundamental na detecção da espécie. No entanto, a decisão final é responsabilidade do nó líder. Este realiza duas operações, a primeira é combinar as probabilidades a

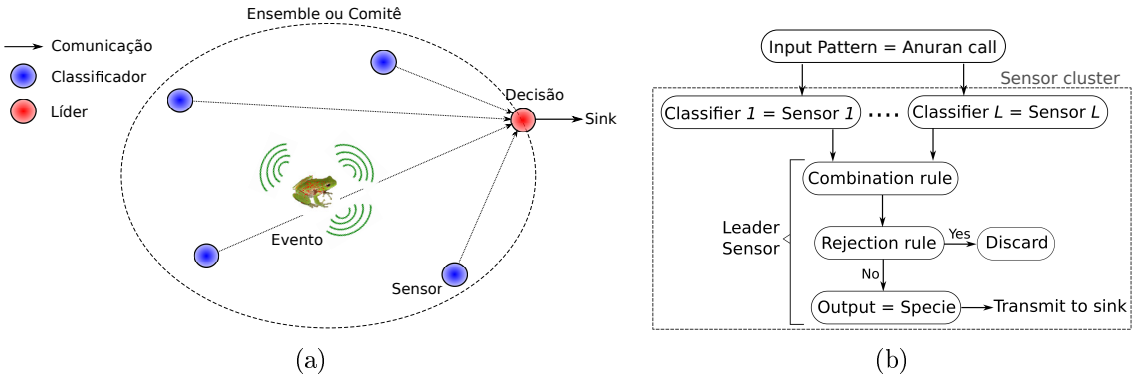


Figura 5.8. (a) Comitê de sensores detectando a chamada do anuro. Considerando esta um evento, a probabilidade das espécies é enviada até o nó líder que toma a decisão final. (b) Relação entre *ensemble* e cluster de sensores, figura extraída de Colonna et al. (2014a).

posteriori de cada sensor para identificar a espécie mais provável, e a segunda é rejeitar casos em que as probabilidades para cada espécie são similares (ou quase uniformes). A operação de rejeição permite eliminar cenários confusos nos quais mais de uma espécie vocaliza ao mesmo tempo. A combinação de vocalizações de espécies diferentes forma novos padrões de sinais desconhecidos pelos classificadores aumentado a taxa de erro.

Através da fusão das decisões, ou classificação colaborativa, obtemos três vantagens: (1) aumento da certeza no reconhecimento das espécies; (2) rejeição de casos confusos que causem uma interpretação errada sobre as espécies presentes no lugar; e (3) diminuição da quantidade de dados armazenados e transmitidos pela rede. Como a que a operação de transmissão é a mais custosa em termos de consumo de energia, o ganho final impacta diretamente nos nós, prolongando a vida útil da rede. Este fato torna a abordagem menos intrusiva e interessante para realizar estudos de longo prazo.

Ao longo desta seção avaliamos quatro técnicas de baixo custo para combinar três tipos de classificadores planos utilizados frequentemente em aplicações com sensores (seção 5.3.4). Investigamos também como as diferentes opiniões dos sensores podem ser utilizadas para descartar cenários confusos em que há várias espécies diferentes ao mesmo tempo (seção 5.3.6). Escolhemos utilizar classificadores planos devido a que estes apresentaram um desempenho similar a nossa abordagem hierárquica, mas consumem menos recursos computacionais dos sensores (menos modelos são avaliados).

5.3.2 Definição do problema

Nos capítulos 4 e 6 foram apresentadas a segmentação automática e a filtragem de ruídos para melhorar o resultado do reconhecimento das espécies. No entanto, os ce-

nários reais possuem uma riqueza acústica[4] e complexidade que dificulta ainda mais o reconhecimento de uma vocalização simples. Nos casos reais da floresta, cada sensor pode receber: (a) uma combinação de vocalizações de diferentes anuros; (b) uma combinação de anuros com outros animais; (c) ruídos diferentes em cada sensor; e (d) sinais com diferentes atrasos e atenuações provocadas pela distância até a fonte sonora.

Dado o contexto de monitoramento ambiental pervasivo, podemos definir o problema como: *melhorar a taxa de reconhecimento das espécies em situações reais, nas quais podem existir diferentes tipos de ruídos, diferentes níveis de atenuações dos sinais e combinações de vocalizações causados por outras espécies presentes na mesma área.* Para que nosso *framework* seja útil como estratégia de monitoramento contínuo *in situ*, devemos considerar o desafio de resolver estes problemas de forma não supervisionada.

Do ponto de vista de aprendizagem de máquina, o problema é identificar quais técnicas de classificação utilizar; enquanto que, do ponto de vista da RSSF, o problema é identificar a melhor técnica de combinação para realizar a fusão das decisões do comitê e desenvolver um critério de rejeição que elimine os possíveis casos confusos. A fusão das decisões está diretamente relacionada com a redução na quantidade de informação transmitida pela rede, enquanto que a rejeição esta relacionada à quantidade informação irrelevante que não devia ser transmitida. Desta forma, a combinação e a rejeição ajudam a diminuir o consumo de bateria e estender a vida útil dos sensores.

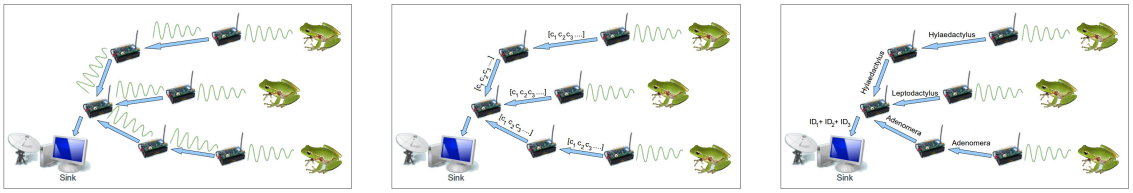
5.3.3 Motivação para utilizar um método colaborativo

No contexto de monitoramento acústico pervasivo de anuros, existem três tipos de cenários (Colonna, 2011): (1) os áudios são capturados e transmitidos até o nó *sink* (Aide et al., 2013); (2) os áudios são representados pelos descritores acústicos e somente estes coeficientes são transmitidos pela rede (Ribas et al., 2012); e (3) os áudios são classificados e somente um número identificador da espécie é enviado até o *sink* (figura 5.9) (Colonna et al., 2014a, Silva and Ruiz, 2015).

A partir destes cenários, podemos definir abordagens híbridas, aplicando técnicas de fusão nos dados coletados, enquanto estes são transmitidos pelos nós da rede (Ribas et al., 2012). Por exemplo, uma abordagem de fusão de descritores acústicos, para o cenário da figura 5.9(b), utiliza os coeficientes de cada membro do comitê e calcula a média destes no nó líder, desta forma o valor médio dos coeficientes viaja pela rede até o *sink* que realiza a classificação final. O mesmo tipo de fusão pode ser aplicado no

[4]A riqueza acústica relaciona-se com a quantidade de animais ou indivíduos que habitam uma região.

primeiro cenário realizando uma combinação linear das amostras dos áudios de cada sensor do comitê.



(a) (b) (c)

Figura 5.9. (a) Transmissão dos áudios completos. (b) Transmissão completa dos descritores acústicos. (c) Transmissão de todas as classificações.

Para cada abordagem, ou combinação possível, existem vantagens e desvantagens. Por exemplo, transmitir a fusão dos áudios do comitê possibilita ouvir ou aplicar diferentes técnicas de processamento de sinais no destino, enquanto que transmitir o resultado das classificações somente permite saber a presença ou ausência da espécie monitorada. Porém, o volume de dados transmitidos no segundo caso é menor, trazendo um economia de energia. Como minimizar o consumo de energia é um dos principais desafios das RSSF (Nakamura et al., 2007a, Figueiredo et al., 2009, Silva and Ruiz, 2015), para que estas possam funcionar por períodos maiores de tempo, notamos que existe uma relação entre: quantidade de dados transmitidos, complexidade do processamento local de cada nó (incluindo o líder) e o objetivo final da aplicação.

A escolha da estratégia depende principalmente do objetivo final. Em nosso caso, o objetivo é inferir as variações das populações no longo prazo identificando a presença dos indivíduos, para tornar o monitoramento uma ferramenta que ajude a diminuir os custos operacionais dos estudos ecológicos. Consequentemente, adotamos a estrutura de fusão de decisões com possibilidade de rejeição (figura 5.8). Desta forma, embarcando o modelo de classificação no sensor é possível ter duas etapas de redução de informação. A primeira etapa de redução permite transformar alguns segundos de áudio em poucos valores de probabilidades em cada membro do comitê, enquanto que a segunda redução é a votação realizada no nó líder. A partir do resultado, somente os casos com maior certeza são encaminhados até o *sink*. Esta abordagem de votação e rejeição permite economizar custos de transmissão de dados e aproveitar a capacidade colaborativa dos nós.

5.3.4 Metodologia

Dadas as restrições de processamento do hardware dos sensores escolhemos embarcar nos nós três técnicas de classificação de baixo custo computacional: Naive Bayes (NB), Análise Discriminante Quadrática (QDA) e Árvore de Decisão (DT) (Theodoridis and Koutroumbas, 2008). Para que cada nó da rede possa reconhecer as espécies, são necessários dois passos:

1. O treinamento, no qual são utilizadas as amostras armazenadas em nossa base para criar e ajustar modelo de classificação; e

2. A previsão, na qual a vocalização é capturada, são extraídas as características e é aplicado o modelo de classificação para obter a probabilidade de cada espécies.

Destes passos, o primeiro, por se mais dispendioso, é realizado *off-line*. Assim, o modelo somente é embarcado nó sensor quanto está pronto. O segundo passo é executado pelos sensores *in situ*.

No modelo de classificação de NB tanto as características quanto as classes são consideradas independentes (seção 2.5.2, página 50). Assim, a partir de uma sílaba da vocalização, representada por um conjunto de características $\mathbf{x}$, a probabilidade de detectar a espécie w_i pode ser estimada por:

$$\mathrm{P}(w_i|\mathbf{x}) = \frac{\mathrm{P}(\mathbf{x}|w_i)\mathrm{P}(w_i)}{\mathrm{P}(\mathbf{x})}. \tag{5.1}$$

No método QDA, as características das espécies são separadas por funções quadráticas (seção 2.5.3, página 51). Este método assume que cada classe possui uma distribuição normal multivariada com matrizes de covariância diferentes[5] (McLachlan, 2004).

Finalmente, no método DT uma árvore de decisão é construída a partir das sílabas de exemplo da base (seção 2.5.4, página 52). A árvore resultante é um modelo hierárquico no qual: cada nó interno indica uma característica, cada ramificação corresponde a uma condição que deve ser testada e cada nó folha é uma classe final. Neste caso, o método de decisão prova as possíveis condições para os valores das características, descendo pela árvore até alcançar uma classe final (Theodoridis and Koutroumbas, 2008).

Os três métodos de classificação mencionados foram embarcados nos sensores de nossas simulações. Definimos uma área de $10 \times 10\,\mathrm{m}^2$ com os nós distribuídos em forma de grade imperfeita com perturbação aleatória ($N \sim (0, 1)$) em cada posição

[5]Para verificar esta hipótese aplicamos o teste estatístico Box's M (Box, 1949).

(figura 5.10). Utilizamos uma base de dados com sílabas de nove espécies de anuros diferentes. Como o objetivo principal deste capítulo é verificar se existe ganho na taxa de reconhecimento causada pela colaboração entre os nós, decidimos realizar a segmentação manualmente, para não ocasionar confusões entre os erros de classificação e os de segmentação.

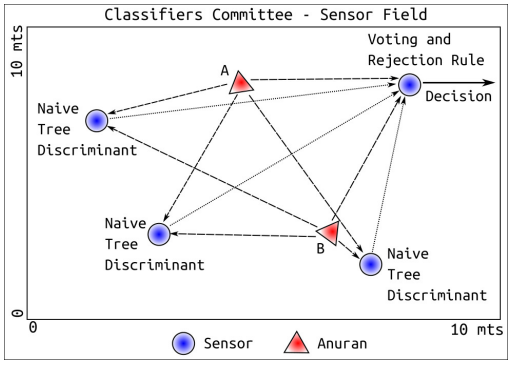


Figura 5.10. Exemplo de cenário simulado com duas espécies de anuros e um comité com quatro sensores. Figura extraída de Colonna et al. (2014a).

A cada iteração da nossa simulação executamos os passos seguintes:

1. Sorteamos aleatoriamente, com probabilidade uniforme, um cenário contendo uma ou duas espécies diferentes com posições aleatórias;

2. Selecionamos aleatoriamente da base uma sílaba para cada espécie escolhida no passo anterior, e deixamos o conjunto restante de sílabas para treinar os modelos de classificação;

3. Cada sílaba sorteada é atenuada de acordo com a distância entre cada sensor e a posição do anuro (passo 1), se o cenário sorteado corresponde a duas espécies as sílabas destas são combinadas linearmente gerando uma mistura de som na entrada do sensor;

4. A partir da mistura das sílabas, os sensores extraem as características (os MFCCs) e realizam a classificação, entregando como resposta ao líder um vetor com as probabilidade de ocorrência de cada espécie; e

5. Finalmente, os vetores de probabilidade são combinados aplicando as regras da seção 5.3.5, e é calculada a entropia para decidir se deve-se rejeitar ou transmitir o resultado final até o nó *sink*.

No passo (3) os sinais são linearmente combinados e atenuados aplicando:

$$\mathbf{x} = \beta_1 x_1 + \beta_2 x_2, \tag{5.2}$$

onde x_1 e x_2 são as sílabas, o coeficiente de atenuação é calculado conforme $\beta = 1/10^{\frac{\alpha d_i}{20}}$, no qual d_i é distância [m] desde o anuro até o sensor e α (dB/m) é a constante de absorção atmosférica (Ribas et al., 2012). Note-se que β_2 é igual a zero no cenário com um único anuro. O *framework* de classificação embarcado nos sensores possui a mesma estrutura apresentada nos capítulos anteriores, no qual são utilizados os MFCCs como conjunto de características extraídas no passo (4).

A métrica de avaliação utilizada é a taxa de erro, obtida a partir da acurácia $e = 1 - \text{Acc}$ de acordo com a equação 2.56 (página 56). Em todos os resultados as taxas médias de erros entre os cenários simulados foram comparadas aplicando o teste *t-Student* com nível de confiança $p < 0,05$ (95%).

5.3.5 Combinação de classificações

Abordar o problema de monitoramento com um único sensor, inclusive utilizando a melhor estratégia de classificação, pode não ser suficiente em situações reais nas quais existem variáveis não previsíveis, tais como as atenuações que mudam os sinais conforme estes se propagam. Uma alternativa para lidar com este problema é usar as informações complementares dos sensores vizinhos. Desta forma a decisão final é a mais relevante dentre todas as opiniões dos nós.

A primeira estratégia de combinação é a votação majoritária simples (MV), ou seja, a decisão final corresponde à espécie mais comum dentre as reconhecidas pelos membros. Assim, cada i^{th} sensor atribui um voto $d_{ij} = 1$ para a classe w_j se esta for reconhecida corretamente, ou $d_{ij} = 0$ caso contrário. Portanto, para estimar a classe final, usamos a seguinte regra:

$$\arg\max_{w_j} \sum_{j=1}^{J} d_{j,L},$$

onde J é o número de classes (ou espécies) e L é o conjunto de sensores. De acordo com esta regra, devemos decidir em favor da classe w_j com maior número de votos. Esta regra é muito popular devido à sua simplicidade e robustez, mas falha quando os classificadores têm precisão diferente.

No cenário distribuído é provável que alguns sensores estejam mais próximos do animal e, consequentemente, recebam sinais de maior amplitude. Neste caso em que a

fonte sonora encontra-se próxima de único sensor, provavelmente a decisão da maioria será errada. Para superar este problema, é razoável dar mais importância, ou peso, aos classificadores que recebem o sinal mais forte. Assim, implementamos uma votação majoritária ponderada (WMV), na qual a potência do sinal recebido é utilizado como peso da votação. Assim, a saída w_j mais provável para a classe j é obtida através da seguinte equação:

$$\arg\max_{w_j} \sum_{j=1}^{J} \mathrm{PW}_i \; d_{j,L},$$

onde $\mathrm{PW}_i = \frac{1}{N}\sum_{n=1}^{N} x_i^2[n]$ é a potência do sinal x recebida pelo sensor i, N é o tamanho do *frame*.

Uma abordagem diferente para combinar os classificadores é utilizar os valores de probabilidade de saída de cada classe. Para fazer isso, usamos duas regras: a regra geométrica, com base na medida de distância de probabilidades de Kullback Leibler (KL), e a regra aritmética, utilizando uma formulação alternativa da distância KL (Theodoridis and Koutroumbas, 2008). De acordo com a regra geométrica (GV), devemos escolher a classe tal que:

$$\arg\max_{w_j} \prod_{i=1}^{L} \mathrm{P}_i(w_j|x),$$

onde P_i é a probabilidade *a posteriori* da j^{th} classe w_j calculada pelo i^{th} nó. Esta regra pode levar a resultados menos confiáveis quando a saída de alguns dos sensores resultar em valores próximos de zero. Uma solução alternativa é substituir o produto da equação anterior pelo somatório, como na seguinte equação:

$$\arg\max_{w_j} \frac{1}{L}\sum_{i=1}^{L} \mathrm{P}_i(w_j|x),$$

de acordo com a regra aritmética (AV).

Embora existam outras técnicas de combinação além das explicadas, nós escolhemos estas pela simplicidade e baixo custo computacional. Na próxima seção, explicamos como descartar casos confusos, encontrados depois de combinar as decisões individualmente, nas quais a incerteza sobre o evento subjacente é maior. Observe-se que nosso objetivo principal é encontrar a melhor técnica de combinação e compará-la com o uso de um sensor isolado.

5.3.6 Rejeição de casos confusos

A combinação (ou votação) dos sensores é realizada pelo nó líder. O resultado desta combinação é um novo vector com as probabilidades *a posteri* de cada espécie $\mathrm{P}(\mathbf{w}) = [\mathrm{P}(w_1|x_i), \mathrm{P}(w_2|x_i), \cdots \mathrm{P}(w_j|x_i)]$. Desta forma, quanto maior for a entropia $\mathrm{H}(\mathbf{w})$ deste conjunto de probabilidades, maior é a incerteza sobre a espécie identificada, i.e., quanto mais uniforme são os valores de probabilidade entre as diferentes classes, maior é a entropia. Por outro lado, quando mais concentrados são estes valores, em uma única classe, a entropia é menor indicando consenso entre as opiniões os sensores. Assim, dado um limiar apropriado é possível compará-lo com o valor de entropia e rejeitar casos confusos. A entropia neste caso é calculada como:

$$\mathrm{H} = -\sum_{j=1}^{\mathbf{w}} \mathrm{P}(w_j) \log \mathrm{P}(w_j). \tag{5.3}$$

Por exemplo, suponha que existem dois cenários $i = \{1, 2\}$, nos quais o nó líder recebe as probabilidades de cada sensor e realiza as combinações aplicando a equação 5.3.5. Após isso os vetores de probabilidades correspondentes são $\mathrm{P}(\mathbf{w}|x_1) = [0.01; 0.97; 0.02; 0]$ e $\mathrm{P}(\mathbf{w}|x_2) = [0.3; 0.3; 0.2; 0.2]$. No primeiro cenário, os sensores concordam claramente com a classe w_2, diferente do segundo cenário no qual nenhuma classe se destaca das restantes. Assim, selecionando um limiar $T_H \approx 0.5$ o primeiro caso é aceito ($\mathrm{H} = 0,06$), enquanto que o segundo é rejeitado ($\mathrm{H} = 0,57$) uma vez que a incerteza geral é maior. Ao utilizarmos este limiar, transformamos o problema de classificação com j classes em um problema binário, utilizando H para inferir se o sinal recebido foi gerado por mais de uma espécie. Desta forma, é necessário encontrar o valor de T_H que diminua a taxa de falsos negativos.

5.3.7 Resultados

Nossa primeira avaliação é comparar a decisão do comitê com um único sensor que constitui nosso *baseline*. Neste caso embarcamos os mesmo modelos de classificação em todos os sensores e aplicamos as técnicas de votação descritas na seção 5.3.5. Para avaliar como o comitê lida com os cenários confusos (mais de uma espécie vocalizando ao mesmo tempo) filtramos os de menor certeza de acordo com o valor de entropia das probabilidades *a posteriori* (equação 5.3). As comparações são apresentadas nas tabelas 5.14, na qual as taxas de erro (e) são utilizadas para avaliar o desempenho da classificação colaborativa. Cada tabela apresenta os resultados de uma técnica de classificação embarcada nos sensores membros do comitê.

As colunas da tabela 5.14 representam: RR a taxa de rejeição; IS a taxa de erro do sensor isolado; MV, WMV, GV e AV as taxas de erro das técnicas de *ensemble* no nó líder; e as colunas identificadas com G(%) representam a porcentagem de ganho na taxa de acurácia comparada com a coluna IS, i.e: a primeira coluna G(%) representa o ganho do comitê aplicando voto majoritário com o sensor isolado (IS) da mesma linha. A coluna RR representa a porcentagem de rejeição para diferentes limiares de entropia. Desta forma, comparando as linhas da tabela podemos considerar a primeira como o cenário sem rejeição e a última como a rejeição máxima permitida em nossos experimentos.

Abordando o problema desta forma lidamos um uma relação entre FP e FN, em que um aumento do limiar de rejeição será acompanhado por um aumento dos FN e uma diminuição dos FP. Assim, quanto mais restrito é o limiar de H, menor é FN, rejeitando mais eventos. Como consequência, existe uma perda na quantidade de amostras que poderiam ter sido classificadas e não foram. Por outro lado, a rejeição permite escolher os melhores eventos para serem classificados, diminuindo o número final de erros cometidos.

Quando consideramos a situação em que nenhum cenário é rejeitado (RR=0% somente primeira linha de cada tabela 5.14), todas as estratégias de combinação envolvendo o classificador QDA superaram o sensor isolado (IS). No caso do NB todas as combinações superam o IS exceto quando utilizamos MV. O pior desempenho foi alcançado pelos modelos de árvores de decisão que superam o IS em apenas dois casos. Em termos gerais, QDA obteve as menores taxas de erro dentre os três classificadores testados. Comparando as estratégias de combinação das três tabelas notamos que a AV é a melhor estratégia, e principalmente quando combinada com QDA. Também notamos que a WMV teve um bom desempenho combinada combinada com QDA.

Analisando somente a primeira linha de cada tabela podemos confirmar que a fusão de decisões melhora o desempenho final de nossa abordagem de monitoramento ambiental de anuros. No entanto, o melhor ganho obtido foi moderado (10,8% QDA com AV). Este ganho moderado, deve-se ao fato dos sensores fornecem poucas informações independentes, uma vez que utilizam os mesmos modelos de classificação, mas com entradas levemente diferentes por causa dos ruídos e das atenuações.

No passo seguinte avaliamos a utilização da entropia das estimativas como métrica para rejeitar cenários confusos (linas com RR>0%). Para alcançar este objetivo, analisamos o desempenho dos sensores para diferentes limiares de T_H provocando porcentagens de rejeição diferentes.

Como esperado, a taxa de erro do IS diminui a medida que os casos de elevada entropia são rejeitados. O mesmo comportamento é observado na maioria das estraté-

Erro utilizando QDA

RR	IS	MV	G(%)	WMV	G(%)	GV	G(%)	AV	G(%)
0%	36,9	34,6	**+8,1**	34,2	**+7,2**	34,1	**+7,7**	33,8	**+10,8**
10%	32,9	31,8	**+6,0**	30,7	**+7,8**	34,1	**-3,1**	30,1	**+9,0**
20%	29,7	28,9	**+2,4**	28,6	**+2,4**	35,1	**-22,0**	26,1	**+9,4**
30%	28,3	28,9	**+1,1**	24,4	**+15,2**	35,1	**-23,7**	23,7	**+18,7**
40%	25,1	28,9	**-11,6**	22,3	**+12,4**	36,5	**-43,4**	20,8	**+20,3**
50%	23,3	22,8	**+5,6**	20,0	**+14,2**	37,9	**-58,8**	19,7	**+18,5**

Erro utilizando NB

RR	IS	MV	G(%)	WMV	G(%)	GV	G(%)	AV	G(%)
0%	44,2	44,1	0	42,4	**+4,3**	42,5	**+4,7**	42,3	**+3,7**
10%	43,2	43,3	0	41,8	**+4,7**	42,5	+2,3	41,4	**+4,6**
20%	41,5	40,6	**+2,4**	41,4	0	41,8	0	39,6	**+4,9**
30%	39,4	40,6	**-2,6**	38,9	+2,6	41,8	**-5,2**	38,6	**+2,6**
40%	37,4	38,9	**-2,7**	37,3	0	41,0	**-10,8**	37,2	-0,4
50%	35,6	38,9	**-8,6**	36,3	-2,9	41,0	**-17,1**	35,9	0

Erro DT

RR	IS	MV	G(%)	WMV	G(%)	GV	G(%)	AV	G(%)
0%	63,2	63,5	-0,8	60,9	**+4,8**	63,9	0	60,2	**+5,0**
10%	61,0	63,3	**-3,3**	59,9	+3,3	62,6	**-1,6**	58,8	**+4,8**
20%	59,2	61,5	**-3,4**	58,5	+1,7	62,7	**-5,7**	56,9	**+4,6**
30%	52,4	58,9	**-11,7**	56,8	**-7,7**	61,1	**-17,3**	55,4	**-5,8**
40%	53,5	58,9	**-9,4**	53,4	0	61,1	**-15,1**	52,0	**+1,9**
50%	51,7	58,9	**-13,7**	50,8	+2,0	61,1	**-19,6**	50,2	**+2,0**

Tabela 5.14. Taxas de erro para os classificadores QDA, NB, e DT isoladamente (IS) ou combinadas usando MV, WMV, GV e AV. A coluna RR identifica a taxa de rejeição. A Coluna G(%) mostra os ganhos de cada técnica comparada com a coluna IS da linha correspondente. Os valores em negrito representam diferenças estatisticamente significativas ($p < 0,05$).

gias de combinação a exceção da GV, na qual a entropia não foi capaz de discriminar confusão de não confusão. Isto acontece porque na GV as probabilidades *a posteriori* tendem a ser equilibradas para baixo devido à multiplicação de valores menores que um. Consequentemente, a combinação geométrica é uma conjunção na qual se os sensores discordam todas as estimativas são diminuídas e valores semelhantes causam um aumento de H reagindo ao limiar de corte. Por outro lado, a entropia mostrou-se eficaz na filtragem de cenários confusos quando foi combinada com AV e WMV. Isto sugere que estas estratégias são melhores nos casos em que existe desacordo entre a opinião dos membros do comitê. Em tais casos, quanto maior for o número de casos rejeitados, maior é ganho do comitê comparado com um sensor isolado.

5.3.8 Conclusões sobre a classificação colaborativa

Nos experimentos relatados na seção 5.3.7 encontramos que nossa melhor estratégia de combinação alcança um ganho de 11% comparada com um sensor isolado, e descobrimos também que o comitê do sensor é capaz de identificar de forma eficaz cenários confusos, aumentando os ganhos sobre o sensor isolado cerca de 20%. Com os resultados obtidos verificamos que o modelo QDA combinado com a estratégia de votação aritmética supera o sensor isolado com um ganho aproximado de 11%. Verificamos também que, ao aplicar entropia nas estimativas de espécies (ou decisões finais do sensor líder) é possível identificar eficazmente cenários confusos obtendo ganhos aproximados a 20% sobre IS com rejeição. Estes resultados foram publicados em Colonna et al. (2014a).

5.4 Considerações finais

Neste capítulo propusemos e avaliamos uma estratégia nova de validação cruzada por indivíduos, exclusivamente adaptada ao contexto de monitoramento bioacústico. A partir dos resultados comprovamos que a k-CV tradicional causa uma superestimação da acurácia dos métodos de reconhecimento. Portanto, verificamos que a validação cruzada baseada em espécimes é a melhor maneira de testar as capacidades de generalização de modelos de reconhecimento. Além disso, mostramos a importância de utilizar as Macro-métricas ao invés das tradicionais Micro-métricas, quando o problema de reconhecimento envolve detecção de eventos acústicos provenientes de diferentes fontes sonoras. Mostramos também as vantagens de decompor e simplificar problemas multi-classes mediante dois tipos de abordagens binárias. Estes resultados foram apresentados na *Conferencia de la Asociación Española para la Inteligencia Artificial*, (CAEPIA 2016), e publicados no *Lecture Notes in Computer Science* (Colonna et al., 2016a). Este trabalho apresenta um visão crítica e deixa uma recomendação para futuros autores de aplicações de índole similar.

Após os experimentos com classificadores planos identificamos as limitações que estes proporcionam para entender as relações entre a taxonomia das diferentes espécies de anuros e seus descritores espectrais de baixo nível. Com o objetivo de entender tais relações de forma mais aprofundada, desenvolvemos duas abordagens de reconhecimento hierárquica que permitem identificar a família, o gênero e a espécie à qual pertence cada amostra. Assim, propomos e avaliamos dois métodos *multi-output* (LCPN e LCPL), concluindo que é possível obter resultados satisfatórios, prevendo o mesmo conjunto de espécies que um classificador plano, e adicionalmente ganhar entendimento sobre o problema. A redução do espaço de decisões também permitiu estabelecer quias

espécies causariam menos erros, e portanto, deveriam ser escolhidas nos programas de monitoramento ambiental bioacústico automáticos com RSSF. Futuramente, técnicas de correção de erro ou classificadores hierárquicos baseados em probabilidades com capacidade de voltar e corrigir a decisão de cada nível deveriam ser investigados e avaliados. Os principais resultados formam apresentados na conferencia *Discovery Science*, 2016, e publicados no *Lecture Notes in Computer Science*.

Nos últimos experimentos apresentados neste capítulo avaliamos quatro estratégias colaborativas diferentes para reconhecer espécies de anuros utilizando um *cluster* de sensores. Além das técnicas de combinação utilizadas para criar a "fusão das decisões" dos sensores membros do comitê, avaliamos o desempenho de três modelos de classificação diferentes e mais uma estratégia de rejeição de casos pouco confiáveis. Para avaliar estes métodos consideramos situações em que uma ou duas espécies de anuros diferentes podem vocalizar ao mesmo tempo (cenários confusos). Com os resultado obtidos mostramos que o comitê de sensores é capas de descartar tais cenários e diminuir o número de falsos positivos da rede.

Uma extensão do método colaborativo seria investigar como os membros do comitê poderiam ser escolhidos dinamicamente ou estudar como a aplicação de métodos de *ensemble* avançados, tais como *bagging*, podem melhor a fusão das decisões. Os resultados obtidos forma apresentados na conferência *International Conference on Pattern Recognition*, 2014, e econtram-se disponíveis nos *Proceedings* da IEEE (Colonna et al., 2014a).

6

Aprimoramento de sinais bioacústicos

Os sons captados pelos sensores acústicos não são sinais "limpos". Estes apresentam diversos ruídos de fundo, causados por fenômenos climáticos naturais ou pela presença de outras espécies distantes. Isto degrada a qualidade dos sinais e se faz necessário aplicar uma técnica para filtrar tais ruídos das gravações. A eliminação dos ruídos, têm por objetivo melhorar os sinais de forma que as espécies sejam facilmente distinguíveis por um especialista ou pelo sistema ACR de reconhecimento automático.

Neste capítulo aplicamos uma técnica de análise e síntese para os sinais bioacústicos conhecida como *Singular Spectrum Analysis* (SSA). Este método nos permite decompor os sinais em diferentes componentes oscilatórias que representam as principais bandas de frequências dos sinais originais. Assim, é possível identificar a composição espectral das gravações, separar diferentes tipos de ruídos, sejam estes branco ou coloridos, criar filtros adaptáveis para cada uma das espécies estudadas e sintetizar o sinal original evitando os componentes não desejados.

Na primeira parte deste capítulo, apresentamos um critério de seleção automático dos componentes principais (PCs) do SSA baseado na teoria da informação, utilizando diferentes quantificadores de entropia. Isto permite automatizar o SSA, separar ruídos de fundo, que não são necessariamente brancos, das vocalizações e entender a acústica natural do lugar onde as gravações foram obtidas. Na segunda parte deste capítulo, apresentamos uma versão robusta do SSA, que chamamos *Robust Singular Spectrum Analysis* (RSSA). Esta nova versão é menos sensível as variações das amplitudes dos ruídos, permite identificar as diferentes contribuições dos PCs e reconhecer facilmente o ruído ambiental de baixa frequência.

6.1 Introdução

O ambiente natural da floresta possui uma riqueza acústica elevada, produto da biodiversidade (Gasc et al., 2013). Os ruídos de fundo nas gravações das vocalizações são comuns, devido às variações climáticas dinâmicas causadas por chuvas frequentes, ventos ou outros fenômenos naturais. Por estes motivos, nas vocalizações dos anuros gravadas *in situ*, os sinais nunca são "limpos". Sendo assim, é necessário aplicar uma técnica de filtragem. A filtragem dos ruídos pode: (1) melhorar a qualidade das gravações e (2) aumentar a acurácia do reconhecimento das espécies (Cai et al., 2007).

Quando os anuros são gravados na floresta, os ruídos ambientais e as vocalizações de outras espécies causam modificações no sinal original. Além disso, desde a emissão do canto até a captação pelo microfone, o som viaja pelo ar sofrendo atenuações e distorções. Neste caso, o ar da floresta é o canal de transmissão. Assim, o sinal recebido pode ser modelado como:

$$y = h_c * x + \xi, \tag{6.1}$$

onde $*$ denota a operação de convolução entre a resposta impulsiva do canal h_c e o sinal original x, e ξ são os ruídos aleatórios aditivos não correlacionados. O canal, pode ser modelado por um filtro com resposta em frequência H_{fc}, chamada função de transferência. No domínio espectral, o efeito do canal de transmissão resultante é $Y_f = H_{fc}X_f + \xi_f$. Assim, conhecendo a função de transferência e os sinais recebidos no nó sensor, é possível recuperar o sinal original aplicando a função inversa do filtro $H_{fc}^{-1}(Y_f - \xi_f) = \hat{X}_f$, onde $\hat{X}_f$ é o sinal estimado livre de ruído e dos efeitos do canal.

O modelo teórico apresentado na equação 6.1 é geralmente simplificado nas abordagens acústicas, assumindo-se que o canal de transmissão é causal e linear, e que somente os ruídos aditivos degradam consideravelmente a qualidade do sinal recebido. O modelo simplificado resulta:

$$y = x + \xi. \tag{6.2}$$

Desta forma, a resposta do filtro aplicado para eliminar os ruídos no domínio espectral resulta simplesmente:

$$\begin{aligned} \hat{X}_f &= H_f^{-1} Y_f \\ &= H_f^{-1}\big(X_f + \xi_f\big), \end{aligned} \tag{6.3}$$

sendo Y_f o sinal recebido pelo sensor e $\hat{X}_f$ o sinal estimado. Assim, o objetivo do filtro H_f é aproximar $\hat{X}_f \to X_f$, preservando as frequências fundamentais do sinal original.

Ao se tentar minimizar a diferença entre o sinal filtrado e o sinal original, implicitamente são minimizados os efeitos do canal de transmissão. No domínio temporal a filtragem bioacústica é equivalente a operação de convolução:

$$\begin{aligned} \hat{x} &= h * y \\ &= h * \big(x + \xi\big). \end{aligned} \tag{6.4}$$

Os cenários dinâmicos da floresta impossibilitam estabelecer uma representação única e invariante no tempo de y. Em outras palavras, a cada novo sinal recebido no nó sensor a variável aleatória ξ é uma nova realização independente da anterior e do sinal original x. No entanto, é possível assumir que em períodos curtos de tempo, tais como os *frames*, os sinais e os ruídos apresentem caraterísticas de variáveis aleatórias estacionárias (Rabiner and Schafer, 2007). No sentido amplo, a estacionariedade significa que os momentos que caracterizam a função de distribuição de probabilidades permanecem invariantes dentro do *frame*. Estes são os momentos de primeira e segunda ordem, que na prática não são triviais de verificar.

A teoria de processamento digital de sinais estabelece diferentes métodos para filtrar ruídos aleatórios (Cai et al., 2007, Yan et al., 2006, Gur and Niezrecki, 2011). Entretanto, os métodos que apresentam melhores resultados são geralmente baseados em diferentes transformações dos sinais (*Transform-Based Processing*[1]), dentre as quais as melhores são *Wavelet* ($\mathcal{W}$), *Fourier* ($\mathcal{F}$) e métodos de subespaços. O *Singular Spectrum Analysis* (SSA), adotado por nós como o método principal de estudo neste capítulo, é um abordagem específica da teoria de subespaços que permite lidar com séries temporais (sinais de uma única dimensão). Assim, neste capítulo investigamos a aplicação do SSA no domínio específico bioacústico e o comparamos com o *soft threshold* da transformada *Wavelet*.

Em nossa proposta para filtrar as vocalizações dos anuros, focamos na utilização da teoria conhecida como *Signal Subspace* (Van Der Veen et al., 1993, Ephraim and Van Trees, 1995, Tan et al., 2007, Tomé et al., 2010). Esta teoria aplica os conceitos da representação dos sinais em autovetores e autovalores (ou componentes principais - PCs) para encontrar as bases da transformação que concentram a maior variância do sinal original (Vaseghi, 2008). Particularmente, para sinais de uma dimensão, como é o caso de sinais de áudio, existe a metodologia de análise conhecida como SSA (Tomé et al., 2011). Este método permite obter as funções de filtros chamadas *eigenfilters*, as quais: (a) adaptam-se ao sinal de entrada, (b) são não paramétricas, e (c) selecionam automaticamente as bandas de frequências dos sinais com maior concentração de

[1] Um conjunto maior de categorias foi definido por (Vaseghi, 2008).

energia espectral (Tomé et al., 2010).

O SSA utiliza a decomposição em valores singulares (SVD) da matriz de autocorrelações do sinal, para criar as bases ortogonais da transformada (seção 2.4, página 39). Nesta teoria, os termos transformação, decomposição ou projeção dos sinais são utilizados como sinônimos. A decomposição SSA permite separar os PCs com diferentes frequências. Tais componentes são ordenadas de acordo com a sua contribuição, que neste caso é a energia espectral de cada um.

A reconstrução final (sinal filtrado) é realizada escolhendo os PCs que se correspondem com os autovalores de maior contribuição, isto é, aqueles autovetores que concentram a maior porcentagem de energia do sinal original. Embora na literatura existam diferentes abordagens para escolher o componentes que farão parte reconstrução, esta escolha ainda é um desafio que depende da natureza dos sinais analisados. Em nossa abordagem, descrita na seção 6.7, apresentamos uma regra para escolher os PCs aplicando diferentes quantificadores de entropia.

Com nossa abordagem pretendemos:

- melhorar a qualidade das gravações sem causar distorções que impeçam ao especialista estudar e comparar espécies que habitam numa determinada região ou calcular índices acústicos para compreender a biodiversidade;

- caracterizar o ruído de fundo da floresta e seu impacto na qualidade dos sinais;

- desenvolver um método adaptável as diferentes espécies e condições do ruído ambiente não supervisionado e não paramétrico; e

- melhorar as técnicas de monitoramento ambiental, evitando erros de classificação causados pelos ruídos de fundo.

A primeira parte deste capítulo apresenta uma metodologia de estudo e caracterização de séries temporais bioacústicas combinando duas ferramentas de análises: a Teoria da Informação e SSA. O objetivo principal de nossa análise é identificar os PCs das vocalizações com menor entropia e utilizá-los para criar versões reconstruídas das vocalizações menos ruidosas. Neste caso, estudamos e comparamos três metodologias de cálculo de entropia: a entropia temporal H_t, a entropia espectral H_f e a recentemente desenvolvida entropia das permutações PE. Propomos um critério novo não supervisionado para escolher os PCs que farão parte da reconstrução dos sinais utilizando os valores de entropia. Este critério pode ser aplicado individualmente ou também pode ser combinado com critérios existentes que utilizam o peso dos autovalores.

Na segunda parte deste capítulo, apresentamos uma versão nova e robusta do método original SSA, menos resiliente aos ruídos ambientais. Esta melhoria chama-se *Robust Singular Spectrum Aanalysis* (RSSA). O RSSA apresenta uma organização diferente dos autovalores ilustrados no espectro singular (*singular spectrum*), mostrando que existe outro grau de contribuição de cada um deles, quando se assume um modelo de autocorrelação menos sensível as mudanças de amplitude dos componentes. O método proposto facilita também descobrir os componentes de ruído ambiental de baixa frequência ou o *trend* dos sinais.

Em nossos experimentos, utilizamos diferentes tipos de ruídos, incluindo ruído branco, quatro tipos de ruídos coloridos, ruído impulsivo e *outliers* gerados a partir de uma distribuição de probabilidades de *Cauchy*. Para todos os casos, avaliamos diferentes níveis de ruído e combinação de bases da reconstrução com SSA e com RSSA. Os detalhes e resultados deste novo método encontram-se na seção 6.9

Por último, comparamos duas técnicas de filtragem, nossa proposta baseada no SSA com critério de entropia com o *baseline* baseado na transformada DWT. O objetivo aqui é identificar o método mais apropriado para nossa abordagem de reconhecimento bioacústico de anuros.

6.2 Motivação de escolha dos métodos de filtragem

A partir da revisões do capítulo 3 identificamos duas técnicas de filtragem frequentemente aplicadas no domínio bioacústico. A maioria das abordagens de reconhecimento de espécies confiam que é possível evitar a filtragem, desde que os descritores acústicos sejam robustos aos diferentes ruído ambientais. Portanto, uma técnica de filtragem torna-se relevante para analisar os ruídos de fundo que compõem as gravações e não somente as vocalizações. No capítulo 4, provamos que a segmentação é sensível aos diferentes tipos de ruídos, sejam estes brancos ou coloridos. Consequentemente, consideramos que é necessário incluir uma etapa de filtro de sinal em nosso ACR para evitar falsos positivos na segmentação. No entanto, a filtragem lida com um *trade-off* entre a quantidade de ruído eliminado e o erro ou distorção na reconstrução das vocalizações. Assim, uma boa técnica de filtragem deveria ser capaz de adaptar-se às diferentes vocalizações e cenários, de forma a maximizar este *trade-off*.

As técnicas de filtragem tradicionais incluem três tipos de filtros: passa-baixa, passa-banda ou passa-alta, que são utilizados para rejeitar ou aceitar bandas específicas de frequências dos sinais. Tais filtros são definidos pelas frequências de corte e não são adaptáveis (Proakis and Manolakis, 1996). Embora estes métodos sejam simples

de implementar, definir tais frequências sem excluir informações úteis para o reconhecimento de espécies é o desafio principal. Como pode ser observado na tabela 4.1, a maioria das espécies estudadas possuem sobreposição nas bandas de frequências, indicando que escolher um tipo de filtro não seria apropriado para todas as espécies, e portanto deveria ser definido um filtro específico para cada uma delas. Além disso, o ruído contido na mesma banda de frequência da vocalização não será removido.

A dinâmica de nossa aplicação requer um filtro que possa adaptar-se às diferentes condições sonoras da floresta e às vocalizações das diferentes espécies animais de forma não supervisionada. As técnicas adaptativas e robustas que aplicam transformações dos sinais incluem: *Spectral Mean Subtraction* (SMS) (Boll, 1979), o filtro de Wiener (Chen et al., 2006a), os filtros de Ephraim and Malah (1984), Ephraim and Van Trees (1995), a decomposição em PCs aplicando SSA (Tomé et al., 2010, 2011), filtragem utilizando a transformada de *Karhunen-Loève* (KLT) (Hermus et al., 2007), *soft* e *hard threshold* utilizando a transformada *Wavelet* (DWT) (Donoho, 1995) e a transformada EMD (Kopsinis and McLaughlin, 2009). Nos trabalhos de Ren et al. (2008) e Gur and Niezrecki (2011), algumas destas técnicas foram aplicadas aos problemas de filtrar vocalizações de aves, macacos, baleias e peixes-boi. No entanto, a maioria dessas técnicas foram desenvolvidas com propósitos demasiado gerais ou estritamente específicos, *e.g.* filtrar somente fala humana, sendo pouco apropriadas para sinais bioacústicos.

Dentre as técnicas apresentadas no capítulo 3 identificamos que as mais utilizadas em problemas bioacústicos são SMS (Cai et al., 2007), DWT (Gur and Niezrecki, 2011) e EMD (Kopsinis and McLaughlin, 2009). Uma filtragem mais agressiva para poder eliminar os ruídos da floresta com SMS causa atenuações nos sinais e aparecem "artefatos" (ou distorções) nas altas frequências (seção 2.3.6). O método que utiliza EMD é computacionalmente custoso devido ao fato de utilizar interpolação cúbica para estimar a envoltória superior e inferior dos sinais. A DWT é a melhor opção, embora para poder aplicá-la com sucesso precisa-se ajustar diversos parâmetros, o que dificulta encontrar uma combinação que seja ótima para todos os casos. Por exemplo, deve-se determinar a função base da transformada, a quantidade de níveis da decomposição, e os limiares do filtro para cada nível. Entretanto, a DWT possui um esquema de baixo custo de processamento executável em tempo real. Portanto, avaliaremos esta transformada e a compararemos com nossa proposta utilizando SSA.

6.3 Comparação entre SSA, SMS e DWT

Além dos motivos explicados na seção anterior, apresentamos aqui três exemplos para comparação visual entre os resultados dos métodos SSA, SMS e DWT em uma vocalização da espécie *Adenomera hylaedactyla*.

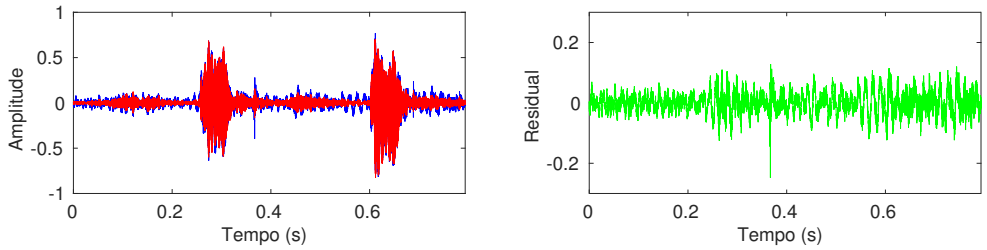
A figura 6.1(a) ilustra a aplicação do método de filtragem proposto que veremos adiante na seção 6.7. Neste exemplo, utilizamos o filtro SSA com o critério da entropia temporal (H_t) para escolher os PCs. A primeira figura 6.1(a) ilustra o sinal original e sua reconstrução, na figura 6.1(b) o sinal foi contaminado com ruído de alta frequência, e por último, na figura 6.1(c) apresenta-se o sinal mais um ruído de baixa frequência. Nos três casos, ilustra-se o sinal contaminado y (azul) e sua versão filtrada $\hat{x}$ (vermelho). As contaminações foram geradas com uma relação SNR $= 3\,\mathrm{dB}$. A coluna da esquerda ilustra o residual obtido como $r = \hat{x} - y$. Graficamente observamos que este método produz ótimos resultados, inclusive nos casos com elevada contaminação.

As figuras 6.2 e 6.3 são dois exemplos dos efeitos da filtragem utilizando os métodos de substração espectral (SMS) e a transformada Wavelet (DWT) com *soft-threshold*. Estas figuras podem ser comparadas visualmente com os exemplos apresentado na seção 6.7.1 (página 172, figura 6.1). As sub-figuras ilustram o sinal contaminado (azul), o sinal filtrado com a técnica respectiva (vermelho) e o residual (verde).

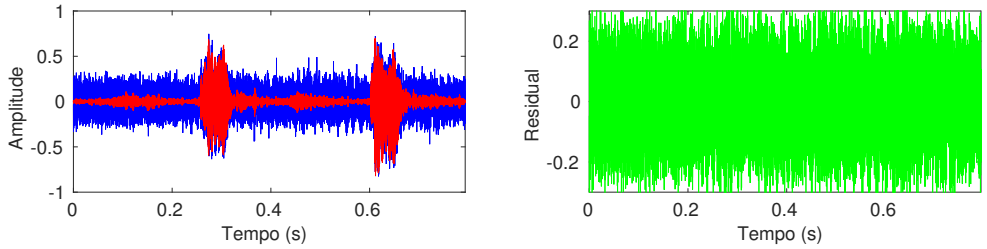
Nos caso do SMS (figura 6.2), notamos que o residual possui uma fração elevada da energia das sílabas, e que o sinal filtrado (vermelho) tem uma distorção elevada comparado ao original (figura 6.2(a)). Adicionalmente, quando a contaminação é causada por um ruído vermelho (figura 4.7), com o SMS aparecem artefatos desconhecidos no sinal filtrado. Portanto, este método não é recomendável neste tipo de cenário. No caso do filtro DWT com *soft-threshold* (figuras 6.3), observamos que a distorção das sílabas foi mínima, mas os componentes de ruído ambiental de baixa frequência permaneceram nos sinais filtrados. No caso extremo com contaminação por ruído vermelho, a DWT não foi capaz de restaurar a forma do sinal original.

6.4 Evidências empíricas e motivação de escolha do SSA

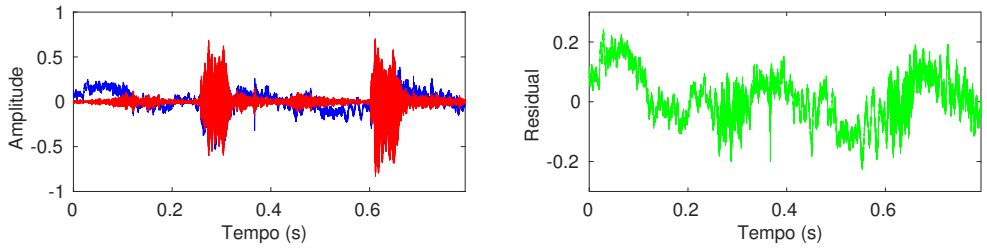
Diferente da transformada de *Fourier*, que permite decompor um sinal em uma combinação linear de funções harmônicas com amplitude constante, o SSA decompõe os sinais em funções oscilatórias. A vantagem das funções oscilatórias do SSA é que estas agrupam conjuntos de frequências fundamentais e seus harmônicos relacionados em um



(a) Gravação original.



(b) Gravação original mais ruído violeta.



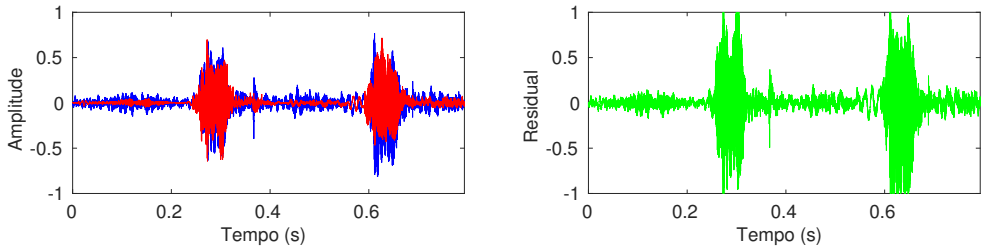
(c) Gravação original mais ruído vermelho.

Figura 6.1. Na coluna da esquerda: (a) gravações originais e versões contaminadas com diferentes ruídos (azul) e suas versões filtradas (vermelho) aplicando SSA. Na coluna da direita o residual de cada caso.

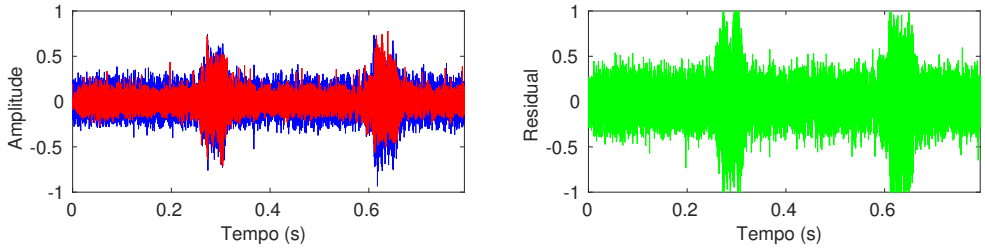
único componente. Além disso, as funções oscilatórias possuem amplitude variável no tempo, semelhantes a um sinal de amplitude modulada. Desta forma, a representação dos áudios mediante SSA é mais compacta, isto é, o sinal completo pode ser descrito por um conjunto menor de componentes.

Ao realizar experimentos prévios, tais como os apresentados nas comparações das figuras 6.1, 6.2 e 6.3, percebemos que alguns PCs extraídos dos registros bioacústicos pelo SSA, podem descrever ruídos de baixa frequência, semelhantes aos ruídos de fundo da floresta (ver residual na figura 6.1(a)). Os componentes de baixa frequência, e seus autovalores associados são classificados entre os primeiros lugares do gráfico de espectro singular se tiverem uma variância elevada (ou energia). Apesar da quantidade de variância retida por eles, estes devem ser evitados durante a etapa de reconstrução.

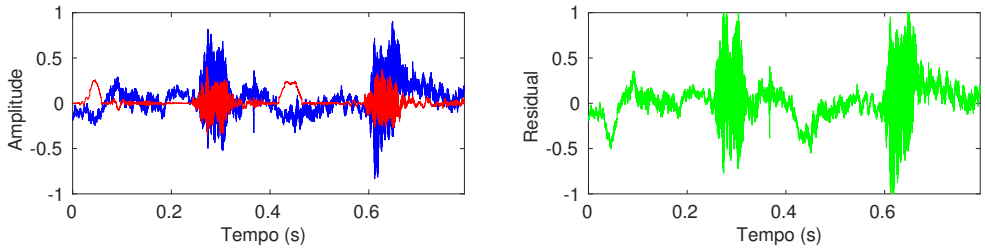
A figura 6.4 apresenta o espectro singular normalizado de três registros de chama-



(a) Gravação original.



(b) Gravação original mais ruído violeta.

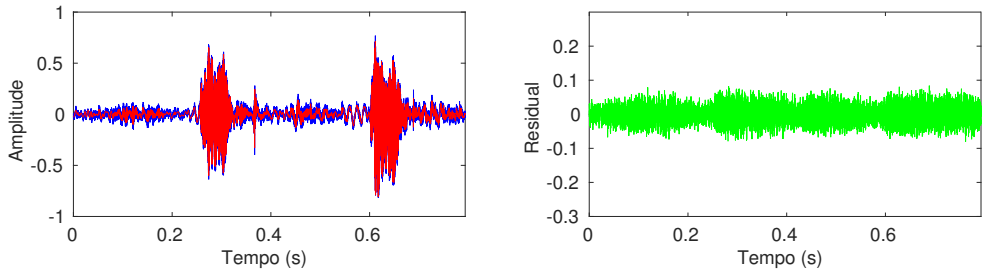


(c) Gravação original mais ruído vermelho.

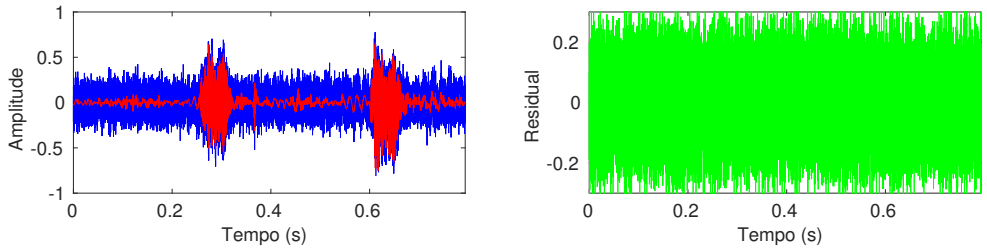
Figura 6.2. Na coluna da esquerda: (a) gravações originais e versões contaminadas com diferentes ruídos (azul) e suas versões filtradas (vermelho) aplicando *Spectral Mean Substraction*. Na coluna da direita o residual de cada caso.

das pertencentes às espécies de anuros Adenomera hylaedactyla, Aplastodiscus perviridis e Hyla minuta, a partir de uma decomposição realizada com $L = 22$. No espectro singular das espécies Adenomera hylaedactyla e Hyla minuta, podemos visualizar e interpretar graficamente quatro possíveis agrupamentos para seus PCs, estes são: $\lambda_{1:2}$, $\lambda_{3:4}$, λ_5 e $\lambda_{6:22}$. Este último agrupamento ($\lambda_{6:22}$) inclui todos os autovalores menores ao valor médio do espectro singular $\overline{\Lambda}$, e portanto são os menos relevantes para obter um reconstrução aproximada $\hat{x} \to x$. Já no espectro singular da espécie Aplastodiscus perviridis, notamos um agrupamento diferente, pelo fato de seus autovalores se encontrarem mais separados verticalmente, os possíveis grupos são $\lambda_{1:2}$, $\lambda_{3:4}$ e $\lambda_{5:22}$.

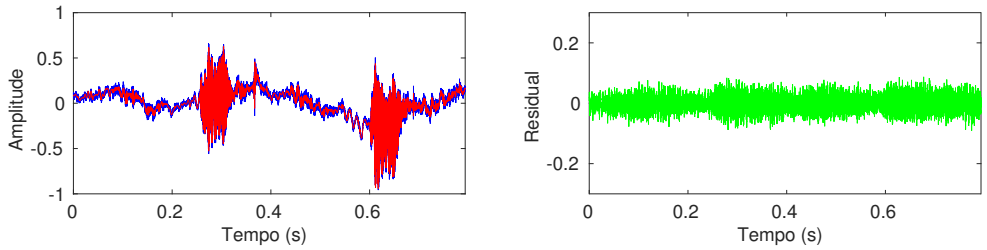
Após a decomposição, a reconstrução pode ser realizada utilizando componentes individuais ou agrupamentos de componentes conforme explicado na seção 2.4 (página 39). Por exemplo, as reconstruções usando apenas o componente λ_5 nas vocali-



(a) Gravação original.



(b) Gravação original mais ruído violeta.



(c) Gravação original mais ruído vermelho.

Figura 6.3. Na coluna da esquerda: (a) gravações originais e versões contaminadas com diferentes ruídos (azul) e suas versões filtradas (vermelho) aplicando DWT com *soft-threshold*. Na coluna da direita o residual de cada caso.

zações das espécies Adenomera hylaedactyla e Hyla minuta são ilustrados por por uma linha preta nas figuras 6.5(a) e 6.5(b). No caso da espécie Aplastodiscus perviridis o componente λ_3 é ilustrado na figura 6.5(c). Como pode-se observar, estes componentes estão presentes tanto na sílabas quanto fora destas, indicando que trata-se de um ruído ambiental com elevada energia que causa distorções nas sílabas. Neste caso, o SSA capturou com exatidão a função oscilatória do ruído. Estas observações são reforçadas pela densidade espectral de potencia (PSD) desses componentes ilustrada na figura 4.7, a qual segue uma lei aproximadamente exponencial proporcional ao ruído vermelho teórico.

A partir destes exemplos, podemos concluir que com SSA é possível separar dos sinais bioacústicos de forma precisa, os componentes espectrais principais e seus harmô-

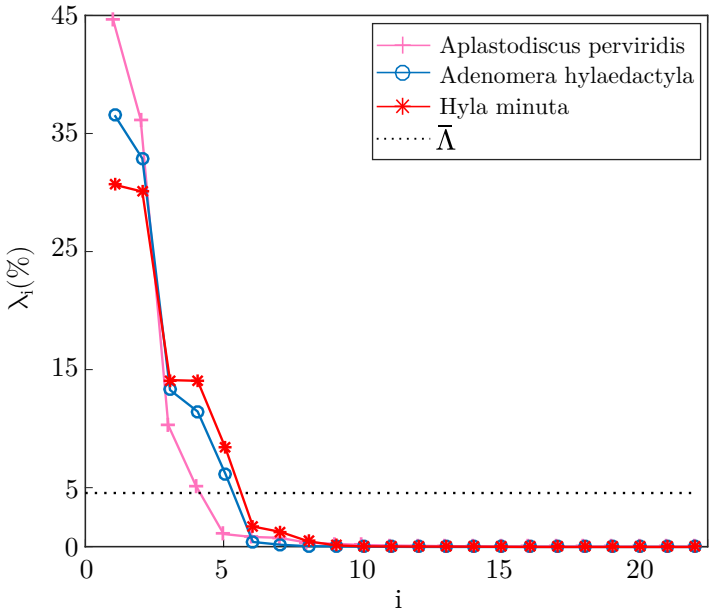
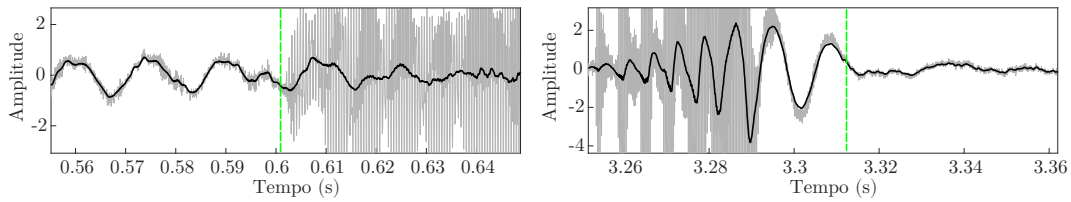
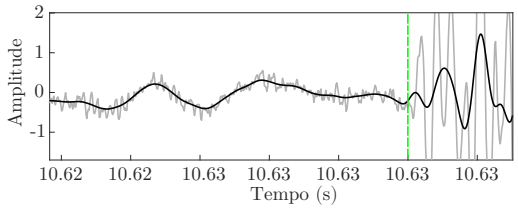


Figura 6.4. Espetros singulares das espécies Adenomera hylaedactyla, Aplastodiscus perviridis, e Hyla minuta representados pelo porcentagem retida da norma da matriz de trajetórias. A média dos autovalores é ilustrada pela linha tracejada preta.



(a) Adenomera hylaedactyla. (b) Hyla minuta.



(c) Aplastodiscus perviridis.

Figura 6.5. Em cinza encontram-se fragmentos de vocalizações das espécies Adenomera hylaedactyla, Hyla minuta e Aplastodiscus perviridis. As linhas pretas ilustram reconstruções utilizando somente os PCs λ_5, λ_5, λ_3 respectivamente. As linhas verdes verticais representam o início ou fim das silabas.

nicos em funções oscilatórias. Além disso, o SSA possibilita realizar uma avaliação visual da contribuição de cada componente. Desta foram, o SSA é útil para analisar tanto as vocalizações quanto o som ambiental.

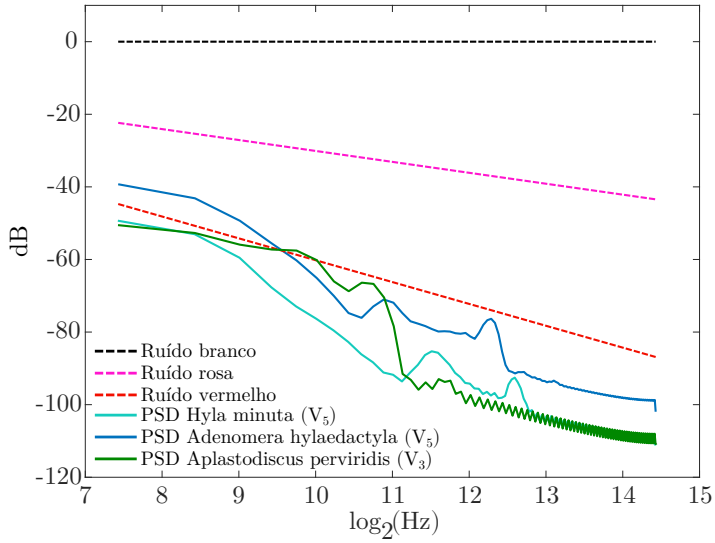


Figura 6.6. Densidade espectral de potencia dos componentes ilustrados na figura 6.5 comparados às curvas teóricas de ruído.

6.5 Descrição do problema de filtragem

A presença de ruídos aleatórios é um dos maiores problemas nos estudos ambientais que utilizam sinais bioacústicos. Os ruídos afetam a qualidade dos sinais coletados e dificultam a análise e o reconhecimento das espécies. Dependendo da localização geográfica do lugar, existem diferentes tipos de ruídos ambientais que alteram o sinal enquanto este é transmitido. O modelo de ruído aditivo considera que o sinal recebido como a soma do sinal original mais os ruídos aleatórios, i.e. $y = x + \xi$ no qual y representa as vocalizações dos anuros em nossa base de dados.

O problema de filtragem é estimar x a partir de y eliminando a maior parte de ξ, sem afetar a qualidade do sinal ou as bandas de frequências principais necessárias para o classificador realizar o reconhecimento. Podemos assumir sem perda de generalidade que as variáveis acústicas são Gaussianas com função densidade de probabilidade $\mathcal{N}(\mu, \sigma^2)$. O ruído ξ pode ser considerado uma variável Gaussiana de média zero decorrelacionado do sinal x.

Se $\hat{x}$ é uma versão filtrada de x, então o objetivo de filtrar ou recuperar o sinal original pode ser expresso em termos do erro quadrático médio (MSE) da forma:

$$\text{MSE} = \frac{\|x - \hat{x}\|_2^2}{N}, \tag{6.5}$$

onde N é o comprimento do sinal. Um valor de MSE baixo significa uma reconstrução melhor, sendo $\hat{x} \approx x$ $(\xi \to 0)$.

Este objetivo também pode ser expresso em termos da distorção. Neste caso, tenta-se diminuir a taxa de distorção da reconstrução em relação ao sinal original, da forma:

$$\text{SDR} = 20 \log_{10} \frac{\|\hat{x}\|_2}{\|\hat{x} - x\|_2}, \tag{6.6}$$

sendo $\|\cdot\|_2$ a norma L2. Assim, quanto menor é o SDR, maior é a distorção quantificada pela relação entre o sinal filtrado $\hat{x}$ e o residual $\hat{x} - x$.

Portanto, o objetivo específico neste capítulo é recuperar $\hat{x}$ a partir de $x+\xi$ usando SSA, para minimizar o MSE e maximizar o SDR quando ξ não é apenas branco, mas também colorido.

Do ponto de vista da teoria de sistemas lineares é invariantes no tempo (LTI), nosso objetivo é encontrar os coeficientes de um filtro discreto $h[n]$ com Resposta de Impulso Finito (FIR) definido pela operação de convolução:

$$\begin{aligned} \hat{x} &= y * h \\ &= \big(x + n\big) * h, \end{aligned} \tag{6.7}$$

tal que $\hat{x}$ seja um a versão filtrada de x sem a necessidade da aplicar uma transformação explicita no sinal y.

No domínio espectral, o problema consiste em separar as frequências das vocalizações considerando as restantes como ruído. As vocalizações, ao contrário dos ruídos, possuem um espectro de frequências mais rico com diferentes formas espectrais e bandas de frequências com maior concentração de energia. Em nosso caso, o ruído é o "barulho de fundo" contido nas gravações, e portanto, possui características e padrões que o diferenciam do ruído branco. Este modelo de ruído branco é geralmente adotado, por possuir energia uniforme em todo o espectro de frequências. Entretanto, para representar as variáveis acústicas exitem os modelos de ruídos coloridos (seção 2.3.1, página 2.3.1), os quais possuem um aumento ou decaimento exponencial da energia espectral conforme as frequências variam. Assim, assumindo uma contaminação de ruído colorido seria possível representar melhor a diversidade de cenários acústicos que a floresta pode apresentar.

6.6 Problema de seleção dos subespaços do sinal

Nos métodos de *Signal Subspace*, o sinal y é representado como uma combinação linear de m_i $(1 \leq i \leq p)$ funções bases ortonormais da forma $\mathbf{y} = M\mathbf{s}$, onde a matriz M, de tamanho $q \times p$ com $p < q$, contem a funções bases e $\mathbf{s}$ é um vetor coluna (Hermus

et al., 2007). Essencialmente, esta representação matricial gera um espaço vetorial de q dimensões no qual os vetores com sinal x ocupam $p < q$ bases, enquanto que os vetores com ruído se expandem ocupando q bases. Em outras palavras, o espaço q-dimensional das bases é subdividido em um espaço de p dimensões ocupado pelo sinal mais um porcentagem da variância dos ruídos e um espaço de $q-p$ dimensões ocupado somente pelos ruídos.

O problema específico de filtrar sinais utilizando as técnicas de *Signal Subspace* consiste em identificar as bases do subespaço p para reconstruir o sinal eliminando a maior porcentagem dos ruídos. Assim o procedimento pode ser descrito pelas tarefas específicas de:

1. criar as bases do espaço q e decompor o sinal original;
2. separar os subespaços p e $q-p$, do sinal e dos ruídos, respectivamente; e
3. recuperar o sinal $\hat{x}$ utilizando para a reconstrução a projeção de x nas bases do subespaço p.

Com este procedimento, podemos identificar dois desafios: o primeiro relacionado com a decomposição do sinal nas bases q e, o segundo, com a determinação do subespaço de reconstrução p. Para o primeiro desafio, adotamos a técnica *Singular Spectrum Analysis* que aplica a decomposição em valores singulares (SVD) da matriz de auto-correlações, para encontrar as bases do espaço q. Para o segundo desafio, utilizamos a entropia das projeções de y, para identificar as bases $q-p$ com comportamentos menos determinístico.

Um exemplo dos problemas ocasionados pelos critérios tradicionais de escolha dos PCs, foi ilustrado nas figuras 6.4 e 6.5. No espectro singular da espécie Adenomera hylaedactyla, o quinto componente possui um autovalor maior que a media dos autovalores ($\lambda_5 \geq \overline{\Lambda}$). Portanto, se aplicarmos alguns critérios de agrupamento tradicionais como "manter os PCs maiores que a média" ou "reter 95% da variância total", o componente ruidoso λ_5 estará presente no sinal reconstruído. Assim, nossa abordagem visa identificar esses componentes usando um quantificador de entropia, para descartá-los do agrupamento antes da reconstrução final.

6.7 Metodologia de filtragem com SSA

O SSA é capaz de descobrir os componentes que contém as principais dinâmicas da série. Todavia, algumas dinâmicas somente podem ser descritas pela combinação de

bases semelhantes (Teixeira et al., 2005). Combinar as bases utilizando técnicas de agrupamento, como por exemplo *k-Means*, não revela as estruturas dos PCs, sejam estas determinísticas ou estocásticas, nem permite extrair informações sobre a complexidade da série. A entropia das permutações (PE, seção 2.2.5), por exemplo, é uma ferramente útil para descrever as estruturas internas (ou correlações) e também para quantificar o grau de aleatoriedade das séries. Portanto, é interessante combinar SSA e PE para analisar e interpretar os PCs da série.

A decomposição da série x aplicando SSA (seção 2.4) requer somente o parâmetro L, mas a reconstrução, requer aplicar um critério mais elaborado para escolher o subconjunto de autovetores. Para a escolha de L, nós propomos uma regra empírica (seção 6.7.2). Em nossa aplicação, o problema de filtragem é definido como um agrupamento não supervisionado de duas classes, "sinal" ou "ruído".

A escolha das bases que formam a reconstrução equivale a aplicar um operador de seleção (P) que anule as colunas com maior quantidade de ruído da matriz V. Neste caso, P é uma matriz identidade que contem uns ("1") $p_{m,m} = \{1\}$ na diagonal principal, mas somente nas linhas e colunas m, selecionando às bases vetoriais que serão mantidas na reconstrução. Consequentemente, os valores da diagonal principal correspondestes as linhas e colunas das bases que serão descartadas são zero $p_{m,m} = \{0\}$. Lembrando que, a decomposição da matriz de trajetórias é uma projeção de X nas bases U (equação 2.40, página 42), e incorporando P, cada coluna de V pode ser mantida ou anulada antes da reconstrução (equação 2.42, página 42).

A reconstrução também pode ser completa, isto é, utilizando todos os componentes SSA. Assim, uma questão fundamental é como escolher tais componentes para obter uma melhor reconstrução, ou como separar e agrupar esses componentes, a fim de interpretar algumas particularidades do sinal original. Os quantificadores de entropia apresentados na seção 2.2 serão utilizados para este propósito.

6.7.1 Regra de filtragem proposta

Durante o estágio de decomposição do SSA, os vetores defasados de X são projetados no subespaço de bases ortogonais U para obter os vetores V. Assim, cada coluna da matriz V é uma versão "projetada" de x realçando os efeitos das autocorrelações em diferentes intervalos de tempo (figura 2.14, página 44). Sabemos que a contribuição de cada PC é quantificada pela porcentagem da variância retida pelos autovalores, mas essas porcentagens não trazem informações sobre a estrutura interna do sinal, apenas a força e a direção das correlações. Mais precisamente, o grau de determinismo (ou aleatoriedade) não é quantificado pelos autovalores. Portanto, propomos comparar as

metodologias H_t, H_f e H_s para obter a entropia de cada coluna de V, a fim de investigar suas contribuições na reconstrução final.

O passos do método completo são:

1. realizar a decomposição de x mediante as projeções $V = X^T U$,
2. representar cada coluna V_i pelo valor de sua entropia $H_i(V_i)$,
3. utilizar o algoritmo 2 (página 2) para agrupar os componentes em dois grupos utilizando os valores de entropia, e
4. com o limiar obtido no agrupamento, aplicar uma regra de decisão, onde as classes "sinal" ou "ruído" são atribuídas ao operador de seleção da forma:

$$p_{m,m} = \begin{cases} 1 & \text{se } \mathrm{H}(V_i) \leq T_{\mathrm{H}} \\ 0 & \text{caso contrário} \end{cases}, \tag{6.8}$$

para $m = \{1, 2, \ldots, L\}$ e $i = m$.

Observa-se assim, que o problema de decisão foi mapeado para um problema de agrupamento binário não supervisionado. Portanto, com base na entropia, podemos analisar a estrutura dinâmica dos componentes desde o ponto de vista da teoria da informação e identificar quais componentes explicam os diferentes comportamentos do sinal. Lembrando que, a metodologia proposta, analisa os componentes do sinal no espaço das projeções ortogonais e as reconstruções.

Um desafio secundário associado à aplicação desta regra é a obtenção limiar T_{H}. O limiar ideal representa um *trade-off* entre a quantidade de ruído descartado e a distorção da reconstrução. O procedimento que propomos para encontrar este limiar é o mesmo do algoritmo 2, útil para segmentar os sinais utilizando a entropia dos *frames*. Neste caso, o algoritmo divide o conjunto $\mathrm{H} = \{\mathrm{H}_1, \mathrm{H}_2, \ldots, \mathrm{H}_L\}$ em dois grupos com separação máxima entre as suas médias, tentando maximizar a distância entre classes. Normalmente, esse procedimento converge em menos de cinco iterações.

6.7.2 Escolha do parâmetro L

Como mencionamos, L é o único parâmetro necessário para realizar a decomposição aplicando SSA. A correta escolha deste valor lida com um problema de separabilidade dos componentes oscilatórios, i.e., a série original pode ser representada como uma soma de sub-séries não correlacionadas com interpretações físicas diferentes como: a tendência, a sazonalidade ou componentes de maior frequência (Golyandina et al.,

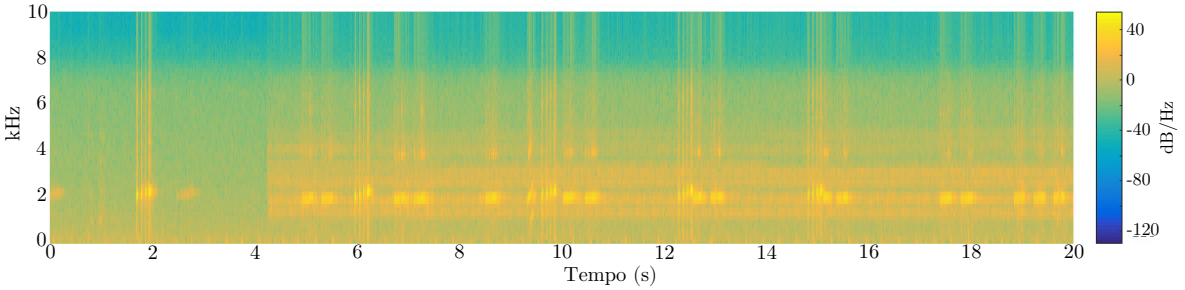
2001). Por exemplo, suponha-se que a x é a combinação linear de duas séries com períodos diferentes $x = x_1 + x_2$, escolhendo o valor correto de L, a reconstrução resulta em $\hat{x} = \tilde{x}_1 + \tilde{x}_2$, na qual $\tilde{x}_1$ e $\tilde{x}_2$ são as componentes principais (Hassani, 2007).

Os resultados teóricos apresentados por Golyandina et al. (2001), indicam que L deve ser grande o suficiente para cobrir o componente de maior frequência e ao mesmo tempo menor que $N/2$ (sendo N o comprimento total da série). Assim, se for possível conhecer *a priori* o período da série (*e.g.*, componente sazonal), é aconselhável utilizar uma janela de tamanho proporcional a este, para obter uma melhor separabilidade. Desta forma, L não pode ser tão pequeno que não consiga separar dois componentes de diferentes frequência, nem tão grande que separe diferentes componentes harmônicos que dificultem a interpretação. Um resultado da caraterização deste parâmetro, que exemplifica a resolução da decomposição desde o ponto de vista de filtros FIR, é apresentado na seção 6.7.4.

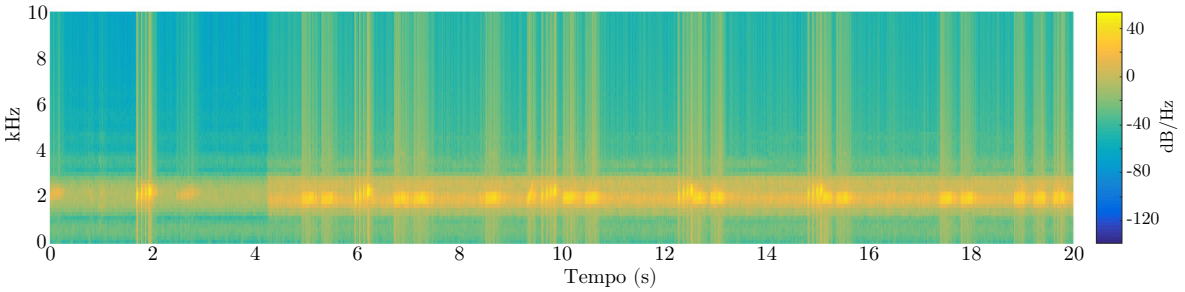
No que diz respeito ao agrupamento, o espectro singular aumenta proporcionalmente ao tamanho de L, tornando a interpretação e inspeção visual inviável. Além disto, existe também o problema de demanda de recursos computacionais, i.e., quanto maior é o valor L maior é a memória necessária para multiplicar e manipular a matrizes S, U e V, tornando o método computacionalmente intratável (seção 2.4.2).

Considerando estas recomendações, e pelo conhecimento do domínio de nosso problema, no qual utilizamos sinais bioacústicos com frequência de amostragem f_s, adotamos como regra prática escolher $L = \lfloor f_s/1000 \rfloor$ (sendo a unidade básica Hz) ou múltiplos desta relação, por exemplo $L = 22$. Assim, os PCs do SVD são as saídas de um banco com L filtros FIR, centralizados nas frequências fundamentais das bases vectoriais U. A largura de banda dos filtros depende do parâmetro L, assim, quanto maior seja este parâmetro, menor será a largura de banda dos filtros, tornando-se mais seletivos (figure 6.11(a), página 189).

O efeito do parâmetro L pode ser observado na figura 6.7. A figura 6.7(a) mostra o espectrograma original da espécie *Hypsiboas cordobae*, com $f_s = 44$kHz. Consequentemente, a frequência máxima do espectrograma é 22 kHz. Pode-se notar que esta espécie possui somente uma banda de frequências principais ($\Delta f = 2 \pm 0.5$kHz). A figura 6.7(b) ilustra o resultado da decomposição e reconstrução utilizando $L = 44$ e $\lambda_{1:2}$. Na reconstrução, podemos observar que a banda de frequências principal foi preservada e a energia das bandas restantes foi reduzida. O resultado, neste caso, é um sinal mais claro que preserva as informações espectrais que caracterizam a espécie. Portanto, a largura de banda dos filtros gerados pelo SSA é razoável. Entretanto, a regra apresentada é empírica e deve ser avaliada para cada caso em particular.



(a) Espectrograma Hypsiboas cordobae.



(b) Espectrograma Hypsiboas cordobae filtrado com SSA.

Figura 6.7. Espectrograma original da espécie Hypsiboas cordobae (a) e seu espectrograma reconstruído utilizando $L = 44$ e os dois maiores autovalores (b).

6.7.3 Avaliações experimentais do filtro SSA

Nesta seção realizamos dois experimentos, um usando um sinal oscilatório sintético contaminado com diferentes tipos de ruídos (branco e coloridos) e um segundo experimento usando três sinais bioacústicos dos anuros gravados em condições reais (incluindo ruído de fundo da floresta). Por fim, no final desta seção, apresentamos como obter os parâmetros de um filtro FIR adaptado para cada vocalização em particular, e como gerar um banco de filtros a partir dos autovetores do SSA.

6.7.3.1 Experimentos com sinais sintéticos

Neste primeiro experimento, nós definimos um sinal sintético com o propósito de realizar avaliações e simulações do comportamento do SSA sob diferentes condições. O sinal foi definido como $y = \sin(8\pi t)\sin(t) + \xi_\alpha$, onde t varia entre $0 \leq t \leq 4\pi$ com frequência de amostragem f_s=50 Hz (figura 6.8(a)). Definimos esta onda modulada em amplitude, com a finalidade de reproduzir um som bioacústico de baixa frequência. O sinal e o ruído foram normalizados para ter variância unitária ($\sigma_x = \sigma_\xi = 1$). Assim, o SNR torna-se SNR = 0 dB. Esta condição nos permite simular um cenário adverso. A PSD de ξ, segue a lei exponencial definida pela equação 2.22 (página 30)

com $\alpha = \{-2, -1, 0, 1, 2\}$.

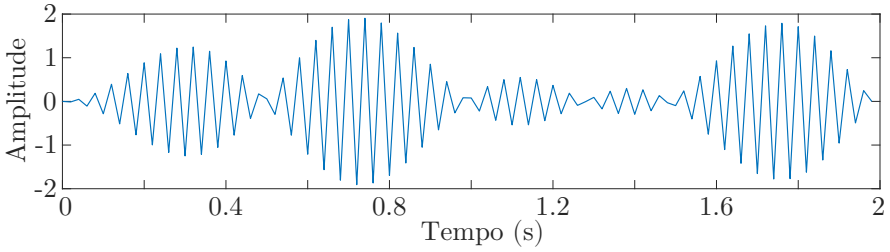
Neste caso, a decomposição foi realizada utilizando $L = 22$. Este valor foi empiricamente escolhido. Para a reconstrução foram utilizados os componentes previstos pelo critério da equação 6.8. A figura 6.8(b) mostra a variação do espectro singular causada pela adição dos diferentes tipos de ruído. Duas figuras ampliadas foram adicionadas (figura 6.8(c) e 6.8(d)), expandindo os intervalos $\lambda_{1:5}$ e $\lambda_{6:L}$ respectivamente.

Na figura 6.8, observamos que o espectro singular do sinal sem contaminação decai rapidamente a zero, mas quando os ruídos são adicionados, a energia dos últimos componentes aumenta. A amplitude crescente dos últimos autovalores, pode levar-nos a escolher erroneamente os PCs para a reconstrução, ao aplicar os critérios de seleção tradicionais. Dentre os diferentes tipos de ruídos observamos que o ruído branco espalhou sua energia de maneira mais "uniforme", aumentando ainda mais a energia dos últimos autovalores. O ruído rosa, o qual decai 1 dB por oitava da sua PSD, e o ruído azul, que aumenta 1 dB por oitava da sua PSD, ambos apresentam uma concentração de energia semelhante nos seus autovalores.

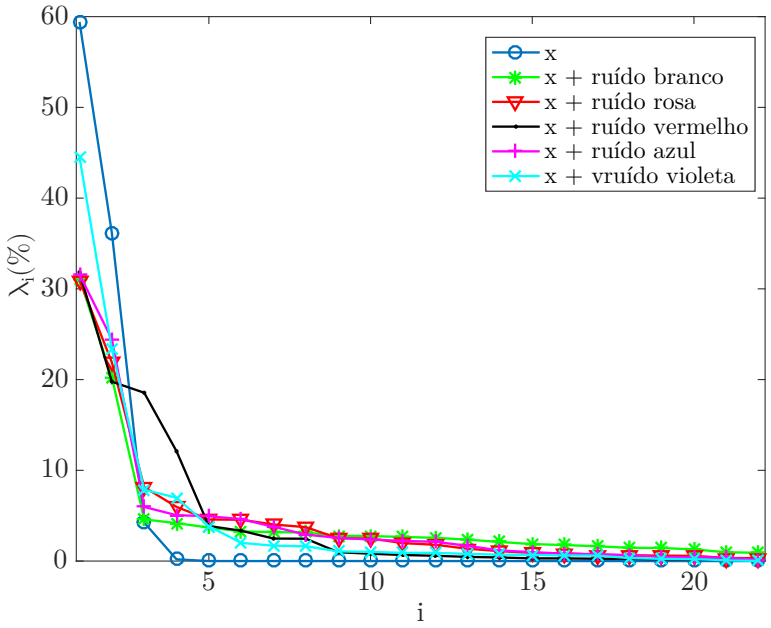
Nos casos de ruídos vermelhos e violetas, percebemos as maiores diferenças. O ruido violeta alterou levemente os primeiros autovetores comparados com os outros ruídos. Podendo-se deduzir que a maior fração de energia deste ruído permaneceu nos primeiros componentes. Finalmente, o ruído rosa causou uma variação no espectro singular, no qual os componentes λ_2 e λ_3 ficaram próximos. Esta forma particular da curva dos autovalores, pode levar a conclusões erradas, assumindo-se que o componente λ_3 é um harmônico do sinal. Assim, diferentes tipos de contaminação podem causar diversos erros na recuperação dos sinais.

Em nossas simulações, realizamos cinquenta iteração com ruído gerado aleatoriamente. A qualidade da reconstrução é mensurada pela média do MSE e do SDR. Os resultados da reconstrução para cada combinação dos ruídos e dos quantificadores de entropia, encontram-se na tabela 6.1. Para cada linha da tabela, aplicamos o teste estatístico *t-test* para verificar a existência de ganhos significativos em comparação com o melhor resultado de cada linha. Assim, o melhor resultado de cada linha é destacado em negrito.

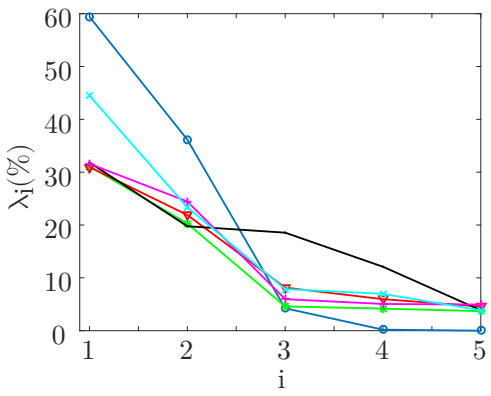
Os resultados da tabela 6.1 mostram que a regra que utiliza a entropia da permutações (H_s) foi superior, selecionando os melhores componentes para a reconstrução. No entanto, percebemos que esses resultados superiores se devem ao fato de que H_s é útil caracterizando as não linearidades presentes nesses ruídos. Um exemplo desse fato, é ilustrado na próxima seção. Observamos também que, quando $\alpha = \{-2, 2\}$, os ruídos com energia mais concentrada em uma determinada banda mais estreita de frequências é mais difícil de separar com SSA, remarcando uma limitação natural deste



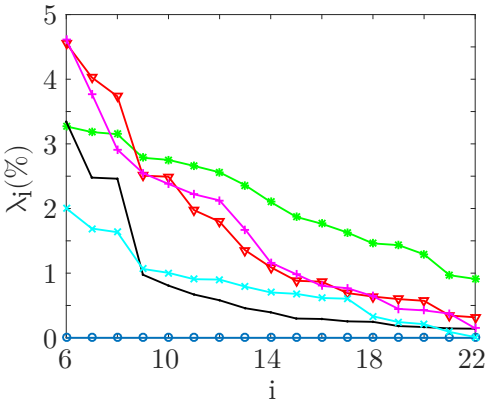
(a) Sinal simulado modulado em amplitude.



(b) Espectro singular do sinal simulado.



(c) Ampliação do intervalo $\lambda_{1:5}$



(d) Ampliação do intervalo $\lambda_{6:L}$

Figura 6.8. Espetro singular do sinal simulado com e sem ruído.

método.

A separação dos componentes realizada pelo SSA, provou ser melhor quando o ruído adicionado é completamente não correlacionado (branco). Além disso, os piores

Tabela 6.1. Qualidade do sinal reconstruído para várias contaminações de ruído. Os melhores resultados são destacados em negrito.

	SDR			MSE		
	H_t	H_f	H_s	H_t	H_f	H_s
$x + \xi_0$	0,05	5,91	**6,92**	1,06	0,27	**0,20**
$x + \xi_1$	-0,30	3,23	**4,39**	1,13	0,55	**0,43**
$x + \xi_2$	-0,47	0,86	**2,45**	1,13	0,83	**0,68**
$x + \xi_{-1}$	0,36	4,07	**5,33**	0,94	0,42	**0,27**
$x + \xi_{-2}$	0,32	2,43	**3,83**	0,94	0,60	**0,42**

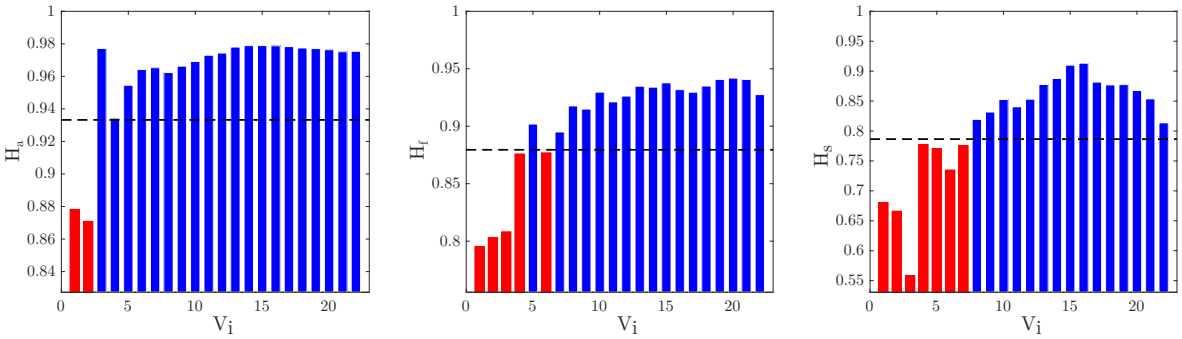
resultados aparecem no caso do ruído vermelho, pois este possui uma alta concentração de energia nas baixas frequências, assimilando-se de um componente aditivo de oscilação lenta com elevada variância. O efeito dos componentes de oscilação lenta torna-se claro na próxima seção, onde apresentamos exemplos detalhados desses fenômenos, usando registros de chamadas de anuros correspondentes a três espécies com seus ruídos de fundo.

6.7.3.2 Experimentos com sinais bioacústicos

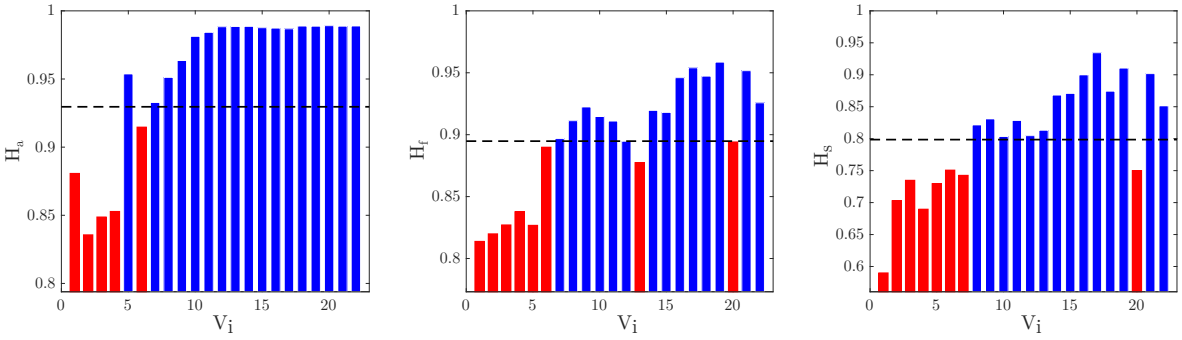
Nesta seção apresentamos três casos de estudo de análise e filtragem de chamadas anuros utilizando SSA. As três gravações pertencem às espécies Adenomera hylaedactyla, Aplastodiscus perviridis e Hyla minuta, coletadas na floresta amazônica em condições reais com ruído de fundo, em formato *.wav*, com frequência de amostragem 44,1 kHz. Um segmento dessas gravações já foi mostrado na figura 6.5 e seus espectros singulares na figura 6.4.

A decomposição SSA foi realizada com 22 PCs ($V_{1:22}$), e para cada gravação foram aplicados os três critérios de entropia (H_t, H_f e H_s). Note que o número de PCs é múltiplo da regra empírica introduzida na seção 6.7.2. Os valores de entropia das colunas de V são mostrados nas barras da figura 6.9. A linha tracejada horizontal representa o limiar de decisão T_H calculado pelo algoritmo 2, assim, para cada gravação, este limiar é único e depende do conjunto dos valores de entropias $H(V_i)$. Consequentemente, aplicando a regra da equação 6.8 obtemos as melhores colunas candidatas de V. Os componentes aceitos pela regra proposta são destacados em vermelho.

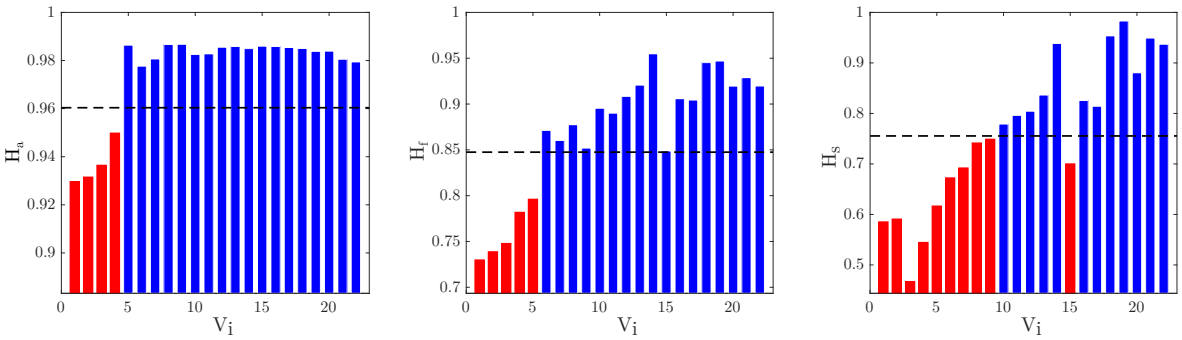
Na figura 6.5, ilustramos o terceiro e quinto PC das espécie Aplastodiscus perviridis, Adenomera hylaedactyla e Hyla minuta, para mostramos como estas componentes relacionam-se à tendência do sinal, sendo semelhantes a uma contaminação por ruído vermelho, presente tanto nas sílabas quanto fora delas. Logo, a partir da figura 6.9 descobrimos que a coluna V_3 e V_5 das mesmas espécies possuem elevada entropia quando



(a) Aplastodiscus perviridis.



(b) Hyla minuta.



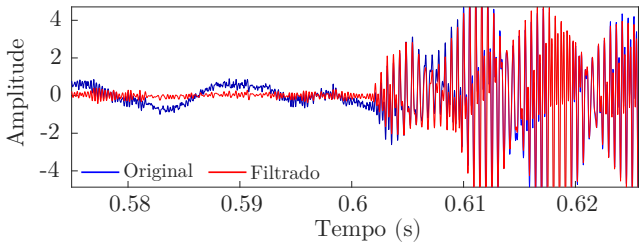
(c) Adenomera hylaedactyla.

Figura 6.9. Diferentes quantificadores de informação das colunas da matriz V. Na coluna esquerda a entropia temporal, na coluna central a entropia espectral e na coluna direita entropia das permutações. A primeira lina corresponde com a espécieAplastodiscus perviridis, a segunda linha com a espécie Hyla minuta e a terceira com Adenomera hylaedactyla. As linhas tracejadas ilustram os limiares de decisão ótimos (T_{H}) para cada caso. As colunas destacas em vermelho foram selecionadas para criar a reconstrução filtrada das chamadas.

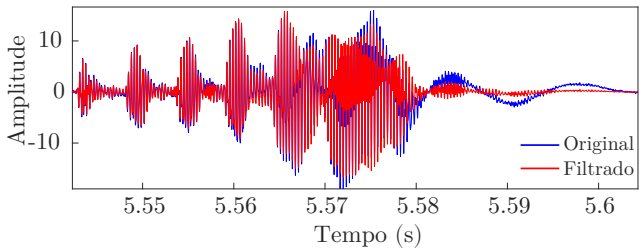
se utiliza o quantificador H_t. Isso significa que o critério de H_t foi capaz de detectar e evitar esses componentes antes da reconstrução, mesmo para os casos em que componentes posteriores correspondentes a autovalores com menor energia foram mantidos, conforme ilustrado na sub-figura esquerda da espécie Hyla minuta.

Os critérios que utilizam H_f e H_s não conseguiram detectar os componentes de ruído de baixa frequência projetados em V. No caso de H_f, isso ocorre porque esses componentes possuem elevada energia concentrada em uma banda estreita de frequência, produzindo valores de entropia baixos. No caso H_s, ocorreu porque os sinais oscilatórios de baixa frequência tendem a aumentar a ocorrência de alguns padrões ordinais mais do que outros, produzindo um histograma concentrado e com menor entropia. No entanto, H_f e H_s revelam contribuições diferentes das projeções V_i. Por exemplo, H_s é um método não-linear, e portanto, aplicá-lo pode ser útil para entender estruturas e relações mais complexas entre as componentes do sinal.

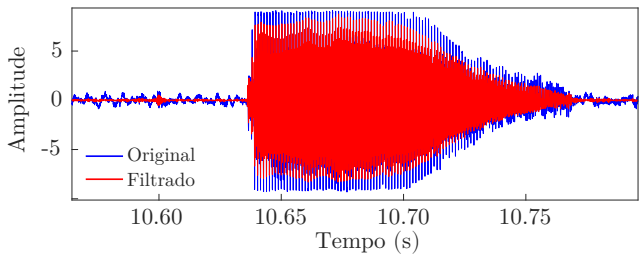
Finalmente, a figura 6.10 ilustra três resultados de reconstrução usando H_t. Nesta figura, a linha azul representa a gravação original e a linha vermelha é a versão filtrada. Aqui é notável a quantidade de ruído removido antes, durante e após as sílabas sem causar elevadas distorções na amplitude da onda. Remarcamos que, avaliamos também qualitativamente essas reconstruções através da opinião de um especialista humano.



(a) Adenomera hylaedactyla componentes $\lambda_{1,2,3,4}$.



(b) Hyla minuta componentes $\lambda_{1,2,3,4,6}$.



(c) Aplastodiscus perviridis componentes $\lambda_{1,2}$.

Figura 6.10. Exemplos de reconstruções usando o critério H_t. Em azul são ilustrados os sinais originais e em vermelho as versões filtradas.

6.7.4 Filtro bioacústico FIR adaptativo (eigenfilter)

Esta seção aborda principalmente como construir um FIR para um determinado sinal bioacústico usando os componentes principais com entropia mínima de SSA. Os filtros construído a partir dos autovetores são chamados de *eigenfilters*. Tomé et al. (2010, 2011) mostraram que a decomposição SSA têm uma equivalência com um banco de L filtros FIR em paralelo, no qual os polinômios que determinam as funções de transferência dos filtros são obtidos a partir dos autovetores respectivos.

A regra baseada em entropia descrita nas seções anteriores é útil para decidir quais componentes projetados do sinal (V_i) proporcionam a melhor reconstrução em termos de MSE e SDR. Observou-se que tais componentes tem um comportamento mais determinístico em relação aos restantes e, na maioria dos casos, estes concentram a maior proporção de energia do sinal original. Esta regra é útil para escolher as colunas de V e também de U. Por exemplo, usando a regra H_f, os componentes 1, 2, 3, 4 e 5 da espécie Aplastodiscus perviridis foram selecionados como as melhores colunas de V para obter uma reconstrução mais limpa (figura 6.9). Consequentemente, as mesmas colunas da matriz U (a base vetorial para a decomposição) são escolhidas durante a etapa de agrupamento e reconstrução (equação 2.42, página 2.42), para projetar o sinal de volta e obter uma nova matriz de trajetória $R = U_{1:p}V_{1:p}^T$. Em outras palavras, uma vez aplicado o critério de entropia, este também é útil para escolher os filtros FIR (ou *eigenfilters*) mais relevantes.

De acordo com Tomé et al. (2011), os coeficientes de um filtro FIR podem ser obtidos como:

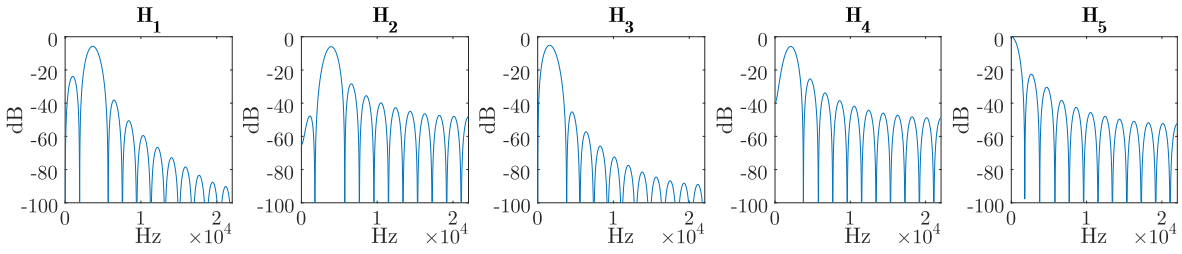
$$h_i = \frac{1}{L}\Big(U_i * \mathrm{flip}(U_i)\Big), \tag{6.9}$$

onde $*$ representa a operação de convolução, U_i é o i-ésimo autovetor e $h_i = \{b_1, b_2, \ldots, b_j\}$ são os coeficientes do filtro. A operação "flip" inverte o vetor U_i. Formalmente, a operação flip de um vetor é calculada como a transposta do próprio vetor vezes a matriz identidade anti-diagonal. Uma vez que os coeficientes h_i foram definidos, a relação entrada-saída do filtro é dada pela equação 2.26 (página 33), a qual é um sistema linear invariante no tempo (LTI). Desta forma, podemos obter um filtro FIR para cada autovetor com resposta impulsiva:

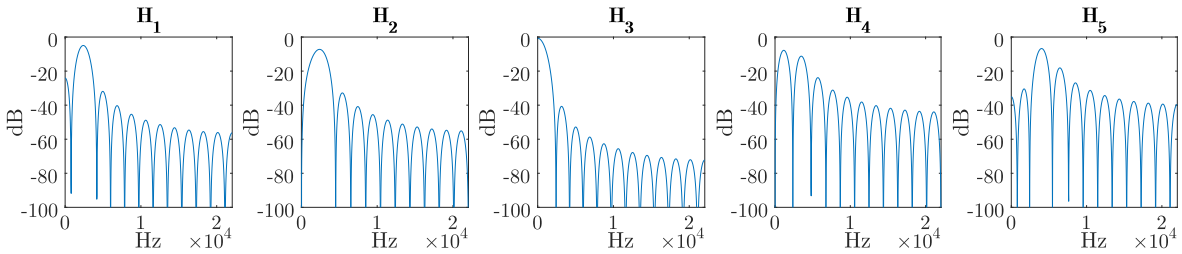
$$h_i[n] = \begin{cases} b_j\delta[n-j] & \text{se } 0 \leq j \leq 2L-1 \\ 0 & \text{caso contrário} \end{cases}, \tag{6.10}$$

onde $M = 2L - 1$ é o grau do filtro. Pelo fato de cada $h_i[n]$ ser a convolução de um autovetor por ele mesmo, a resposta impulsiva dos filtros é simétrica com fase linear.

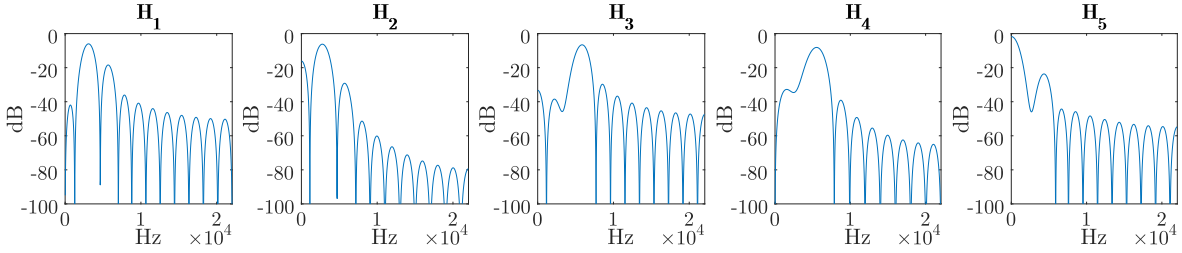
A principal vantagem de obter os coeficientes dos filtros desta forma é que cada um deles se adapta ao sinal de entrada. Assim, cada vocalização de cada espécie pode ter seu próprio conjunto de filtros determinados peles PCs. A figura 6.11 ilustra a resposta em frequência ($h_i[n] \xrightarrow{\mathcal{F}} H_i(f)$) dos cinco filtros principais das espécie analisadas. Como podemos observar, as respostas têm vários lóbulos formando ondulações. Cada um destes lóbulos possui a forma de um filtro passa-banda com diferente ganho e largura de banda. Por exemplo, no filtro H_1, na sub-figura 6.11(a), existe um lóbulo maior centralizado em 3,64 kHz que se corresponde com o valor central da faixa principal das frequência do espectrograma da figura 6.13(a). Este fato estabelece uma das propriedades principais do SSA, que afirma que os primeiros autovalores correspondem aos filtros centrados nas faixas de frequências com maior concentração de energia espectral. Então, dado um sinal bioacústico, os *eigenfilters* são construídos e ordenados automaticamente de acordo com sua energia espectral.



(a) Adenomera hylaedactyla.



(b) Aplastodiscus perviridis.



(c) Hyla minuta.

Figura 6.11. Magnitude das respostas em frequência dos cinco primeiros *eigenfilters* em dB.

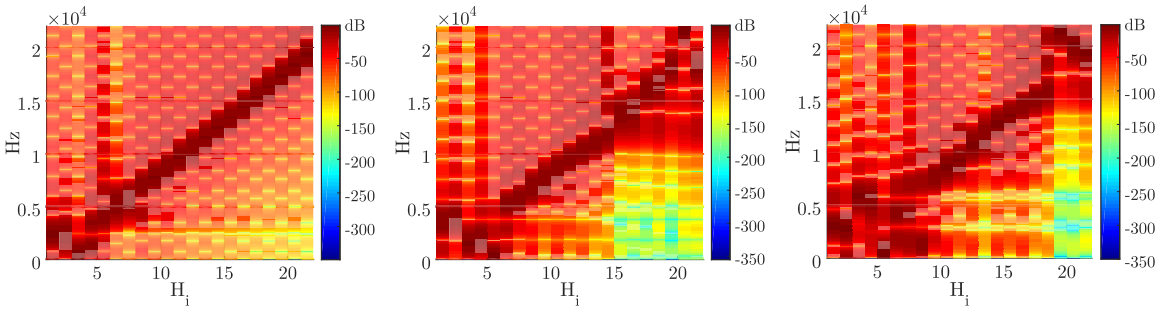
É possível observar que a resposta do filtro H_1, na figura 6.11(a), produz uma atenuação de aproximadamente 17 dB entre as frequências dos lóbulos principal e o secundário, e também que as altas frequências são severamente atenuadas. Isso ocorre porque as gravações contêm ruídos da floresta, que geralmente concentram sua energia em baixas frequências. Além disso, a resposta do filtro H_5 desta mesma espécie, claramente possui as características de um filtro passa-baixa, permitindo que a maior parte do ruído ambiental permaneça na reconstrução. As respostas dos filtros H_3 e H_4, tem seus lóbulos centrados na segunda faixa de frequência principal da vocalização desta espécie. A segunda faixa principal de frequências desta espécie pode ser observada na figura 6.13(a).

Além disso, comparando as respostas H_1 e H_2 da espécie Adenomera hylaedactyla, percebemos que os lóbulos principais estão centrados na mesma frequência. No entanto, H_1 atenua mais as altas frequências do que as baixas frequências e H_2 apresenta o comportamento oposto. Essas observações levam a duas conclusões importantes. Primeiro, para obter um sinal bioacústico reconstruído que preserve a faixa de frequência principal, precisamos de pelo menos dois componentes principais V_1 e V_2. E em segundo lugar, existe um grau de correlação entre o ruído de fundo de baixa frequência e o sinal projetado em V_1. Isto é ilustrado pela pequena diferença entre os lóbulos principais e secundários deste filtro.

Portanto, observando as respostas das magnitudes dos *eigenfilters*, também é possível escolher os componentes do SSA que desejamos manter na reconstrução, com base na escolha de filtros que preservem as frequências principais dos sinais. Esta observação pode ser um critério adicional para escolher os PCs do SSA. Análises e observações similares podem ser transferidas para respostas dos filtros das espécies Aplastodiscus perviridis e Hyla minuta.

Uma visão mais geral, usando a vista superior das respostas da magnitude dos filtros H_i para as três espécies analisadas, são ilustradas na figura 6.12. Por exemplo, na figura 6.12(b) as primeiras cinco colunas correspondem à vista superior das respostas dos filtros apresentadas na figura 6.11(a). As seguintes colunas representam as respostas dos filtros restantes $H_{6:L}$. A partir desta representação gráfica, fica claro qual é a frequência do lóbulo principal para cada filtro *eigenfilter*. Lembramos que a ordem dos filtros, representada pelo eixo horizontal, é a mesma ordem dos autovalores ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L$). Uma particularidade que pode ser observada nesta representação, é que a frequência central dos lóbulos principais dos filtros aumenta conforme aumentam as frequência dos PCs. Isto ocorre pois, nos registros bioacústicos da floresta os sons com altas frequências são mais raros e possuem uma menor energia.

Pro fim, sabemos que filtrar o sinal usando SSA com os *eigenfilters* H_i é equi-



(a) Aplastodiscus perviridis. (b) Adenomera hylaedactyla. (c) Hyla minuta.

Figura 6.12. Respostas em frequência de todos os *eigenfilters*.

valente a aplicar cada um destes individualmente (em paralelo) e somar todas as saídas (Tomé et al., 2010), da forma:

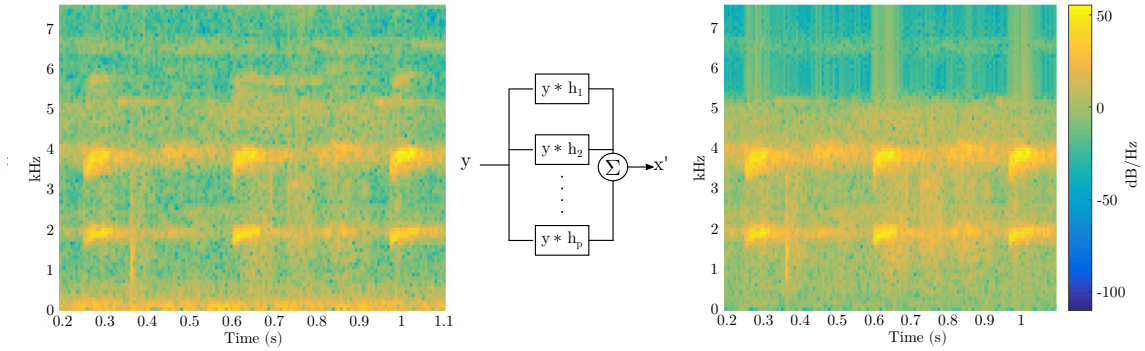
$$\hat{x} = \sum_i h_i * y, \quad \forall\, i \mid \mathrm{H}_i < T_H, \tag{6.11}$$

onde i identifica os filtros com entropia mínima e $y = x + \xi_\alpha$. Isso significa que, para obter $\hat{x}$ com SSA, precisamos aplicar um banco de filtros conforme detalhado na figura 6.13(a). Finalmente, um exemplo de uma vocalização filtrada da espécie Adenomera hylaedactyla com $h_{1:4}$, é representado no espectrograma da direita. Aqui, podemos observar a proporção do ruido que foi eliminado nas altas frequências e nas baixas frequências. O efeito da filtragem no domínio temporal é ilustrado na figura 6.13(b). O banco de filtros FIR ilustrado aqui, é o método que deve se embarcar no nó sensor para o monitoramento da espécie escolhida. Porém, mais de um banco de filtros pode ser embarcado para monitorar mais de uma espécie.

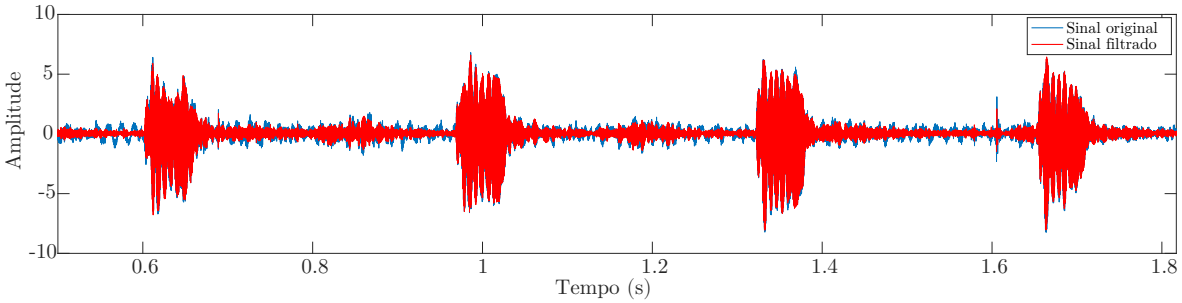
6.7.5 Avaliação da segmentação utilizando filtro

Na introdução deste capítulo levantamos a hipóteses de que filtrar os sinais, além de melhorar a qualidade das gravações também, poderia melhorar a segmentação dos mesmos, evitando os falsos positivos causados pelos ruídos de alta amplitude.

Para avaliar esta hipótese, utilizamos uma base de vocalizações com as mesmas espécies e anotações introduzidas na tabela 4.3 (página 109). Na tabela 6.2, comparamos a segmentação sem filtro contra a segmentação com os critérios de filtragem propostos, usando SSA e entropia. A quantidade de PCs do SSA foi fixado em 22. O método DWT foi utilizado com o propósito de comparação com os seguintes parâmetros: família de funções bases *Daubechies 8*, sete níveis de decomposição, e *soft thresholding* com critério *SURE*. Esta combinação de parâmetros para a DWT foi recomendada por Gur



(a) Espectrograma do sinai original (esquerda) e do sinal filtrado (direita) aplicando um banco FIR construído a partir dos quatro PCs do SSA selecionados com o critério H_t.



(b) Forma de onda original (azul) e filtrada (vermelho) após aplicar o banco de filtros FIR.

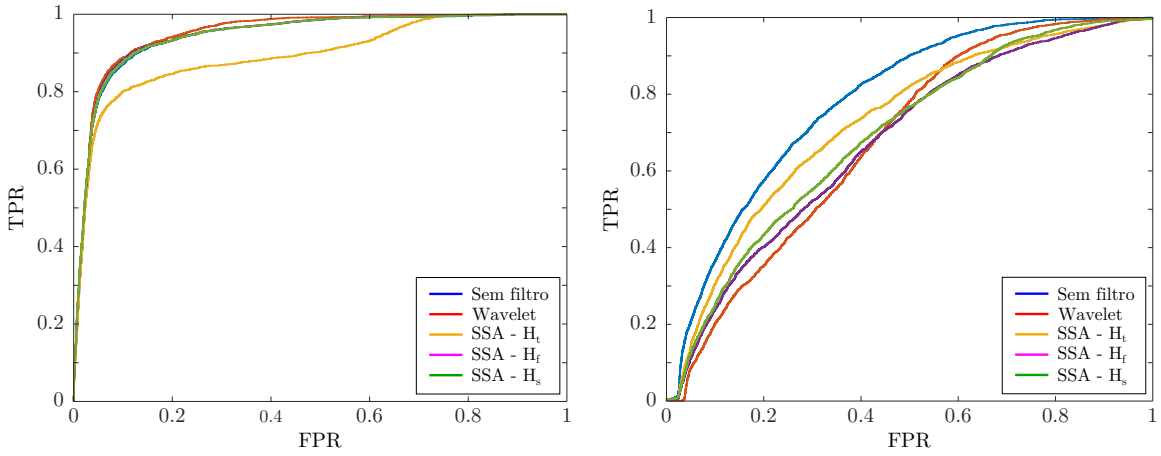
Figura 6.13. Efeitos do banco de filtros adaptativos FIR (*eigenfilters*) no domínio espectral (a) e no domínio temporal (b) em uma gravação da espécie Adenomera hylaedactyla.

and Niezrecki (2011).

Tabela 6.2. Resultados da segmentação em sílabas dos sinais utilizando a energia dos mesmos com e sem filtragem prévia. AUC representa a área da curva ROC.

	Sem filtro	Wavelet	SSA com H_t	SSA com H_f	SSA com H_s
AUC - E	0,9451	**0,9515**	0,8945	0,9467	0,9460
AUC - H_f	**0,7828**	0,6789	0,7263	0,6759	0,6913

Observamos nos resultados da tabela 6.2 que a segmentação melhorou utilizando o filtro baseado em DWT, quando o segmentador utiliza a energia do sinal. A energia do sinal relaciona-se diretamente com os valores de amplitude da onda, e portanto, concluímos que DWT filtra melhor as variações de amplitude do que SSA. As figuras 6.14(a) e 6.14(b) ilustram as curvas ROC dos dois métodos de segmentação avaliados.



(a) Segmentação com energia do sinal.

(b) Segmentação com a entropia espectral do sinal.

Figura 6.14. Curvas ROC da segmentação avaliada *frame-a-frame*.

6.7.6 Avaliação da classificação utilizando filtro

O filtro utilizado como primeira etapa do sistema (figura 1.1), modifica o comportamento final do sistema. O filtro pode mudar a segmentação como foi mostrado na seção anterior, e portanto, o resultado da classificação também pode variar. Para avaliar a variação final dos resultados realizamos um experimento completo, incluindo filtragem dos sinais, segmentação automática, extração de caraterísticas (LLDs) e classificação, utilizando a base de vocalizações introduzida na tabela 5.2. Os resultados deste experimento são apresentados na tabela 6.3.

Para a classificação escolhemos o método plano kNN com $k = 3$ vizinhos e decomposição 1AA, com foi avaliado na seção 5.1.5 (tabela 5.3, página 136). Este método apresentou a melhor relação entre complexidade de classificação e resultados. A segmentação foi gerada da mesma forma que foi apresentada na seção 5.1.5, isto é, utilizando a abordagem por *frames* com a entropia espectral. Uma variável artificial aleatória Gaussiana com 10% da variância do sinal original foi adicionada antes da segmentação, para quebrar as correlações fracas e melhorar a segmentação (figura 4.3). O LLDs utilizados foram 22 coeficientes MFCCss gerados a partir de 44 filtros triangulares.

A tabela 6.3 apresenta os resultados do sistema ACR com e sem filtragem. Os filtros avaliados foram: DWT e SSA com os critérios H_t, H_f e H_s. Como podemos observar, em geral os resultados com filtro foram inferiores aos resultado sem filtro. Isto significa que os MFCCs são robustos aos ruídos ambientais e, o mais importante, que os ruídos ambientais capturados por estes coeficientes ajudaram ao classificador a

separar as sílabas. Com estes resultados, podemos concluir a importância dos ruídos ambientais. Estes carregam informação sobre o lugar onde os indivíduos foram gravados. Portanto, uma técnica de decomposição tal como SSA combinada com a entropia ou a complexidade estatística é útil na análise de tais ambientes acústicos.

Comparando os métodos de filtragem avaliados notamos que a decomposição DWT foi menos agressiva e afetou menos o desempenho da classificação do que SSA. A segunda vantagem importante do DWT é a velocidade de processamento. Esta transformada pode ser executada em tempo real com baixo custo de memória, sendo ideal para um sensor com *hardware* restrito, enquanto que o SSA não é sequencial (nem pode ser executado em tempo real) e demanda mais memória e tempo de processamento. Entretanto, as bases da transformação SSA são adaptáveis ao sinal de entrada sendo melhor para estudar a composição dos sons ambientais.

Tabela 6.3. Resultados da classificação das espécies com e sem filtragem dos sinais. Utilizou-se kNN com $k = 3$ e validação cruzada por indivíduos. Os resultados são apresentados pelas Macro-métricas.

	Sem filtro	Wavelet	SSA - H_t	SSA - H_f	SSA - H_s
Prec	0,69	0,67	0,63	0,66	0,64
Rec	0,72	0,65	0,60	0,63	0,61
Macro-F1	0,70	0,66	0,61	0,64	0,62

6.7.7 Considerações do filtro SSA

Nesta seção, mostramos como o SSA constrói as bases ortonormais do subespaço de decomposição do sinal acústico com base na matriz de autocorrelações. Mostramos então, que existe uma correspondência entre essas bases vetoriais do subespaço e as principais faixas de frequências do sinal analisado. Portanto, o SSA pode ser interpretado como um método de decomposição espectral, em que os PCs adaptam-se às frequências das vocalizações, como foi mostrado na seção de construção dos *eigenfilters*. Esta é a principal razão pela qual escolhemos o SSA para decompor e filtrar os sinais bioacústicos.

Aqui também introduzimos um método que combina quantificadores de teoria da informação junto com um algoritmo de agrupamento binário, para selecionar os melhores componentes do sinal quando este é projetado no subespaço gerado pela decomposição SSA. Avaliamos assim a reconstrução em termos dos MSE e SDR, mostrando que nossa regra baseada em entropia é útil para gerar uma reconstrução "limpa". Resumindo, o método apresentado fornece-nos uma decomposição e reconstrução ótima e

adaptável às vocalizações das diferentes espécies e discrimina melhor os componentes relacionados ao som ambiental de fundo.

A entropia dos componentes projetados permite discriminar qual destes é similar a um ruído sem estrutura. Este é um critério alternativo aos tradicionais que apenas levam em consideração o peso dos autovalores (λ), e não a estrutura interna do sinal. Isto também é válido para escolher as bases do subespaço úteis para criar os filtros adaptativos (*eigenfilters*) e o banco de filtros FIR ótimos.

Através das simulações com um sinal sintético e ruídos coloridos, como os encontrados nos fenômenos naturais, mostramos que utilizar H_s é o melhor candidato para recuperar um sinal de uma condição de ruído grave (SDR $\leq 0\,\mathrm{dB}$), minimizando o MSE e o SDR. No exemplo com sinais acústicos reais, o critério H_s conseguiu detectar um agrupamento menos óbvio dos PCs, provavelmente com padrões determinísticos ocultos e correlações não lineares.

Apesar desses bons resultados com H_s, este quantificador é sensível aos ruídos determinísticos de fundo, inclusive os de baixa amplitude, gerando reconstruções menos claras do sinal de interesse. Portanto, nós apontamos H_a como o melhor critério em nosso contexto de aplicação. No entanto, com qualquer um dos três critérios de entropia, o SSA pode ser ajustado automaticamente para obter uma filtragem não supervisionada. Adicionalmente, a regra da entropia pode ser combinada com os critérios tradicionais que utilizam o peso dos autovalores para gerar um critério final mais especializado e robusto.

No que diz respeito ao impacto dos filtros na segmentação e na classificação, observamos que o SSA obteve um desempenho inferior comparado com o filtro *Wavelet*. Nossa proposta resultou em um critério muito estrito que eliminou informação útil para o classificador. Entretanto, nenhum dos quatro filtros avaliados na tabela 6.3 melhorou o resultado da classificação. Isto ocorre pois o classificador "aprendeu" parte do som ambiental como característica útil para separar as classes.

O esquema de filtragem FIR é ótimo para o sinal da espécie no qual foi criado. Com isto, podemos criar um banco de filtros ótimos usando exemplos de indivíduos da espécie que se deseja monitorar, e embarcar os coeficientes FIR no nó sensor. Neste caso, se o interesse for monitorar mais de uma espécie, deve-se criar um banco de filtros para cada uma destas, e embarcar no nó sensor cada um desses filtros. Por outro lado, se o que se deseja é uma técnica mais geral, não específica apenas para um subconjunto de espécies, então recomendamos optar por embarcar um filtro DWT, cujo requerimento de memoria é linear e o processamento pode ser realizado em tempo real. Além disso, o SSA também é recomendado para o estudo dos sinais já armazenados no nó *sink* por causa da memória requerida, como foi explicado na seção 2.4.2 (página 46).

6.8 Metodologia de análise utilizando a complexidade estatística generalizada

Nesta seção, apresentamos uma analise da complexidade estatísticas dos PCs do SSA. O objetivo desta análise é ganhar conhecimento sobre a estrutura de correlações dos componentes, caracterizando cada uma destes no plano de Entropia-Complexidade (HxC) introduzido na seção 2.2.6 (página 28). Conforme foi observado na seção anterior, principalmente na figura 6.9, a H_s discrimina e agrupa um conjunto maior de componentes com estrutura interna não linear independente da amplitude do sinal. Assim, decidimos investigar a complexidade estatística (C_s) de tais componentes para gerar uma descrição completa do sistema físico envolvido, considerando transmissão e contaminação do sinal.

6.8.1 Complexidade estatística dos componentes do SSA

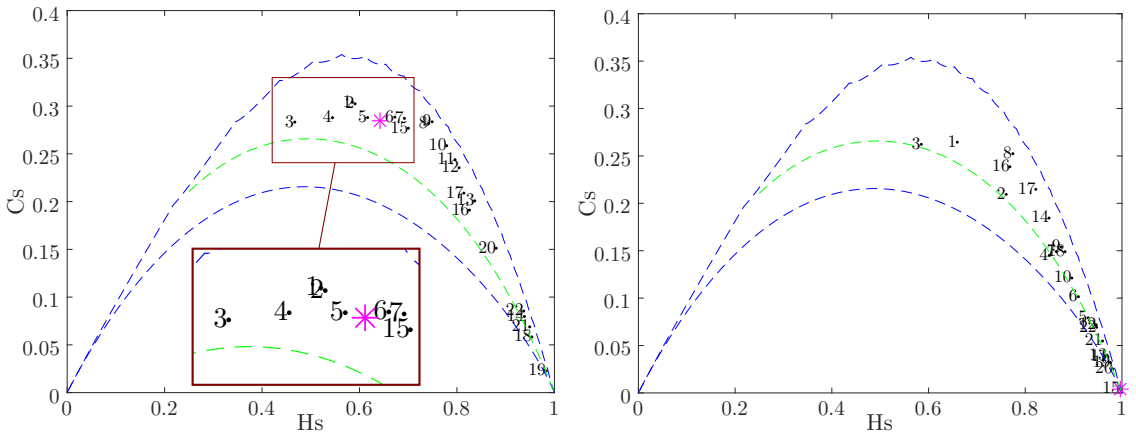
Existem duas possibilidades para combinar C_s e SSA. A primeira é gerar diferentes agrupamentos e calcular a complexidade das reconstruções $\hat{x}$. Esta forma de análise é computacionalmente mais custosa, pois é necessário gerar todas as possíveis combinações de agrupamentos, sendo um problema exponencial. A segunda opção é utilizar as colunas da matriz V que representam subséries de tempo ou projeções da matriz trajetória original X. Como foi mostrado na figura 2.14 (página 44), as colunas de V também possuem uma estrutura temporal relacionada com o sinal original e, portanto, são séries temporais.

O grau de determinismo e a complexidade estatística de cada componente é apresentado aqui. A entropia dos componentes pode ser quantificada aplicando a metodologia PE e, a partir desta, obter a complexidade de cada componente, da forma:

$$C_s = Q[P, P_e] \times H_s[P], \tag{6.12}$$

onde $Q[P, P_e]$ é a divergência, que também pode ser interpretada em termos de distancia, entre o histograma dos padrões ordinais Π de cada V_i e o histograma P_e do ponto de equilibro do sistema. Neste caso, P_e é um histograma de referência dado por uma distribuição uniforme que representa a entropia de um componente completamente aleatório (ruído puro). Para o cálculo de Q, utilizamos a divergência de *Jensen-Shannon* definido na equação 2.20 (página 29). Desta forma, cada componente projetado do sinal $V = X^T U$ é representado por um ponto dentro do plano de Entropia-Complexidade (HxC), caracterizando a dinâmica deste.

As figuras 6.15(a) e 6.15(b) mostram o plano de Entropia-Complexidade estatística, onde cada coluna de V_i é representada pelo par ordenado [H_s,C_s]. O número associado a cada ponto do plano corresponde ao índice do i-ésimo autovalor (λ_i). Neste exemplo, utilizamos a decomposição da vocalização original da espécie Adenomera hylaedactyla (esquerda) e a decomposição da mesma vocalização adicionando ruído aleatório branco com SNR = -3 dB (direita). Em ambos casos foi utilizado $m = 4$ e $\tau = 1$. A figura 6.16 apresenta um exemplo utilizando $m = 5$ e $\tau = 1$.



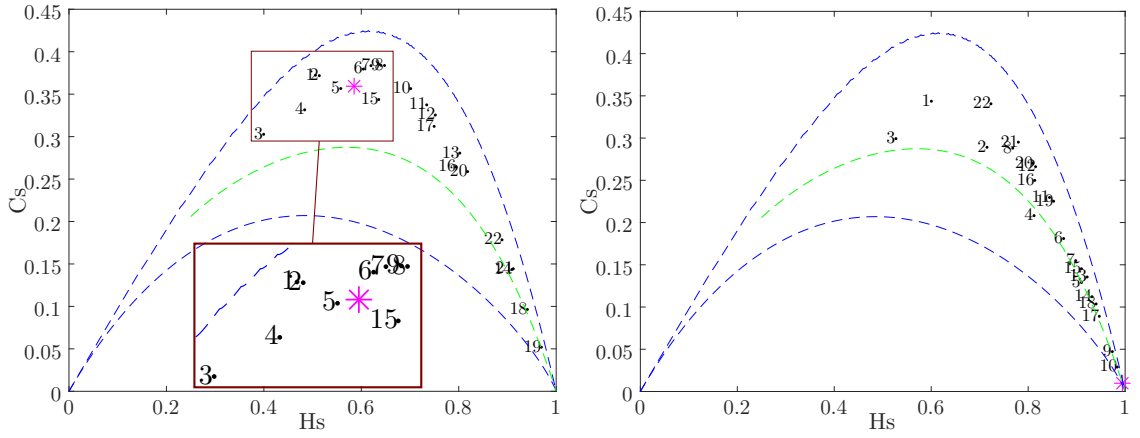
(a) Série original * e os componentes V_i.

(b) Série original mais ruído branco * e os componentes V_i.

Figura 6.15. Planos de Entropia-Complexidade estatística generalizada dos componentes V_i da vocalização da espécie Adenomera hylaedactyla utilizando H_s com $m = 4$ e $\tau = 1$. Áudio original (a) e áudio mais ruído Gaussiano branco com SNR = -3 dB (b).

No plano de Entropia-Complexidade, as linhas azuis indicam os limites inferior e superior considerando todas as possíveis combinações de histogramas e divergências. Em outras palavras, todos as séries estão contidas dentro da região demarcada pelos limites. A linha verde representa a posição dos ruídos coloridos. A posição de cada ponto no plano indica uma relação entre determinismo e aleatoriedade. Assim, os pontos situados próximos do canto inferior direito possuem um comportamento mais estocástico, quando comparados aos pontos que encontram-se mais a esquerda. Pontos com valores elevados de C_s, identificam componentes com distribuições de probabilidade muito divergentes da distribuição uniforme que representa os ruídos. Com estas interpretações, em nossa aplicação de filtro seria ideal encontrar os componentes que minimizam H_s e ao mesmo tempo possuam um valor elevado de C_s.

Observando as figuras 6.15(a) e 6.16(a), podemos notar que os componentes com grau maior de determinismo, os que foram identificadas na figura 6.9(c) com nosso algoritmo de agrupamento binário, encontram-se na região de maior complexidade



(a) Série original * e os componentes V_i.

(b) Série original mais ruído branco * e os componentes V_i.

Figura 6.16. Planos de Entropia-Complexidade estatística generalizada utilizando H_s com $m = 5$ e $\tau = 1$ dos componentes de V. Áudio original (a) e áudio mais ruído Gaussiano branco com SNR=-3 dB (b).

do plano HxC. Além disso, notamos que os mesmos componentes situam-se em uma região próxima ao sinal original (*). Assim, se os componentes com menor entropia ($i = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 15\}$) são escolhidos para a reconstrução usando o critério H_t, a maior fração da complexidade estatística do sistema físico de vocalização do anuro será preservada pelo filtro. Ao mesmo tempo, isto significa que o SSA consegue decompor o áudio original em componentes com diferentes grau de complexidade estatística quando analisada com H_s.

Comparando as figuras 6.15(b) e 6.16(b), percebemos que o aumento dos ruídos provocou um deslocamento do sinal original até a região considerada mais aleatória ($H_s \to 1$). No entanto, os PCs que foram identificados com menor entropia no agrupamento da figura 6.9(c) (λ_1, λ_2 e λ_3}, permaneceram com maior complexidade e situados a esquerda do sinal contaminado (*). Isto significa que, além de possuir um grau maior de determinismo, possuem também elevada amplitude, pois o aumento dos ruídos não foi suficiente para quebrar as correlações fortes e mudar substancialmente sua dinâmica. Estes componentes foram mais resilientes ao aumento dos ruído, e portanto, são ótimos candidatos para gerar uma melhor reconstrução do sinal contaminado.

Como foi concluído no capítulo 4, a adição de ruídos brancos ajuda a descorrelacionar o sinal, quebrando as autocorrelações fracas, concluindo que as sílabas segmentadas tinham um maior grau de determinismo. Assim, com estas conclusões e com as observações das figuras 6.15 e 6.16, podemos inferir quais PCs das vocalizações carregam a maior informação das sílabas. Lembramos também que H_s é um método não

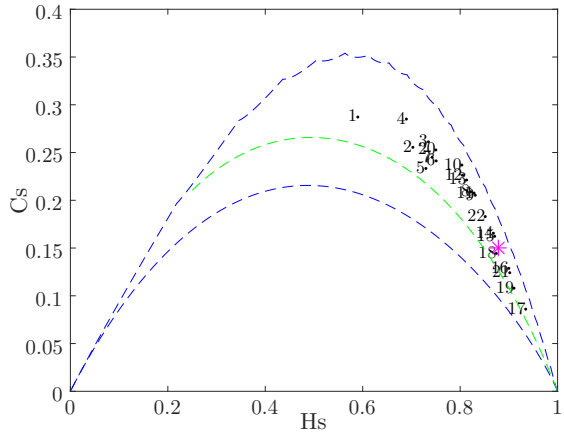
linear, e portanto, a localização dos componentes no plano HxC fornece informações adicionais sobre a não linearidade da estrutura interna dos componentes separadas pelo SSA.

Finalmente, as figuras 6.17 e 6.18 apresentam os planos HxC das espécies Hyla minuta e Aplastodiscus perviridis obtidos com $m = 4$ e $m = 5$ respectivamente. Para cada caso, ilustramos a decomposição da gravação original e uma versão contaminada com ruído branco a -3 dB. Novamente, podemos identificar os componentes mais determinísticos pela posição no plano. Também podemos identificar visualmente os componentes que foram descorrelacionadas e os que resultaram menos afetadas pelo ruído. Reforçamos a observação que, os componentes identificados pela critério de filtro H_t, representam ruído coloridos, pois estes se deslocaram no plano HxC até posições extremas de aleatoriedade, quando foram descorrelacionados por efeito do ruído branco. Isto é mais um indicativo que o filtro SSA-H_t é ótimo para nossa aplicação.

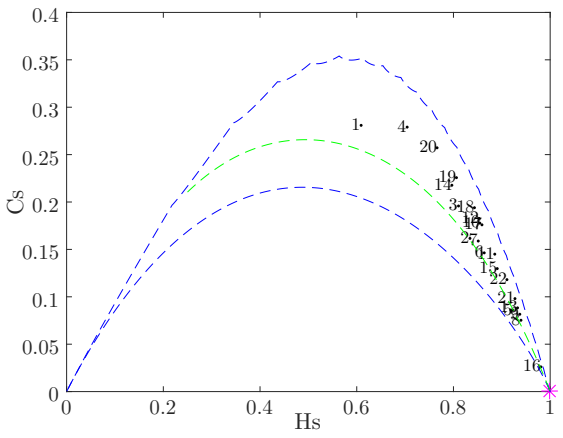
6.8.2 Considerações sobre a complexidade dos componentes acústicos

A representação no plano HxC dos componentes obtidas pelo SSA apresentou ser útil para identificar propriedades dos sinais. Na primeira observação, notamos que após contaminar o sinal original com ruído branco, a complexidade estatística mudou drasticamente, trasladando o ponto no * até as regiões de maior entropia. Embora a qualidade do sinal se degrade, um subconjuto dos componentes que o conformam mantém suas propriedades determinísticas, minimizando sua entropia. A observação na variação das posições dos componentes, permite-nos estabelecer as propriedades físicas destes e a influência que o ambiente da floresta pode causar nas mesmas.

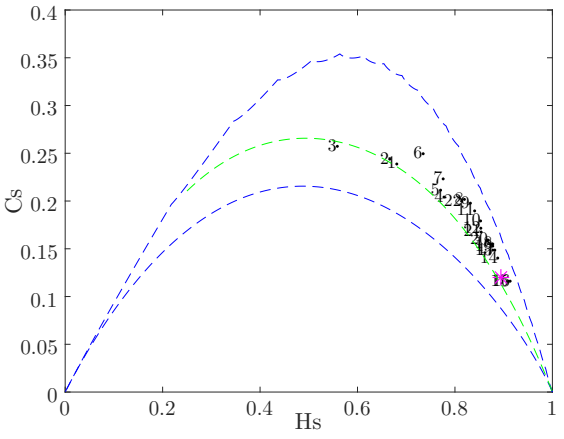
A segunda observação indica que, os componentes projetadas dos sinais no subespaço criado pelo SSA, não se relacionam apenas com a energia deles, mas também com a "força" das correlações lineares, que identificam o terminismo nas vocalizações produzidas. Assim, a partir da posição no plano HxC, poderia se estabelecer um novo critério para escolher e combinar os melhores componentes. Porém, um critério visual não permite desenvolver um método não supervisionado. A posição das vocalizações ou dos seus componentes no plano, poderiam ser utilizadas para identificar as diferentes espécies ou indivíduos.



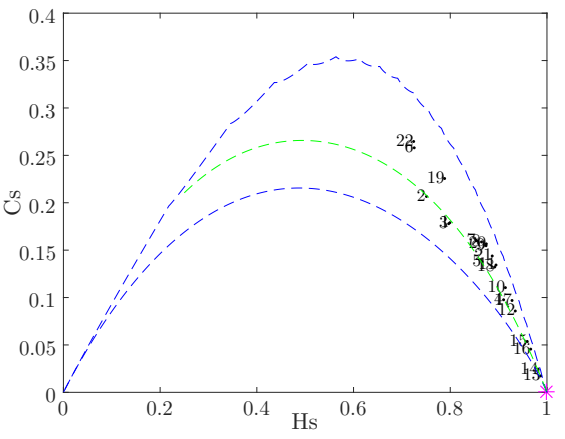
(a) Série original Hyla minuta.



(b) Série Hyla minuta mais ruído branco.



(c) Série original Aplastodiscus perviridis.

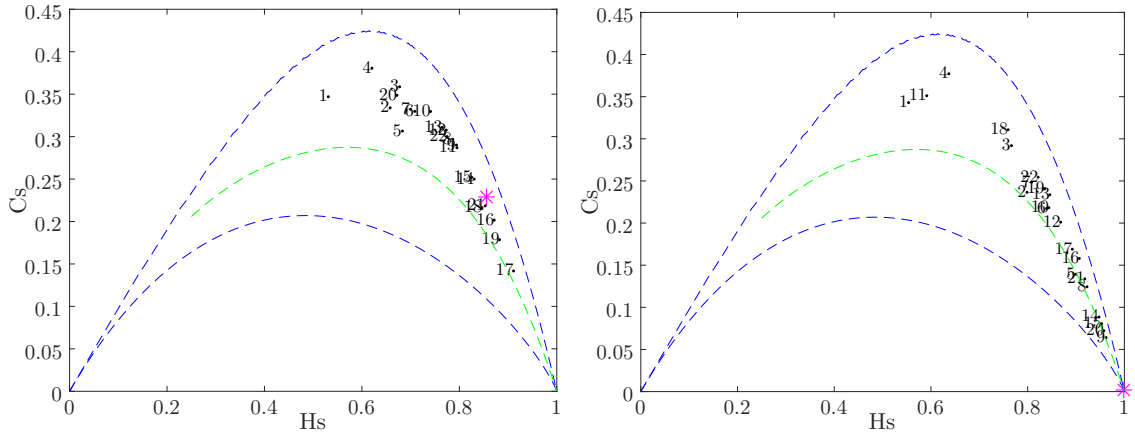


(d) Série Aplastodiscus perviridis mais ruído branco.

Figura 6.17. Planos HxC dos componentes V_i das espécies Hyla minuta e Aplastodiscus perviridis usando H_s com $m = 4$ e $\tau = 1$. Na coluna esquerda o áudio original (a) e a direita o áudio mais ruído Gaussiano branco com SNR = -3 dB (b).

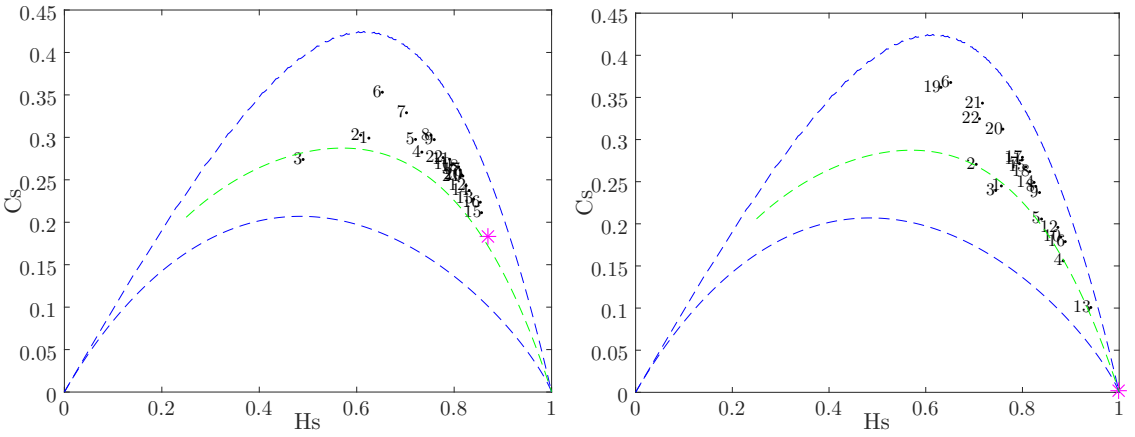
6.9 SSA Robusto

Nesta seção apresentamos uma variante robusta ao método SSA. Como foi explicado anteriormente, os ruídos aleatórios, principalmente os ruídos brancos, distribuem sua energia em todas os PCs do SSA. Em outras palavras, os ruídos contaminam todas as projeções do sinal no seu próprio subespaço. Na seção anterior, apresentamos alguns critérios para identificar e eliminar as projeções do sinal que possuem um comportamento menos determinístico, similar a um ruído sem correlação, para reconstruir o sinal "limpo", sem esses componentes. No entanto, parte dos ruídos ainda permanecem na reconstrução, pois os componentes que melhor descrevem o sinal também possuem uma fração da energia de dos ruídos. Para amenizar este problema, propomos uma



(a) Série original Hyla minuta.

(b) Série Hyla minuta mais ruído branco.



(c) Série original Aplastodiscus perviridis.

(d) Série Aplastodiscus perviridis mais ruído branco.

Figura 6.18. Planos HxC dos componentes V_i das espécies Hyla minuta e Aplastodiscus perviridis usando H_s com $m = 5$ e $\tau = 1$. Na coluna esquerda o áudio original (a) e a direita o áudio mais ruído Gaussiano branco com SNR = -3 dB (b).

versão mais robusta do SSA.

O SSA pode ser considerado uma técnica de filtragem ótima quando os ruídos que contaminam os sinais são Gaussianos (Chen and Sacchi, 2013). Entretanto, quando estes ruídos são gerados a partir de distribuições não Gaussianas, por exemplo Cauchy, o possuem *outliers* tipo ruído impulsivo, o SSA torna-se subótimo. Nestes casos, é necessário uma técnica robusta que seja capaz de recuperar os sinais mesmo quando a contaminação não é Gaussiana. Por estes motivos, desenvolvemos uma versão robusta do SSA, que chamamos *Robust SSA* (RSSA). Conseguimos a robustez substituindo o coeficiente de autocorrelação linear utilizado para obter a matriz de autocorrelações ($S = XX^T$) pelo coeficiente de autocorrelação *Quadrant (sing)*.

Em nossos experimentos, utilizamos sinais sintéticos e sinais bioacústicos das chamadas dos anuros para mostrar que o RSSA possui: (1) menor sensibilidade ao ruído aditivo gaussiano comparado com o SSA, (2) robustez ao ruído impulsivo (*outliers*) e a ruídos gerados por distribuições de Cauchy, e (3) menor SDR na reconstrução dos sinais bioacústicos. Além disso, mostramos que o RSSA revela uma contribuição diferente de cada PC no espectro singular, simplificando a detecção do *trend* do sinal (ou componentes de baixa frequência). Com a modificação proposta, não foi necessário adicionar parâmetros de ajuste extra, permanecendo o RSSA essencialmente não paramétrico.

6.9.1 Coeficiente de autocorrelação robusto

O coeficiente de correlação de Pearson (r_{xy}) é uma medida da força e direção da relação linear entre duas variáveis x e y. Este coeficiente pode ser definido como a covariância entre x e y dividida pelo produto dos desvios padrões amostrais (Wilcox, 2012).

O coeficiente de autocorrelação r_{xx} tem a mesma interpretação do r_{xy}, mas em vez de utilizar duas variáveis diferentes o cálculo é aplicado entre x e uma versão deslocada da mesma variável τ unidades de tempo, da forma:

$$r_{xx}(\tau) = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} \left(\frac{x_i - \bar{x}}{s} \right) \left(\frac{x_{i+\tau} - \bar{x}}{s} \right), \tag{6.13}$$

onde N é o comprimento do sinal, e $\bar{x}$ é a média amostral e s o desvio padrão amostral. Por exemplo, se $\tau = 1$ então r_{xx} quantifica a força da associação entre x_i e x_{i+1}. Neste caso, o valor máximo de τ deve satisfazer a condição $\tau \ll N$. A média e a variância amostrais são severamente afetadas pela ocorrência de *outliers*, e consequentemente a autocorrelação também é afetada (Wilcox, 2012). Portanto, um método mais robustos e menos propensos a erros é necessário.

Métodos estatísticos robustos foram desenvolvidos para evitar problemas quando as suposições dos métodos paramétricos clássicos não são satisfeitas ou quando a presença de ruídos não gaussianos afeta as estimativas (Maronna et al., 2006). No que diz respeito ao coeficiente de autocorrelação robusto, podemos encontrar alternativas que utilizam métodos paramétricos e não-paramétricos, tais como: o *Correlation Median Estimator* (r_{COMED}) ou *Quadrant* ($\operatorname{sign} r_Q$), respectivamente (Shevlyakov and Smirnov, 2011). Também pode-se obter um coeficiente de correlação robusto através de *Robust Principal Variables*, por exemplo, utilizando o *median absolute deviation correlation coefficient* (r_{MAD}), o coeficiente de correlação trucado (ou *trimmed correlation coefficiente* r_{T}) e o coeficiente de correlação da mediana (r_{Med}). Outra alternativa, inclui o coeficiente de correlação usando regressão robusta (r_{REG}). A definição destes coefici-

entes podem ser consultados em Shevlyakov and Smirnov (2011), Shevlyakov and Oja (2016), Kharin and Voloshko (2011)

Todos estes coeficientes são alternativas viáveis ao coeficiente de Pearson padrão (equação 6.13), uma vez que estimam relações lineares e, adicionalmente, fornecem uma melhor estimativa da relação entre as variáveis quando as distribuições marginais se desviam da hipótese Gaussiana (Wilcox, 2012). Entre estas alternativas, o coeficiente de autocorrelação r_Q produz um estimador robusto notável não-paramétrico e com baixa complexidade computacional, como comprovado por Shevlyakov and Smirnov (2011).

O r_Q é o coeficiente de correlação amostral obtido pela diferença dos sinais (positivo e negativo) a partir da variável e sua mediana:

$$r_Q(\tau) = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} \big(\text{sign}(x_i - \tilde{x})\ \text{sign}(x_{i+\tau} - \tilde{x})\big), \tag{6.14}$$

onde $\tilde{x}$ é a mediana e sign$(\cdot)$ é definida como:

$$\text{sign}(z) = \begin{cases} +1, & \text{if } z > 0 \\ 0, & \text{if } z = 0 \\ -1, & \text{if } z < 0 \end{cases}. \tag{6.15}$$

A função sign(x) discretiza os valores originais de x em três níveis independentes da amplitude de x, tornando-se menos sensível à presença de *outliers*. A robustez do coeficiente $r_Q(\tau)$ decorre da utilização da mediana de x. Assim, propomos o método RSSA baseado neste coeficiente.

6.9.2 Limitações do SSA e vantagens de nossa proposta robusta

O sinal reconstruído pelo SSA é obtido mediante a minimização dos mínimos quadrados (LS) da diferença entre o sinal e sua projeção usando SVD. Assim, a reconstrução SSA garante a menor SDR e, ao mesmo tempo, o maior nível de ruído residual quando é assumida uma distribuição Gaussiana dos dados. No entanto, uma desvantagem do SSA é que o desempenho do estimador de mínimos quadrados dependente do posto da matriz (*matrix rank*), ou do sinal neste caso (Hassani and Thomakos, 2010, Hassani et al., 2014, Kalantari et al., 2016). Ou seja, temos que selecionar quais PCs do SSA farão parte da reconstrução e descartar os restantes, sem alterar o posto da matriz dos dados.

Ademais, outro desvantagem crítica do SSA é sua sensibilidade ao desvio da normalidade dos sinais. De acordo com Chen and Sacchi (2013): "SSA é eficiente para atenuar o ruído gaussiano, mas não pode eliminar o ruído errático", e segundo Hassani et al. (2014): "a existência de *outliers* muda o posto da matriz dos dados, aumentando as dimensões recorrentes lineares e resultando num maior número de autovalores significativos, o que produz impactos negativos na fase de reconstrução". Similares limitações são verificadas mais adiante nesta seção.

Devido às capacidades do SSA para detectar padrões regulares, cada vez que um padrão está ausente ou distorcido por efeito dos ruídos, o sinal reconstruído, bem como os PCs, sofrem variações indesejáveis. Esta sensibilidade é uma característica inerente do coeficiente de autocorrelações do sinal (Wilcox, 2012), o qual constitui uma parte fundamental do método SSA.

Para superar este inconveniente, propomos uma modificação de SSA, onde o cálculo original da matriz de autocorrelações é substituído por uma matriz robusta baseada no coeficiente robusto *Quadrant (sing)*. Chamamos esta variante proposta de "Robust Singular Spectrum Analysis" (RSSA). O termo "robusto" tem sido utilizado em muitos contextos no processamento de sinais (Zoubir et al., 2012). Mas, nós adotamos este termo para designar o tratamento de correlações suspeitas ou espúrias causadas por ruídos e *outliers*, que causam desviações nas estruturas de autocorrelação dos sinais originais. A proposta RSSA atinge:

1. uma melhor reconstrução em termos de SDR, inclusive nos casos em que a contaminação é Gaussiana,

2. robustez à presença de *outliers* ou ruídos com distribuição não Gaussiana, e

3. uma curva de autovalores diferente, que revela novas contribuições dos PCs, ou seja, um novo espectro singular.

Nós conseguimos estas melhorias sem adicionar nenhum parâmetro extra no SSA original, somente substituímos do coeficiente de autocorrelação robusto (equação 6.15). Desta forma, mantemos a mesma complexidade computacional do SSA $(O(LK)^3)$.

Considerando o modelo de contaminação apresentado na equação 6.2, o problema de filtragem robusto é definido como: eliminar ou diminuir os efeitos de ξ quando este é Gaussiano, não Gaussiano ou um conjunto de *outliers*. Assim, o objetivo geral permanece inalterado como foi definido na seção 6.5, isto é, minimizar o SDR e o MSE da reconstrução. Portanto, a seguir, vamos comparar o SSA e RSSA usando a taxa de distorção para os casos Gaussiano (não existem *outliers*) e não Gaussiano, por exemplo, quando $\xi \sim Cauchy$ ou ξ representa ruído impulsivo.

6.9.3 Avaliações do RSSA

Realizamos dois tipos de experimentos para mostrar as principais diferenças e vantagens do RSSA em relação ao SSA. Nas primeiras duas avaliações utilizamos um sinal artificial e adicionamos a este ruído Gaussiano e não Gaussiano. Na segunda avaliação, analisamos duas vocalizações de anuros com ruído de fundo da floresta.

6.9.3.1 Simulação com ruído Gaussiano

O sinal de teste simulado é ilustrado na figura 6.19(a) (linha verde). Este sinal foi definido como $x = \sin(8\pi t)\sin(t)$, onde t varia entre $0 \leq t \leq 4\pi$, com frequência de amostragem igual a 50 Hz. Este sinal modulado em amplitude, foi definido com o propósito de reproduzir um sinal bioacústico de baixa frequência e aplicarmos normalização para ter variância unitária ($\sigma_x = 1$). Após a normalização, foi adicionada uma contaminação $\xi \sim \mathcal{N}(0, \sigma_n)$. Assim, quanto maior é a variância da contaminação, mais severo são os efeitos de distorção causados no sinal original. A variação do parâmetro σ_ξ nos permite simular diversas situações de degradamento do sinal quantificados pela SNR (equação 2.23, página 32).

Na figura 6.19, apresentamos uma instância de simulação para a condição SNR = 0 dB. Neste exemplo, realizamos uma decomposição com $L = 20$ autovalores e escolhemos os sete primeiros componentes principais ($\lambda_{1:7}$) para realizar a reconstrução com os dois métodos, SSA e RSSA. Os detalhes das reconstruções mostram que o $\hat{x}_{RSSA}$ (linha vermelha) permaneceu mais próximo do sinal limpo x (linha verde) comparado com o $\hat{x}_{SSA}$, principalmente nos pontos em que ξ tem picos de amplitude elevada. Isso exemplifica o fato do r_Q ser menos sensível às variações de amplitudes do que o r_{xx}. Também, neste exemplo, o SDR do RSSA foi aumentado em 5.11 dB em comparação ao SDR do SSA.

Ao se aumentar o ruído, em ambos métodos a qualidade da reconstrução é degradada. Como foi apresentado na seção anterior, a adição de ruído altera também o espectro singular, e portanto, a contribuição relativa dos PCs. As figuras 6.20(a) e 6.20(b) ilustram a mudança dos autovalores dos sinais limpos e ruidosos, respectivamente. Aqui, os autovalores foram normalizadas de acordo com a equação 2.41 (página 42) para obter a porcentagem de contribuição de cada PCs.

O espectro singular de x obtido pelo SSA rapidamente atinge zero (figura 6.20(a). Neste exemplo sintético, SSA tem uma representação mais compacta do que RSSA quando a distribuição dos dados não é alterada. Além disso, o espectro singular do RSSA (figura 6.20(b)) revela um outro grau de contribuição de cada componente, produzindo uma decomposição ligeiramente diferente, distribuindo a energia de sinal

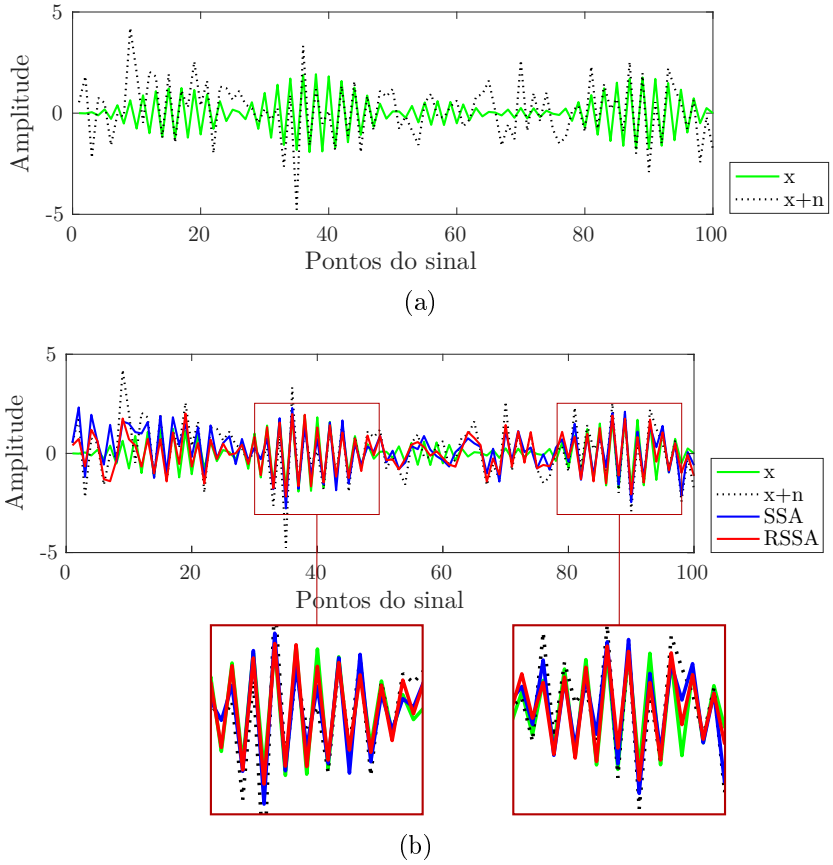
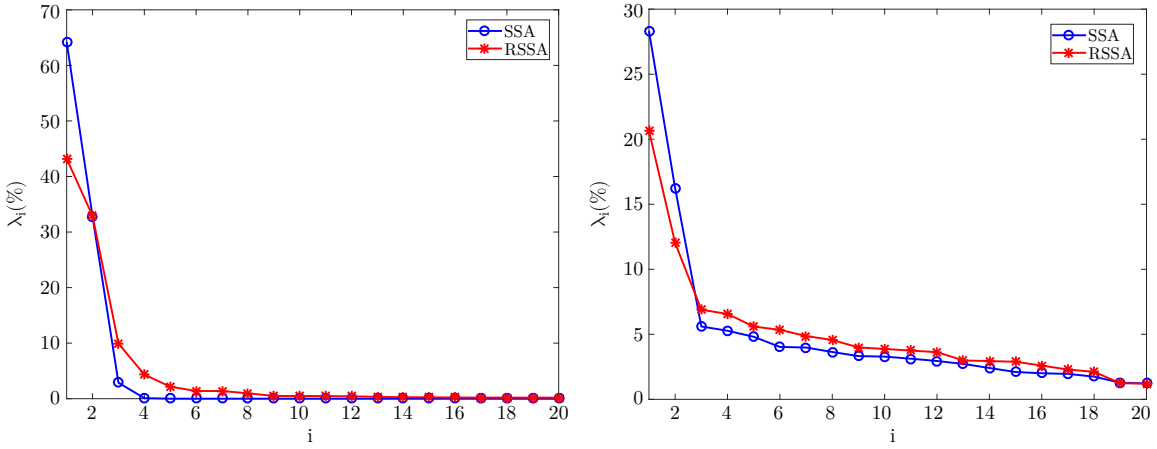


Figura 6.19. Sinal de amplitude modulada x (green line) e sua versão contaminada com ruído branco a SNR = 0 dB (a). Reconstruções (filtragem) utilizando as duas abordagens SSA e RSSA com sete PCs (b). A parte inferior da figura (b) apresenta uma amplificação dos detalhes das reconstruções. Neste caso, o RSSA resultou 5.11 dB melhor do que o SSA.

entre um conjunto maior de bases. Apesar disso, a reconstrução com RSSA é próxima de SSA quando o sinal de entrada permanece inalterado, mas em contraste, quando $x \to \hat{x}$, o RSSA obtém uma melhor reconstrução.

O espectro singular do RSSA torna-se mais suave do que no SSA, como consequência natural da baixa sensibilidade do coeficiente r_Q. No entanto, ambos espectros sofrem variações quando σ_ξ aumenta. A variação dos autovalores como consequência do aumento dos ruídos tinha sido ilustrado na figura 2.17, no entanto, aqui confirmamos que RSSA possui a mesma propriedade. A figura 6.20(b) mostra estas variações para um ruído aditivo branco com SNR = 0 dB. Como esperamos, o ruído aditivo não correlacionado distribui sua energia uniformemente entre todas as bases e projeções do sinal, modificando a contribuição de cada PC do RSSA.

Além deste exemplo, nós comparamos o desempenho do RSSA contra o SSA,



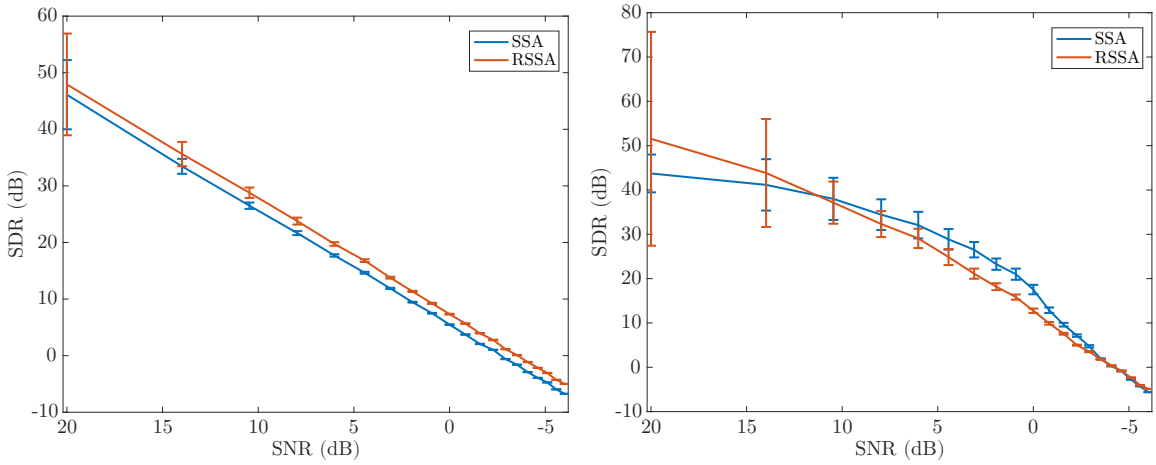
(a) Espectro singular de $x(t)$. (b) Espectro singular de $x(t) + n(t)$.

Figura 6.20. Mudanças dos espectros singulares antes e depois da adição de ruído a SNR $= 0\,\text{dB}$.

simulando diversas condições de ruído. Nesta simulação, variamos a quantidade de ruído adicionado no intervalo $\sigma_\xi \in [0, 2\sigma_x]$. Assim, para cada valor de SNR, realizamos duzentas repetições independentes de contaminação, decomposição e reconstrução. Os resultados médios e os seus intervalos de confiança de 95% são apresentados na figura 6.21. Na figura 6.21(a), avaliamos um agrupamento fixo arbitrário para a reconstrução escolhendo os oito autovalores principais ($\lambda_{1:8}$). Neste caso, observamos que RSSA tem um ganho aproximadamente constante em comparação com SSA em termos de SDR quando SNR varia. A partir do ponto SNR $\leq 15\,\text{dB}$, os intervalos de confiança não se sobrepõem, mostrando que o desempenho do RSSA é significativamente melhor, concluindo que, quanto maior for σ_ξ, melhor é o SDR do RSSA.

Os resultados de reconstrução mostrados na figura 6.21(b), foram obtidos aplicando o critério de agrupamento automático da média introduzido na seção 2.4.1 (equação 2.45). Consequentemente, os grupos formados por ambas as técnicas podem variar de acordo com o nível de σ_ξ. Três observações interessantes surgem da figura 6.21(b): (1) existe um ponto crítico a partir do qual um método se torna vantajoso em comparação ao outro (SNR=12 dB); (2) a variância da reconstrução usando RSSA foi maior quando o SNR aumentou, mostrando uma pior estabilidade para valores baixos de σ_ξ; (3) para condições de ruído severas (SNR $\leq -3\,\text{dB}$), onde o sinal é completamente indistinguível do ruído, ambos os métodos têm um desempenho inaceitável.

Finalmente, comparando as figuras 6.21(a) e 6.21(b) notamos que, para cada grupo específico de autovalores, a qualidade da reconstrução tem um comportamento diferente dependendo de σ_ξ. Portanto, na seguinte seção, avaliamos o efeito do agrupamento usando todas as combinações de autovalores sucessivos.



(a) SDR da reconstrução agrupamento fixo $\lambda_{1:8}$. (b) SDR da reconstrução utilizando agrupamento automático da média.

Figura 6.21. Relação entre o SDR e o SNR em dB quando varia-se σ_ξ. As barras de erro verticais indicam um intervalo de confiança de 95%.

6.9.3.2 Avaliação do agrupamento sequencial das componentes principais sob contaminação normal

Nossos objetivos visam melhorar o SDR da reconstrução para uma ampla gama de PCs quando a distribuição dos sinais é alterada pelo efeito dos ruídos. Analisando apenas uma uma condição (SNR $\approx 0\,\text{dB}$) não é suficiente para avaliar o desempenho completo do RSSA. Portanto, nesta seção, testamos todos os grupos possíveis sequenciais $\lambda_{1:i}, i \in [1, L]$, e quantificamos o ganho G_{SDR} do RSSA em relação ao SSA. A eficiência foi avaliada em dB como:

$$\text{G}_{\text{SDR}} = \text{SDR}_{\text{RSSA}} - \text{SDR}_{\text{SSA}}, \tag{6.16}$$

onde SDR_{RSSA} e SDR_{SSA} são obtidos usando o mesmo conjunto de PCs. Assim, se G_{SDR} for positivo para um determinado grupo $\lambda_{1:i}$, isso significa que o RSSA foi mais eficiente do que SSA.

Cada curva da figura 6.22 representa o G_{SDR} para todos os grupos de autovalores. Novamente, cada ponto deste gráfico foi obtido como a média de duzentas simulações. O eixo horizontal (linha tracejada preta) é o limiar acima do qual os ganhos do RSSA são superiores, isto é: os agrupamentos que geram uma curva sobre esta linha possuem um melhor SDR_{RSSA}, caso contrário, SDR_{SSA} é melhor.

Na figura 6.22, notamos que todos os grupos compreendidos entre $\lambda_{1:i}, i \in [5, 19]$ do RSSA têm melhor desempenho do que SSA, independentemente da SNR. Além disso, estas curvas mostram um comportamento mais regular, com uma tendência aproximadamente horizontal, em comparação aos agrupamentos $\lambda_{1:i}, i \in [1, 4]$. Além

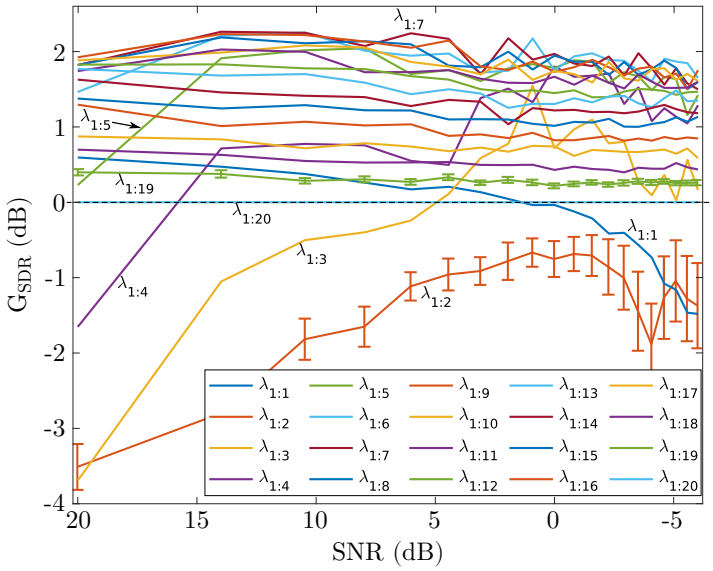


Figura 6.22. Ganhos de RSSA em relação ao SSA representado por SDR como função do SNR para todos os agrupamentos possíveis $\lambda_{1:L}$.

disso, os agrupamentos $\lambda_{1:i}, i \in [1, 4]$ possuem ganho SDR_{SSA} maior quando o SNR é baixo, como era esperado. Esse fato confirma que, para um sinal que possui uma matriz de autocorrelação de posto baixo e contaminação Gaussiana, o SSA permanece eficiente. Mas quando o SNR aumenta, as curvas cruzam o limiar indicando que o SDR_{RSSA} tem melhor desempenho. Embora, numa situação real, o posto da matriz de autocorrelações e o tipo de contaminação podem ser desconhecidos *a priori*, pode se optar por avaliar os dois métodos e comparar as reconstruções.

As únicas exceções foram os agrupamentos $\lambda_{1:1}$ e $\lambda_{1:2}$. No primeiro caso, o SDR_{RSSA} foi melhor indicando que a compactação da informação retida pelo RSSA foi maior no PC. O segundo caso particular mostra que, dado o número ótimo de componentes oscilatórios, que neste caso são dois, o SSA foi superior ao RSSA para todos os níveis de contaminação por ruído branco.

A figura 6.22 inclui os intervalos de confiança para as curvas $\lambda_{1:2}$ e $\lambda_{1:9}$ somente. Os intervalos das curvas restantes foram removidos da figura para melhorar a clareza e evitar uma excessiva contaminação visual. Todavia, os intervalos incluídos na figura fornecem um ideia do comportamento geral da variabilidade do SDR, quando aumenta-se o agrupamento e o SNR. Isto é, os intervalos diminuem quando o agrupamento de componentes aumenta, tornando-se a reconstrução mais estável. Apesar da omissão na figura, nós garantimos que todos os intervalos dos agrupamentos $\lambda_{1:i}, i \in [5, 19]$ permaneceram sempre acima do limiar indicado.

Finalmente, para o caso base $\sigma_\xi = 0$, esperamos que o RSSA tenha um desempe-

nho de reconstrução próximo ao SSA, pois a distribuição original dos dados permanece inalterada. Para testar esta hipótese, realizamos uma simulação com as mesmas configurações que o caso anterior, mas apenas para um único valor de σ_ξ. Neste caso, substituímos o SDR pelo erro quadrático médio (MSE) da reconstrução para evitar computar o logaritmo de pequenos números durante o cálculo de SDR. A figura 6.23 apresenta a média do MSE e seus intervalos de confiança como função do agrupamento $\lambda_{1:i}, i \in [1, L]$. As pequenas diferenças entre o MSE do SSA e o MSE do RSSA confirmam a hipótese de que o RSSA não perde informações relevantes durante a decomposição. Assim, mostramos o RSSA pode ser tão eficiente quanto SSA quando o ruído é Gaussiano.

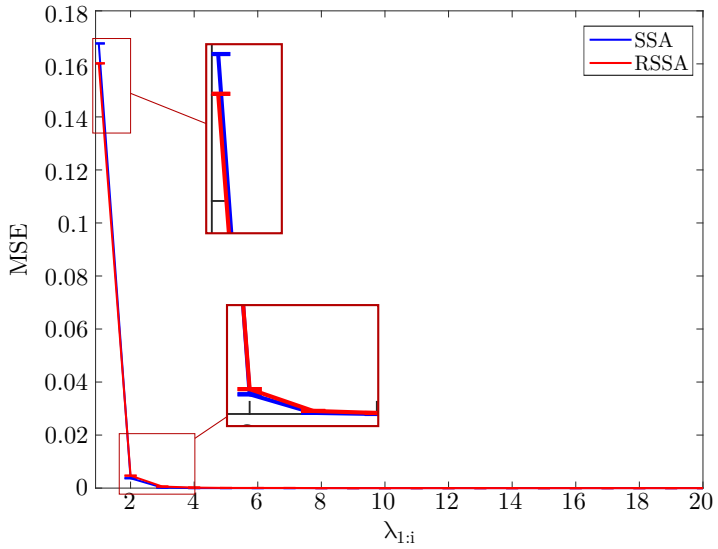


Figura 6.23. MSE de $\hat{x}$ para a condição base $\xi = 0$.

6.9.3.3 Avaliação do agrupamento sequencial das componentes principais sob contaminação de Cauchy

Para testar a robustez do RSSA, contaminamos y com ruído não Gaussiano utilizando uma distribuição de Cauchy$\sim \mathcal{C}(x_0, \gamma)$, onde x_0 e γ são os parâmetros de localização e escala respectivamente (Aysal and Barner, 2007). Esta distribuição de probabilidades descreve um fenômeno de ressonância com média, variância e momentos superiores indefinidos, e mediana zero. Assim, a contaminação foi gerada de acordo com uma PDF com caudas laterais dadas por $\xi \sim \mathcal{C}(0, 1)$, as quais são superiores às caudas de uma distribuição normal. Depois disso, a variância amostral de ξ foi reescalada para gerar vários níveis de SNR.

O procedimento de avaliação experimental neste caso é igual ao descrito na seção anterior. A figura 6.24 mostra os ganhos em termos de SDR comparando RSSA contra SSA. Novamente, observamos que o RSSA é mais robusto que o SSA para um elevado agrupamento de autovalores ($\lambda_{1:i}, i \in [5, 19]$). Isto sugere que ao se utilizar, por exemplo, uma regra típica para a reconstrução automática, como reter o 95% autovalores, o RSSA pode ser uma ótima escolha.

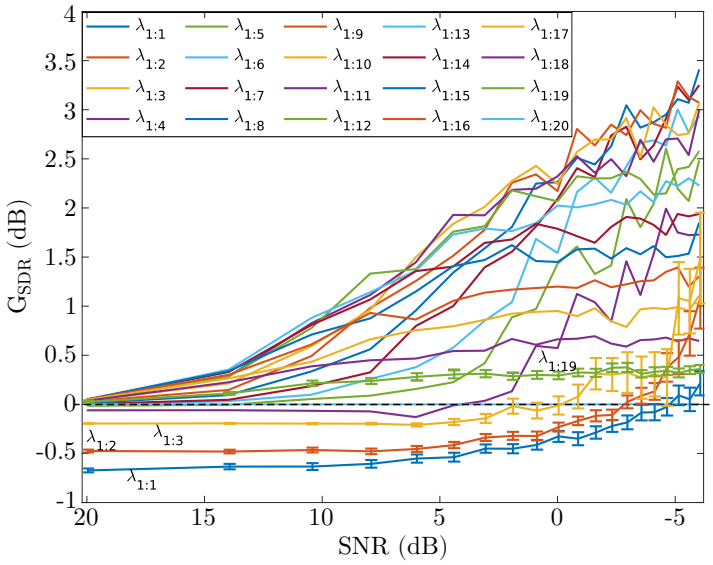


Figura 6.24. Ganhos do RSSA sobre SSA representado pela SDR contra a SNR com ruído gerado a partir de uma distribuição de Cauchy.

6.9.3.4 Avaliação do agrupamento sequencial das componentes principais sob contaminação impulsiva (*outliers*)

O segundo teste de robustez realizado é a tolerância aos ruídos impulsivos. Aqui, o ruído impulsivo (ou ruído de pico) é a ocorrência esparsa de impulsos instantâneos com alta energia e curta duração, denotados por δ. Este tipo de contaminação é um sinal não-estacionário com tempos aleatórios de ocorrência entre cada impulso (Vaseghi, 2008). Tipicamente, esses impulsos são denotados por $\pm\delta_k$, onde k representa a posição temporal de cada impulso (Hassani et al., 2014).

Em nossas simulações definimos os valores de pico máximos e mínimos de acordo com $\delta(\cdot) = \pm 2\sigma_x$. Assim, o sinal contaminado resulta $y = x \pm \delta_k$. A diferença entre os tempos de ocorrência de cada impulso Δk é uma variável aleatória uniforme e a densidade de ruído total pode ser obtida como a razão entre a quantidade total de impulsos K dividida pelo comprimento total do sinal N (equação 2.24, página 32). Este tipo de ruído é completamente descorrelacionado do sinal acústico e também não correla-

cionado contra si mesmo, ou seja $r_{xx} = 0$. Este fenômeno pode aparecer em sensores acústicos devido a descargas elétricas produzidas por interferência eletromagnética ou por causa dos sons com volume elevado do ambiente.

As avaliações foram realizadas considerando a densidade do ruído impulsivo no lugar da SNR. A figura 6.25 mostra o G_{SDR} como função da porcentagem $d_\delta(\%)$. Esta figura mantém as mesmas considerações da figura 6.22, isto é: as curvas acima do limiar $G_{SDR} = 0$ indicam que RSSA obteve um desempenho superior para o agrupamento especificado por $\lambda_{1:i}$. Neste caso, as curvas $\lambda_{1:i}, i \in [4, 15]$, e seus intervalos de confiança, permanecem acima deste limiar para todos os valores de densidade d_δ, mostrando que RSSA é mais robusto aos ruídos impulsivos. As curvas $\lambda_{1:1}$ e $\lambda_{1:2}$ começam mostrando uma superioridade do SSA, mas o G_{SDR} é invertido quando a densidade de ruído excede o valor $d_\delta \geq 0,35\%$. Observamos também que as curvas $\lambda_{1:i}, i \in [16, 20]$ convergem para um ganho quase nulo quando as condições de ruído são severas ($d_\delta \geq 0,9\%$). Finalmente, comparando os intervalos de confiança percebemos que os agrupamentos com um número maior de PCs resultam mais estáveis.

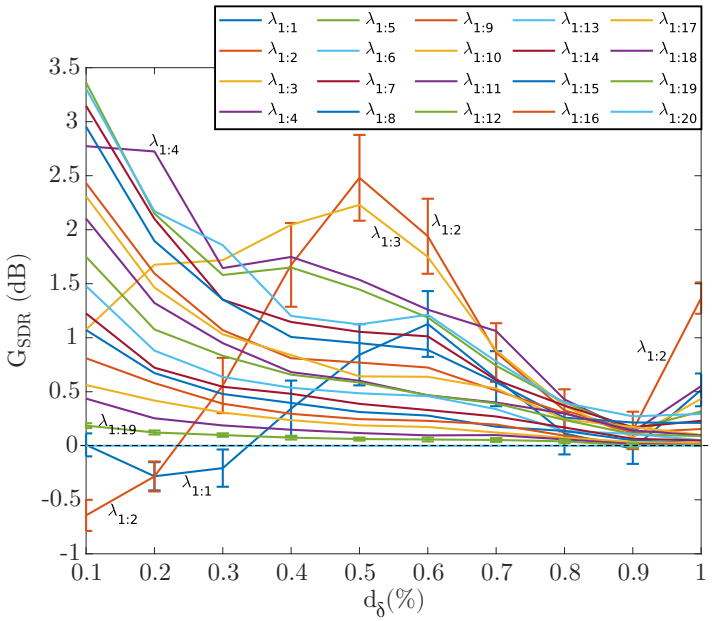


Figura 6.25. Ganhos de RSSA sobre SSA representado por SDR contra a densidade de ruído impulsivo em porcentagem.

6.9.3.5 Avaliações do RSSA em sinais bioacústicas

Nesta seção, empregamos dois registros de chamadas de anuros, incluindo o ruído de fundo da floresta, para avaliar a robustez de nossa proposta em um cenário realista. Uma sílaba de cada uma das gravações pertencentes às espécie Adenomera hylaedactyla

e Hyla minuta é ilustrado na figura 6.26. Nas figuras, pode se comparar visualmente as diferenças entre o sinal original e as reconstruções com RSSA e SSA. Neste caso, a decomposição foi realizada com $L = 20$ componentes e seus espectros singulares são ilustrados na figura 6.27, onde pode se observar as diferenças entre os espectros singulares gerados com os dois métodos. Além disso, a "média" de $\overline{\Lambda}$ foi adicionada nos espectros para dar uma ideia sobre um dos critérios de agrupamento mais utilizados descrito anteriormente.

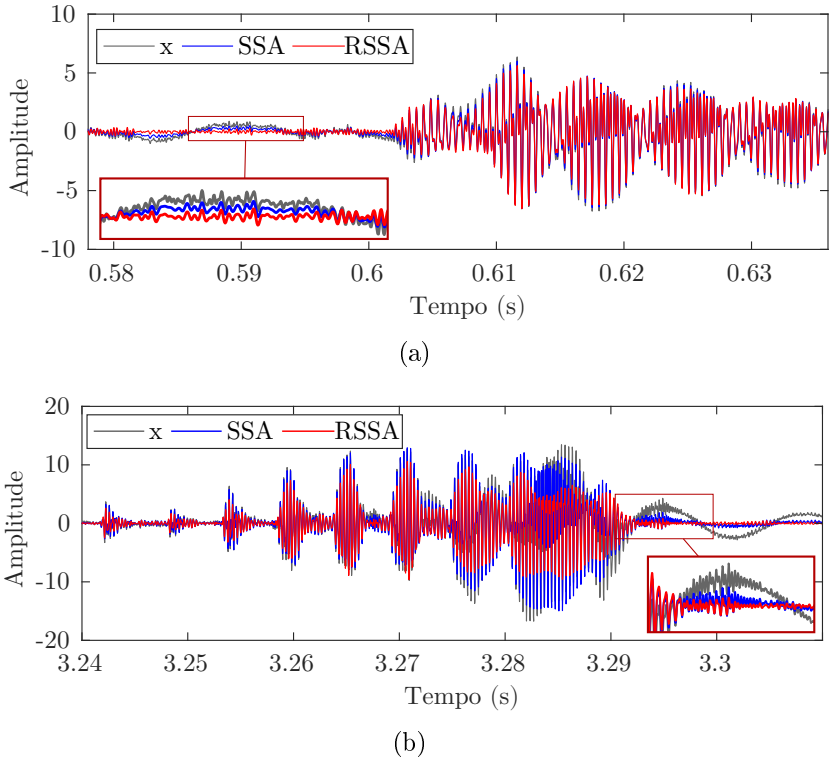


Figura 6.26. Reconstrução das vocalizações das espécies Adenomera hylaedactyla e (b) Hyla minuta usando a base $\lambda_{1:4}$ no caso SSA e $\lambda_{2:5}$ com RSSA. Aqui, observamos que o ruído aditivo de amplitude foi melhor removido pelo RSSA.

Nas figuras 6.27(a) e 6.27(b), o componente de baixa frequência (*trend*) causado pelos ruídos de fundo, aparece representado pelo primeiro componente singular nas duas gravações quando analisado com RSSA, enquanto que com SSA, o mesmo componente encontra-se na quinta posição. Esta observação torna-se mais evidente na figura 6.27(b). Os componentes de baixa frequência podem ser visualizados com mais detalhes nas reconstruções das figuras 6.28(a) e 6.28(b). Além da exatidão com a qual RSSA conseguiu identificar o *trend* do sinal, também observamos que o ranking dos valores singulares do espectro singular foi alterado, e que o componente associado ao *trend* encontra-se na primeira posição. Portanto, o novo agrupamento simplifica a

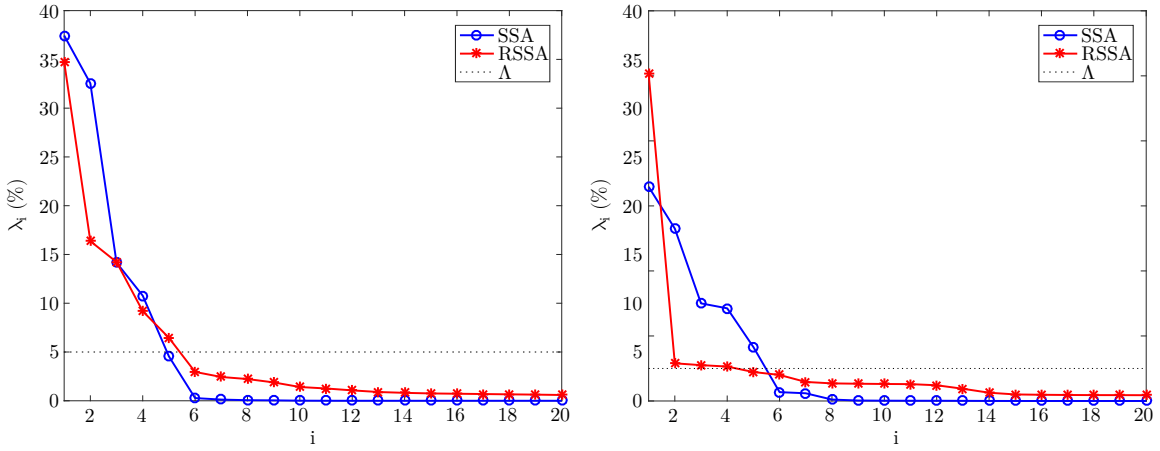


Figura 6.27. Espectros singulares das espécies (a) Adenomera hylaedactyla e (b) Hyla minuta com SSA e RSSA.

identificação desse componente e pode ser facilmente removido dos registros bioacústicos.

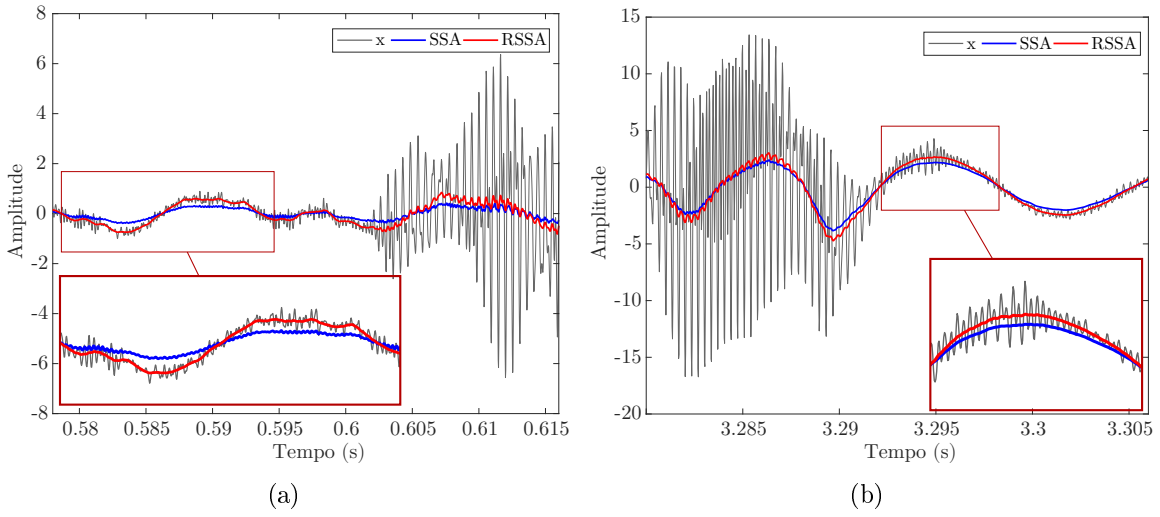


Figura 6.28. Reconstrução de (a) Adenomera h. e (b) Hyla m. utilizando somente λ_5 para SSA e λ_1 para RSSA. Com RSSA o componente oscilatório de baixa frequência foi melhor aproximado.

6.9.4 Conclusões sobre o RSSA

Nesta seção, apresentamos uma abordagem nova para transformar o método SSA tradicional em um método de decomposição robusto. Nosso método fornece um ajuste adequado dos PCs quando o sinal subjacente é contaminado com ruído Gaussiano, não

Gaussiano e com *outliers*. Os resultados das simulações sugerem que SSA é subótimo na presença de ruídos Gaussianos, porém continua sendo superior ao RSSA para pequenos agrupamentos das bases (e.g. $\lambda_{1:2}$). Isto mostra que um método robusto não é sempre necessário quando o posto da matriz de autocorrelações é baixo. No entanto, as perdas de eficiência do SSA podem ser muito maiores do que são assumidas previamente. Neste caso, RSSA apresenta melhor desempenho para agrupamento maiores das bases (e.g. $\lambda_{4:19}$), principalmente quando a variância dos ruídos aumenta consideravelmente.

Os dados coletados em uma ampla gama de aplicações frequentemente incluem ruídos e *outliers* que causam distorções nos padrões dos sinais. Os resultados apresentados aqui apoiam o uso do RSSA para problemas de análise e filtragem bioacústica. Nos experimentos o RSSA mostrou melhor desempenho, principalmente quando os sinais foram coletados em um ambiente real como a floresta tropical, devido às condições de ruído adversas e possíveis interferências. As principais vantagens do RSSA são:

- Baixa sensibilidade aos ruídos aditivos que afetam a forma de onda dos sinais;
- Habilidade para identificar e ressaltar o ruído ambiental de baixa frequência ou *trend*; e
- Reconstrução mais limpa do que o SSA.

Além disso, usando dados sintéticos e reais, mostramos que o RSSA produz um SDR baixo em comparação com o SSA para uma ampla gama de valores de SNR. O conceito de empregar um coeficiente de autocorrelação alternativo em SSA pode ser estendido a outros tipos de coeficientes, capazes de detectar outra estrutura de decomposição dos sinais de forma não paramétrica.

Finalmente, em nossos experimentos, percebemos que a qualidade da reconstrução do RSSA depende dos autovetores escolhidos durante a etapa de agrupamento, do mesmo modo que o SSA. Neste seção, nós aplicamos uma inspeção visual do espectro singular, uma abordagem utilizando a média dos autovalores e uma avaliação exaustiva considerando todos os grupos sequenciais dos componentes. No entanto, futuramente deveriam ser avaliados critério baseados em entropia.

6.10 Considerações finais

Neste capítulo, exploramos diferentes filtros de ruído para os sinais bioacústicos emitidos pelos anuros. As técnicas utilizadas foram: a subtração espectral (*Spectral Subtraction*), o *soft-threshold* utilizando a transformada *Wavelet* e *Signal Subspace*.

O método SSA foi escolhido para desenvolver nossa proposta de análise, sínteses e filtragem dos sinais bioacústicos. O SSA possui a característica de criar funções oscilatórias de diferentes frequências que podem ser interpretadas com bandas espectrais das vocalizações. Diferentemente dos filtros tradicionais, o SSA cria as funções da transformada a partir do sinal de entrada, tornando a decomposição sempre ótima. Verificamos também que, através da decomposição SVD da matriz de autocorrelações é possível identificar estruturas determinísticas e não determinísticas das vocalizações e sons da floresta.

A transformação SVD permite identificar os PCs mais significativos dos sinais ao longo das dimensões com maior covariância. Além disso, as projeções dos sinal são ortogonais, o que significa que cada um destas pode ser processada de forma independente. Sabemos também, que quando a matriz de autocorrelações possui posto maior do que o número de PCs, as últimas projeções do SSA representam os ruídos descorrelacionados. Com esta ideia, propomos uma nova metodologia para escolher os autovalores e autovetores. A regra proposta baseia-se em escolher os componentes que possuam valores baixos de entropia. Assim, conseguimos identificar os componentes mais determinísticos que compõem as vocalizações para realizar a reconstrução, eliminando a maioria dos ruídos sem causar distorções nos sinais.

Neste capítulo, apresentamos também a relação entre os *eigenfilters* e as bases do subespaço SSA. Mostramos que os autovetores do sinal possuem uma equivalência com o coeficientes dos filtros FIR, e que a partir desses vetores é possível construir um banco de filtros ótimos para cada sinal. Assim, o sinal recuperado possui a maior concentração de energia espectral no sinal original. Quando combinamos a teoria FIR com nosso critério de entropia, surge uma nova ferramenta de análise, que além de capturar a energia das frequências, também captura o determinismo dos componentes. Portanto, podem escolher os filtros do banco FIR para eliminar os componentes com comportamento estocástico.

Embora os coeficientes dos filtros sejam ótimos, parte da energia dos ruídos e principalmente aqueles com baixa frequência, também afetam os componentes com maior energia dos sinais. Por este motivo, propomos e desenvolvemos um novo método de decomposição robusto baseado no cálculo da matriz de autocorrelações robusta. O método proposto chamou-se RSSA, e foram mostradas suas vantagens nos casos com contaminações Gaussianas e não Gaussianas, incluindo ruídos impulsivos. No caso do RSSA, avaliamos somente o critério da energia retida pelos autovalores, para gerar as reconstruções pelo fato de ser um método naturalmente robusto aos ruídos. Embora, futuramente o critério de entropia possa ser utilizado para identificar os componentes robustos mais determinísticos dos sinais.

Finalmente, avaliamos e comparamos os filtros SMS, DWT e SSA aplicados ao problema de reconhecimento das espécies. Na primeira avaliação mostramos como o filtro modifica os resultados da segmentação bioacústica utilizando a metodologia de avaliação que foi proposta no capítulo 4. Posteriormente, avaliamos o impacto que os filtros causam na taxa de reconhecimento das espécies usando a proposta de classificação apresentada no capítulo 5. Com estes resultados, concluímos que ao se utilizar filtros (pouco ou muito agressivos), algumas sílabas foram perdidas durante a segmentação, fato que causou uma diminuição na taxa de classificação. Por outro lado, notamos que os LLDs mudaram antes e depois da filtragem, o que também afetou a classificação.

Como a maioria dos indivíduos da nossa base de dados foram gravados em locais próximos, em dias consecutivos e na mesma faixa horária, inferimos que o classificador também aprendeu condições acústicas do ambiente. Esta observação é coerente com a diminuição na taxa de classificação, quando é aplicado o filtro para eliminar os ruídos ambientais.

Além dos resultados após aplicar os filtros, devem-se considerar outros aspectos de implementação prática em uma RSSF. Utilizar filtros baseados em transformações dos sinais aumenta a complexidade computacional dos métodos e o custo de memória destes. Se realizamos uma escala crescente de custo computacional teríamos: (1) a transforma Wavelet, (2) a transformada de Fourier e (3) SSA. No entanto, os sinais filtrados com SSA apresentaram melhores resultados. Uma possibilidade para contornar o problema de consumo de memória do SSA é criar o banco de filtros FIR *offline*, e embarcar somente aqueles coeficientes FIR das espécies que se deseje monitorar. Desta forma, evita-se realizar o SVD no próprio nó sensor e obtém-se a vantagem de ter um método ótimo para cada uma das espécies monitoradas.

7

Conclusões

Nesta tese apresentamos uma abordagem para: análise, síntese e classificação de sinais bioacústicos e abordamos os desafios específicos relacionados com a segmentação bioacústica não supervisionada, filtragem dos ruídos ambientais e aprimoramentos dos sinais, extração e avaliação de descritores acústicos em diferentes tarefas, reconhecimento de espécies de anuros e classificação colaborativa utilizando sensores acústicos distribuídos. Cada um dos desafios abordados pertence às diferentes etapas de nosso ACR para monitoramento ambiental bioacústico pervasivo e abrangente, o qual poderá ser empregado para identificar variações nas populações animais e estimar indicadores ambientais.

Durante o processo de revisão dos trabalhos relacionados identificamos diferentes aspectos dos sistemas de reconhecimento de espécies que deveriam ser aprimorados. Identificamos também, que a maioria das técnicas existentes abordam principalmente o problema de classificação das espécies e despreza as dificuldades inerentes à aplicação *in situ*. Por este motivo, conduzimos nossos experimentos a fim de obter um método de classificação eficiente e autônomo considerando as implicações práticas.

Nossa abordagem de monitoramento bioacústico foi além da simples classificação das espécies. No capítulo 6 apresentamos um estudo sobre as caraterísticas físicas dos sinais bioacústicos coletados nas floresta. Mostramos como identificar componentes dos sinais com estrutura interna determinística e analisamos tais componentes no plano de Entropia-Complexidade. Além da análise, mostramos como obter um banco de filtros FIR otimizado para as frequências do coaxar de cada espécie.

Resumindo, no decorrer deste trabalho contribuímos para melhorar problemas fundamentais de segmentação, filtragem, classificação centralizada, classificação multi-rótulo e classificação distribuída, visando: (1) diminuir a quantidade de dados transmitidos e processados pelos sensores da rede; (2) diminuir o impacto negativo dos ruídos

ambientais na taxa de classificação; e (3) aproveitando as informações correlacionadas dos sensores vizinhos para aperfeiçoar o resultado final.

Além disso, avaliamos a *Permutation Entropy* e suas variantes no contexto acústico como descritor de baixo nível. Em relação à filtragem dos ruídos adotamos a técnica *Singular Spectrum Analisys* e propomos uma metodologia baseada em diferentes quantificadores de entropia para escolher as componentes da reconstrução. No que diz respeito à utilização da rede de sensores, avaliamos diferentes estratégias de decomposição de problemas multi-classe em problemas binários e avaliamos quatro métodos de votação e rejeição criando um *Ensemble Learning* para aumentar a acurácia final do monitoramento.

Portanto, concluímos que é necessário combinar técnicas de processamento e análise de sinais com conceitos da teoria da informação e métodos de aprendizagem de máquina para criar um sistema para reconhecimento bioacústico de anuros.

7.1 Contribuições e considerações específicas

As contribuições deixadas neste trabalho podem ser divididas de acordo com a segmentação, a filtragem e a classificação. Analisando os resultados obtidos nos capítulos anteriores concluímos que nossas contribuições ajudam a melhorar a eficácia e eficiência dos método de monitoramento ambiental bioacústico, principalmente aqueles que devem ser embarcados em sensores de monitoramento ambiental, e por tal motivo não podem ser alterados ou atualizados frequentemente. Note-se que utilizamos anuros como objeto principal de estudo devido as características biológicas destes animais, mas isso não impede que o método seja aplicado ou adaptado ao monitoramento de outras espécies animais. Além disso, as estratégias desenvolvidas auxiliam na diminuição do consumo de bateria dos sensores, permitindo estender a vida útil da rede.

7.1.1 Considerações sobre os LLDs

Nesta tese adotamos principalmente medidas de teoria da informação com descritores acústicos de baixo nível. Esta escolha foi baseada no principio de que a entropia consegue caraterizar melhor as dinâmicas das séries temporais. Comparações com os descritores tradicionais, tais como E e ZCR, mostraram que os LLDs baseados em entropia são úteis para tarefas gerais, como a segmentação não supervisionada, ou para escolher as componestes mais determinísticas da decomposição SSA.

Um diferencial deste trabalho foi explorar a aplicabilidade da metodologia de transformação simbólica *Permutation Entropy* e suas variantes (WPE e PME) no con-

texto bioacústico. Concluímos que estas metodologias conseguem capturar correlações fortes e fracas dos sinais, que correspondem aos padrões principais das vocalizações e também aos ruídos de fundo. Isto nos ajudou a interpretar os ruídos de fundo com mais detalhes, e concluir que os cenários acústicos da floresta possuem padrões determinísticos que diferem sustancialmente dos ruídos brancos.

No que diz respeito aos descritores acústicos utilizados para a classificação, nós adotamos os coeficientes Mel. Em trabalhos anteriores realizados por nós (Colonna et al., 2012), foi mostrado que os MFCCss são robustos aos diferentes ruídos ambientais e produzem as melhores taxas de reconhecimento das espécies. Aqui, mostramos que estes também são úteis nas abordagens de classificação hierárquicas multi-rótulos, para identificar diferentes granularidades do espaço de caraterísticas dos classificadores. Além disso, durante as avaliações da filtragem prévia à classificação, descobrimos que estes coeficientes capturam também informações relevantes do fundo acústico das gravações. Esses detalhes ajudaram a reconhecer indivíduos que foram gravados em condições acústicas similares, *i.e.*, como o mesmo ruído de fundo.

7.1.2 Considerações sobre a segmentação bioacústica

A segmentação foi modelada como um problema de classificação binária não supervisionado, *i.e.*, detectar e separar fragmentos dos sinais bioacústicos com ou sem sinal útil para o reconhecimento. Neste caso realizamos uma análise comparativa utilizando diferentes descritores acústicos baseados em teoria da informação para detectar as mudanças que caracterizam o começo e o final das sílabas. A vantagem principal de se utilizar entropia é que não há necessidade de conhecer todos os padrões das vocalizações *a priori*, basta simplesmente saber se o sinal subjacente possui características determinísticas ou aleatórias.

Para simular condições ambientais adversas, em nossos experimentos de segmentação e filtragem, contaminamos as gravações de nossa base com diferentes tipos e níveis de ruídos aleatórios (branco, azul, vermelho, rosa, violeta e impulsivos). Desta forma, mostramos que, inclusive nas condições mais desfavoráveis, o método de segmentação proposto é robusto e eficaz. Além disso, propomos um algoritmo de segmentação que permite encontrar o limiar ótimo de separação entre as classes "sinal" e "ruído". Das comparações realizadas, concluímos também que os descritores baseados em entropia melhoram significativamente seu desempenho quando adicionamos ruídos artificias descorrelacionados aos sinais. Isto sucede sempre que a variância da variável aleatória adicionada seja inferior à variância do sinal original.

Mostramos no capítulo 4 quanto a segmentação impacta no resultado final da

classificação. Para quantificar o impacto desenvolvemos um sistema multinível e um conjunto de equações que consideram a confiabilidade total do sistema de reconhecimento como função dos *tp*, *fn*, *tn* e *fn* finais. Desta forma, possibilitamos avaliar o impacto da segmentação na taxa de reconhecimento das espécies e avaliar o desempenho final do ACR como um sistema de classificação multinível.

Adicionalmente, desenvolvemos uma técnica de segmentação incremental, capaz de se adaptar as mudanças graduais dos sinais, com custo de processamento linear ($\mathcal{O}(n)$) e custo de memória constante ($\mathcal{O}(1)$). Para isto, adaptamos os descritores temporais E e ZCR para uma abordagem com pesos exponenciais, gerando uma relação custo-benefício entre o valores históricos da série temporal e seu valor atual, útil para decidir quando os sinais devem ser segmentados. O melhor benefício desta técnica é o custo reduzido de memória para ser aplicado num nó sensor com hardware limitado.

No que diz respeito as limitações dos métodos de segmentação, destacamos como principal problema do método incremental o ajuste manual dos limiares aplicados aos valores dos descritores E e ZCR. A técnica automática desenvolvida para separar os *frames* dos sinais é um processo iterativo, que precisa dos valores dos descritores correspondentes aos *frames* anteriores. Por este motivo, a transformação da técnica em incremental foi aplicada com limiares fixos, consequentemente, o ajuste manual dos limiares causou uma perda na eficácia do método. No entanto, este problema não existe nas situações em que os áudios podem ser armazenados e processados posteriormente. Além disso, a abordagem incremental possui uma demora para detectar o começo e o final das sílabas, relacionada ao tempo de adaptação às mudanças do método. Este problema pode ser solucionado incluindo uma memória auxiliar que armazene os valores passados mais próximos, porém, isso requer elevar o custo do *hardware*. O filtro da moda utilizado para evitar as micro-segmentações das sílabas, também recebeu um ajuste manual. No entanto, se for possível conhecer *a priori* a duração da sílabas segmentadas, este filtro poderia ser melhorado.

7.1.3 Considerações sobre os métodos de classificação

No capítulo 5, avaliamos o desempenho dos métodos: kNN, SVM, DT e QDA. Estes métodos, foram utilizados nas configurações planas com decomposição 1AA e 1A1. Dentre eles, destacamos que os melhores resultados foram obtidos com kNN e SVM com kernel polinomial. Destacamos que, neste caso, o modelo de classificação vetorial gerado pelo SVM é a melhor opção para embarcar em um nó sensor. Além disso, dentre as abordagens de decomposição avaliadas, a 1AA obteve os melhores desempenhos.

A partir da estrutura taxonômica das espécies, propomos um novo método de

classificação bioacústica. Neste caso, desenvolvemos uma estrutura de classificação hierárquica multi-rótulo, capaz de reconhecer a família, o gênero e a espécie a qual pertence cada amostra. As principais motivações por traz deste método são: aumentar a taxa de reconhecimento decompondo o problema em instâncias mais abrangentes do mesmo, para reduzir a complexidade da função de classificação quando o número de espécies que se deseja reconhecer aumenta consideravelmente. Uma das principais vantagens de nosso método hierárquico, é que após o reconhecimento das classes superiores (as famílias), certos caminhos da árvore não precisam ser avaliados (figura 5.5, página 142), simplificando assim a tomada de decisão.

Para a construção do método hierárquico utilizamos os classificadores mencionados anteriormente em duas configurações: um classificador por nível (LCPL) e um classificador por nó da árvore hierárquica (LCPN), figura 5.6 (página 144). O método LCPL, permite estudar o problema de decisão como uma sobreposição das diferentes partições do espaço dos descritores acústicos. O método LCPN, permite simplificar as funções de classificação conforme as decisões do níveis superiores acontecem. Dentre estas duas abordagens, a segunda permite relacionar melhor a estrutura taxonômica das espécies com os valores de seus descritores acústicos. Isto possibilita encontrar similaridades sonoras entre as espécies e planejar estratégias de monitoramento diferentes dependendo do conjunto de espécies que habitam uma região.

Os métodos LCPL e LCPN são similares ao sistema multinível avaliado no capítulo de segmentação. Consequentemente, as abordagens de classificação hierárquicas sofrem as mesmas desvantagens dos sistemas de cascata de classificadores, sendo a propagação dos erros para os níveis seguintes da hierarquia a principal delas. Entretanto, no caso hierárquico não foi necessário avaliar a interação entre os erros e acertos dos classificadores da mesma forma que foi avaliado na seção 4.6.2 (página 120), basta neste caso avaliar o desempenho final do reconhecimento, que pode ser quantificado diretamente usando a matriz de confusão.

Futuramente, algumas estratégias para corrigir as decisões dos níveis superiores da hierarquia baseadas nos resultados dos níveis inferiores, podem ser adotadas para mitigar o efeito de propagação dos erros.

7.1.4 Considerações sobre a filtragem e o aprimoramento dos sinais

No capítulo 6 abordamos o problema de filtragem dos ruídos ambientas. A decisão de incluir uma etapa de filtro dentro do ACR de monitoramento, foi tomada após observar o impacto negativo causado por estes ruídos na qualidade sonora das gravações.

Devido às características regulares de repetição das sílabas ao longo do tempo, escolhemos a técnica de decomposição e reconstrução de sinais SSA. A partir das revisões bibliográficas identificamos os critérios clássicos de escolha das componentes principais utilizadas para a reconstrução, e contribuímos com uma nova proposta aplicando os conceitos de teoria da informação, tais como a entropia H_t, H_f e H_s. Com estes, conseguimos um método não supervisionado e não paramétrico.

Novamente experimentamos adicionamos diferentes ruídos aos sinais para provar a efetividade do método. A partir disso, encontramos evidências empíricas que apontam à entropia temporal (H_t) como um critério útil para separar e discriminar as componentes mais ruidosas. Mostramos também, uma análise mais detalhadas das componentes dos sinais bioacústicos aplicando a complexidade estatística e construindo os planos HxC. Como esta análise, foi possível mostrar quais componentes são mais afetadas pelo aumento da variância dos ruídos e quais componentes possuem estruturas internas de correlações não lineares mais determinísticas.

A partir do método SSA, mostramos como obter um banco de filtros FIR ótimos para cada espécie. Os filtros FIR possuem um esquema computacional de tempo linear e memória reduzida $\mathcal{O}(2L)$, onde L é a quantidade de componentes principais do SSA. Este custo reduzido permite a implementação em nó sensores de baixo custo.

Apesar dos critérios de entropia serem úteis para separar as componentes ruidosas, existe uma fração dos ruídos que ainda permanece nas componentes selecionadas, e portanto, aparecem na reconstrução do sinal. Para amenizar este problema, desenvolvemos um novo método chamado SSA Robusto (RSSA). Este novo método baseia-se no cálculo de coeficiente de autocorrelação robusto r_Q. O RSSA, além de ser eficiente filtrando os ruídos ambientais, também mostrou-se ótimo nos casos de contaminação Gaussiana, não Gaussiana e contaminação impulsiva, principalmente quando o grau da contaminação é severo (SNR$\leq$0 dB).

A principal limitação de aplicar SSA como técnica de filtragem é que, os filtros criados são específicos para cada espécie, e portanto não podem ser generalizados para um conjunto de espécies. A solução para este problema é criar um banco de filtros para cada espécie, porém numa situação onde as amostras das espécies não estão disponíveis *a priori*, deve-se optar por um método mais geral, como por exemplo, a DWT. Uma segunda limitação é que o SSA somente decompões eficientemente sinais que foram linearmente combinados. Assim, uma possível extensão futura de nosso filtro é criar a versão não linear do SSA. Outra alternativa possível é criar uma versão incremental, para atualizar os coeficientes dos filtros FIR instantaneamente.

7.1.5 Considerações sobre as RSSF e combinação de classificadores

O objetivo final de nossa abordagem de classificação bioacústica é contribuir com uma solução que possibilite monitorar automaticamente populações de anuros por longos períodos, com mínimo esforço humano na coleta e processamento dos dados. Portanto, é fundamental discutir as possibilidades de implementação utilizando redes de sensores acústicos.

Com as decomposições 1AA 1A1, cada modelo de classificação é treinado para reconhecer uma única espécie. Assim, cada modelo especializado pode ser embarcado em um nó da rede, para posteriormente e distribuí-los numa determinada região. Os sensores individuais irão classificar os eventos acústicos e transmitir o resultado do reconhecimento para o nó *sink*, onde uma operação de fusão de decisões avaliará os resultados para chegar a uma conclusão final. Entretanto, uma limitação das abordagens 1AA 1A1 é ter que utilizar m sensores ou $^{m(m-1)}/_{2}$, respectivamente. Consequentemente, o número de sensores, e os custos de transmissão dos dados, aumenta conforme se adicionam novas espécies ao estudo de monitoramento ambiental. Por este motivo, desenvolvemos uma abordagem de classificação multiclasse colaborativa.

No contexto das RSSF, pode-se pensar em diversas soluções que incluam sensores com mais ou menos capacidade de processamento, com ou sem capacidade de armazenamento e transmissão sem fio, ou em um conjunto disperso ou concentrado de sensores. Não obstante, dentre todas as possibilidades tecnológicas, focamos no método de processamento dos resultados das classificações como uma abordagem de "fusão de decisões com rejeição". Logo, trabalhamos na camada de aplicação da rede, realizando uma analogia entre os conceitos de *ensemble learning* e comitê de sensores.

Desta maneira, mostramos que utilizar um comitê de sensores, com classificação multiclasse local e combinados com um nó líder que realiza a votação e toma a decisão final, melhora consideravelmente o resultado do monitoramento. Principalmente em situações próximas à realidade, nas quais mais de uma espécie pode estar presente ao mesmo tempo.

Entretanto, existem algumas limitações físicas e tecnológicas da abordagem colaborativa, por exemplo, transmitir o resultado das classificações, ao invés dos áudios completos, limita a verificação dos resultados por um especialista humano no lado receptor. Além disso, a distribuição dos sensores deve ser planejada, de forma que maximize a cobertura alcança pelos microfones sem excluir sinais correlacionados correspondentes às vocalizações. A lógica no desenvolvimento da rede, incluindo a formação dinâmica dos *clusters*, a recepção dos dados e o alcance da transmissão, são diferentes consi-

derações que tornam a abordagem distribuída mas complexa. No que diz respeito ao *hardware*, a utilização de um comitê de sensores aumenta os custos do monitoramento e a quantidade de pontos de coleta que devem ser revisados, quando comparado com sensores isolados.

7.2 Direções futuras

Além das possibilidades futuras específicas de cada método trabalhado, mencionadas na seção anterior, identificamos três direções futuras mais abrangentes.

A primeira direção futura aponta os novos métodos de aprendizagem que máquina, baseados em *Deep Learning*, como os mais promissores. Uma vez que estes métodos não requerem LLDs manualmente extraídos, as representações das vocalizações no espaço de caraterísticas se consegue de forma ótima e automática (Xie, 2017). Além de facilitar a tarefa de extração de *features*, estes métodos escalam para grandes volumes de dados (*Big Data*), sendo atrativos para classificar os dados provenientes dos nós das RSSF.

A segunda direção futura aponta as estimativas das populações como próximo passo, para realizar a ligação entre o reconhecimento acústico e os índices de variação das populações animais. Atualmente, existem métodos estatísticos que são utilizados após a aplicação dos *suveys* acústicos manuais para inferir o estado das populações animais (Royle and Dorazio, 2008). Entretanto, precisa-se de uma modelagem que permita obter tais índices usando as observações dos sensores. Isto não é uma tarefa trivial, uma vez que os classificadores automáticos estão sujeitos a diferentes tipos de erros.

Por último, mas não menos importante, apontamos o desenvolvimento de *hardware* específico para a tarefa de classificação bioacústica como possível direção futura. Recentemente, foi desenvolvida uma plataforma de monitoramento específico para uma espécie de ave baseada na tecnologia *Field-Programmable Gate Array* (FPGA) (Hervás et al., 2017). O principal ganho, neste caso, é a otimização do *hardware* para os métodos de classificação, diminuindo o consumo de energia e aumentando a vida útil das baterias. Além de FPGA, existem outras alternativas de plataforma para desenvolvimento, um exemplo é o sensor desenvolvido por Lattanzi et al. (2016) para obter automaticamente um índice de riqueza acústica. Estes trabalhos são recentes, e indicam que as tecnologias emergentes para RSSF estão sendo desenvolvidas e melhoradas dia-a-dia.

As direções futuras apontadas indicam que atualmente está-se abrindo uma opor-

tunidade única nas áreas da ciência relacionadas ao monitoramento bioacústico não intrusivo, para preservar a biodiversidade e a qualidade do habitat em que vivemos.

7.3 Publicações

7.3.1 Publicações principais

- As primeiras comparações dos descritores acústicos, baseados na teoria da informação utilizando *Permutation Entropy*, aplicados ao problema de segmentação foram apresentadas no *XVIII Conference on Nonequilibrium Statistical Mechanics and Nonlinear Physics* (Colonna et al., 2014b).

- As avaliações comparativas completas entre seis descritores acústicos baseadas principalmente na entropia do sinal e mais um algoritmo ótimo de segmentação, foi submetido para revisão no *Expert System With Applications* da *Elsevier*.

- O método de segmentação incremental de baixo custo computacional foi publicado no periódico *Expert Systems with Applications*, da Elsevier (Colonna et al., 2015).

- A filtragem de sinais bioacústicos com SSA juntamente com os critérios da teoria de informação para escolher as bases da reconstrução, serão submetidos ao *Digital Signal Processing* da *Elsevier*.

- A modificação *Robust SSA* (RSSA) foi submetida ao periódico *Digital Signal Processing* da *Elsevier*.

- O método de avaliação do erro de classificação utilizando validação cruzada por indivíduos foi apresentado na *Conference of the Spanish Association for Artificial Intelligence* (CAEPIA 2016), o artigo foi publicado como um capítulo de livro no *Lecture Notes on Computer Science* (LNCS), da série *Advances in Artificial Intelligence*, pela Springer (Colonna et al., 2016a).

- A primeira versão do método de classificação hierárquica foi apresentado na conferência *Discovery Science* (DS 2016) e publicado nos *Proceedings* da LNCS, pela Springer (Colonna et al., 2016b).

- Uma versão estendida de nosso método hierárquico multi-label, incluindo dois tipos diferentes de classificadores hierárquicos e as comparações com um classificador plano, foi submetida para o *Machine Learning Journal*, da *Springer*.

- O método de fusão de decisões (ou ensemble) do comitê de sensores para reforçar o resultado da classificação em cenários confusos foi apresentado no *22nd International Conference on Pattern Recognition* (ICPR 2014) e encontra-se disponível nos *Proceedings* da IEEE Explorer (Colonna et al., 2014a).

7.3.2 Publicações em colaboração

Nosso método de reconhecimento acústico e adaptações deste, foram aplicadas com sucesso em outros domínios.

- Uma adaptação do ACR para reconhecimento acústico de sons de motosserra, aplicado ao problema de detecção de desmatamento ilegal na Amazônia, encontra-se disponível nos *proceedings* da *ACM/IEEE International Conference on Information Processing in Sensor Networks* (15th IPSN 2016) (Colonna et al., 2016c).

- Ainda no contexto de reconhecimento de desmatamento ilegal foram publicados dois artigos produto da Co-orientação de um aluno de graduação. Como título "Sensor Acústico para Detecção de Desmatamento Ilegal na Floresta Amazônica" o artigo foi publicado nos anais do IX Simpósio Brasileiro de Computação Ubíqua e Pervasiva (SBCUP 2017) (Seabra et al., 2017b). Este trabalho foi premiado com menção honrosa.

- O segundo trabalho com o título "Detecção de Desmatamento Ilegal na Floresta Amazônica Baseada em Processamento de Áudio" foi selecionado como finalista no 36º Concurso de Trabalhos de Iniciação Científica (CTIC 2017) (Seabra et al., 2017a).

- Recentemente desenvolvemos uma metodologia para classificação dos sinais bioacústicos de anuros utilizando a teoria de subespaços. O método foi baseado nos princípios de funcionamento do SSA e foi aceito para publicação no *IEEE International Workshop on Machine Learning for Signal Processing* (2017).

- Finalmente, no contexto de reconhecimento de séries temporais com métodos de aprendizagem de máquina, publicamos o artigo "Experimental Evaluation on Machine Learning Techniques for Human Activities Recognition in Digital Education Context" (Leitão et al., 2016).

Referências Bibliográficas

Acevedo, M. A., Corrada-Bravo, C. J., Corrada-Bravo, H., Villanueva-Rivera, L. J., and Aide, T. M. (2009). Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics*, 4(4):206–214.

Acevedo, M. A. and Villanueva-Rivera, L. J. (2006). Using automated digital recording systems as effective tools for the monitoring of birds and amphibians. *Wildlife Society Bulletin*, 34(1):211–214.

Adam, T. B., Salam, T. S., and Gunawan, T. S. (2013). Wavelet cesptral coefficients for isolated speech recognition. *Indonesian Journal of Electrical Engineering*, 11(5):2731–2738.

Ahlén, A. and Sternad, M. (1991). Wiener filter design using polynomial equations. *IEEE Transactions on Signal Processing*, 39(11):2387–2399.

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., and Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1(e103):1–19.

Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., and Cayirci, E. (2002). Wireless sensor networks: a survey. *Computer Networks*, 38(4):636–641.

Allen, M. R. and Smith, L. A. (1997). Optimal filtering in singular spectrum analysis. *Physics letters A*, 234(6):419–428.

Amatriain, X. (2004). *An Object-Oriented Metamodel for Digital Signal Processing with a focus on Audio and Music*. PhD thesis, Departament de Tecnologia, Universitat Pompeu Fabra.

Amigó, J. M., Zambrano, S., and Sanjuán, M. A. F. (2008). Combinatorial detection of determinism in noisy time series. *Europhysics Letters (EPL)*, 83(6):1–6.

Armitage, D. W. and Ober, H. K. (2010). A comparison of supervised learning techniques in the classification of bat echolocation calls. *Ecological Informatics*, 5(6):465–473.

Arntzen, J. W., Abrahams, C., Meilink, W. R. M., Iosif, R., and Zuiderwijk, A. (2017). Amphibian decline, pond loss and reduced population connectivity under agricultural intensification over a 38 year period. *Biodiversity and Conservation*, 26(6):1411–1430.

Aysal, T. C. and Barner, K. E. (2007). Meridian filtering for robust signal processing. *IEEE Transactions on Signal Processing*, 55(8):3949–3962.

Bal, M., Liu, M., Shen, W., and Ghenniwa, H. (2009). Localization in cooperative wireless sensor networks: A review. In *13th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 438–443.

Balageas, D., Fritzen, C.-P., and Güemes, A. (2010). *Structural health monitoring*. John Wiley & Sons.

Ballón, M., Bertrand, A., Lebourges-Dhaussy, A., Gutiérrez, M., Ayón, P., Grados, D., and Gerlotto, F. (2011). Is there enough zooplankton to feed forage fish populations off peru? an acoustic (positive) answer. *Progress in Oceanography*, 91(4):360–381.

Bandt, C. and Pompe, B. (2002). Permutation entropy: a natural complexity measure for time series. *Physical Review Letters*, 88(17):1–5.

Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K.-H., and Frommolt, K.-H. (2010). Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*, 31(12):1524–1534.

Benitez, D., Gaydecki, P. A., Zaidi, A., and Fitzpatrick, A. P. (2001). The use of the hilbert transform in ecg signal analysis. *Computers in biology and medicine*, 31(5):399–406.

Bernarde, P. S. and Macedo, L. C. (2008). Impacto do desmatamento e formação de pastagens sobre a anurofauna de serapilheira em rondônia. *Iheringia. Série Zoologia*, 98(4):454–459.

Bertrand, A. (2011). Applications and trends in wireless acoustic sensor networks: A signal processing perspective. In *18th Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, pages 1–6.

Bhandari, G. M., Kawitkar, R. S., and Borawake, M. P. (2014). *Audio Segmentation for Speech Recognition Using Segment Features*, volume 249 of *Advances in Intelligent Systems and Computing*, chapter ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol II, pages 209–217. Springer.

Boll, S. F. (1979). A spectral subtraction algorithm for suppression of acoustic noise in speech. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 200–203.

Borchani, H., Larrañaga, P., Gama, J., and Bielza, C. (2016). Mining multi-dimensional concept-drifting data streams using bayesian network classifiers. *Intelligent Data Analysis*, 20(2):257–280.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, 36(3/4):317–346.

Bridges, A. S. and Dorcas, M. E. (2000). Temporal variation in anuran calling behavior: implications for surveys and monitoring programs. *Copeia*, 2000(2):587–592.

Buckley, L. B. and Jetz, W. (2008). Linking global turnover of species and environments. *Proceedings of the National Academy of Sciences*, 105(46):17836–17841.

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

Carey, C. and Alexander, M. A. (2003). Climate change and amphibian declines: is there a link? *Diversity and Distributions*, 9(2):111–121.

Carey, C., Heyer, W. R., Wilkinson, J., Alford, R. A., Arntzen, J. W., Halliday, T., Hungerford, L., Lips, K. R., Middleton, E. M., Orchard, S. A., and Rand, A. S. (2001). Amphibian declines and environmental change: Use of remote-sensing data to identify environmental correlates. *Conservation Biology*, 15(4):903–913.

Cettolo, M., Vescovi, M., and Rizzi, R. (2005). Evaluation of bic-based algorithms for audio segmentation. *Computer Speech & Language*, 19(2):147–170.

Chen, J., Benesty, J., Huang, Y., and Doclo, S. (2006a). New insights into the noise reduction wiener filter. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1218–1234.

Chen, J., Kwong, K., Chang, D., Luk, J., and Bajcsy, R. (2006b). Wearable sensors for reliable fall detection. In *27th Annual International Conference of the Engineering in Medicine and Biology Society (EMBS).*, pages 3551–3554.

Chen, K. and Sacchi, M. D. (2013). Robust singular spectrum analysis for erratic noise attenuation. Technical report, University of Alberta, Edmonton, Canada. geoConvection 2013, `https://goo.gl/ZCDCvt`.

Chen, W.-P., Chen, S.-S., Lin, C.-C., Chen, Y.-Z., and Lin, W.-C. (2012). Automatic recognition of frog calls using a multi-stage average spectrum. *Computers and Mathematics with Applications*, 64(5):1270–1281.

Cheng, J., Sun, Y., and Ji, L. (2010). A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. *Pattern Recognition*, 43(11):3846–3852.

Cheng, S.-S. and Wang, H.-M. (2003). A sequential metric-based audio segmentation method via the bayesian information criterion. In *European Conference on Speech Communication and Technology (Interspeech)*, pages 945–948.

Chu, W. and Blumstein, D. T. (2011). Noise robust bird song detection using syllable pattern-based hidden markov models. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 345–348.

Clemins, P. J. (2005). *Automatic classification of animal vocalizations.* PhD thesis, Faculty of the Graduate School, Marquette University ,Milwaukee, Wisconsin.

Clemins, P. J., Johnson, M. T., Leong, K. M., and Savage, A. (2005). Automatic classification and speaker identification of african elephant (loxodonta africana) vocalizations. *The Journal of the Acoustical Society of America*, 117(2):956–963.

Cole, E. M., Bustamante, M. R., Reinoso, D. A., and Funk, W. C. (2014). Spatial and temporal variation in population dynamics of andean frogs: Effects of forest disturbance and evidence for declines. *Global Ecology and Conservation*, 1(0):60–70.

Collins, J. P. and Storfer, A. (2003). Global amphibian declines: sorting the hypotheses. *Diversity and distributions*, 9(2):89–98.

Colonna, J. G. (2011). Uma abordagem para classificação de anuros baseada em vocalizações. Master's thesis, Universidade Federal do Amazonas.

Colonna, J. G., Cristo, M. A. P., and Nakamura, E. F. (2014a). A distribute approach for classifying anuran species based on their calls. In *22nd International Conference on Pattern Recognition*, pages 1242–1247.

Colonna, J. G., Cristo, M. A. P., Nakamura, E. F., and Rosso, O. A. (2014b). Permutation entropy applied to bioacoustic signal segmentation. In *XVIII Conference on Nonequilibrium Statistical Mechanics and Nonlinear Physics*, pages 1–1.

Colonna, J. G., Cristo, M. A. P., Salvatierra, M., and Nakamura, E. F. (2015). An incremental technique for real-time bioacoustic signal segmentation. *Expert Systems with Applications*, 42(21):7367–7374.

Colonna, J. G., Gama, J., and Nakamura, E. F. (2016a). *How to Correctly Evaluate an Automatic Bioacoustics Classification Method*, volume 9868 of *Lecture Notes in Computer Science (LNCS)*, chapter Advances in Artificial Intelligence, pages 37–47. Springer.

Colonna, J. G., Gama, J., and Nakamura, E. F. (2016b). *Recognizing Family, Genus, and Species of Anuran Using a Hierarchical Classification Approach*, volume 9956 of *Lecture Notes in Computer Science (LNCS)*, chapter Discovery Science, pages 198–212. Springer.

Colonna, J. G., Gatto, B. B., Nakamura, E. F., and Santos, E. M. d. (2016c). Poster abstract: A framework for chainsaw detection using One-Class and WSNs. In *15th International Conference on Information Processing in Sensor Networks (IPSN)*, pages 1–2.

Colonna, J. G., Ribas, A. D., Santos, E. M. d., and Nakamura, E. F. (2012). Feature subset selection for automatically classifying anuran calls using sensor networks. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to algorithms*. MIT press.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Cowling, M. and Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, 24(15):2895–2907.

Curado, N., Hartel, T., and Arntzen, J. W. (2011). Amphibian pond loss as a function of landscape change-a case study over three decades in an agricultural area of northern france. *Biological Conservation*, 144(5):1610–1618.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Dayou, J., Han, N. C., Mun, H. C., Ahmad, A. H., Muniandy, S. V., and Dalimin, M. N. (2011). Classification and identification of frog sound based on entropy approach. In *International Conference on Life Science and Technology*, volume 3, pages 184–187.

Delacourt, P. and Wellekens, C. J. (2000). Distbic: A speaker-based segmentation for audio data indexing. *Speech Communication*, 32(1):111–126.

Deng, L., Seltzer, M. L., Yu, D., Acero, A., Mohamed, A., and Hinton, G. E. (2010). Binary coding of speech spectrograms using a deep auto-encoder. In *European Conference on Speech Communication and Technology (Interspeech)*, pages 1692–1695.

Deng, L. and Yu, D. (2014). Deep learning: methods and applications. *Foundations and Trends in Signal Processing*, 7(3-4):197–387.

Depraetere, M., Pavoine, S., Jiguet, F., Gasc, A., Duvail, S., and Sueur, J. (2012). Monitoring animal diversity using acoustic indices: implementation in a temperate woodland. *Ecological Indicators*, 13(1):46–54.

Diaz, J. M., Colonna, J. G., Soares, R. B., Figueiredo, C. M. S., and Nakamura, E. F. (2012). Compressive sensing for efficiently collecting wildlife sounds with wireless sensor networks. In *21st International Conference on Computer Communications and Networks (ICCCN)*, pages 1–7.

Dong, X., Towsey, M., Truskinger, A., Cottman-Fields, M., Zhang, J., and Roe, P. (2015). Similarity-based birdcall retrieval from environmental audio. *Ecological Informatics*, 29(1):66–76.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627.

Durisic, M., Tafa, Z., Dimic, G., and Milutinovic, V. (2012). A survey of military applications of wireless sensor networks. In *Mediterranean Conference on Embedded Computing (MECO)*, pages 196–199.

Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(6):1109–1121.

Ephraim, Y. and Van Trees, H. L. (1995). A signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 3(4):251–266.

Esfahanian, M., Zhuang, H., and Erdol, N. (2013). Using local binary patterns as features for classification of dolphin calls. *The Journal of the Acoustical Society of America*, 134(1):105–111.

Espi, M., Fujimoto, M., Kinoshita, K., and Nakatani, T. (2015). Exploiting spectro-temporal locality in deep learning based acoustic event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(26):2–12.

Eterovick, P. C., Carnaval, A. C. O. d. Q., Borges-Nojosa, D. M., Silvano, D. L., Segalla, M. V., and Sazima, I. (2005). Amphibian declines in Brazil: an overview. *Biotropica*, 37(2):166–179.

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

Fadlallah, B., Chen, B., Keil, A., and Príncipe, J. (2013). Weighted-permutation entropy: A complexity measure for time series incorporating amplitude information. *Physical Review E*, 87(2):1–7.

Fagerlund, S. (2007). Bird species recognition using support vector machines. *EURASIP Journal of Applied Signal Processing*, 2007(1):1–8.

Fagerlund, S. and Laine, U. K. (2014). Classification of audio events using permutation transformation. *Applied Acoustics*, 83(1):57–63.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

Feldman, M. (1994). Non-linear system vibration analysis using Hilbert transform–I. Free vibration analysis method 'Freevib'. *Mechanical Systems and Signal Processing*, 8(2):119–127.

Figueiredo, C. M. S., Nakamura, E. F., and Loureiro, A. A. F. (2009). A hybrid adaptive routing algorithm for event-driven wireless sensor networks. *Sensors*, 9(9):7287–7307.

Finch, T. (2009). Incremental calculation of weighted mean and variance. Technical report, University of Cambridge, Computing Service, England.

Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. In *International Conference on Multimedia and Expo (ICME)*, volume 1, pages 452–455.

Freitas, A. A. and Carvalho, A. C. P. L. F. (2007). A tutorial on hierarchical classification with applications in bioinformatics. In *Research and Trends in Data Mining Technologies and Applications*, pages 175–208.

Frost, D. R. (2016). Amphibian species of the world: an online reference. Electronic Database accessible at `http://goo.gl/3WRZhx`. American Museum of Natural History, New York, USA.

Fukane, A. R. and Sahare, S. L. (2011). Different approaches of spectral subtraction method for enhancing the speech signal in noisy environments. *International Journal of Scientific & Engineering Research*, 2(5):1–6.

Fürnkranz, J. (2001). Round robin rule learning. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 146–153.

Gama, J. and Gaber, M. M. (2007). *Learning from data streams*. Springer.

Ganchev, T. D., Jahn, O., Marques, M. I., Figueiredo, J. M., and Schuchmann, K. (2015). Automated acoustic detection of vanellus chilensis lampronotus. *Expert Systems with Applications*, 42(15-16):6098–6111.

Garcia, N., Marcias-Toro, E., Vargas-Bonilla, J. F., Daza, J. M., and López, J. D. (2014). Segmentation of bio-signals in field recordings using fundamental frequency detection. In *3rd International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 86–92.

Gasc, A., Sueur, J., Pavoine, S., Pellens, R., and Grandcolas, P. (2013). Biodiversity sampling using a global acoustic approach: contrasting sites with microendemics in new caledonia. *PLOS ONE*, 8(5):1–10.

Gerhardt, H. C. (1975). Sound pressure levels and radiation patterns of the vocalizations of some north american frogs and toads. *Journal of Comparative Physiology A*, 102(1):1–12.

Ghaderi, F., Mohseni, H. R., and Sanei, S. (2011). Localizing heart sounds in respiratory signals using singular spectrum analysis. *IEEE Transactions on Biomedical Engineering*, 58(12):3360–3367.

Giannakopoulos, T., Pikrakis, A., and Theodoridis, S. (2008). A novel efficient approach for audio segmentation. In *19th International Conference on Pattern Recognition (ICPR)*, pages 1–4.

Giannoulis, D., Stowell, D., Benetos, E., Rossignol, M., Lagrange, M., and Plumbley, M. D. (2013). A database and challenge for acoustic scene classification and event detection. In *European Signal Processing Conference (EUSIPCO)*, pages 1–5.

Gibbs, J. P., Whiteleather, K. K., and Schueler, F. W. (2005). Changes in frog and toad populations over 30 years in new york state. *Ecological Applications*, 15(4):1148–1157.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix computations*, volume 3. Johns Hopkins University Press.

Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. A. (2001). *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Graps, A. (1995). An introduction to wavelets. *IEEE Computational Science and Engineering*, 2(2):50–61.

Gunasekaran, S. and Revathy, K. (2010). Content-based classification and retrieval of wild animal sounds using feature selection algorithm. In *2nd International Conference on Machine Learning and Computing (ICMLC)*, pages 272–275.

Gur, B. M. and Niezrecki, C. (2007). Autocorrelation based denoising of manatee vocalizations using the undecimated discrete wavelet transform. *The Journal of the Acoustical Society of America*, 122(1):188–199.

Gur, M. B. and Niezrecki, C. (2011). A wavelet packet adaptive filtering algorithm for enhancing manatee vocalizations. *The Journal of the Acoustical Society of America*, 129(4):2059–2067.

Haddad, C. (2005). Guia sonoro dos anfíbios anuros da Mata Atlântica.

Han, C. N., Dayou, J., Ho, C. M., Muniandy, S. V., Ahmad, A. H., and Dalimin, M. N. (2015). Investigation on the possibility of using entropy approach for classification and identification of frog species. *Jurnal Teknologi*, 75(1):225–231.

Han, N. C., Muniandy, S. V., and Dayou, J. (2011). Acoustic classification of australian anurans based on hybrid spectral-entropy approach. *Applied Acoustics*, 72(9):639–645.

Hansen, P. C. and Jensen, S. H. (2007). Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis. *EURASIP Journal on Advances in Signal Processing*, 2007(1):1–24.

Harma, A. (2003). Automatic identification of bird species based on sinusoidal modeling of syllables. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1–4.

Harma, A. and Somervuo, P. (2004). Classification of the harmonic structure in bird vocalization. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 701–704.

Hassani, H. (2007). Singular spectrum analysis: Methodology and comparison. *Journal of Data Science*, 5(2):239–257.

Hassani, H., Mahmoudvand, R., Omer, H. N., and Silva, E. S. (2014). A preliminary investigation into the effect of outlier(s) on singular spectrum analysis. *Fluctuation and Noise Letters*, 13(4):1–23.

Hassani, H. and Thomakos, D. (2010). A review on singular spectrum analysis for economic and financial time series. *Statistics and its Interface*, 3(3):377–397.

Heinicke, S., Kalan, A. K., Wagner, O. J. J., Mundry, R., Lukashevich, H., and Kuhl, H. S. (2015). Assessing the performance of a semi-automated acoustic monitoring system for primates. *Methods in Ecology and Evolution*, 6(7):753–763.

Hemakumar, G. and Punitha, P. (2014). Automatic segmentation of kannada speech signal into syllables and sub-words: Noised and noiseless signals. *International Journal of Scientific & Engineering Research*, 5(1):1707–1711.

Hermus, K., Wambacq, P., and Van hamme, H. (2007). A review of signal subspace speech enhancement and its application to noise robust speech recognition. *EURASIP Journal on Applied Signal Processing*, 2007(1):195–195.

Hervás, M., Alsina-Pagés, R. M., Alías, F., and Salvador, M. (2017). An FPGA-Based WASN for remote real-time monitoring of endangered species: A case study on the birdsong recognition of botaurus stellaris. *Sensors*, 17(6):1331–1356.

Howard, L. and López, G. E. (2009). Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. In *European Conference on Speech Communication and Technology (Interspeech)*, pages 2323–2326.

Hu, W., Bulusu, N., Chou, C. T., Jha, S., Taylor, A., and Tran, V. N. (2009). Design and evaluation of a hybrid sensor network for cane toad monitoring. *ACM Transactions on Sensor Networks (TOSN)*, 5(1):4–28.

Huang, C. J., Yang, Y. J., Yang, D. X., and Chen, Y. J. (2009). Frog classification using machine learning techniques. *Expert Systems with Applications*, 36(2):3737–3743.

Ingelrest, F., Barrenetxea, G., Schaefer, G., Vetterli, M., Couach, O., and Parlange, M. (2010). Sensorscope: Application-specific sensor network for environmental monitoring. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):17–32.

IUCN (2016). Geographic patterns. `http://goo.gl/nq2qt7`. The IUCN Red List of Threatened Species.

Jaafar, H. and Ramli, D. A. (2013). Automatic syllables segmentation for frog identification system. In *9th International Colloquium on Signal Processing and its Applications (CSPA)*, pages 224–228.

Jaafar, H., Ramli, D. A., and Rosdi, B. A. (2014). Comparative study on different classifiers for frog identification system based on bioacoustic signal analysis. In *International Conference on Communications, Signal Processing and Computers*, pages 172–176.

Jaber, G. (2013). *An approach for online learning in the presence of concept change*. PhD thesis, Université Paris.

Johnson, M. T., Yuan, X., and Ren, Y. (2007). Speech signal enhancement through adaptive wavelet thresholding. *Speech Communication*, 49(2):123–133.

Joly, A., Goëau, H., Bonnet, P., Spampinato, C., Glotin, H., Rauber, A., Fisher, R. B., and Müller, H. (2014). Are species identification tools biodiversity-friendly? In *Proceedings of the 3rd ACM International Workshop on Multimedia Analysis for Ecological Data*, pages 31–36. ACM.

Joy, J., Peter, S., and John, N. (2013). Denoising using soft thresholding. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2(3):1027–1031.

Kalan, A. K., Mundry, R., Wagner, O. J. J., Heinicke, S., Boesch, C., and Kuhl, H. S. (2015). Towards the automated detection and occupancy estimation of primates using passive acoustic monitoring. *Ecological Indicators*, 54:217–226.

Kalantari, M., Yarmohammadi, M., and Hassani, H. (2016). Singular spectrum analysis based on l1-norm. *Fluctuation and Noise Letters*, 15(1):1–26.

Kasdin, N. J. (1995). Discrete simulation of colored noise and stochastic processes and $\frac{1}{|f|^{\alpha}}$ power law noise generation. *Proceedings of the IEEE*, 83(5):802–827.

Kaur, M. and Kaur, A. (2013). A review: Different methods of segmenting a continuous speech signal into basic units. *International Journal Of Engineering And Computer Science*, 2(11):3184–3186.

Kharin, Y. S. and Voloshko, V. A. (2011). Robust estimation of ar coefficients under simultaneously influencing outliers and missing values. *Journal of Statistical Planning and Inference*, 141(9):3276–3288.

King, J. R. and Jackson, D. A. (1999). Variable selection in large environmental data sets using principal components analysis. *Environmetrics*, 10(1):67–77.

King, V. (1969). A study of the mechanism of water transfer across frog skin by a comparison of the permeability of the skin to deuterated and tritiated water. *The Journal of physiology*, 200(2):529–538.

Kopsinis, Y. and McLaughlin, S. (2009). Development of emd-based denoising methods inspired by wavelet thresholding. *IEEE Transactions on Signal Processing*, 57(4):1351–1362.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31(3):249–268.

Krizhevsky, A., Ilya, S., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates Inc.

Labate, D., Foresta, F. L., Morabito, G., Palamara, I., and Morabito, F. C. (2013). Entropic measures of eeg complexity in alzheimer's disease through a multivariate multiscale approach. *IEEE Sensors Journal*, 13(9):3284–3292.

Laiolo, P. (2010). The emerging significance of bioacoustics in animal species conservation. *Biological Conservation*, 143(7):1635–1645.

Lakshminarayanan, B., Raich, R., and Fern, X. (2009). A syllable-level probabilistic framework for bird species identification. In *International Conference on Machine Learning and Applications (ICMLA)*, pages 53–59.

Lambert, K. T. A. and McDonald, P. G. (2014). A low-cost, yet simple and highly repeatable system for acoustically surveying cryptic species. *Austral Ecology*, 39(7):779–785.

Lasseck, M. (2014). Large-scale identification of birds in audio recordings. In *Working notes of CLEF 2014 conference*, pages 1–11.

Lattanzi, E., Freschi, V., Dromedari, M., and Bogliolo, A. (2016). An acoustic complexity index sensor for underwater applications. *IEEE Sensors Journal*, 16(11):4043–4050.

Lau, D. L., Arce, G. R., and Gallagher, N. C. (1998). Green-noise digital halftoning. *Proceedings of the IEEE*, 86(12):2424–2444.

Lee, C.-H., Chou, C.-H., Han, C.-C., and Huang, R.-Z. (2006). Automatic recognition of animal vocalizations using averaged mfcc and linear discriminant analysis. *Pattern Recognition Letters*, 27(2):93–101.

Lee, H., Pham, P., Largman, Y., and Ng, A. Y. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems*, pages 1096–1104.

Leitão, G., Colonna, J. G., Ribeiro, E., Barreto, R., Araujo, T., Martins, A., Koster, A., and Koch, F. (2016). *Experimental Evaluation on Machine Learning Techniques for Human Activities Recognition in Digital Education Context*, volume 606 of *Communications in Computer and Information Science*, chapter International Workshop on Social Computing in Digital Education, pages 124–139. Springer.

Lellouch, L., Pavoine, S., Jiguet, F., Glotin, H., and Sueur, J. (2014). Monitoring temporal change of bird communities with dissimilarity acoustic indices. *Methods in Ecology and Evolution*, 5(6):495–505.

Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *8th ACM SIGMOD*

Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 2–11.

Loizou, P. C. (2013). *Speech Enhancement: Theory and Practice, Second Edition*. Taylor & Francis.

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

Lowen, S. B. and Teich, M. C. (1990). Power-law shot noise. *IEEE Transactions on Information Theory*, 36(6):1302–1318.

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., and Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, 84(8):2200–2207.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, A. J., and Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255.

Magaña-Espinoza, P., Aquino-Santos, R., Cárdenas-Benítez, N., Aguilar-Velasco, J., Buenrostro-Segura, C., Edwards-Block, A., and Medina-Cass, A. (2014). Wisph: A wireless sensor network-based home care monitoring system. *Sensors*, 14(4):7096.

Mainwaring, A., Culler, D., Polastre, J., Szewczyk, R., and Anderson, J. (2002). Wireless sensor networks for habitat monitoring. In *Proceedings of the 1st ACM International Workshop on Wireless sensor networks and Applications (WSNA)*, pages 88–97.

Mammone, R. J., Xiaoyu, Z., and Ramachandran, R. P. (1996). Robust speaker recognition: a feature-based approach. *IEEE Signal Processing Magazine*, 13(5):58–71.

Mansour, A., Leblond, I., Hamad, D., and Artigas, L. F. (2013). Wireless sensor networks for ecosystem monitoring & port surveillance. In *2nd Symposium on Wireless Sensor and Cellular Networks (WSCN)*, pages 2–11.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: theory and practice*. John Wiley & Sons.

Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D., and Tyack, P. L. (2013). Estimating animal population density using passive acoustics. *Biological Reviews*, 88(2):287–309.

Márquez, R., Riva, I., Matheu, B., and Matheu, E. (2002). Sounds of frogs and toads of Bolivia.

Marty, C.and Gaucher, P. (1999). Sound guide to the tailless amphibians of French Guiana.

McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons.

Modic, R., Lindberg, B., and Petek, B. (2003). Comparative Wavelet and MFCC speech recognition experiments on the Slovenian and English Speechdat2. In *Proceedings of ISCA tutorial and research workshop on non-linear speech processing*, pages 1–3.

Morettin, P. A. (1999). *Ondas e Ondaletas. Da Análise de Fourier à Análise de Ondaletas*. EdUSP, São Paulo.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nakamura, E. F. (2007). *Fusão de Dados em Redes de Sensores sem Fio*. PhD thesis, Universidade Federal de Minas Gerais, UFMG, Brasil.

Nakamura, E. F., Figueiredo, C. M., Nakamura, F. G., and Loureiro, A. A. F. (2007a). Diffuse: A topology building engine for wireless sensor networks. *Signal Processing*, 87(12):2991–3009.

Nakamura, E. F., Loureiro, A. A. F., Boukerche, A., and Zomaya, A. Y. (2014). Localized algorithms for information fusion in resource constrained networks. *Information Fusion*, 15(1):2–4.

Nakamura, E. F., Loureiro, A. A. F., and Frery, A. C. (2007b). Information fusion for wireless sensor networks: Methods, models, and classifications. *ACM Computing Surveys*, 39(3):1–55.

Nakamura, E. F., Ramos, H. S., Villas, L. A., Oliveira, H. A. B. F., Aquino, A. L. L., and Loureiro, A. A. F. (2009). A reactive role assignment for data routing in event-based wireless sensor networks. *Computer Networks*, 53(12):1980–1996.

Neal, L., Briggs, F., Raich, R., and Fern, X. Z. (2011). Time-frequency segmentation of bird song in noisy acoustic environments. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2012–2015.

Obrist, M. K., Pavan, G., Sueur, J., Ride, K., Llusia, D., and Márquez, R. (2010). *Bioacoustic approaches in biodiversity inventories*, volume 8, chapter 5, pages 69–98. Abc Taxa.

Oliveira, A. G., Ventura, T. M., Ganchev, T. D., Figueiredo, J. M., Jahn, O., Marques, M. I., and Schuchmann, K. (2015). Bird acoustic activity detection based on morphological filtering of the spectrogram. *Applied Acoustics*, 98(1):34–42.

Oppenheim, A. V. and Schafer, R. W. (2010). *Discrete-Time Signal Processing*. Pearson.

Parlitz, U., Berg, S., Luther, S., Schirdewan, A., Kurths, J., and Wessel, N. (2012). Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics. *Computers in biology and medicine*, 42(3):319–327.

Plaszczynski, S. (2007). Generating long streams of $\frac{1}{f^\alpha}$ noise. *Fluctuation and Noise Letters*, 07(01):1–13.

Potamitis, I. (2014). Automatic classification of a taxon-rich community recorded in the wild. *PLOS ONE*, 9(5):1–11.

Potamitis, I., Ntalampiras, S., Jahn, O., and Riede, K. (2014). Automatic bird sound detection in long real-field recordings: Applications and tools. *Applied Acoustics*, 80(1):1–9.

Powers, D. M. W. (2007). Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. Technical Report SIE-07-001, School of Informatics and Engineering, Flinders University.

Proakis, J. G. and Manolakis, D. G. (1996). *Digital Signal Processing: Principles, Algorithms, and Applications (3rd Ed.)*. Prentice-Hall, Inc.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers, Inc.

R2015a, M. D. (2015). dsp.colorednoise system object.

Rabiner, L. and Schafer, R. (2007). *Introduction to Digital Speech Processing*. Now Publishers Inc.

Rahman, M. and Bhuiyan, A. (2012). Continuous bangla speech segmentation using shortterm speech features extraction approaches. *International Journal of Advanced Computer Sciences and Applications*, 3(1):11.

Ramírez, J., Segura, J. C., Benítez, C., De La Torre, A., and Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech communication*, 42(3):271–287.

Rein, S. and Reisslein, M. (2011). Low-memory wavelet transforms for wireless sensor networks: A tutorial. *IEEE Communications Surveys & Tutorials*, 13(2):291–307.

Ren, Y., Johnson, M. T., and Tao, J. (2008). Perceptually motivated wavelet packet transform for bioacoustic signal enhancement. *The Journal of the Acoustical Society of America*, 124(1):316–327.

Ribas, A. D., Colonna, J. G., Figueiredo, C. M. S., and Nakamura, E. F. (2012). Similarity clustering for data fusion in wireless sensor networks using k-means. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Rickwood, P. and Taylor, A. (2008). Methods for automatically analyzing humpback song units. *The Journal of the Acoustical Society of America*, 123(3):1763–1772.

Rilling, G., Flandrin, P., and Goncalves, P. (2003). On empirical mode decomposition and its algorithms. In *EURASIP Workshop on Nonlinear Signal and Image Processing*, pages 8–11.

Romero, F., Alonso, F. J., Cubero, J., and Galáin-Marín, G. (2015). An automatic ssa-based de-noising and smoothing technique for surface electromyography signals. *Biomedical Signal Processing and Control*, 18(1):317–324.

Root-Gutteridge, H., Bencsik, M., Chebli, M., Gentle, L. K., Terrell-Nield, C., Bourit, A., and Yarnell, R. W. (2014). Identifying individual wild eastern grey wolves (*Canis lupus lycaon*) using fundamental frequency and amplitude of howls. *Bioacoustics*, 23(1):55–66.

Rosso, O. A., Blanco, S., Yordanova, J., Kolev, V., Figliola, A., Schürmann, M., and Başar, E. (2001). Wavelet entropy: a new tool for analysis of short duration brain electrical signals. *Journal of Neuroscience Methods*, 105(1):65–75.

Rosso, O. A., De Micco, L., Larrondo, H. A., Martín, M. T., and Plastino, A. (2010). Generalized statistical complexity measure. *International Journal of Bifurcation and Chaos*, 20(3):775–785.

Rosso, O. A., Larrondo, H. A., Martin, M. T., Plastino, A., and Fuentes, M. A. (2007). Distinguishing noise from chaos. *Physical Review Letters*, 99(15):1–4.

Royle, J. A. and Dorazio, R. M. (2008). *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Academic Press.

Royle, J. A. and Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87(4):835–841.

Rudnick, D. L. and Davis, R. E. (2003). Red noise and regime shifts. *Deep Sea Research Part I: Oceanographic Research Papers*, 50(6):691–699.

Rybach, D., Gollan, C., Schluter, R., and Ney, H. (2009). Audio segmentation for speech recognition using segment features. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4197–4200.

Sahidullah, M. and Saha, G. (2012). Comparison of speech activity detection techniques for speaker recognition. *Computing Research Repository (CoRR)*, abs/1210.0297:1–7.

Seabra, W. J. G., Colonna, J. G., and Nakamura, E. F. (2017a). Detecção de desmatamento ilegal na floresta Amazônica baseada em processamento de áudio. In 36º *Concurso de Trabalhos de Iniciação Científica (CTIC)*, pages 1–10.

Seabra, W. J. G., Colonna, J. G., and Nakamura, E. F. (2017b). Sensor acústico para detecção de desmatamento ilegal na floresta Amazônica. In *IX Simpósio Brasileiro de Computação Ubíqua e Pervasiva (SBCUP)*, pages 1–10.

Selavo, L., Wood, A., Cao, Q., Sookoor, T., Liu, H., Srinivasan, A., Wu, Y., Kang, W., Stankovic, J., Young, D., and Porter, J. (2007). Luster: wireless sensor network for environmental research. In *Proceedings of the 5th international conference on Embedded networked sensor systems*, pages 103–116. ACM.

Sezgin, M. and Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–168.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Shannon, G., Lewis, J. S., and Gerber, B. D. (2014). Recommended survey designs for occupancy modelling using motion-activated cameras: insights from empirical wildlife data. *PeerJ*, 2:e532.

Shen, J., Hung, J., and Lee, L. (1998). Robust entropy-based endpoint detection for speech recognition in noisy environments. In *5th International Conference on Spoken Language Processing (ICSLP)*, pages 232–235.

Shevlyakov, G. and Oja, H. (2016). *Robust Correlation: Theory and Applications*. Wiley Series in Probability and Statistics. Wiley.

Shevlyakov, G. and Smirnov, P. (2011). Robust estimation of the correlation coefficient: an attempt of survey. *Austrian Journal of Statistics*, 40(1-2):147–156.

Shimamura, T. and Kobayashi, H. (2001). Weighted autocorrelation for pitch extraction of noisy speech. *IEEE Transactions on Speech and Audio Processing*, 9(7):727–730.

Silla, C. N. and Kaestner, C. A. A. (2013). Hierarchical classification of bird species using their audio recorded songs. In *International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1895–1900.

Silla Jr, C. N. and Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(22):31–72.

Silva, C. A. and Ruiz, L. B. (2015). Mannanuro: Classification and identification of anuran amphibians using wireless multimedia sensor network. *IOSR Journal of Computer Engineering*, 17(5):39–45.

Silva, F. R. D. (2010). Evaluation of survey methods for sampling anuran species richness in the neotropics. *South American Journal of Herpetology*, 5(3):212–220.

Sinn, M., Keller, K., and Chen, B. (2013). Segmentation and classification of time series using ordinal pattern distributions. *The European Physical Journal Special Topics*, 222(2):587–598.

Skowronski, M. D. and Harris, J. G. (2006). Acoustic detection and classification of microchiroptera using machine learning: Lessons learned from automatic speech recognition. *The Journal of the Acoustical Society of America*, 119(3):1817–1833.

Slaby, A. (2007). Roc analysis with matlab. In *29th International Conference on Information Technology Interfaces (ITI)*, pages 191–196.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Somervuo, P., Harma, A., and Fagerlund, S. (2006). Parametric representations of bird sounds for automatic species recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2252–2263.

Soriano, M. C., Zunino, L., Rosso, O. A., Fischer, I., and Mirasso, C. R. (2011). Time scales of a chaotic semiconductor laser with optical feedback under the lens of a permutation information analysis. *IEEE Journal of Quantum Electronics*, 47(2):252–261.

Sueur, J., Gasc, A., Grandcolas, P., and Pavoine, S. (2012). *Global estimation of animal diversity using automatic acoustic sensors*, chapter 2, pages 99–117. Sensors for ecology. CNRS.

Sueur, J., Pavoine, S., Hamerlynck, O., and Duvail, S. (2008). Rapid acoustic survey for biodiversity appraisal. *PLOS ONE*, 3(12):1–9.

Swiston, K. A. and Mennill, D. J. (2009). Comparison of manual and automated methods for identifying target sounds in audio recordings of pileated, pale-billed, and putative ivory-billed woodpeckers. *Journal of Field Ornithology*, 80(1):42–50.

Tan, A. W. C., Rao, M. V. C., and Sagar, B. S. D. (2007). A signal subspace approach for speech modelling and classification. *Signal Processing*, 87(3):500–508.

Taylor, A., Watson, G., Grigg, G., and Hamish, M. (1996). Monitoring frog communities: an application of machine learning. In *Proceedings of the eighth annual conference on Innovative applications of artificial intelligence*, pages 1564–1569.

Teixeira, A. R., Lang, E. W., Gruber, P., and Martins da Silva, A. (2005). On the use of clustering and local singular spectrum analysis to remove ocular artifacts from electroencephalograms. In *International Joint Conference on Neural Networks (IJCNN)*, pages 2514–2519.

Theodoridis, S. and Koutroumbas, K. (2008). *Pattern Recognition, Fourth Edition*. Academic Press, 4th edition.

Theodoridis, S., Pikrakis, A., Koutroumbas, K., and Cavouras, D. (2010). *Introduction to Pattern Recognition: A Matlab Approach*. Academic Press.

Thuiller, W. (2004). Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, 10(12):2020–2027.

Tiwari, A., Ballal, P., and Lewis, F. L. (2007). Energy-efficient wireless sensor network design and implementation for condition-based maintenance. *ACM Transactions on Sensor Networks*, 3(1):1–23.

Tkacenko, A., Vaidyanathan, P. P., and Nguyen, T. Q. (2003). On the eigenfilter design method and its applications: A tutorial. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 50(9):497–517.

Tomé, A. M., Teixeira, A. R., Figueiredo, N., Santos, I. M., Georgieva, P., and Lang, E. W. (2010). SSA of biomedical signals: A linear invariant systems approach. *Statistics and its Interface*, 3(1):345–355.

Tomé, A. M., Teixeira, A. R., Teixeira, A., Miguel, G., Georgieva, P., and Lang, E. W. (2011). Linear invariant systems theory for signal enhancement. *Electrónica e Telecomunicações*, 5(3):290–294.

Towsey, M., Wimmer, J., Williamson, I., and Roe, P. (2014a). The use of acoustic indices to determine avian species richness in audio-recordings of the environment. *Ecological Informatics*, 21(1):110–119.

Towsey, M., Zhang, L., Cottman-Fields, M., Wimmer, J., Zhang, J., and Roe, P. (2014b). Visualization of long-duration acoustic recordings of the environment. *Procedia Computer Science*, 29(1):703–712.

Vaca-Castaño, G. and Rodriguez, D. (2010). Using syllabic mel cepstrum features and k-nearest neighbors to identify anurans and birds species. In *Workshop on Signal Processing Systems (SIPS)*, pages 466–471.

Van Der Veen, A.-J., Deprettere, E. F., and Swindlehurst, A. L. (1993). Subspace-based signal analysis using singular value decomposition. *Proceedings of the IEEE*, 81(9):1277–1308.

Vaseghi, S. V. (2000). *Advanced Digital Signal Processing and Noise Reduction*. Wiley.

Vaseghi, S. V. (2008). *Advanced Digital Signal Processing and Noise Reduction*. Wiley.

Vasseur, D. A. and Yodzis, P. (2004). The color of environmental noise. *Ecology*, 85(4):1146–1152.

Veisi, I., Pariz, N., and Karimpour, A. (2007). Fast and robust detection of epilepsy in noisy eeg signals using permutation entropy. In *7th International Symposium on BioInformatics and BioEngineering*, pages 200–203.

Ventura, T. M., Oliveira, A. G., Ganchev, T. D., Figueiredo, J. M., Jahn, O., Marques, M. I., and Schuchmann, K.-L. (2015). Audio parameterization with robust frame selection for improved bird identification. *Expert Systems with Applications*, 42(22):8463–8471.

Verteletskaya, E. and Simak, B. (2011). Noise reduction based on modified spectral subtraction method. *IAENG International Journal of Computer Science*, 38(1):82–88.

Vié, J., Hilton-Taylor, C., and Stuart, S. (2009). *Wildlife in a Changing World: An Analysis of the 2008 IUCN Red List of Threatened Species*. World Conservation Union.

Voss, R. F. and Clarke, J. (1978). "1/f noise" in music: Music from 1/f noise. *The Journal of the Acoustical Society of America*, 63(1):258–263.

Wang, H., Elson, J., Girod, L., Estrin, D., Yao, K., and Vanderberge, L. (2003). Target classification and localization in habitat monitoring. In *International Conference on Speech and Signal Processing*, pages 597–600.

Wax, M. and Kailath, T. (1985). Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(2):387–392.

Weninger, F. and Schuller, B. (2011). Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 337–340.

Wilcox, R. (2012). *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, third edition edition.

Williams, S. (2001). Multiple determinants of australian tropical frog biodiversity. *Biological Conservation*, 98(1):1–10.

Williams, S. E., Bolitho, E. E., and Fox, S. (2003). Climate change in australian tropical rainforests: an impending environmental catastrophe. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1527):1887–1892.

Wimmer, J., Towsey, M., Roe, P., and Williamson, I. (2013). Sampling environmental acoustic recordings to determine bird species richness. *Ecological Applications*, 23(6):1419–1428.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Xie, J. (2017). Multi-label classification of frog species via deep learning. *PeerJ Preprints*, 5:1–7.

Xie, J., Towsey, M., Truskinger, A., Eichinski, P., Zhang, J., and Roe, P. (2015a). Acoustic classification of australian anurans using syllable features. In *10th International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, pages 2–7.

Xie, J., Towsey, M., Yasumiba, K., Zhang, J., and Roe, P. (2015b). Detection of anuran calling activity in long field recordings for bio-acoustic monitoring. In *10th International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, pages 2–7.

Xie, J., Towsey, M., Zhang, J., and Roe, P. (2015c). Image processing and classification procedure for the analysis of australian frog vocalisations. In *Proceedings of the 2Nd International Workshop on Environmental Multimedia Retrieval*, pages 15–20.

Xing, G., Wang, X., Zhang, Y., Lu, C., Pless, R., and Gill, C. (2005). Integrated coverage and connectivity configuration for energy conservation in sensor networks. *ACM Transactions on Sensor Networks*, 1(1):36–72.

Yan, Z., Niezrecki, C., Cattafesta III, L. N., and Beusse, D. O. (2006). Background noise cancellation of manatee vocalizations using an adaptive line enhancer. *The Journal of the Acoustical Society of America*, 120(1):145–152.

Yen, G. and Fu, Q. (2002). Automatic frog call monitoring system: a machine learning approach. In *Proceedings of SPIE*, volume 4739, pages 188–199.

Yuan, C. L. T. and Ramli, D. A. (2013). *Frog Sound Identification System for Frog Species Recognition*, volume 109 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, chapter Context-Aware Systems and Applications, pages 41–50. Springer.

Zhao, F. and Guibas, L. J. (2004). *Wireless sensor networks: an information processing approach*. Morgan Kaufmann.

Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C., and Shamma, S. (2011). Linear versus mel frequency cepstral coefficients for speaker recognition. In *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 559–564.

Zoubir, A. M., Koivunen, V., Chakhchoukh, Y., and Muma, M. (2012). Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts. *IEEE Signal Processing Magazine*, 29(4):61–80.

Zunino, L., Olivares, F., and Rosso, O. A. (2015). Permutation min-entropy: An improved quantifier for unveiling subtle temporal correlations. *Europhysics Letters (EPL)*, 109(1):1–6.

Zunino, L., Soriano, M. C., and Rosso, O. A. (2012). Distinguishing chaotic and stochastic dynamics from time series by using a multiscale symbolic approach. *Physical Review E*, 86(4):1–10.