COPPE/UFRJ

CLASSIFICAÇÃO DE INSTRUMENTOS MUSICAIS EM CONFIGURAÇÕES MONOFÔNICAS E POLIFÔNICAS

Jorge Costa Pires Filho

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

Orientadora: Mariane Rembold Petraglia

Rio de Janeiro

Setembro de 2009

CLASSIFICAÇÃO DE INSTRUMENTOS MUSICAIS EM CONFIGURAÇÕES MONOFÔNICAS E POLIFÔNICAS

Jorge Costa Pires Filho

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

Aprovada por:

Prof. Mariane Rembold Petraglia, Ph.D.

Prof. Luiz Pereira Calôba, D.Sc.

Dr. Sergio Rodrigues Neves, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2009

Pires Filho, Jorge Costa

Classificação de Instrumentos Musicais em Configurações Monofônicas e Polifônicas/Jorge Costa Pires Filho. – Rio de Janeiro: UFRJ/COPPE, 2009.

XIV, 190 p. 29, 7cm.

Orientadora: Mariane Rembold Petraglia

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2009.

Referências Bibliográficas: p. 169 – 174.

1. Classificação de Instrumentos. 2. Separação de Fontes. 3. Processamento Digital de Sinais. I. Petraglia, Mariane Rembold. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Aos meus pais, Jorge e Diva, à minha esposa, Ellen, e ao meu filho,
Rodrigo.

Agradecimentos

Agradeço a todas as pessoas que possibilitaram a elaboração deste trabalho: meus pais, minha esposa e meu filho, pelo suporte pessoal e compreensão pelas horas que lhes foram roubadas; à minha orientadora Mariane Rembold Petraglia pela inspiração e confiança; aos meus amigos Diego Barreto Haddad pelo apoio irrestrito e exemplo e Sérgio Rodrigues Neves pela ajuda e incentivo, sem os quais não teria realizado a dissertação; ao Laboratório de Processamento de Sinais - LPS, que garantiu a aquisição de duas das três bases de dados usadas nessa dissertação, em especial aos Profs. Luiz Wagner Pereira Biscainho e Paulo Antônio Andrade Esquef; ao laboratório de Processamento Analógico e Digital de Sinais - PADS e ao Instituto de Pesquisas da Marinha - IPqM, em particular ao Capitão-de-Fragata Jorge Amaral Alves do Grupo de Guerra Eletrônica, pelo grande apoio que me foi concedido.

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

CLASSIFICAÇÃO DE INSTRUMENTOS MUSICAIS EM CONFIGURAÇÕES MONOFÔNICAS E POLIFÔNICAS

Jorge Costa Pires Filho

Setembro/2009

Orientadora: Mariane Rembold Petraglia

Programa: Engenharia Elétrica

Este trabalho apresenta um conjunto de técnicas para extração e classificação de características de sinais de áudio provenientes de gravações musicais visando reconhecer o tipo de fonte geradora, ou seja, o instrumento musical. Mesclando aspectos teóricos e práticos, discute-se e afere-se o desempenho das técnicas correntes e são propostas contribuições para melhorar a capacidade de discriminação dos sinais. São abordados classificadores como máquinas de vetor suporte, discriminantes lineares, K-vizinhos mais próximos, e algumas técnicas de extração de características como coeficientes de predição linear, frequências de linhas espectrais e coeficientes cepstrais. É mostrado um estudo de classificação hierárquica assim como comparações com outros trabalhos. Por fim, apresenta-se propostas para classificação dos instrumentos musicais de gravações polifônicas e monofônicas com o uso de algortimos de separação de fontes.

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

CLASSIFICATION OF MUSICAL INSTRUMENTS IN MONOPHONIC AND POLYPHONIC CONFIGURATIONS

Jorge Costa Pires Filho

September/2009

Advisor: Mariane Rembold Petraglia

Department: Electrical Engineering

This work presents a set of techniques for extraction of features and classification of audio signals from recorded music, aiming at recognizing the source, i.e., the musical instrument. Mixing theoretical and practical aspects, the performance of current techniques is evaluated, and contributions are proposed for improving the signal discrimination. Within the scope of the dissertation, classification techniques such as Support Vector Machine, Linear Discrimination and K-Nearest-Neighbors as well as techniques for feature extraction such as Linear Prediction coefficients, Line Spectral Frequencies and Cepstral Coefficients are discussed. A preliminary study on a hierarchic classification is shown and compared against other methods presented in the literature. At last, applications of techniques for musical instruments classification with algorithms for sources separation from polyphonic and monophonic signals are proposed.

Sumário

Lista de Abreviaturas

ADSR	*Attack, Decay, Sustain, Release.*
BW	*Band Width.*
CDA	*Canonical Discriminant Analysis.*
CQT	*Constant Q Transform.*
DCT	*Discrete Cosine Transform.*
DEAM	Detector de Envoltória AM.
DFT	*Discrete Fourier Transform.*
DLG	Discriminante Linear Generalizado.
FFT	*Fast Fourier Transform.*
FRBS	Agrupamento - *Flutes, Reeds, Brass, Strings.*
FS	Fluxo Espectral.
GMM	*Gaussian Mixtures Models.*
HMM	*Hidden Markov Models.*
IMF	Início-Meio-Fim.
K-NN	*K-Nearest-Neighbors* .
LDA	*Linear Discriminant Analysis.*
LPC	*Linear Prediction Coefficients.*
LSF	*Lines Spectral Frequencies.*
MFCC	*Mel-Frequencie Cepstral Coefficients.*
MFPC	Agrupamento - Metais, Flautas, Palhetas e Cordas.
MFPPC	Agrupamento - Metais, Flautas, Palhetas, Percussão e Cordas.
MIMO	*Multiple Input Multiple Output.*
MIS	*Music Instruments Samples.*

MISO	*Multiple Input Single Output.*
MUMS	*McGill Master Samples.*
PCA	*Principal Component Analysis.*
QDA	*Quadratic Discriminant Analysis.*
RBF	*Radial Basis Function.*
RMS	*Root Mean Square.*
RWC	*Real World Computing.*
SC	*Spectral Centroid.*
SCF	Separação Cega de Fontes.
SDR	*Signal-to-Distortion Ratio.*
SF	Separação de Fontes.
SIMO	*Single Input Multiple Output.*
SIR	*Signal-to-Interference Ratio.*
SISO	*Single Input Single Output.*
SNR	*Signal-to-Noise Ratio.*
SPC	Agrupamento - Sopro, Percussão e Cordas.
STFT	*Shor Time Fourier Transform.*
SVM	*Support Vector Machine.*
TP	Temporal do Pulso.
ZCR	*Zero-Crossing Rate.*

Parte I

Introdução

Capítulo 1

Apresentação

1.1 Tema e Motivação

Um sistema de classificação de instrumentos musicais a partir de sinais gravados representa uma sub-área de estudo do processamento de sinais de áudio.

Pode-se, de antemão, identificar alguns tópicos necessários ao desenvolvimento do tema proposto, tais como: definição de nota musical, extração de características das notas, taxonomia dos instrumentos musicais e classificadores.

Normalmente um sistema de classificação de instrumentos musicais pode ser estruturado como uma sequência de blocos que correspondem aos diversos problemas (relativamente) independentes supracitados.

Assim, um sistema de reconhecimento de instrumentos musicais envolve necessariamente os seguintes estágios: pré-processamento para filtrar informações desnecessárias e destacar os aspectos que deverão ser contemplados; técnicas de extração de características relevantes desses trechos, visando maximizar a discriminação dos instrumentos distintos; técnicas de agrupamento de amostras similares, visando minimizar o erro de classificação; e de classificação, visando identificar o agrupamento associado ao vetor de características extraídas de um trecho de uma dada amostra.

Cada um dos blocos referidos na Figura 1.1 por si representa uma linha própria de pesquisa. No presente trabalho pretende-se percorrer todas as etapas conceituais envolvidas na classificação, de forma a não se perder de vista a visão geral do sistema.

As aplicações do trabalho incluem, por exemplo: aplicações comerciais que

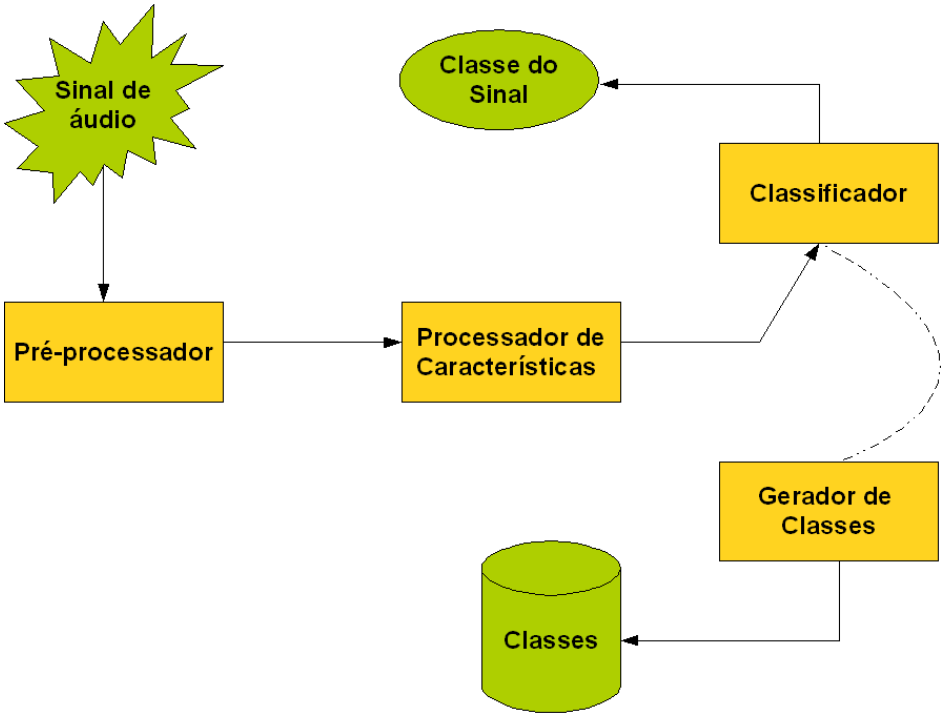

Figura 1.1: Sistema padrão de classificação.

visam catalogar discotecas através de um processo automático (rotulando cada música de acordo com a presença dos instrumentos musicais que a compõem, facilitando assim uma busca seletiva); a transcrição automática de música [1], quando o processo de classificação, depois de determinado o momento de ocorrência de uma nota musical, é capaz de identificar, além do instrumento que a emitiu, o *pitch* e a duração da nota, facilitando o seu registro correto no trecho musical; ou a codificação de áudio em alto nível, ao se usar uma modelagem da fonte sonora, tendo esta sido previamente identificada pelo processo de classificação, para reproduzir total ou parcialmente a nota, evitando, assim, uma codificação de baixo nível, ou seja, uma codificação que exija manipulação direta das amplitudes do sinal [2].

1.2 Objetivo do Trabalho

O presente trabalho tem como principal objetivo obter um método capaz de reconhecer automaticamente instrumentos musicais a partir das notas por eles produzidas. Pode-se assumir que o escopo do presente trabalho é identificar qual é o instrumento musical associado a um sinal e avaliar sua capacidade para classificar o instrumento musical presente numa sequência monofônica ruidosa ou contami-

nada com sinal interferente, ambas oriundas de misturas instantâneas, e classificar os instrumentos musicais presentes em sequências polifônicas oriundas de misturas instantâneas ou convolutivas. Uma das preocupações deste trabalho foi comparar resultados obtidos por diferentes classificadores. Assim, para se traçar uma avaliação de desempenho utilizaram-se como paradigmas os resultados apresentados por diversos autores, sumarizados em [3]. Isso permite avaliar o quão bom é o desempenho que se obtém com cada classificador combinado com uma dada forma de obtenção do vetor de características. O uso de bases de dados obtidas de formas distintas serve para validar os métodos empregados. Portanto, espera-se que o sistema de classificação que obtiver o melhor resultado numa base de dados seja uma das soluções a apresentar os melhores resultados na outra base de dados, devendo apresentar taxas de acertos "consistentes" e, portanto, revelando sua capacidade de generalização.

A opção por abordar a classificação de instrumentos musicais a partir de notas isoladas nesse estudo pode ser justificada por diversos motivos. Primeiramente, ela pode ser adaptada tanto para classificar trechos monofônicos de uma música (polifônica) quanto para outros sinais de áudio oriundos de uma única fonte. No mais, a identificação de instrumentos a partir de notas isoladas, apesar de não ser a mais apropriada para resolver o problema na sua concepção mais geral (sinais de música contendo sobreposição no tempo e na frequência de vários instrumentos musicais), não é restritiva caso se queira identificar sinais que já tenham passado por um processo de separação de fontes. Uma desvantagem inerente a essa abordagem é a dependência de um algoritmo que consiga separar a partir de uma música polifônica o sinal oriundo de cada instrumento, e/ou de um algoritmo extrator de notas ou de pequenos trechos oriundos de um único instrumento musical. Essa dissertação apresenta uma possibilidade do uso desse classificador para o problema de classificação de instrumentos musicais em sequências polifônicas havendo superposição temporal. Tal algoritmo necessita de um separador de fontes e um extrator de notas, os quais devem estar presentes numa fase preliminar. Portanto, esses algoritmos influenciam a taxa de acerto do sistema de classificação posterior. Por óbvio, este pré-processamento poderá funcionar como um agente contaminador, caso ocorra uma separação de fontes mal feita ou uma extração de nota equivocada.

Outra restrição desse trabalho se encontra nas características do ambiente de gravação do banco de dados. Idealmente, escolhe-se uma câmara anecóica para obtenção dos sinais de referência, que apresenta características distintas das normalmente encontradas em espaços reais que envolvem maior ou menor grau de reverberação. De forma que, bancos de dados distintos gravados em ambientes e sensores com discriminações e/ou resoluções diferentes podem resultar em padrões distintos para uma mesma nota produzida pelo mesmo instrumento, podendo afetar, assim, o desempenho do sistema de classificação.

A utilização de notas isoladas não se mostra um problema na possível adaptação dessa abordagem para outros problemas, por exemplo, em Guerra Eletrônica, haja vista que os pulsos emitidos pelos radares na faixa de frequência de 1 GHz a 40 GHz, num cenário padrão contendo em torno de dezenas de emissores pulsados, apresentam uma baixa taxa de sobreposição no tempo. Isto permite a um processo de separação cega identificar o número de emissores presentes neste cenário, o que torna o problema de classificação do radar a partir da envoltória do pulso similar ao de identificação de uma fonte sonora a partir das notas isoladas.

1.3 Abordagens da Literatura

No atual contexto de reconhecimento de instrumentos musicais, ainda não há consenso quanto à melhor abordagem para sinais polifônicos (os quais apresentam simultaneamente sons de diversos instrumentos musicais). Atualmente, a maior parte dos estudos desta área contempla o caso monofônico, seja em notas isoladas, seja em trechos de música solo.

Antes de iniciarmos o presente trabalho, foi feito um levantamento de trabalhos de diversos autores na área de classificação de instrumentos musicais, em que foram examinadas a taxa de acerto média obtida e a abordagem utilizada. Essa avaliação objetivou angariar uma noção tanto das dificuldades enfrentadas quanto dos desempenhos obtidos, e qual o estado da arte. Os resultados desses trabalhos nem sempre podem ser diretamente comparados, por terem sido elaborados a partir de restrições e com objetivos ligeiramente distintos e, eventualmente, de banco de dados diferentes.

A seguir apresentaremos um breve resumo dos principais trabalhos levantados nessa área, em ordem cronológica.

Em 1998, Keith D. Martin et al. apresentaram um artigo [4] onde uma técnica estatística de reconhecimento de padrão é aplicada para classificação de notas de instrumentos musicais. Foram usadas 1023 notas isoladas, compreendendo as escalas completas para um total de 14 instrumentos (violino, viola, violoncelo, contrabaixo, flauta, piccolo, clarinete, oboé, corne inglês, fagote, trompete, trombone, trompa e tuba). As amostras foram obtidas da *McGill Masters Samples*, MUMS [5]. Foram usadas as macro-famílias cordas, madeiras e metais. As notas foram divididas em dois grupos de 70% e 30% para treinamento e teste, respectivamente. Foram construídos classificadores *maximum a posteriori* baseados em modelos gaussianos derivados diretamente da análise múltiplo-discriminante de Fisher. Nesse estudo foi constatada uma melhora no desempenho da classe madeiras quando reagrupada em subclasses mais homogêneas, ou seja, quando as flautas foram separadas do conjunto formado pelas palhetas e metais (*flutes/reeds-brass*). Posteriormente foi feita uma nova separação das palhetas dos metais (*reeds/brass*). As famílias dos instrumentos obtiveram uma taxa média de acerto de 90%, e para o reconhecimento dos instrumentos individuais foi obtida uma taxa média de acerto de 70%. Nesse estudo também se verificou que era possível uma melhora para 93% na taxa média de acerto das famílias de instrumentos musicais ao se usar somente as 10 melhores características que distinguem as subclasses que formam o agrupamento em questão. Inicialmente os instrumentos são divididos em duas classes ou grupos: beliscado e sustentado. Posteriormente os instrumentos pertencentes ao grupo beliscado foram divididos em instrumentos individualizados, todos da família das cordas. Abaixo do grupo sustentado ficaram três classes, ou seja, alguns instrumentos da família das cordas e os instrumentos da família dos metais e das madeiras.

Em 1999, Janet Marques e Pedro J. Moreno apresentaram um relatório técnico [6] do Laboratório de Pesquisa de Cambridge contendo um estudo preliminar para classificar instrumentos musicais com o objetivo de avaliar a capacidade de identificação do instrumento presente numa música solo. O modelo proposto pelo estudo enfocou o uso em um sistema de anotação de arquivos de áudio. Portanto, testaram-se 8 instrumentos musicais, onde um CD foi usado para treinamento e o

outro para teste. Os instrumentos usados foram gaita, clarinete, flauta, cravo, órgão, piano, trombone e violino. Para tanto, usaram-se segmentos do sinal de duração igual a 0,2 s para a extração das características. O sistema proposto nesse trabalho obteve uma taxa de acerto de 70% na determinação do instrumento que originou o trecho de som. Foram experimentados vários tipos de características e diferentes tipos de algoritmos de classificação. Usou-se para extração das características *Linear Prediction Coefficients* (LPC), *Cepstral Coefficients* FFT e *Mel-Frequencies Cepstral Coefficients* (MFCC)[1] . Os algoritmos de classificação usados foram modelo de misturas gaussianas (*Gaussian Mixture Model* (GMM)) e máquina de vetor suporte. Nesse estudo foi também avaliado que o desempenho do GMM era melhor para os coeficientes mel-cepstrais, seguido pelos coeficientes cepstrais e LPC, respectivamente. Também se procurou avaliar a taxa de acerto obtida pelo classificador para as classes que representam os instrumentos acima enumerados.

Eronen e Klapuri publicaram um artigo [7] em que apresentam um sistema para reconhecimento de instrumentos musicais a partir do *pitch* obtido das notas oriundas de diversos instrumentos musicais. Nesse trabalho, foram usadas características espectrais e temporais para analisar as propriedades do som, a partir de 1498 amostras da MUMS que cobriam a tessitura de cada um de 30 instrumentos musicais escolhidos. Nesse conjunto de instrumentos estavam compreendidas as seguintes famílias de instrumentos musicais: cordas, madeiras e percussão. Todas essas famílias tiveram seus instrumentos tocados com diferentes técnicas de execução. Algumas características usadas foram: tempo de subida, i.e., a duração do ataque; tempo de decaimento; tempo entre o fim do ataque e o máximo valor rms (associado à energia); e mais outras dezenas de características. Foram usados segmentos de 10 ms com um fator de superposição de 50%. Os autores avaliaram a classificação hierárquica (abordada na Seção 6.3.2) contra a classificação não-hierárquica, chegando à conclusão de que a classificação não-hierárquica é vantajosa em termos de taxa de acerto[2]. Usaram-se dois agrupamentos hierárquicos, um excluindo a classe metais+palhetas, conforme o agrupamento definido por Martin, e outro contendo-a. Já para a avaliação sem agrupamentos hierárquicos definiu-se o número de classes

[1] Os dois últimos a partir da DFT.

[2] Quando são preservados o vetor de características e o classificador.

em função do número de nós existentes para a respectiva altura (nível) da árvore. A taxa de acerto para a família dos instrumentos foi de 94% e para os instrumentos individualmente foi de 80%.

Em 2001, Agostini et al. apresentaram um trabalho [8] onde um conjunto de características é avaliado para o reconhecimento de instrumentos musicais. Além da avaliação das características tentou-se alcançar uma representação compacta do sinal. Foram usadas somente características espectrais do sinal sonoro, limitadas a um número determinado. A partir de 27 instrumentos musicais foram obtidas 1007, notas e sem emprego de qualquer estrutura hierárquica. As classes definidas foram seis, piano-*staccato*, cordas beliscadas em *rock*, cordas beliscadas em modo não *rock*, cordas sustentadas, madeiras sustentadas e metais sustentados. Foram testados os seguintes classificadores, que aqui aparecem organizados em ordem decrescente da taxa de acerto: *Quadratic Discriminant Analysis* (QDA), *Support Vector Machines* (SVM), *Canonical Discriminant Analysis* (CDA), e *K-Nearest Neighbours* (KNN), com taxas de acertos para instrumentos individuais de 92,81%, 69,71%, 66,74% e 65,74%, respectivamente. A taxa de acerto média obtida pelo QDA para as famílias de instrumentos definidas anteriormente foi de 96,87%. O *kernel* usado para a SVM foi o *Radial Basis Function* (RBF), e a melhor solução para o algoritmo de K-vizinhos mais próximos foi 1-NN com norma 1 para a métrica de distância. As características mais relevantes de um total de 9 tipos de características foram não-harmonicidade, centróide espectral e energia contida na primeira parcial. Além dessas características, são calculadas, entre outras, a taxa de cruzamento por zeros, a energia contida da segunda até a quarta parcial e a largura da banda. Para cada uma das 9 características são calculados o desvio-padrão e a média.

Em 2003, Kitahara et al. apresentaram um artigo [9] no qual afirmam que a relação entre *pitch* e timbres até então não vinha sendo bem explorada para a identificação dos instrumentos musicais. Foi avaliada a dependência das características com o *pitch*, respectivamente a partir das funções que usam a média (f_0-dependentes) e a covariância (f_0-normalizadas) do *pitch*. Os sons dos instrumentos musicais são primeiramente analisados pela distribuição normal multivariável f_0-dependente e então, a identificação do instrumento é feita usando uma função discriminante baseada na regra de decisão de Bayes. São usadas características espectrais, temporais e de

modulação e características de componentes não-harmônicos, resultando num total de 129 características. Cada instrumento musical possui amostras na taxa de 44,1 kHz com 16 bits. Para cada trecho de 10 ms, obtido de uma das amostras a ser analisada, é aplicada uma *Short-time Fourier Transform* (STFT), com uma janela de *Hanning* de 4096 pontos; os picos espectrais são extraídos do espectro de potência do sinal. A partir dos picos, são obtidas a f_0 e a estrutura harmônica. Usou-se posteriormente *Principal Component Analysis* (PCA), para reduzir o espaço de dimensão 129 para dimensão 79. Depois usou-se *Linear Discriminant Analysis* (LDA), conseguindo-se uma redução para um espaço de dimensão 18, no caso de 19 instrumentos. São extraídas 40 características espectrais, 35 características temporais, 32 características de modulação e 22 características de componentes não-harmônicos. São usados 6.247 tons solo de 19 instrumentos, obtendo-se taxas de acerto de cerca de 90% e 80% para família e instrumento, respectivamente.

Em 2004, Krishna e Sreenivas publicaram um artigo [3] que propõe o uso de *Line Spectral Frequencies* (LSF), como características representativas de segmentos obtidos a partir de notas isoladas, vista a sua efetividade para reconhecimento de voz. Posteriormente é feita a classificação, usando os modelos de misturas gaussianas e K-NN. Nesse trabalho são avaliados também o uso das características MFCC, e *Linear Prediction Cepstral Coefficients* (LPCC). Foram utilizados 14 instrumentos e agrupamento hierárquico contendo 4 classes, a saber: palhetas, metais, cordas e flautas. Foram usadas 2 bases de dados distintas: a *UIowa's MIS* [10] e *C Music corporation's RWC* [11]. Foram obtidas taxas de acerto de 95% e 90% para família e instrumento, respectivamente.

Exceto para Marques [6], todos os outros resultados reportados se referem a sistemas classificadores que utilizam notas isoladas.

1.4 Organização do Texto

O presente trabalho foi dividido em cinco partes: introdução, visão geral do sistema, resultados, conclusão e apêndice.

Nesta introdução foram apresentados o tema, a motivação, o objetivo deste trabalho, um resumo de alguns dos principais trabalhos na área, um esboço da

metodologia usada e uma breve descrição de como os instrumentos musicais são comumente agrupados, apresentando também alguns agrupamentos hierárquicos alternativos.

Na visão geral do sistema são apresentadas e detalhadas as arquiteturas empregadas, as etapas da cadeia de processamento do sinal, que incluem os módulos de segmentação, pré-processamento, extração de características e classificação. Ainda nessa parte são descritos os modelos usados pelos codificadores e classificadores.

Nos resultados são feitas as avaliações dos codificadores, classificadores, agrupamentos hierárquicos, estimativa da taxa de acerto do classificador, classificação dos instrumentos em sinais polifônicos, e a avaliação da robustez do classificador frente à inserção de ruído branco e sinal interferente. No entanto, devido ao fato de que uma avaliação exaustiva de todas as combinações entre o pré-processamento, os vetores de características e os classificadores levaria a um custo elevado (por ser de natureza combinatorial), preferiu-se adotar uma estratégia sequênciada, onde somente alguns métodos de codificação e classificação são avaliados.

Por fim, apresentam-se as duas últimas partes, ou seja, a conclusão e os apêndices, estes contendo uma descrição mais detalhada dos bancos de dados, e um estudo de caso que avalia o desempenho do sistema de classificação obtido quando comparado a um outro sistema de classificação (oriundo de um outro trabalho feito na área) a partir das mesmas amostras para identificar as mesmas classes.

Capítulo 2

Instrumentos Musicais e suas Classificações Hierárquicas

O estudo dos instrumentos musicais é conhecido como organologia. O propósito de um instrumento musical é produzir música e, para tanto, os materiais empregados e a forma do objeto, bem como o modo de produzir o som, são elementos importantes para a construção e a classificação do instrumento musical. Existem diversas formas de classificar os instrumentos musicais, segundo diferentes perspectivas, e cada uma delas se presta melhor para uma dada finalidade. A mais comum se baseia na forma pela qual o som é produzido.

No contexto das orquestras sinfônicas, por exemplo, é comum dividir os instrumentos musicais em cordas, sopros (subdivididos em metais e madeiras) e percussão, o que vem a ser uma classificação híbrida, misturando a forma como o som é produzido e o material empregado na confecção do instrumento musical. As madeiras podem ser subdivididas em palhetas (lâminas que com a passagem do ar vibram produzindo o som do instrumento) e sem palhetas (flautas). Dessa forma podemos classificar os instrumentos por essa regra em: cordas, metais, palhetas, flautas e percussão.

Em tese, qualquer objeto pode servir para produzir sons e ser utilizado na música, mas costuma-se utilizar o termo "instrumento musical" para designar objetos que são elaborados especificamente com o propósito de produzir música. A seguir apresentaremos um breve resumo das principais características presentes num instrumento musical, pois esses elementos em alguns casos são determinantes na formação

da taxonomia dos instrumentos musicais e poderão servir de inspiração a qualquer novo procedimento de classificação automática dos instrumentos musicais.

Antes de continuarmos é necessário definirmos a notação que será empregada para descrever as notas musicais usadas nesse trabalho. As notas, independentemente das oitavas, são sete, a saber [12]: 'C' - dó, 'D' - ré, 'E'- mi, 'F' - fá, 'G' - sol, 'A'- lá, 'B' - si, podendo representar até 12 semitons com o uso dos acidentes (bemol para abaixamento e sustenido para elevação), o que pode ser expresso em uma das duas escalas abaixo:

- Escala sustenido - {C, C#, D, D#, E, F, F#, G, G#, A, A#, B}
- Escala bemol - {C, Db, D, Eb, E, F, Gb, G, Ab, A, Bb, Cb}

Por sua vez, as oitavas costumam ser numeradas em algarismos arábicos crescentes com a frequência fundamental da nota. Tipicamente adota-se como referência Lá 4 em 440 Hz e uma distribuição de temperamento igual, o que significa que a frequência fundamental associada a cada nota é dada pela seguinte expressão:

$$\text{frequência} = 440 \times 2^{\text{oitava}-4+\frac{\text{tom}-10}{12}}. \tag{2.1}$$

Esse capítulo será dividido em duas seções, uma contendo uma descrição dos componentes que existem em um instrumento musical e outra contendo os agrupamentos hierárquicos tradicionalmente encontrados nos estudos sobre organologia.

2.1 Componentes

Instrumento musical é qualquer artefato capaz de produzir música. É definido como tal todo artefato que de fato consegue controlar com precisão pelo menos algumas das características do som produzido, tais como: altura (grave, médio e agudo), duração (do som e/ou do silêncio), intensidade e timbre.

Por existir uma gama enorme de instrumentos musicais, a generalização das características que descrevem o funcionamento de um instrumento musical é difícil; porém, alguns elementos constitutivos permanecem presentes, tais como: corpo vibratório, corpo, elementos de estímulos e controle, caixa de ressonância e acessórios. A seguir são descritas as principais características do som, controlados por um instrumento musical, bem como seus principais elementos constitutivos. As descrições

apresentadas desses elementos (características do som e elementos constitutivos), foram obtidas com base no material disponível em [13, 14].

2.1.1 Características do Som

2.1.1.1 Altura

A altura é o elemento que nos permite distinguir um som grave de um som agudo. Assim, quanto maior for a frequência fundamental percebida (*pitch*), mais agudo será o som e maior será a sua altura. Os instrumentos musicais podem ser divididos quanto à altura do som produzido em: de altura determinada e de altura indeterminada.

Um instrumento é dito de altura definida ou determinada quando as notas desse instrumento podem ser afinadas de acordo com escalas definidas. A maioria dos instrumentos musicais de cordas e sopro têm altura definida. Alguns instrumentos de percussão, como o vibrafone, o glockenspiel e o xilofone, também possuem altura definida.

Não é necessário que o instrumento musical durante a execução consiga variar a frequência das notas para que a altura seja considerada como definida, no entanto é necessário que as notas produzidas por ele possam ser afinadas com precisão em relação a outros instrumentos. Assim, há alguns instrumentos musicais que possuem altura definida (tons-tons), apesar das suas notas não poderem ser alteradas durante a execução.

Um instrumento musical é dito de altura indefinida ou indeterminada quando as notas produzidas pelo instrumento não podem ser precisamente afinadas. Isso ocorre porque esses instrumentos (não-harmônicos) possuem em seu timbre uma grande quantidade de parciais não harmônicos, tornando a afinação difícil ou impossível. A maioria dos instrumentos musicais de altura não definida está entre os instrumentos de percussão, como tambores, pratos, gongos e sinos. Existem alguns instrumentos de cordas (berimbau) e sopros com altura indefinida.

Instrumentos musicais de altura indefinida, em geral, podem ser utilizados sem que haja problemas de afinação em músicas de qualquer tonalidade, porque é possível definir o seu registro (posteriormente definido), embora não se possa fazer o mesmo com sua altura. Assim, um tamborim possui um registro mais agudo que

uma caixa e um bumbo mais grave que ambos.

A tessitura é a extensão de notas que um instrumento ou voz pode alcançar, identificada através do nome e da oitava da nota mais grave e da mais aguda associada a essa extensão. Por exemplo, a extensão útil de um saxofone contralto vai de Db2 (ré bemol da segunda oitava) até Ab4 (lá bemol da quarta oitava). A tessitura do piano vai do A0 até o C7.

Os registros são as denominações dadas às três regiões em que a tessitura de um instrumento musical ou voz pode ser dividida: grave, médio e agudo. Assim, cada uma dessas regiões (registro) conserva características próprias, podendo em alguns casos ocorrer diferenças significativas do timbre de região para região. Da mesma forma, pode ser impossível executar todas as notas de uma escala em determinadas regiões para um dado instrumento musical. Do mesmo modo, certos efeitos sonoros de alguns instrumentos musicais podem ter a sua execução limitada em um de seus registros.

Para a perfeita execução do instrumento e composição musical é indispensável o conhecimento da tessitura e do registro instrumental. Caso contrário, um compositor poderia querer escrever uma melodia para um instrumento específico com notas impossíveis de serem executadas por esse instrumento. A tessitura só faz sentido para instrumentos que possibilitam variações controladas de altura, o que não é inteiramente verdade para o registro, o qual pode indicar a região de alturas predominantes mesmo em instrumentos cuja altura é indefinida.

2.1.1.2 Intensidade e Duração

A intensidade é diretamente proporcional à potência, medida relacionada ao quadrado da amplitude do sinal. Normalmente a percepção da intensidade do som leva a distinguí-lo como sendo forte, médio (*mezzo*) ou baixo (*piano*).

A duração representa o tempo transcorrido em que o sinal sonoro existe. O período de tempo em que se dá a ausência do sinal sonoro (a pausa) representa um aspecto igualmente importante.

2.1.1.3 Timbre

O timbre é o elemento sonoro que faz com que um ouvinte seja capaz de distinguir a mesma nota quando produzida por diferentes instrumento musicais. É o que comumente se define como a "cor" do som. Assim, facilmente somos capazes de distinguir a nota produzida por um oboé quando a comparamos com a mesma nota produzida por um violino. Essa percepção se dá pelas relações entre as intensidades dos harmônicos que cada instrumento gera ao produzir uma nota.

2.1.2 Elementos Constitutivos

2.1.2.1 Corpo Vibratório

Também chamado de elemento produtor do som, é a parte do instrumento musical responsável pela altura do som emitido. É ele que vibra ao entrar em contato com o estímulo excitante, produzindo assim uma onda sonora. Pode ser parte do instrumento (cordas, palhetas) ou ser o próprio instrumento. Assim, em princípio, quanto maior a frequência da excitação tanto maior será a altura percebida. Nos instrumentos aerófonos é o próprio ar que entra em vibração ao passar por uma aresta, como em uma flauta.

2.1.2.2 Corpo

É a parte do instrumento que mantém unidas as demais partes do instrumento, como no agogô. Em diversos instrumentos o corpo possui funções também na produção ou controle do som, como nos casos dos corpos do violino ou do violão, que servem como caixas de ressonância e também ajudam no tensionamento das cordas, permitindo que o instrumentista tenha controle sobre a altura das notas.

2.1.2.3 Caixa de Ressonância

É uma câmara cheia de ar, que funciona como um amplificador da intensidade do sinal sonoro. Apresenta formatos variados, permitindo um reforço em determinadas frequências e uma atenuação em outras, determinando em grande parte o timbre do instrumento. A caixa de ressonância pode tanto fazer parte do corpo do instrumento (por exemplo piano, um violão ou um tambor) quanto estar incorporada ao

próprio elemento produtor de som (por exemplo agogô).

2.1.2.4 Elementos de Estímulo e Controle

É responsável por controlar a forma como os sons são produzidos, afinados ou modificados, ou por gerar os estímulos ao elemento produtor de som, fazendo com que o elemento produtor de som entre em vibração. Estes elementos englobam uma variedade de objetos ou mecanismos especificamente destinados para gerar esses estímulos. Entre outros, temos arcos, trastes, plectros, baquetas, martelos, bocais, foles, teclados, válvulas, chaves ou pedais.

2.1.2.5 Acessórios

Alguns instrumentos permitem o uso de acessórios com a finalidade de alterar a forma de execução ou modificar algumas características do som produzido. Podem-se citar: caixas de ressonância alternativas, abafadores (que diminuem a intensidade sonora), surdinas (que abafam e modificam o som produzido), suportes ou alças (que servem para facilitar a execução em posições não convencionais).

2.2 Agrupamentos Hierárquicos

O estudo detalhado dos sistemas de classificação daria material para um livro, portanto se encontrando além da pretensão deste tópico. Contudo, a definição do agrupamento hierárquico é de extrema importância devido ao fato de que, dependendo de como é feito esse agrupamento, o sistema de reconhecimento automático desses agrupamentos pode encontrar maior ou menor facilidade, o que afeta diretamente sua taxa de acerto.

A fim de melhor explanar a natureza e as possibilidades dos sistemas de classificação hierárquica dos instrumentos musicais, se apresentarão alguns deles: um sistema nativo, o sistema grego (Aristides Quintilianus) [13]; o sistema mais usual (Hornsbostel e Sachs) [14, 15] e o sistema elaborado por André Schaeffner [13].

2.2.1 Sistema Grego

Desde a antiguidade o homem elabora sistemas de classificação para os instrumentos musicais que constrói. Já na Grécia antiga, Aristides Quintilianus, que viveu por volta do Século III d.C., foi autor de um tratado musical denominado *Perì musikês*, no qual tentava organizar os instrumentos musicais em famílias de instrumentos [13].

Naquela obra, apresentaram-se dois esquemas de classificação. O primeiro se baseava numa distinção dos instrumentos musicais quanto à forma de produção do som. Assim, os gregos classificavam os instrumentos em 2 classes: cordas e sopros. Os instrumentos de percussão, embora conhecidos, eram desprezados por serem considerados inferiores.

O segundo esquema de Quintilianus baseava-se em aspectos da morfologia humana, e classificava o instrumento musical em masculino, feminino ou misto. Essa teoria tentava responder à questão de como instrumentos musicais sem vida conseguiam causar efeitos emocionais em homens e mulheres.

Portanto, o sistema grego dividia os instrumentos em duas classes quanto à forma do som ser produzido, e em três classes quanto ao gênero do instrumento, gerando um total de seis combinações, ou nove, se incluirmos os instrumentos de percussão:

1. Cordas-Homem
2. Cordas-Misto
3. Cordas-Mulher
4. Sopro-Homem
5. Sopro-Misto
6. Sopro-Mulher
7. Percussão-Homem
8. Percussão-Misto
9. Percussão-Mulher

2.2.2 Sistema Hornbostel e Sachs

Concepções com características universais para os instrumentos musicais só apareceram bem mais tarde, por volta do Século XIX. Um primeiro sistema foi formulado por Victor-Charles Mahillon em 1880, que a partir dos conceitos usados

pelos gregos antigos e teóricos europeus da Idade Média, elaborou uma classificação em forma de árvore, onde os ramos representam instrumentos musicais da sua classe. Para elaborar essa classificação o elemento usado foi o tipo de vibração causado pelo material usado no corpo vibratório, a partir do qual o som é produzido. Essa abordagem gerou as seguintes famílias para os instrumentos musicais: a) autofones; b) membranofones; c) cordofones e; d) aerofones. Esse sistema apresentou como restrição basicamente a incompletude, por se restringir aos instrumentos europeus e o modo de utilização de alguns instrumentos. Essa concepção gera algumas inconsistências como no caso dos instrumentos de teclado e mecânicos [16].

O sistema de Mahillon foi ampliado por Curt Sachs e Erich von Hornbostel, dando origem ao chamado sistema Hornbostel-Sachs de classificação. Além de mudarem o nome da classe autofones para idiofones, eles alteraram a forma de subdivisão de suas classes e introduziram um código decimal baseado no código que Melvil Dewey criou para a classificação de livros em bibliotecas. A seguir apresentamos as famílias de instrumentos musicais conforme a classificação mais usual, baseada no sistema de Hornbostel e Sachs.

1. Cordofones
2. Idiofones
3. Membranofones
4. Aerofones

As subdivisões dentro das quatro famílias de instrumentos musicais podem ser encontradas no artigo [15] de 1914. Além destas, também são previstos no seu sistema numérico algarismos reservados que permitem uma expansão das divisões para cada classe.

2.2.2.1 Idiofones

Os idiofones são aqueles que produzem som ao serem percutidos, provocando a vibração de todo o instrumento musical. Alguns exemplos são:

- Agogô;
- Bateria (pratos), Bloco sonoro;
- Caneca, Carrilhão, Castanhola, Celesta, Chocalho;
- Glockenspiel, Gongos;
- Matraca;
- Pratos;
- Reco-reco;

- Sino, Sinos tubulares;
- Triângulo;
- Vibrafone;
- Xilofone.

2.2.2.2 Membranofones

Os membranofones são aqueles em que o som é produzido quando são percutidos sobre uma membrana esticada que entra em vibração. As membranas podem ser de origem animal, sintéticas ou até mesmo de tecido. Alguns exemplos são:

- Bateria;
- Caixa, Cuíca;
- Djembê;
- Pandeireta (pele), Pandeiro (pele);
- Repinique, Surdo;
- Tambor, Tamborim, Tom-tom;
- Zabumba.

2.2.2.3 Cordofones

Os cordofones, ou instrumentos de cordas são aqueles em que o som é obtido pela vibração das cordas. As cordas podem ser dedilhadas, percutidas ou colocadas em vibração com um arco (friccionadas). Alguns exemplos são:

- Alaúde;
- Baixo, Balalaica, Bandolim, Banjo, Berimbau ou Urucungo;
- Cavaquinho, Cembalo, Cítara, Clavicórdio, Contrabaixo, Cravo;
- Dulcímer;
- Espineta;
- Guitarra, Guitarra inglesa;
- Harpa;
- Kantele - (Derivado da Cítara), Koto;
- Lira;
- Piano;
- Rabeca;
- Saltério, Sanfona, Sangen, Siamise, Sitar;
- Ukulele;
- Violão, Viola, Viola caipira, Viola da gamba, Violino, Violoncelo.

2.2.2.4 Aerofones

Os aerofones, ou instrumentos de sopro, são aqueles que produzem som quando o ar ao ser neles introduzido entra em vibração, excitando os componentes do instrumento musical. Alguns exemplos são:

- Acordeão;
- Bombardino;
- Clarinete, Clarone, Corne-inglês;
- Escaleta;
- Fagote, Flauta (contralto, doce, baixa) Flautim, Flugelhorn;
- Gaita;
- Órgão, Oboé, Ocarina;
- Pífaro;
- Saxofone (contralto, baixo, barítono, tenor);
- Trompa, Trompete (de pistões, de chaves), Trombone, Tuba.

2.2.3 Sistema de André Schaeffner

Em 1932, André Schaeffner publicou na França um trabalho chamado *"D'une nouvelle classification méthodique des instruments de musique"*, republicado como *"Origine des instruments de musique"*, em 1936. Nesse trabalho, ele apresenta um novo esquema de classificação dos instrumentos musicais, que os divide em grupos segundo o tipo de material pela qual o som é produzido [13]. Dessa forma, todos os instrumentos musicais podem ser agrupados em duas grandes classes, uma em que o som advém da vibração dos materiais sólidos, os Gaiafones [16], e outra em que o som advém da vibração dos materiais gasosos, os Aerofones.

Este método apresenta vantagens em relação aos demais, por exemplo, ao evitar as possíveis confusões com instrumentos que se encontram no limite entre instrumentos de cordas e de percussão (como o piano); nesse esquema, ambas as categorias são enquadradas na mesma classe.

Segue um exemplo simplificado da estrutura do sistema de classificação proposto por Schaeffner, que continua subdividindo suas classes além das aqui exemplificadas.

I: Gaiafones
  I.A: Não-Tensionáveis - Sem tensão (exemplo: xilofones);
  I.B: Flexíveis - Linguafones ou lamelofones (exemplo: kalimba)
  I.C: Tensionáveis - Cordofones (exemplo: piano, violino)

II: Aerofones
II.A: Com ar ambiente - (exemplo: acordeão)
II.B: Contendo Cavidades Livres - (exemplo: tambores)
II.C: Contendo uma coluna de ar - (exemplo: flautas)

2.2.4 Outros Agrupamentos Tradicionais

2.2.4.1 Eletrofones

Os eletrofones, ou instrumentos musicais elétricos/eletrônicos, representam uma categoria introduzida por Galpin em 1937 na sua obra, *A Textbook of European Musical Instruments*, para permitir a representação dos instrumentos musicais que produzem sons através de componentes que se utilizam da energia elétrica. Esta categoria é comumente acrescentada ao sistema de Mahillon e Hornbostel e Sachs. Alguns exemplos de instrumentos musicais pertencentes a essa categoria são:

- Ondas Martenot
- Órgão Eletrônico
- Piano Digital
- *Sampler*
- Sintetizador
- Teremim

2.2.4.2 Teclados

Os instrumentos de teclas são agrupados nessa categoria pelo modo de tocar, nesse caso, são consierados como pertencentes a uma categoria a parte, no entanto, também podem ser classificados nas diversas categorias anteriormente relacionadas nessa dissertação (como por exemplo, pelo modo de produção do som). Alguns exemplos são:

- Acordeão (Sopro)
- Celesta (Percussão)
- Clavicórdio (Cordas)
- Cravo (Cordas)
- Piano Digital (Instrumento Eletrônico)
- Órgão (Sopro)
- Piano (Cordas)

2.2.4.3 Em função da Altura

É comum uma associação entre o timbre da voz humana cantada com a escala que os instrumentos musicais alcançam quando comparados a outros instrumentos. Portanto, podemos dizer que os instrumentos musicais podem ser:

1. Baixo
2. Barítono
3. Tenor
4. Contralto
5. Soprano

Assim, dependendo da escala em que o instrumento atua, ele pode ser enquadrado numa das classes acima e mesmo ser adjetivado por elas, como o saxofone.

Parte II

Visão Geral do Sistema de Classificação

Capítulo 3

Segmentação e Pré-processamento

Para efeitos de convenção, neste trabalho entende-se nota musical como sendo o sinal acústico associado ao som (tipicamente, com altura definida) produzido por um determinado instrumento musical. Nesse capítulo, se apresentarão as transformações efetuadas sobre as notas antes de se extrair as características pertinentes ao processo de classificação. Para tal, serão abordados os seguintes tópicos: caracterização da nota musical, obtenção da envoltória da potência da nota musical, segmentação da nota musical, e obtenção dos momentos e escalamento dinâmico.

3.1 Caracterização da Nota Musical

Para a identificação de instrumentos musicais a partir de notas isoladas, cumpre destacar quais fatores podem afetar o padrão apresentado pela nota musical. O desafio será contemplado num cenário simplificado, onde inexistem interferências (como a presença de outras fontes sonoras), bem como reverberações decorrentes do ambiente acústico. Portanto, as bases de dados usadas nesse trabalho possuem gravações de notas musicais de diversos instrumentos feitas em ambientes preparados acusticamente (sem interferências nem reverberações[1]). Assim, dada uma nota, restam os seguintes elementos que a modificam[2]: .

[1]Exceto a base de dados MUMS, que possui tempo de reverberação igual a 0,4 s

[2]O timbre é uma característica perceptiva associada ao processo de identificação de um instrumento musical podendo ser modificado (em maior ou menor grau) pelos elementos citados nessa lista.

1. o instrumento;
2. o modelo (marca, fabricante) do instrumento;
3. o músico;
4. as variações[3] aplicadas à nota, como o *tremolo*[4], *vibrato*[5], *pizzicato*[6], *staccato*[7], etc;
5. o *pitch* com que a nota é produzida;
6. a intensidade, nível dinâmico, com que a nota é produzida.

Melhor dizer que há duas tarefas em questão: 1) identificar o início e o fim da nota; e 2) localizar temporalmente trechos (segmentos) de interesse ao longo da duração da nota.

3.2 Obtenção da Envoltória da Potência da Nota musical

Uma vez que um sinal de áudio tipicamente oscila em torno do zero[8], é conveniente analisar a forma de onda correspondente à potência instantânea (Figura 3.1) ou a uma versão retificada do sinal. Tais transformações facilitam, por exemplo, a atribuição dos instantes inicial e final da nota bem como uma envoltória a cada nota musical.

A seguir apresentaremos algumas formas de se obter a envoltória do sinal.

---

[3]Articulações ou variações no modo de execução da nota.

[4]O *tremolo* é um efeito musical que representa variações periódicas no volume (amplitude) da nota musical.

[5]O *vibrato* é um efeito musical que representa uma variação periódica no *pitch* (frequência) da nota musical. O *vibrato* pode ser classificado tanto pela quantidade de variação no *pitch*, quanto na velocidade em que o *pitch* varia.

[6]O *pizzicato* é um modo de execução normalmente empregado nos instrumentos de corda, que consiste em pinçar as cordas com os dedos.

[7]O *staccato* é classificado como sendo uma articulação, ou seja, consiste em executar as notas musicais inserindo silêncio (intervalo) entre elas. Essa técnica é o oposto do legato, que une as notas de forma a não haver entre elas inserção de silêncio.

[8]Supondo o sinal sem *offset*.

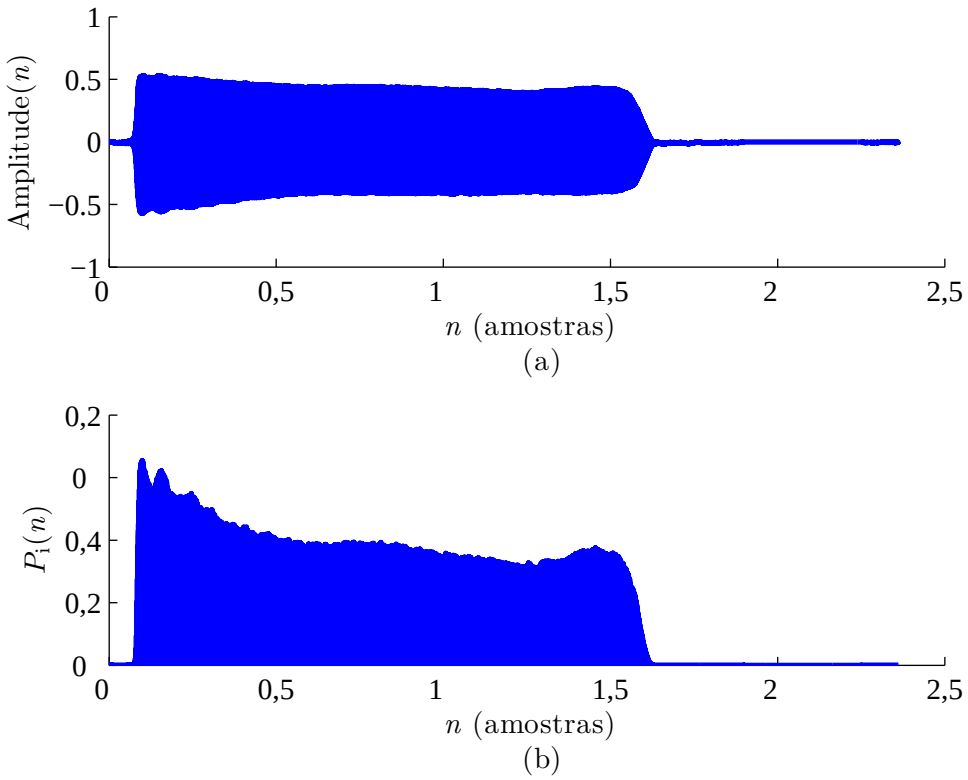

Figura 3.1: (a) $P_{\mathrm{i}}(n)$ nota (A4) de um Clarinete Bb; (b) potência instantânea, P_{i}(n).

3.2.1 Detector de Envoltória AM (DEAM)

A primeira forma mais intuitiva de obter uma envoltória para a nota musical baseou-se no algoritmo que aproxima um circuito detector de envoltória AM [17] (DEAM).

Primeiramente detectam-se os picos do sinal $P_i(n)$. A partir do primeiro pico, inicia-se uma exponencial definida por uma taxa de decaimento previamente estabelecida de forma empírica. O método empírico empregado foi estabelecido a partir da base de dados MIS, avaliando a taxa de decaimento da parte final da nota de maior *pitch* para todos os instrumentos dessa base de dados. Posteriormente usou-se a maior taxa de decaimento entre todas avaliações obtidas. Em seguida, no instante de tempo associado ao próximo pico detectado, comparam-se os valores da exponencial e da intensidade do pico: caso o pico seja mais intenso que a exponencial, preserva-se o pico, iniciando a partir daí uma nova exponencial decrescente; em caso contrário, preserva-se a exponencial decrescente originada a partir do pico anterior. Neste trabalho, a taxa de decaimento da exponencial decrescente foi ob-

tida empiricamente, observando diretamente as notas contidas no banco de dados [10]. Essa abordagem mantém o número de amostras do sinal antes de passar pelo algoritmo descrito. A função recursiva usada foi:

$$P_{\mathrm{i}}(n) = P_{\mathrm{i}}(n-1)e^{-0{,}002}, \tag{3.1}$$

onde $P_{\mathrm{i}}(n-1)$ representa a exponencial presente no instante $n-1$.

3.2.2 Método do Máximo

Uma segunda maneira de se obter um padrão parecido com o método DEAM é segmentar a nota em N *frames* de mesmo tamanho e, para cada *frame*, achar e armazenar a amostra com o maior valor. Por conveniência denominar-se-á tal procedimento de "método do máximo". Tal esquema apresenta um padrão de resposta bem próximo ao produzido pelo DEAM. No entanto, a resolução temporal da curva de envoltória é reduzida por N. Essa redução pode causar uma suavização da curva, ocasionando perda de informação que é utilizada pelo sistema de classificação em estudo. Sendo assim, faz-se necessário avaliar o impacto dessas perdas na estimação da envoltória sobre o desempenho do processo de identificação automática dos instrumentos musicais.

Na Figura 3.2 mostra-se primeiro uma estimativa de envoltória obtida pelo DEAM (usando a Eq. (3.1)) e, em seguida, outra estimativa produzida pelo método do máximo [2].

3.2.3 Método do Filtro

Um terceiro procedimento para se obter a envoltória do sinal é através do uso de um filtro passa-baixas com frequência de corte e ordem previamente estabelecidas[9] representada na Figura 3.3. A saída do filtro passa-baixas, excitado por $P_i(n)$, produz uma estimativa da envoltória do sinal de entrada. No exemplo apresentado na Figura 3.3, a implementação, aqui denominada "método do filtro", se baseou num filtro FIR de ordem 1023 projetado por janela Hamming com frequência de corte igual a fs/1000 (onde fs é a frequência de amostragem). Essa abordagem gera um

[9]Pode-se simplesmente usar-se a média de cada janela (método da média).

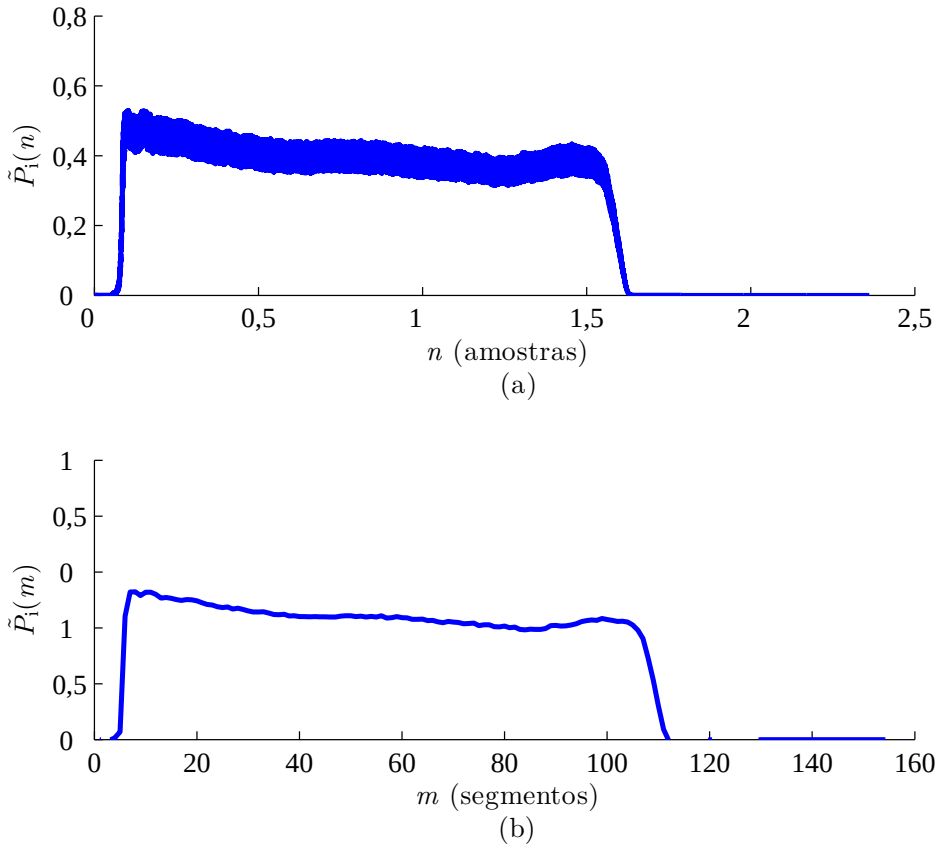

Figura 3.2: Envoltórias da nota de um Clarinete: (a) métodos DEAM; (b) método do Máximo.

número final de amostras, após a convolução, superior ao número inicial, dado pelo comprimento do sinal L somado à ordem do filtro passa-baixas.

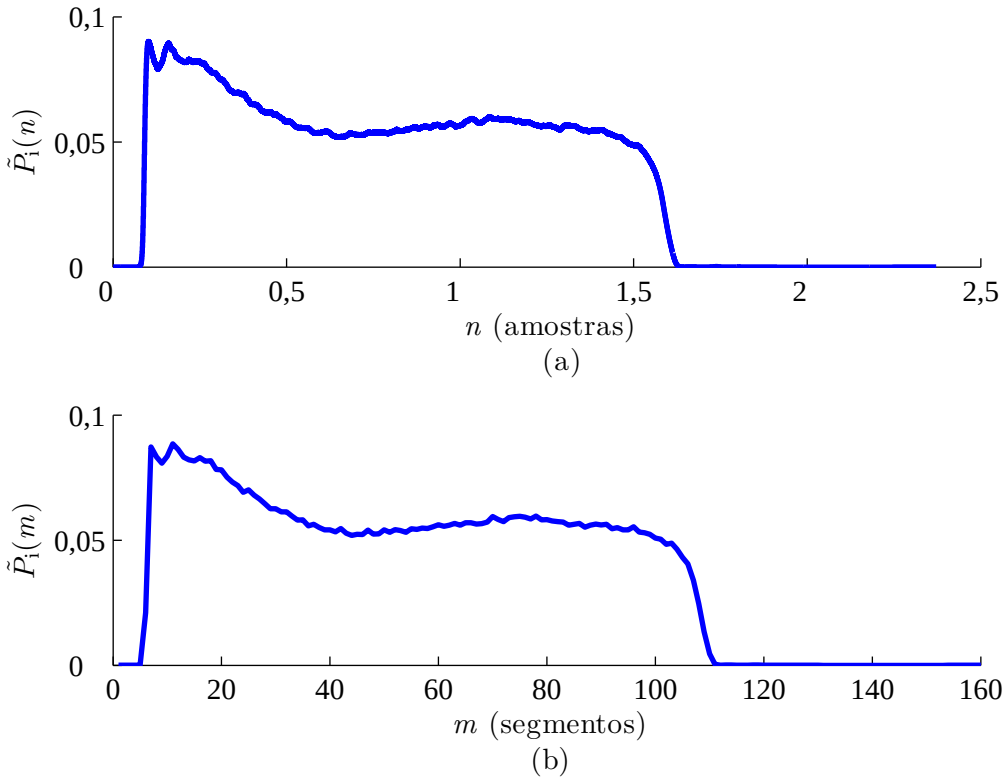

Figura 3.3: Envoltórias da nota de um Clarinete (a) pelo método do filtro e (b) pelo método da média.

3.3 Segmentação da Nota Musical

Existem duas tarefas iniciais a serem resolvidas na elaboração de um sistema de reconhecimento de instrumentos musicais a partir de notas isoladas: 1) identificar o início e o fim da nota e 2) localizar temporalmente trechos (segmentos) de interesse ao longo da duração da nota.

Pode-se, numa abordagem simplificada, dividir qualquer nota em 3 trechos (segmentos), onde o primeiro representa o surgimento da nota, o segundo (intermediário) representa a sustentação da nota, e o terceiro representa o encerramento da nota. A partir destes segmentos, o desafio é extrair adequadamente a informação mais útil para se identificar o instrumento que a originou. Assim, serão avaliadas al-

gumas estratégias, as quais tentarão delimitar de forma aproximada tais segmentos. A importância da determinação dos segmentos se deve ao fato de sua localização poder ser afetada diferentemente pelo instrumentista, o que pode ser desejável ou não, a depender do nível de discriminação que se pretende chegar. Por exemplo, quando o músico prolonga a nota do instrumento (aumentando a duração do segmento intermediário), ou quando a abafa repentinamente (ocasionando uma redução do segmento final), provoca voluntariamente padrões diversos para a mesma nota proveniente da mesma fonte (instrumento) num dado cenário (ambientação acústica). Essas alterações podem gerar confusões caso se queira identificar somente o instrumento. Nesses casos podemos dizer que certas características (duração, por exemplo) desses segmentos, intermediário e final, sofreram uma maior modificação pelo instrumentista do que os segmentos iniciais da nota. Logo, pode-se pensar que determinados segmentos são mais interessantes do que outros para a extração de características não-volitivas do instrumento, isso porque eles são mais imunes à intervenção do instrumentista[10]. Portanto uma escolha adequada do segmento para a extração de características é uma estratégia que deve ser considerada, principalmente quando se quer obter características para reconhecimento de fontes comuns, independentemente das inflexões expressivas (tais como o *vibrato*) que o instrumentista possa vir a inserir no som produzido. Isto é importante, já que essas alterações podem em princípio ser um fator complicador para a discriminação dos instrumentos musicais.

3.3.1 Segmentação pelo Modelo ADSR

3.3.1.1 Definições Originais

Uma divisão clássica de um sinal acústico associado ao som (nota) de um instrumento musical pode ser feita em quatro segmentos, ou seja, Ataque, Decaimento, Sustentação e Relaxação [2], conhecido como ADSR (*Attack, Decay, Sustain*

[10]Dependendo do tipo de instrumento, pois em alguns instrumentos o instrumentista possui controle sobre a natureza do ataque (suave, incisivo agressivo), o que pode modificar o padrão da nota.

e Release)[11]. Cada segmento pode ser melhor determinado no padrão da envoltória da potência instantânea do sinal, conforme pode se ver na Figura 3.4.

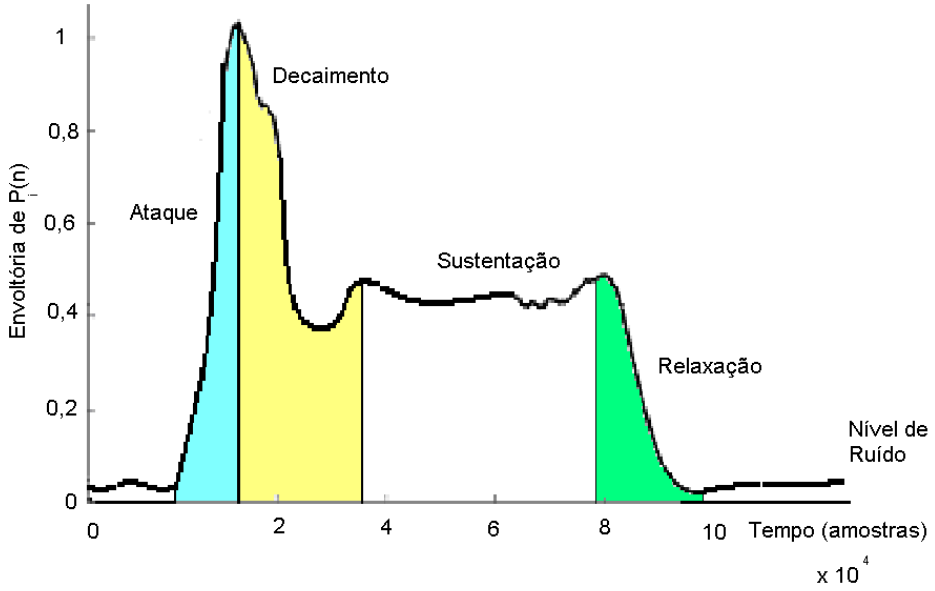

Figura 3.4: Modelo ADSR.

No entanto, nem todos os instrumentos produzem notas contendo todos os tipos de segmentos, assim, somente os segmentos de ataque e relaxação se encontram necessariamente presentes em todos os instrumentos [2].

O segmento de ataque normalmente corresponde à subida do sinal, em termos de potência, indo do nível do ruído de fundo até um máximo inicial[12]. Representa, portanto, o intervalo de tempo em que ocorre o transitório inicial. É senso comum asseverar que retirar o ataque do sinal implica uma maior dificuldade para se conseguir uma diferenciação auditiva do timbre de alguns instrumentos musicais por parte dos ouvintes [18]. Daí advém uma justificativa natural para tentar o ataque na fase de extração de características.

O segmento de decaimento representa o intervalo de tempo decorrido entre o instante do máximo inicial e aquele em que se atinge o nível de sustentação[13]. O seg-

[11]Esse modelo foi concebido por Vladimir Ussachevsky em 1965, quando liderava a *Columbia-Princeton Electronic Music Center.*

[12]Normalmente representado pelo máximo global, como o instante desse pode ocorrer dentro do trecho de sustentação, neste trabalho adotou-se o máximo inicial como o primeiro máximo local.

[13]Aproximadamente o intervalo de tempo compreendido entre o instante do primeiro máximo e

mento de relaxação ocorre quando novamente o volume do sinal começa a diminuir até atingir o nível de ruído de fundo ou zero. Já o segmento de sustentação ocorre entre o término do trecho de decaimento e o início do segmento de relaxação [2].

3.3.1.2 Definições Alternativas

Uma abordagem aproximada seria definir o segmento de ataque como o intervalo compreendido entre o momento em que o sinal ultrapassa o ruído de fundo até o instante em que ocorre o primeiro pico. Logo, para que o primeiro pico não venha ser obtido das pequenas flutuações durante a subida da envoltória do sinal, é necessário suavizá-la suficientemente para que esses pequenos picos desapareçam. Para isso verificou-se que o procedimento mais adequado para a obtenção da envoltória foi o método do máximo (ver Seção 3.2.2). O segmento de decaimento é determinado pelo intervalo circunscrito entre o primeiro pico obtido anteriormente e o instante associado ao primeiro vale após esse pico[14]. O segmento de relaxação é definido como o intervalo entre o instante associado ao primeiro pico (máximo local) com valor superior ao limiar de ruído de fundo, a partir do final da nota, e o instante em que o sinal volta a ficar menor que o limiar de ruído de fundo (no sentido do início ao final da nota). Por fim, o segmento de sustentação é obtido pelo intervalo situado entre o final do segmento de decaimento e o início do segmento de relaxação.

Como visto anteriormente, as definições aqui propostas para os trechos de uma nota diferem das do modelo ADSR convencional. Portanto, nesse trabalho chamaremos o método proposto de modelo ADSRm. A modificação proposta se deve basicamente ao fato de os sinais reais apresentarem em alguns casos comportamento não previsto pelo modelo ADSR que, se fosse aplicado, descaracterizaria o significado dos segmentos (e.g., um ataque com duração até o meio da nota). O modelo ADSRm obtém segmentos similares aos do ADSR, sendo capaz de determinar também, quando for o caso, a ausência do decaimento e da sustentação.

---

o instante de término do primeiro vale.

[14]Note-se que o método descrito difere um pouco quanto ao momento do início do segmento de sustentação apresentado na Figura 3.4, em que o início do trecho de sustentação começa a partir do primeiro pico que sucede o primeiro vale.

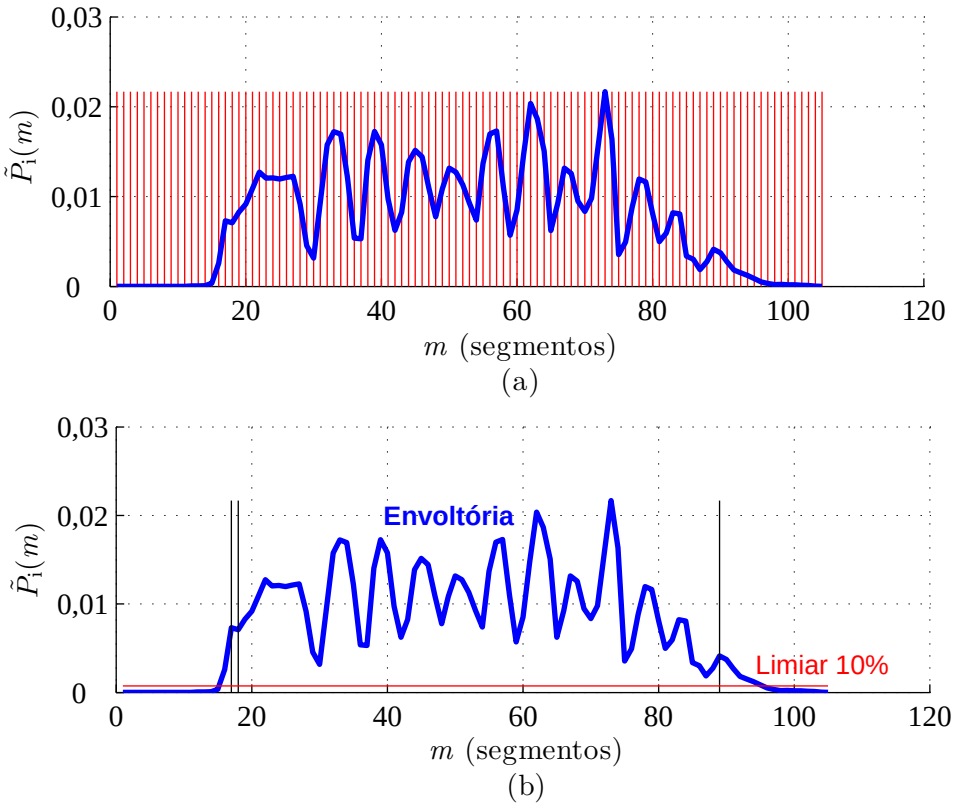

Figura 3.5: Segmentação segundo o modelo ADSRm de uma nota C4 de uma flauta contralto. As linhas vermelhas verticais sólidas correspondem em (a) aos inícios de cada frame analisado. Já em (b), as linhas pretas verticais, da esquerda para a direita, correspondem aos inícios dos segmentos de decaimento, sustentação e relaxação, respectivamente.

3.3.1.3 Exemplos

Nas Figuras 3.5, 3.6, 3.7 apresentamos resultados de segmentação obtidos para 3 instrumentos, respectivamente: Flauta Contralto sem *vibrato*, Saxofone Contralto sem *vibrato* e Violino *pizzicato*.

No caso da flauta, ilustrado pela Figura 3.5, pode-se constatar que o segmento de decaimento é bastante curto. Já no caso do saxofone contralto, apresentado pela Figura 3.6, o segmento de decaimento é melhor destacado. Pode-se também observar que neste caso todos os segmentos aparecem na forma prevista pelo modelo ADSR.

Finalmente, no caso da envoltória da nota do violino, apresentada na Figura 3.7, pode-se perceber que a nota da envoltória do violino só apresenta os segmentos de ataque e relaxação. Conforme já comentado, determinadas notas de instrumentos

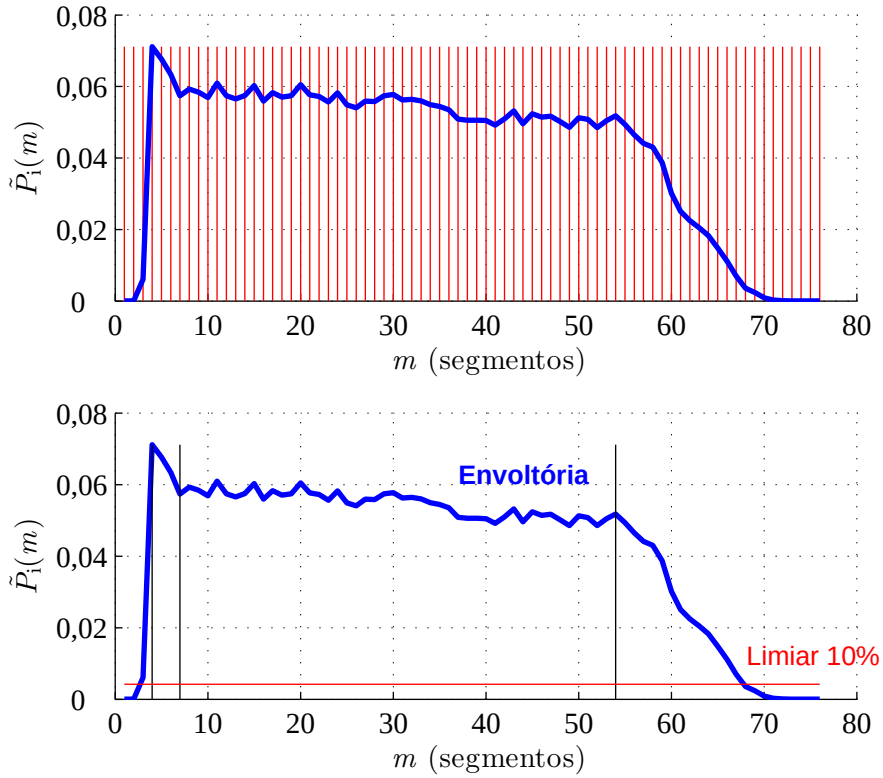

Figura 3.6: Segmentação segundo o modelo ADSRm para a nota C4 de um saxofone. Vide Figura 3.4 para a definição dos elementos gráficos envolvidos.

podem não apresentar todos os segmentos previstos pelo modelo ADSR[15].

3.3.2 Segmentação por Limiares

Uma forma de se localizar o início e o fim da nota é usarmos limiares sobre a envoltória ou potência da nota. Tipicamente o segmento em questão é caracterizado pelo intervalo de tempo em que o sinal apresenta intensidade em níveis superiores a um determinado limiar, cujo valor normalmente é escolhido para destacar a nota do ruído de fundo. No entanto, essa mesma abordagem pode ser usada para se obter segmentos de interesse contidos na nota, ou seja, início, meio e fim. Para isso basta acrescentarmos um segundo limiar, cujo valor é superdimensionado em relação ao anterior. Neste tipo de abordagem, o objetivo é determinar os segmentos de interesse, a partir dos quais se irão extrair as características representativas da

[15]Os trechos marcados nas figuras 3.5 a 3.7 foram obtidos automaticamente pelo algortimo implementado para o modelo ADSRm

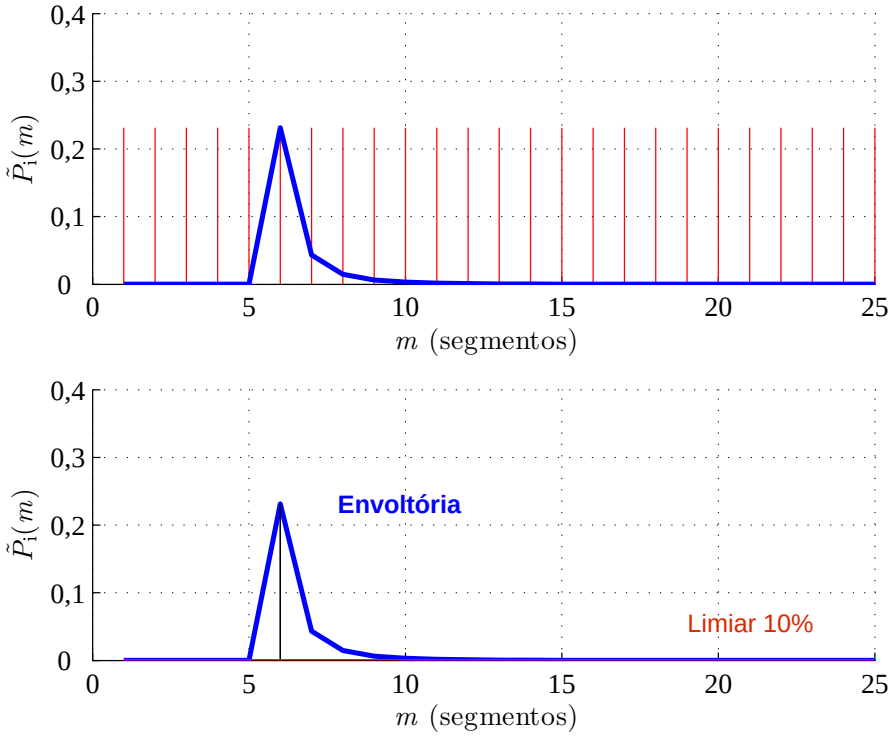

Figura 3.7: Segmentação segundo o modelo ADSRm para a nota C4 de um violino.

nota.

3.3.2.1 Segmentação com 1 Limiar

Como proposta inicial, pensou-se em destacar somente um segmento intermediário da nota através de um limiar. Neste caso, o segmento de interesse é obtido do intervalo compreendido pelo instante em que a potência da nota[16] pela primeira vez ultrapassa o limiar até o instante em que a potência da nota pela última vez cruza esse mesmo limiar. Assim, o limiar deve ser definido para um valor acima do nível do ruído de fundo, pois caso contrário se poderá ter amostras do segmento contendo somente ruído de fundo. Infelizmente, nessa abordagem, o controle do segmento extraído da nota é restrito, ou seja, será impossível garantir a extração de um segmento intermediário delineado por níveis distintos de potência, por um intervalo de tempo predeterminado, ou mesmo um segmento de sustentação, conforme pode ser visto na Figura 3.8, a qual ilustra essa situação.

No presente trabalho, usaram-se valores de limiar entre 10% e 90% da média

[16] O mesmo procedimento pode ser aplicado à envoltória da nota.

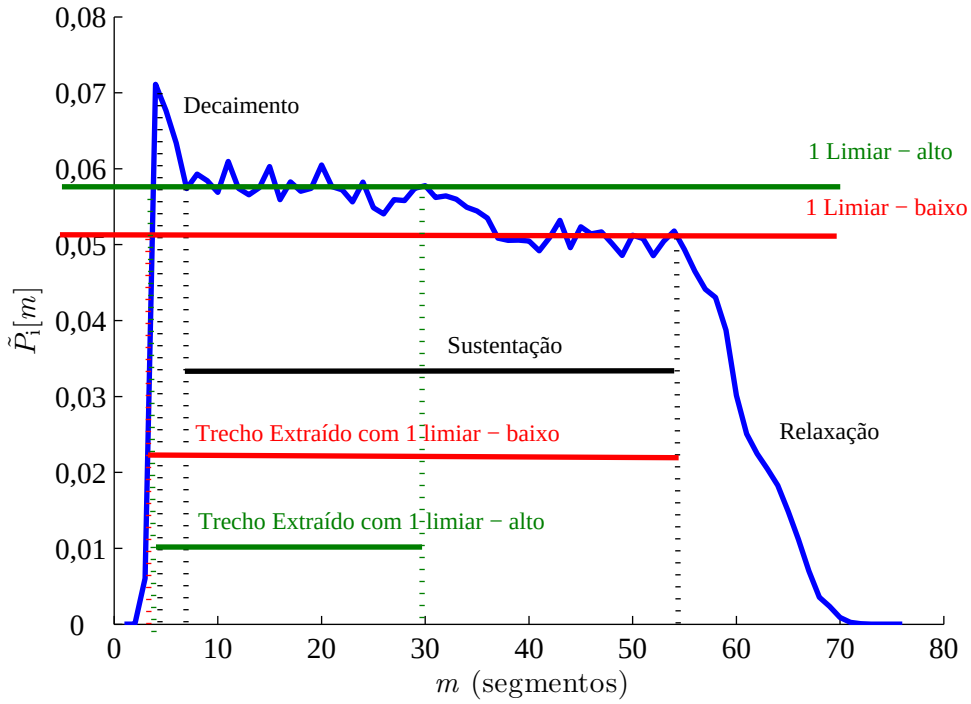

Figura 3.8: Segmentação a partir de um limiar sobre a envoltória da nota C4 de um saxofone contralto.

da potência instantânea do sinal, visando à obtenção de um segmento intermediário mais estável (em frequência e possivelmente em amplitude) da nota, sobre o qual serão medidos elementos caracterizadores, tais como parametrização por codificadores como LSF, LPC, MFCC e CEPSTRUM. Devido à sua simplicidade, essa técnica é comumente utilizada e, na prática, apresenta bons resultados [3].

Como visto, a abordagem de segmentação através de um único limiar se restringe a selecionar apenas um segmento. Tal desvantagem pode ser contornada pelo acréscimo de mais limiares. Isso resolve o problema parcialmente, restando-se ainda determinar de forma mais precisa os valores desses limiares em correspondência aos segmentos que se pretende extrair.

3.3.2.2 Segmentação com 2 limiares - Modelo IMF

Como uma alternativa ao método de segmentação com um limiar, elaborou-se para esse trabalho um critério de segmentação baseado em 2 limiares, possibilitando

uma possível correspondência com alguns segmentos previstos no modelo ADSR[17].

No caso de adotarmos 2 limiares tais que:

- limiar 1 (inferior) define o nível máximo do ruído de fundo, ou o nível em que se considera o sinal presente; e
- limiar 2 (superior) determina o nível máximo que o início ou o final da nota pode atingir;

poderemos obter 3 segmentos, abaixo definidos:

- o primeiro segmento (início), será definido pelo intervalo que vai do instante em que o sinal[18] cruza pela primeira vez o limiar 1 até o instante em que o sinal cruza pela primeira vez o limiar 2 a partir do **início** do sinal, ou seja, a **subida** do sinal;
- o segundo segmento será definido pelo intervalo que vai do instante em que o sinal cruza a primeira vez o limiar 2 até o instante em que o sinal cruza pela última vez o limiar 2, ou seja, a partir do **meio** do sinal, aqui denominado como segmento **estacionário**[19];
- e o terceiro segmento do sinal será definido pelo intervalo compreendido entre o instante em que o sinal cruza pela última vez o limiar 2 até o instante em que o sinal cruza pela última vez o limiar 1 **final** do sinal, ou seja, a **descida** do sinal.

A Figura 3.9 mostra um exemplo em que 2 limiares são aplicados à envoltória do sinal de uma nota, de forma a ilustrar o critério de segmentação anteriormente descrito. Pode-se notar que é possível associar de modo aproximado o primeiro segmento à subida do sinal, o segundo segmento ao trecho mais estacionário (onde em tese teríamos poucas variações de amplitude e pouca variação na frequência

[17]Critérios de segmentação usando dois limiares podem obter 3 segmentos aproximadamente equivalentes aos segmentos de ataque, sustentação e relaxação do modelo ADSR.

[18]Neste contexto o termo "sinal" poderá ser entendido como envoltória do sinal ou potência instantânea do sinal.

[19]Na verdade este segmento representa um trecho intermediário do sinal, onde normalmente se encontra o segmento de sustentação.

fundamental)[20] do sinal, e o último segmento à descida do sinal, conforme se pode observar na Figura 3.9.

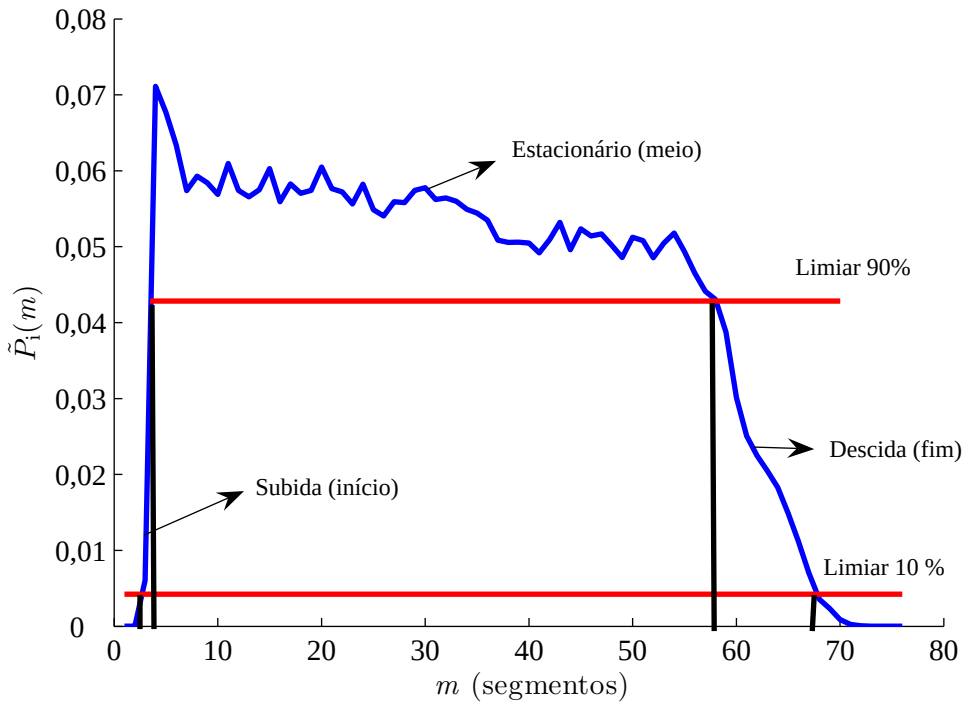

Figura 3.9: Segmentação IMF a partir de 2 limiares sobre a envoltória da nota C4 de um saxofone contralto. Usou-se 10% e 90% da média da potência instantânea do sinal para definirmos o limiar 1 e o limiar 2, respectivamente, em vez de 10% e 90% do maior pico, como proposto em [19].

3.3.3 Segmentação pelo *Pitch*

Um aspecto que se deve destacar é que a nota produzida por um instrumento musical não apresenta um padrão senoidal puro. De fato, quando dizemos que o tom da referida nota é Lá da quarta oitava, queremos dizer que a percepção auditiva de um ouvinte sobre a altura (*pitch*) desta nota é aproximadamente a mesma de escutar

[20] Caso ocorra variações na amplitude, como é o caso do tremolo, ainda assim teremos a frequência fundamental apresentando pouca variação. Assim, estamos usando o termo estacionário num contexto amplo e não puramente estatístico, ou seja, o segmento onde a fundamental e a amplitude apresentam variações menores que os demais segmentos da nota.

um sinal senoidal na frequência de 440 Hz. No entanto, quando analisamos o espectro de frequência da nota de um instrumento musical, tipicamente observamos diversas frequências, conforme pode ser visto no espectrograma apresentado na Figura 3.10, referente ao Lá de quarta oitava (440 Hz) produzida pelo instrumento Clarinete Si bemol. A figura também apresenta a magnitude do espectro e a envoltória de potência do sinal.

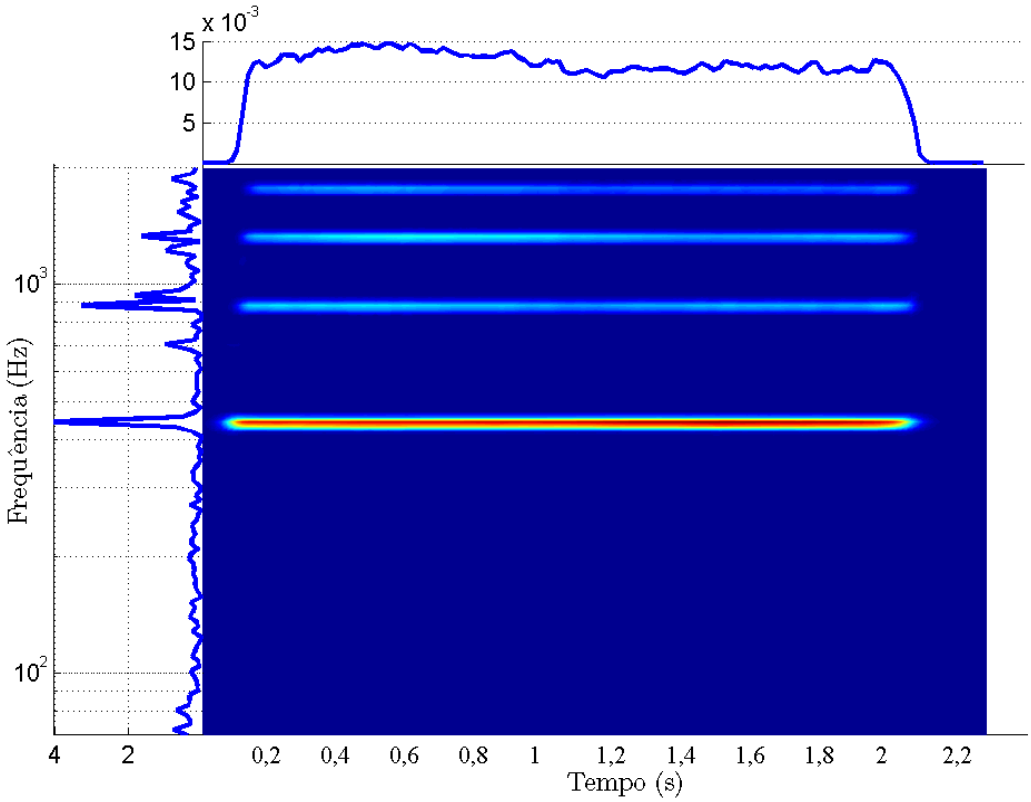

Figura 3.10: Espectrograma CQT (*constant Q spectral transform*) do Lá de quarta oitava (440 Hz) tocado em um Clarinete Si bemol [20, 21].

Uma outra forma de segmentar a nota é perceber que, na execução de notas isoladas com altura fixa, o *pitch* se manterá aproximadamente constante no segmento que sucede o ataque da nota. Assim, é de se esperar que existam um segmento associado ao processo de estabilização do *pitch*, outro de manutenção do *pitch*, e um trecho final de supressão do *pitch*. Portanto, para obtermos o segmento de manutenção do *pitch*, basta usar um estimador de *pitch* para identificar o período em que o *pitch* da nota se mostra mais estável. Uma pista que ajuda a determinar o segmento de estabilização do *pitch* da nota é sua maior duração em relação a outros segmentos obtidos nesse processo.

Para tanto, precisaremos de algoritmos estimadores de *pitch*. Foram avaliados vários algoritmos estimadores de *pitch*, que apresentaram resultados similares. Portanto, apresentaremos somente o estimador de *pitch* a partir da função de autocorrelação da nota [22].

O tamanho do segmento foi determinado pela menor frequência audível, uma vez que a menor frequência proporcionará o maior comprimento de onda, que deverá caber dentro da janela usada para estimarmos o *pitch*. Como a menor frequência da base de dados é o Dó da primeira oitava, ou seja, 32,7 Hz, e esta frequência está próxima do limite inferior da audição humana (20Hz), preferiu-se arredondar para baixo (30Hz) esse limite, deixando-o próximo ao limite inferior da audição humana e independente do limite inferior da tessitura de qualquer instrumento sob consideração neste trabalho. Já o limite superior foi a maior frequência encontrada na base de dados, que é de 3.951,07 Hz, portanto, nesse caso preferiu-se a nota imediatamente acima ($C8, 4.186,01$ Hz ≈ 4.200 Hz) de B7 para a restrição superior[21].

Os resultados obtidos para o Saxofone Contralto A4 e Trompa B2 são apresentados nas Figuras 3.11 e 3.12, respectivamente.

No caso de estimarmos o *pitch* para a nota B2 de uma trompa, podemos novamente observar que os métodos 1 e 2 se aproximam do *pitch* real. O que se pode concluir é que, apesar de notarmos que é possível destacar o momento em que o *pitch* é alcançado dentro da nota e quando ele se encerra, observamos também que a estacionariedade do *pitch* é rapidamente obtida, assim podemos dizer que o *pitch* se estabiliza ainda durante o ataque e se prolonga até quase o final da relaxação.

Portanto, pode-se afirmar que boa parte do segmento associado à subida e parte do segmento associado à descida do sinal ainda apresentam estacionariedade do *pitch*.

Portanto, o método para obtermos segmentos a partir do *pitch* será descartado neste trabalho, visto que o segmento intermediário deste modelo praticamente destaca a nota inteira. No entanto, isto não significa que, caso venhamos a extrair as características desse segmento, não possamos obter bons resultados. Logo, esse

[21] Poderia-se, sem acréscimo no esforço computacional, usar o limite superior da audição humana (20 kHz), mas não estamos contemplando essas frequências.

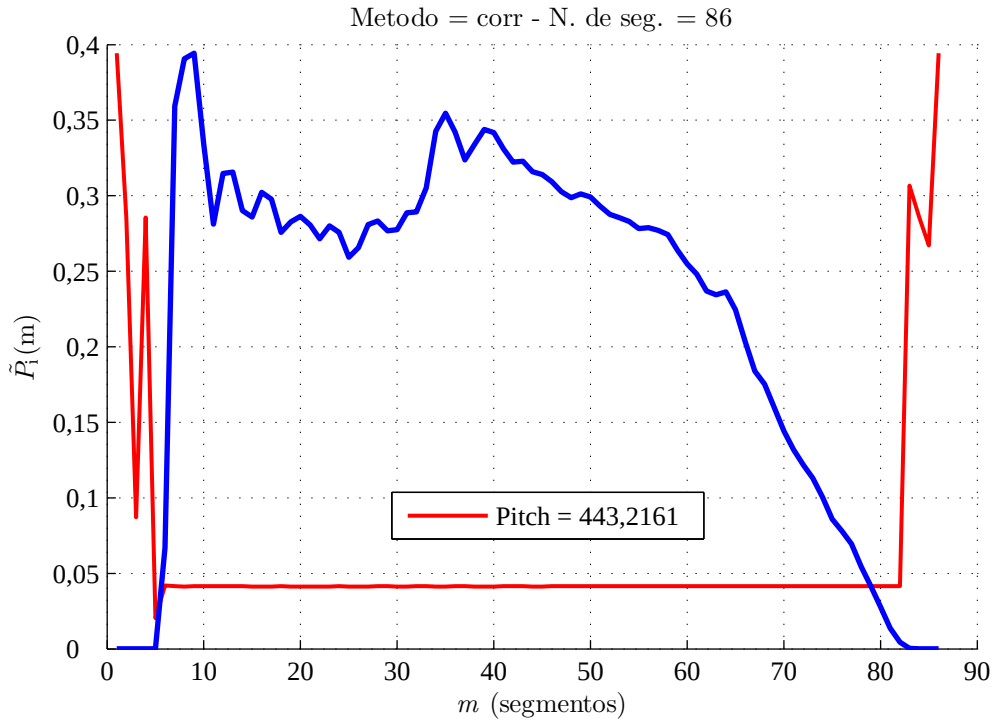

Figura 3.11: Envoltória da nota de um saxofone contralto - A4 e seu *pitch* estimado. A curva de *pitch* se encontra fora de escala, e somente está representada juntamente com a curva de envoltória para indicar os instantes em que o *pitch* se estabiliza.

é um método que, embora não segmente sempre a nota em 3 trechos[22] poderá ser futuramente investigado.

3.4 Obtenção dos Momentos e Escalamento Dinâmico

Inicialmente temos que levar em conta que não se pretende identificar diferenças das notas em relação a intensidade sonora (nível dinâmico) em que ela foi produzida. Também devemos levar em conta que notas provenientes de bases de dados distintas provavelmente apresentarão valores de intensidades sonoras diferentes, já que dependem do *setup* de gravação utilizado. Pode-se, por outro lado, contra-

[22] Os segmentos associados à subida e descida do sinal para a maioria dos casos avaliados praticamente inexistem.

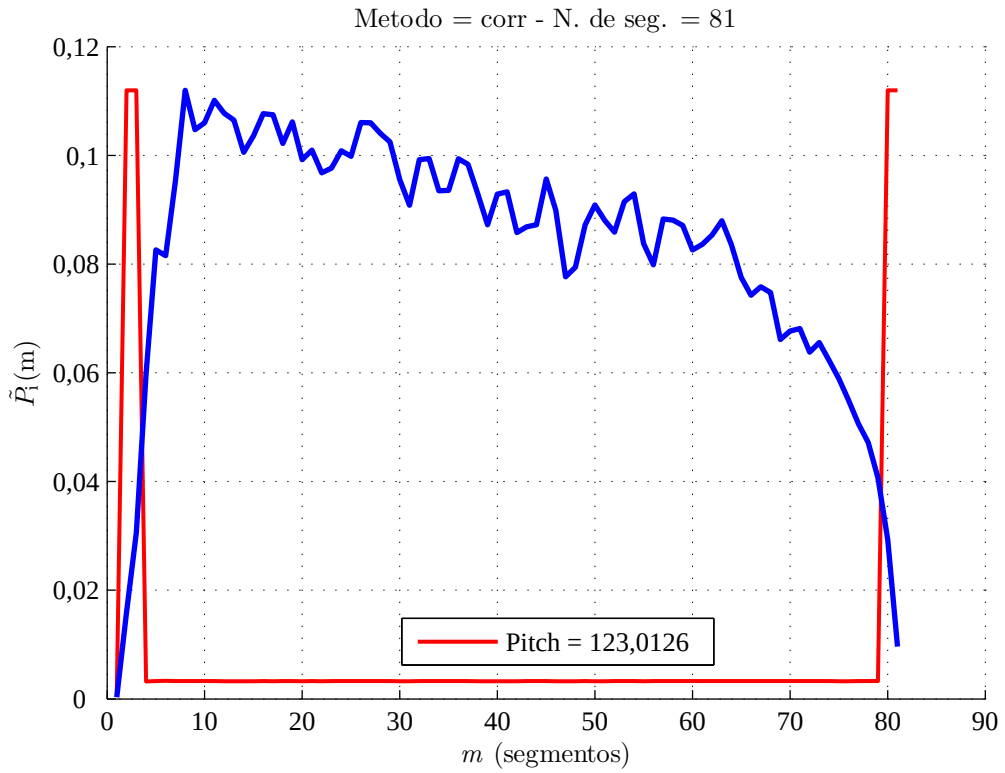

Figura 3.12: Envoltória da nota de uma trompa - B2 e seu *pitch* estimado.

argumentar que poderíamos deixar que tais diferenças de dinâmica fossem resolvidas pelo classificador. Contudo, observou-se que determinados classificadores, como a SVM e as Redes Neurais, podem eventualmente apresentar problemas de convergência na ausência de escalamento das amostras num dado segmento. Assim, para evitar problemas de convergência e possíveis confusões nas superfícies de separação, que poderiam ficar demasiadamente especializadas com os níveis dinâmicos encontrados na base de dados, optou-se por fazer um escalamento dinâmico de todas as amostras conforme:

$$\hat{x} = \frac{\vec{x} - \bar{x}}{\sigma} \tag{3.2}$$

Assim, para um dado segmento extraído por um dos métodos descritos anteriormente, é feito o escalamento do segmento, e são calculados os valores m_1 (média), m_2 (variância), m_3, e m_4 conforme [23]

$$m_1 = E[\vec{x}] = \bar{x} = \frac{1}{N}\sum x_i, i = 1 \ldots N \tag{3.3}$$

$$m_2 = E[(\vec{x} - \bar{x})^2] = \sigma^2, \sigma = \sqrt{\frac{1}{N-1}\sum (x_i - \bar{x})^2}, i = 1 \ldots N \tag{3.4}$$

$$m_3 = E[(\vec{x} - \bar{x})^3] = \frac{1}{N}\sum(x_i - \bar{x})^3, i = 1 \ldots N \tag{3.5}$$

$$m_4 = E[(\vec{x} - \bar{x})^4] = \frac{1}{N}\sum(x_i - \bar{x})^4, i = 1 \ldots N \tag{3.6}$$

sendo armazenados os valores σ (desvio padrão), m_3 e m_4 para compor o vetor de características. Note-se que o escalamento faz com que o segmento escolhido tenha média zero e desvio-padrão unitário, sendo que, para o valor armazenado do desvio-padrão, usou-se a fórmula não polarizada. Esses parâmetros serão doravante denominados descritores estatísticos.

Finalmente, devemos levar em conta que existe uma interdependência entre a tríade escolhida (segmentação, extração de características e classificação) e o resultado obtido (taxa de acerto) do sistema de reconhecimento de notas. Logo, dizer que uma forma de segmentação é melhor ou preferível à outra é uma assertiva algo problemática, uma vez que esse resultado é dependente dos demais elementos da tríade. Assim, uma comparação no desempenho do segmento escolhido, em última análise, somente é possível caso os demais elementos da tríade se mantenham inalterados.

Capítulo 4

Extração de Características

Este capítulo pretende descrever diversas formas de características representativas dos segmentos de sinais de áudio, obtidos através de um dos métodos descritos no capítulo anterior. Abordamos neste capítulo: descritores temporais, características obtidas sobre a envoltória da nota musical (inspiradas em medidas obtidas sobre um pulso radar) [19]; descritores de áudio usuais (descritores específicos) [2]; coeficientes de predição linear (LPC); *Line Spectral Frequencies* (LSF); coeficientes Cepstrais (CEPSTRUM); coeficientes Mel-Cepstrais (MFCC); e formação do vetor de características.

Essas características representativas são usadas pelo módulo extrator de características, que é responsável pela obtenção de um conjunto de características representativas do segmento analisado. O vetor de características é obtido por meio da concatenação de características estatísticas (desvio-padrão e momento de terceira ordem) dos segmentos, conforme visto na Seção 3.4, com as novas características abordadas nesse capítulo. Este vetor será utilizado pelo processo de classificação no sistema de reconhecimento automático de instrumentos musicais, o qual é objeto dessa dissertação.

4.1 Descritores Temporais

A ideia de se usar essas características obtidas a partir da envoltória da potência instantânea da nota em sinais de áudio veio, originalmente, de uma abordagem utilizada no campo da Guerra Eletrônica (*Warfare* [19, 24]) para o problema de re-

conhecimento automático do pulso radar. Nessa área existe um problema similar ao da identificação dos instrumentos musicais através de notas isoladas, que consiste em identificar e classificar de forma individualizada pulsos de Rádio Frequência (RF), originados por radares distintos. Espera-se que o pulso de RF gerado por um radar preserve características específicas desse Radar, em função de particularidades como o circuito gerador do pulso de RF, a válvula, o amplificador do pulso e a antena transmissora. Todos esses elementos impactam no padrão do pulso de RF em função das escolhas feitas durante o projeto do RADAR. Abaixo apresentamos na Figura 4.1, o modelo temporal de um pulso (Modelo TP), com a descrição dessas características representadas pelos seus respectivos rótulos (A,B, ... etc), extraídas da envoltória da potência instantânea da nota.

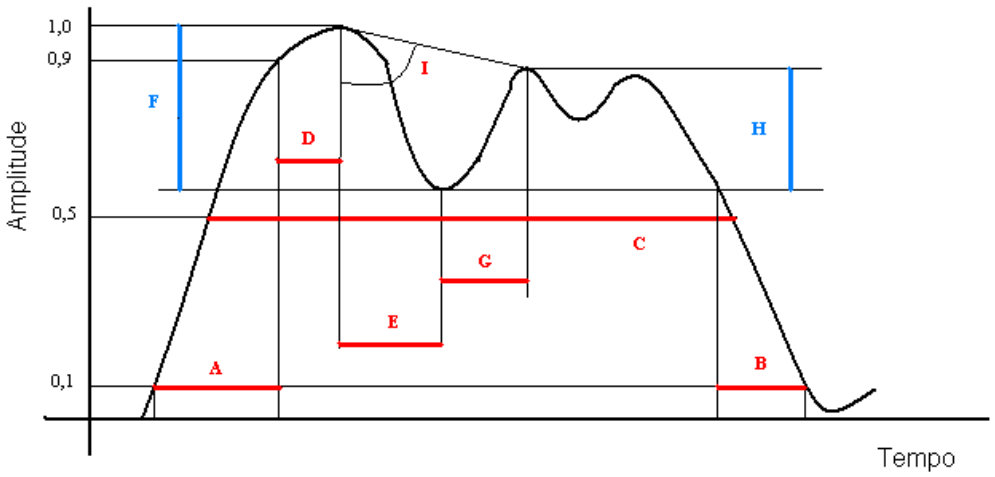

Figura 4.1: Modelo TP - Parâmetros de um pulso de RF típico.

A - Tempo de subida: é definido como o período de tempo entre os pontos de 10% e 90% da intensidade do pico máximo de amplitude marcados sobre o *leading edge* (bordo de ataque) do pulso;

B - Tempo de descida: é definido como o período de tempo entre o ponto com a amplitude correspondendo ao primeiro vale[1] e o ponto de 10% da intensidade

[1]Normalmente se adota 90% da intensidade do pico máximo, no entanto tal ponto quando marcado sobre o *tralling edge* (borda posterior) nos sinais de áudio não representava a descida do sinal, assim, fez-se essa adequação.

do pico máximo da envoltória, marcados sobre o *tralling edge* do pulso;

C - Largura do pulso: é definida como o período de tempo entre os pontos com 50% da amplitude do pico máximo marcados sobre o *leading edge* e o *tralling edge* do pulso, respectivamente;

D - Tempo dos 90% do pico: é definido como o período de tempo entre o primeiro pico e o ponto de 90% do pico máximo de amplitude marcado sobre o *leading edge* do pulso;

E - Tempo entre pico e vale: é definido como o período de tempo entre o primeiro pico e o primeiro vale da modulação do pulso;

F - Ripple percentual: é definido como a diferença percentual de amplitude entre o vale mais baixo e o pico mais alto;

G - Tempo entre vale e segundo pico: é definido como o período de tempo entre o primeiro vale e o segundo pico de modulação do pulso;

H - Percentual entre o vale e o segundo pico: é definido como a diferença percentual de amplitude entre o primeiro vale e o segundo pico;

I - *Droop*: é definido como o ângulo em radianos entre a linha que liga o segundo pico ao primeiro pico e a vertical.

Conforme se pode observar, exceto o *droop* (que é uma medida angular) todas essas características correspondem a medidas temporais ou de amplitude do sinal.

4.2 Descritores Específicos

Alguns trabalhos de classificação de instrumentos musicais utilizam determinados descritores específicos para áudio definidos no MPEG-7 [2] como medidas discriminadoras para o processo de classificação. Esses descritores apresentam correlação com aspectos da percepção humana. Portanto, são também chamados de descritores perceptuais [25], apesar de serem parâmetros objetivos e não subjetivos, como deveriam ser para que pudessem ser assim denominados.

A seguir apresentam-se alguns desses parâmetros e suas definições.

1. A taxa de cruzamento pelo zero (ZCR) representa um indicador da presença de componentes periódicas no sinal, sendo frequentemente utilizado nas aplicações de processamento de voz. É definido por:

$$\mathbf{ZCR} = \frac{\sum_n |\text{sign}(F(n)) - \text{sign}(F(n-1))|}{2N_a}, \tag{4.1}$$

onde N_a é o número de amostras no *frame*, e $F(n)$ é o valor da n-ésima amostra do *frame*.

2. A raiz da média quadrática (RMS) das amostras em um *frame*, definida por

$$\mathbf{RMS} = \sqrt{\frac{\sum_n F(n)^2}{N}}, \tag{4.2}$$

está associada à distribuição da energia ao longo dos *frames*.

3. O centróide espectral mede a frequência média ponderada em um dado *frame*; no seu cálculo, as frequências são ponderadas por suas respectivas amplitudes, dividindo-se o resultado pela soma das amplitudes:

$$\mathbf{SC} = \frac{\sum_k P(f(k))f(k)}{\sum_k P(f(k))}, \tag{4.3}$$

onde $f(k)$ é a k-ésima frequência do espectro do *frame* e $P(f(k))$ é o valor de amplitude associado a essa frequência.

4. A largura do centróide é calculada pelo módulo da diferença entre o centróide espectral e cada frequência, sendo ponderado pelas respectivas magnitudes:

$$\mathbf{BW} = \frac{\sum_k |SC - f(k)|P(f(k))}{\sum_k P(f(k))} \tag{4.4}$$

5. O fluxo espectral representa uma medida da mudança local espectral; no seu cálculo, considera-se o quadrado da diferença entre as magnitudes normalizadas de distribuições espectrais consecutivas:

$$\mathbf{FS} = \sum_k |P(f(k)) - P(f(k-1))|^2. \tag{4.5}$$

4.3 Coeficientes de Predição Linear

A parametrização LPC é muito utilizada em modelos fonte-filtro de produção de fala e música. No caso da voz, por exemplo, cujo modelo de produção é mostrado

na Figura 4.2, a fonte $u(n)$ é um sinal de excitação que representa a vibração produzida no ar ao ser forçado através das cordas vocais. Tal excitação passa então por um filtro $H(z)$ que modela as ressonâncias produzidas pelo trato vocal com função de transferência:

$$H(z) = \frac{G}{1 + \sum_{i=1}^{N_r} a_i z^{-i}}, \tag{4.6}$$

onde G é o ganho do modelo fonte-filtro, N_r é a ordem do filtro e a_i, $i = 1 \ldots N_r$, representam os seus coeficientes.

A saída do filtro $o(n)$ origina o sinal de voz de interesse. A modelagem do som produzido por um instrumento musical é análoga [26].

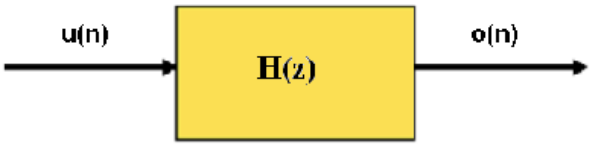

Figura 4.2: Modelo de fonte-filtro para produção de voz e música.

Uma vez que o sistema mostrado na Figura 4.2 modela convenientemente a produção da fala ou do som originado por um instrumento musical, espera-se que no caso dos instrumentos musicais, seus coeficientes forneçam um vetor de características capaz de modelar a tendência espectral, de forma a permitir uma discriminação dos instrumentos musicais.

A estimação dos coeficientes de predição linear consiste em encontrar um conjunto de coeficientes que minimizam o erro quadrático médio do seguinte preditor *forward*, aplicado em uma sequência:

$$\hat{s}(n) = \sum_{k=1}^{N} -a_k s(n-k) \tag{4.7}$$

onde o erro de predição $e(n)$ é a diferença entre o valor estimado $\hat{s}(n)$ e o valor real $s(n)$[2].

[2]A predição pode ser feita utilizando-se outros métodos de predição, tais como: método da autocorrelação, covariância, Burg, etc...

4.4 *Line Spectral Frequencies*

Pela análise LPC, o preditor da Equação (4.7) pode ser visto como a saída de um filtro gerador só-pólos $H(z) = 1/A(z)$ excitado por $u(n)$, onde

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-N}, \tag{4.8}$$

sendo N a ordem do filtro. Para obtermos os coeficientes LSFs, são elaborados dois polinômios, um simétrico e outro antisimétrico, que são definidos a partir de $A(z)$, respectivamente, por

$$P(z) = A(z) + z^{-(N+1)} A(z^{-1}) \tag{4.9}$$

$$Q(z) = A(z) - z^{-(N+1)} A(z^{-1}). \tag{4.10}$$

As raízes de P(z) e Q(z) se localizam na circunferência unitária e suas fases definem os valores das LSFs.

4.5 Características Cepstrais

Diversas aplicações em processamento de sinais utilizam-se de técnicas não-lineares, tais como a análise cepstral. Conceitualmente, o *cepstrum* complexo de um sinal $u(n)$ é dado por:

$$\hat{u}(n) = Z^{-1} \{\ln Z \{u(n)\}\} \tag{4.11}$$

onde a transformação Z normalmente é a DFT conforme pode ser visto na Figura 4.3 [27].

Na prática, antes da computação do *cepstrum*, a sequência $u(n)$ é multiplicada por uma janela de suavização (e.g., janela de Hamming).

Os primeiros componentes do *cepstrum* guardam informação sobre a envoltória da magnitude do espectro de um sinal, enquanto que os picos localizados no final do *cepstrum* correspondem à parte coerente (determinística) do espectro, tais como os picos senoidais de um sinal harmônico. Assim, os coeficientes de ordem mais alta do cepstrum podem ser relacionados à excitação quasi-harmônica em um modelo de produção de fala [27].

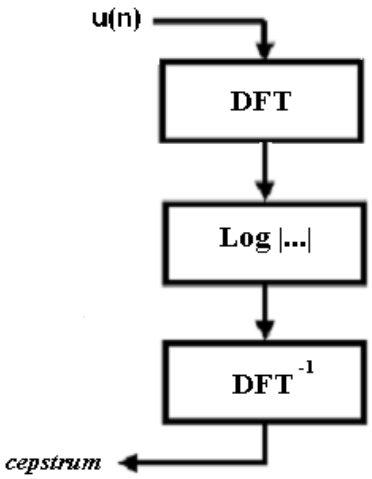

Figura 4.3: Modelo *cepstrum* para entrada u(n).

4.6 *Mel Cepstral Features*

Uma das contribuições da psico-acústica foi a descoberta que o sistema auditivo humano realiza uma análise espectral de sinais sonoros na qual a resolução frequêncial é não-uniforme. Esta descoberta levou à proposição de escalas auditivas (e.g., Mel, Bark e ERB), no lugar de escalas físicas (em Hz), para a análise perceptual de sinais acústicos.

O ponto de referência entre as escalas mel e Hz foi definido como sendo 1000 mels para o *pitch* de um tom senoidal puro de frequência igual a 1 kHz, com potência 40 dB acima do limiar mínimo da audição humana.

Na Figura 4.4 vemos o mapeamento aproximado entre as escalas mel e Hz, que é analiticamente fornecida por:

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{4.12}$$

onde f é a frequência em Hz. Para as frequências abaixo de 1000 Hz a relação é aproximadamente linear, e acima desse valor a relação é logarítmica.

Um outro fenômeno psico-acústico de interesse é o mascaramento dentro das chamadas bandas críticas [28].

Com o intuito de incorporar a escala mel e o conceito de banda crítica, introduziu-se o efeito da banda crítica dentro da escala mel, de forma que, ao invés de usarmos o logaritmo da magnitude das frequências, passou-se a utilizar o loga-

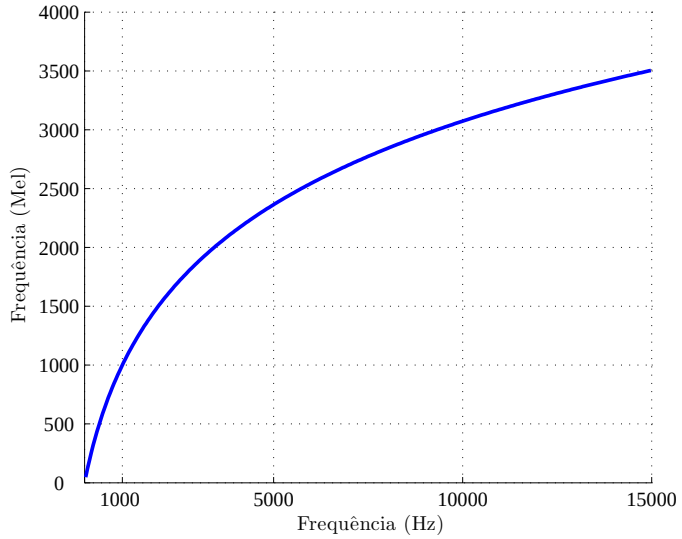

Figura 4.4: Mapeamento entre as escalas Hz e Mel, segundo a Eq. (4.12).

ritmo da energia total das bandas críticas em torno das frequências mel. Para isso utiliza-se um banco de filtros triangulares com resposta unitária na sua frequência central. No mais, as frequências centrais dos filtros são espaçadas linearmente na escala Mel. Seus limites (inferior e superior da banda de passagem) coincidem com as frequências centrais dos filtros triangulares adjacentes [1].

A principal diferença entre o *cepstrum* e os coeficientes *mel-frequency cepstral* é que no primeiro a análise espectral é feita com as bandas linearmente espaçadas, enquanto que no segundo as bandas de frequência são igualmente espaçadas conforme a escala mel, aproximando-se mais do processo de análise realizado no sistema humano de audição.

Finalmente, para se calcular os MFCC, divide-se o sinal $s(n)$ em janelas. Para cada janela estima-se a magnitude do espectro (na escala Hz), utilizando-se o módulo da DFT. Posteriormente cada espectro tem sua magnitude multiplicada com cada filtro triangular. Ao fim desse processo, faz-se o agrupamento dos valores obtidos em cada canal. Obtém-se assim um coeficiente para cada canal. O vetor feito do logaritmo destes coeficientes, é mapeado novamente para o domínio do tempo usando a DCT. A Figura 4.5 representa as etapas necessárias para obtenção do vetor MFCC[3].

[3]Figura baseada no livro [2].

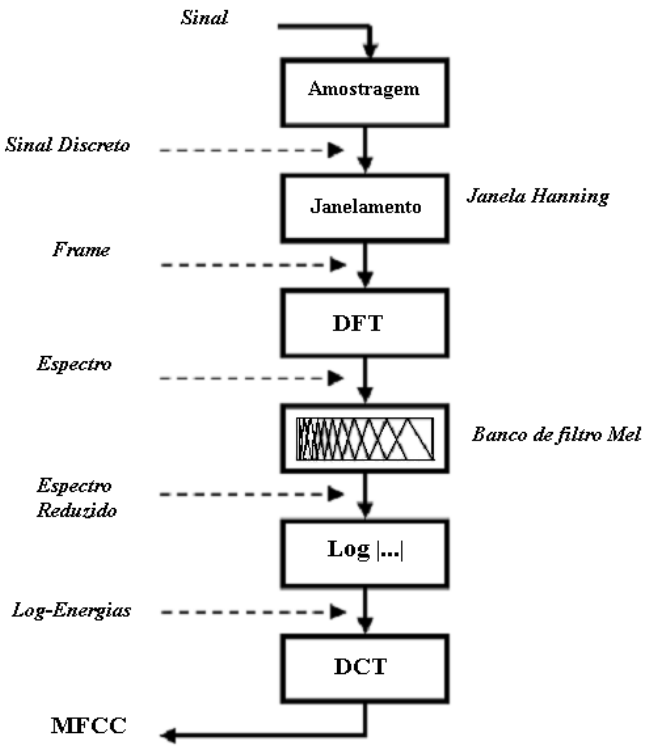

Figura 4.5: Esquema de obtenção do vetor MFCC.

A escala mel normalmente garante uma melhor representação do som. Por esse motivo os coeficientes MFC têm um uso amplo em diversas aplicações de áudio, como por exemplo na compressão e extração de características para sistemas de reconhecimento automático, tanto para a fala quanto para instrumentos musicais [3, 6].

4.7 Vetor de Características

Conforme foi descrito nas seções anteriores, poderemos ter descritores temporais, descritores específicos de áudio ou coeficientes provenientes de um dos codificadores (LPC, LSF, CEPSTRUM ou MFCC). O número de coeficientes que cada codificador fornecerá para o vetor de característica, assim como o número de elementos estatísticos, serão objetos de estudo no Capítulo 7. Já os demais descritores possuem número fixo de coeficientes, ou seja, 9 para o descritor temporal e 5 para o descritor específico de áudio.

Ao final da codificação será montado um vetor de características. Nesse vetor

serão acrescentados os elementos descritos nesse capítulo e os descritores estatísticos elencados ao final do capítulo anterior. Assim, o vetor de características poderá apresentar uma das seguintes formas:

- descritores temporais + descritores específicos de áudio + descritores estatísticos;
- codificadores + descritores estatísticos;
- codificadores + descritores específicos de áudio + descritores estatísticos;
- codificadores + descritores temporais + descritores estatísticos;
- codificadores + descritores temporais + descritores específicos de áudio + descritores estatísticos.

Além disso, devemos levar em conta que os descritores temporais usarão o modelo TP, portanto a envoltória da potência do sinal (média RMS) será normalizada. Já os codificadores irão codificar um segmento específico, o qual, conforme comentado no Capítulo 3, será escalonado. Portanto, o segmento onde se fará parte da extração de características poderá ser:

- subida (modelo IMF);
- ataque (modelo ADSR);
- intermediário (modelo IMF);

Todos os segmentos iniciais tiveram dois padrões distintos processados pelo sistema de reconhecimento automático, um que sofreu uma transformação antes de calcularmos os coeficientes e um outro que não. No primeiro padrão aplica-se a DCT com intuito de aproveitarmos a propriedade de que a DCT de um sinal impulsivo é aproximadamente uma reta horizontal (conforme pode ser observado na Figura 4.6); já no segundo padrão não se aplica a DCT.

Ao aplicarmos a DCT no segmento inicial, onde se espera que exista maior incidência de sinais impulsivos e não periódicos, podemos garantir que a saída resultará num segmento com uma menor variação da amplitude e um certo nível de estacionariedade como se estivéssemos no segmento de sustentação do modelo ADSR. Dessa

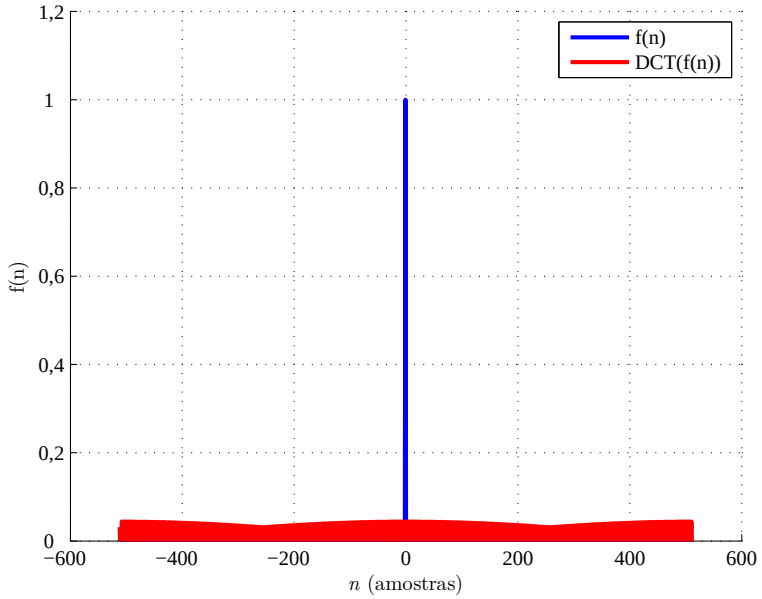

Figura 4.6: Sinal impulsivo e sua DCT.

forma espera-se facilitar a predição linear. Pode-se constatar essa menor variação da amplitude nas Figuras 4.7 e 4.8, onde mostramos a DCT do segmento de ataque e do segmento de subida, respectivamente, de uma nota C4 do instrumento Saxofone Contralto.

Neste momento, podemos finalmente elaborar a Figura 4.9, a qual apresenta um quadro resumo dos segmentos e dos descritores que serão empregados nessa dissertação.

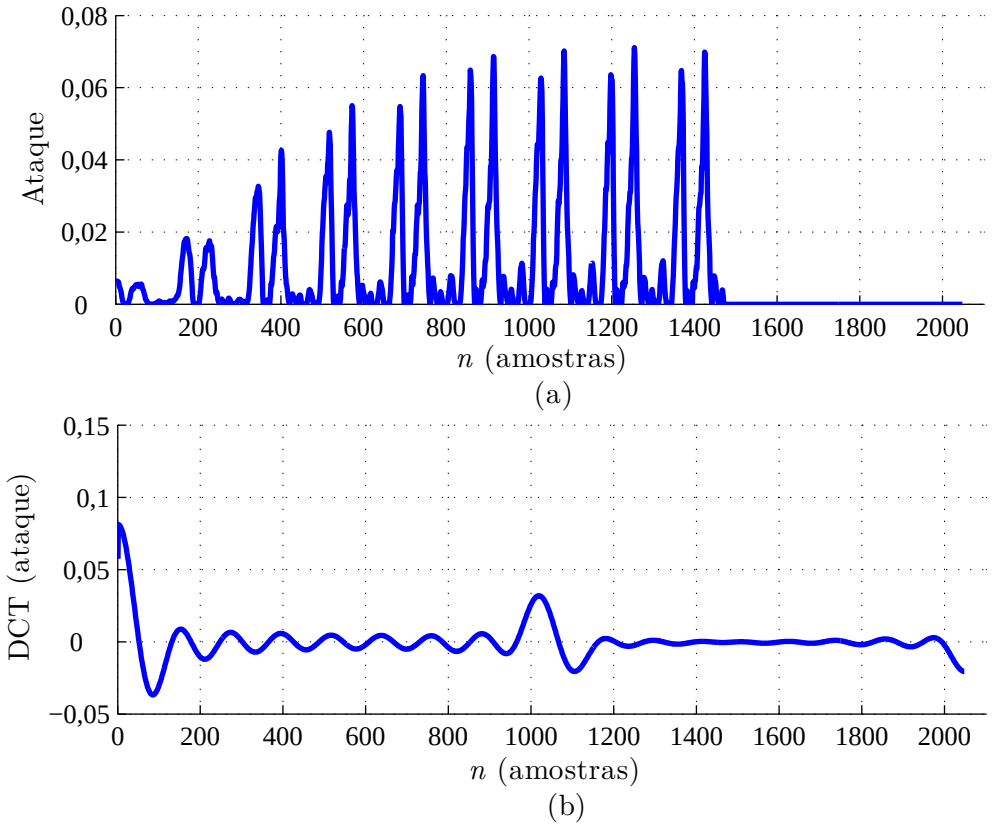

Figura 4.7: (a) o segmento de ataque de uma nota C4 de um Saxofone Contralto; (b) DCT do segmento (a).

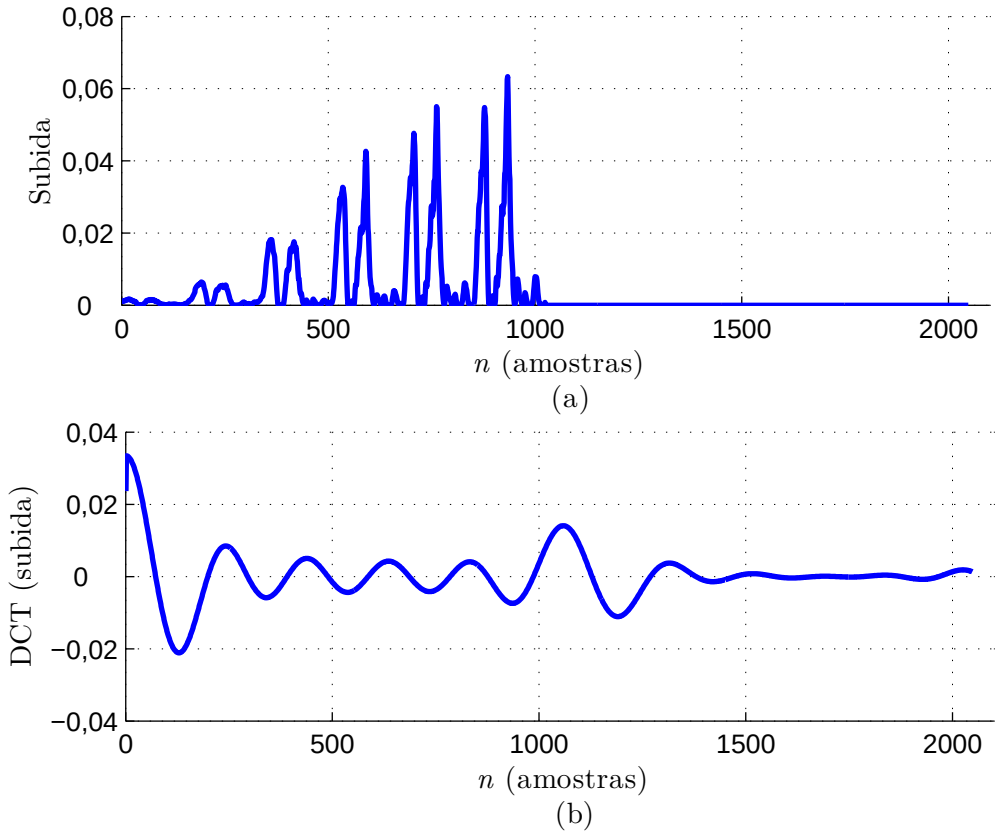

Figura 4.8: (a) o segmento de subida de uma nota C4 de um Saxofone Contralto; (b) DCT do segmento (a).

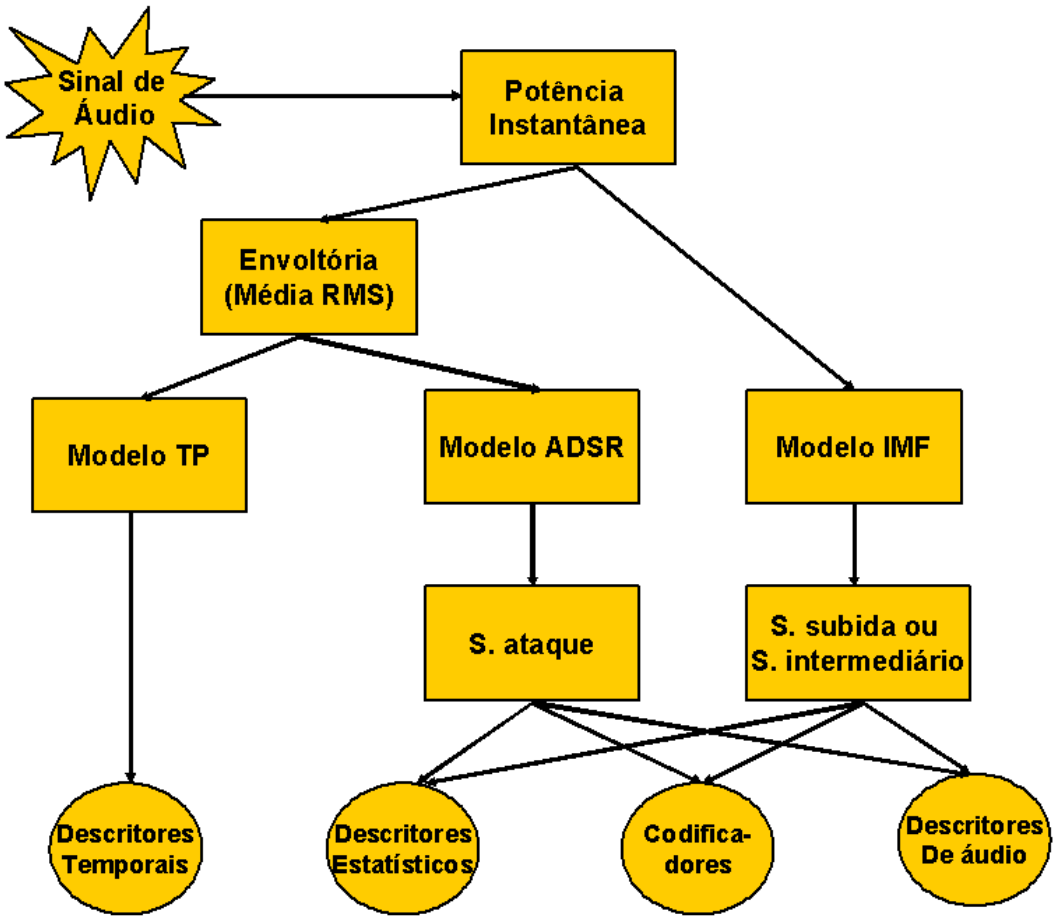

Figura 4.9: Quadro resumo da codificação.

Capítulo 5

Métodos de Classificação

Este capítulo tem como escopo apresentar os métodos de classificação que foram empregados nessa dissertação e as transformações não-lineares aplicadas aos métodos de classificação.

Existem diversos métodos que usualmente são empregados para a etapa de classificação, e que, de forma geral, podem ser classificados em um dos seguintes grupos:

1. Métodos conexionistas (Redes Neurais);
2. Métodos probabilísticos (Modelo de Misturas Gaussianas);
3. Métodos baseados em distância (K-Vizinhos mais próximos);
4. Métodos baseados em hiperplanos separadores (Máquina de vetor suporte).

Para reconhecimento automático de padrões ou no processo automático de reconhecimento de padrões são encontrados diversos métodos de classificação, tais como: Redes Neurais [29, 30], *Hidden Markov Models* (HMM [31]), Modelo de Misturas Gaussianas (GMM [6, 32]), Máquina de Vetor Suporte (SVM [8, 33]), Discriminantes Lineares [33], K-Vizinhos mais próximos (K-NN [34]), etc.

O presente trabalho não almeja avaliar todos os possíveis métodos de classificação. Assim, serão abordados 3 métodos de classificação: K-vizinhos mais próximos, SVM e Discriminantes Lineares. O primeiro método foi escolhido por se tratar de um classificador que normalmente é encontrado em trabalhos de reconhecimento de padrões. Sua popularidade deve-se ao fato de ser um método simples e, por esse motivo, normalmente encontrado como referência nos trabalhos para a demonstração dos resultados obtidos. O segundo método foi escolhido em função

da sua popularidade no emprego para trabalhos de reconhecimento automático de instrumentos musicais. Nesse quesito existem 2 métodos que se destacam: a SVM e o GMM. A escolha da SVM em detrimento do GMM foi arbitrária. Já a escolha do terceiro método se deu visando apresentar uma abordagem diferente para reconhecimento de instrumentos musicais, a fim de que se possa fazer uma contribuição alternativa nessa etapa para o sistema de reconhecimento automático de instrumentos musicais. Assim, foi elaborado um classificador por discriminantes lineares, por apresentar uma complexidade intermediária entre o SVM e o 1-NN e por ter baixíssimo emprego em reconhecimento de instrumentos musicais.

É comum que, a partir das variáveis de entrada obtidas do vetor de características, aqui neste capítulo chamado de vetor de entrada, delimitado por um domínio, conhecido como espaço de entrada (de dimensão N), se faça uma transformação não-linear (sobre o vetor de entrada) que mapeia a imagem num espaço de dimensão maior (de dimensão M), conhecido como espaço de características. Após essa transformação não-linear é feita a classificação, que mapeia a relação entre o conjunto definido pelo espaço de características e o conjunto delimitado pelo espaço de saída, conforme pode se ver na Figura 5.1

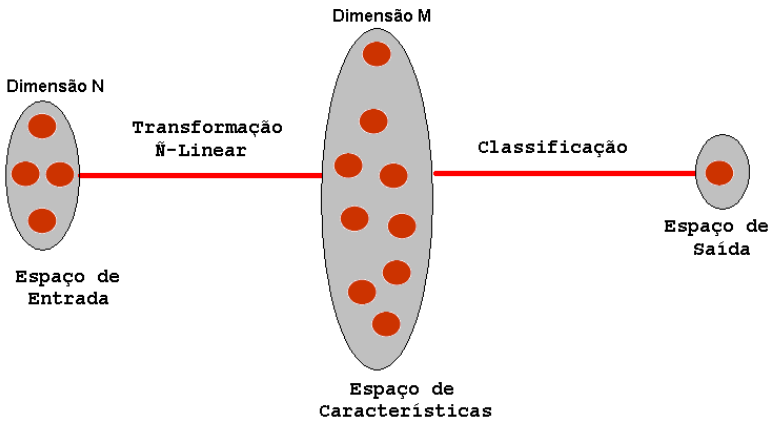

Figura 5.1: Mapeamento dos espaços envolvidos na classificação.

Portanto, o presente capítulo irá apresentar as principais características dos métodos K-vizinhos mais próximos, uma implementação alternativa e generalizada de Discriminantes Lineares, conhecida como *Generalized Linear Discriminant* [35], e aqui denominada de DLG (Discriminantes Lineares Generalizados) e Máquina de Vetor Suporte.

5.1 K-Vizinhos mais Próximos

O algoritmo dos K-vizinhos mais próximos (K-NN) é um método baseado em distância [35]. Assim, este método estima a classe mais provável de uma dada amostra a ser classificada segundo alguma métrica de distância a um conjunto de treinamento formado por amostras cujas classes são previamente conhecidas. Percorre-se o conjunto de treinamento, calculando a distância de cada uma de suas amostras em relação à amostra a classificar. Obtém-se então as K menores distâncias associadas à amostra que se deseja classificar, ou seja, os K-vizinhos mais próximos. A classe é atribuída àquela que foi mais frequente nos K-vizinhos. Caso K seja igual a 1, o algoritmo é reduzido à busca do vizinho que apresenta a menor distância, ou seja, o vizinho mais próximo (1-NN). Nesse trabalho usou-se a métrica de Minkowski [36] de ordem p, para medir a distância entre uma amostra do conjunto de treinamento e a amostra que se pretende classificar. Para p=2, a métrica de Minkowski equivale à distância Euclidiana entre a amostra $\mathbf{X}$ e a amostra $\mathbf{M}^j$ do conjunto de treinamento. A métrica de Minkowski é definida por:

$$D_x^j = \sqrt[p]{\sum_{i=1}^{n}(x_i - \mathbf{M}_i^j)^p} \tag{5.1}$$

onde x_i é o elemento i do vetor de características da amostra $\mathbf{X}$ e $\mathbf{M}_i^j$ é o elemento i do vetor de características da amostra $\mathbf{M}^j$ do conjunto de treinamento M.

Além dessa métrica também foi avaliada uma métrica conhecida como *city-block*, que mede os valores absolutos da diferença entre as amostras.

5.2 Discriminantes Lineares

O discriminante linear almeja encontrar um hiperplano que separe duas classes. Assim, o seu objetivo é achar, a partir de um conjunto de treinamento, o vetor $\vec{w}$ que define um hiperplano separador, por meio da minimização do quadrado do erro de classificação dado por:

$$\epsilon = t_{\vec{x}} - \tilde{y}(\vec{x}) \tag{5.2}$$

onde $t_{\vec{x}}$ (que pode assumir os valores -1 e 1) é a classe da amostra $\vec{x}$, e $\tilde{y}$ é uma função estimadora da classe. Assim, espera-se que se $\vec{w}'\vec{x} > 0$, a amostra $\vec{x}$ pertença à classe

1, caso contrário pertencerá à classe -1. Portanto, a classe da amostra é determinada por:

$$\vec{y}(\vec{x}) = \text{sign}(w'\vec{x}) \tag{5.3}$$

Para viabilizar a minimização por métodos que utilizam a direção do gradiente, substituiu-se a função sinal na Equação (5.3) pela função tangente hiperbólica. A mudança se justifica, uma vez que esta função, assim como a função sinal, possui sua imagem limitada pelos valores -1,1, sendo, ao contrário da função sinal, totalmente diferenciável em todo o seu domínio. Redefine-se, então, a classe da amostra $\vec{x}$ por:

$$\tilde{y}(\vec{x}) = \tanh(w'\vec{x}) \tag{5.4}$$

O algoritmo utilizado para minimizar o erro quadrático foi o *Least Mean Squares* (LMS) modificado por uma normalização [37]. Portanto o passo de iteração para se obter a convergência pode ser facilmente obtido, sendo dado por:

$$\vec{w}_{k+1} = \vec{w}_k - \frac{\mu \frac{\partial \vec{f}}{\partial \vec{w}}}{\Sigma + \Gamma} \tag{5.5}$$

O algoritmo de gradiente utilizado para a atualização da estimativa é dado pela equação acima e por:

$$\frac{\partial \vec{f}}{\partial \vec{w}} = -2\epsilon(1 - \tilde{y}^2)\vec{x} \tag{5.6}$$

$$\Sigma = (1 - \alpha)\Sigma + \alpha x'x \tag{5.7}$$

onde α assume o valor 0,05 (arbitrário), a matriz Σ é inicializada como uma matriz zero e Γ é uma matriz diagonal que apresenta valores da diagonal idênticos e próximos a zero, a fim de que a Equação (5.5) não apresente divisão por zero.

5.2.1 Transformação no Espaço das Características

Também foi investigado o efeito de uma extensão do espaço de características [33], consistindo na incorporação das potências, até um inteiro k, de cada parâmetro do vetor de características. Desta forma, se M é a dimensão do vetor de características associado a uma amostra, após a extensão kM será a nova dimensão tanto deste vetor de características transformado, agora definido pela Equação (5.8)

$$\vec{x}_p = [\ (\vec{x}^T)\ \ (\vec{x}^T)^2\ \ \ldots\ \ (\vec{x}^T)^{k-1}\ \ (\vec{x}^T)^k\ ]^T \tag{5.8}$$

quanto do hiperplano separador, agora dado por

$$\vec{w}_p = [\ (\vec{w}_1^T)\ \ (\vec{w}_2^T)\ \ \ldots\ \ (\vec{w}_{k-1}^T)\ \ (\vec{w}_k^T)\ ]^T \tag{5.9}$$

Nesse caso, a nova função estimadora da classe passa a ser

$$\tilde{y}'(\vec{x}) = \tanh(\sum_{i=1}^{k} \vec{w}_i^T \vec{x}^i) \tag{5.10}$$

Esta transformação não-linear foi usada em particular com o classificador DLG, de forma que a partir dela pode-se separar classes que antes não eram linearmente separáveis. Como se verá mais adiante, ela provocou um aumento na taxa de acerto das classes.

5.3 Máquina de Vetor Suporte

Esta seção pretende fazer uma breve descrição de como funciona uma Máquina de Vetor Suporte, SVM, (do inglês *Support Vector Machine*)[1]. A SVM implementa discriminantes lineares (hiperplanos) num espaço obtido por uma transformada do espaço de entrada, diferenciando uma classe, a positiva, de outra, a negativa (uma classificação binária, na sua forma tradicional).

A SVM visa construir um hiperplano que maximize a margem de separação entre os exemplares positivos e os negativos. Esse objetivo é obtido através de uma abordagem baseada na Teoria Estatística de Aprendizagem [38], implementando aproximadamente o método de minimização do risco estrutural[2]. Na Figura 5.2 encontramos um exemplo de um hiperplano separador ótimo. Neste exemplo, as cruzes pertencem a uma classe e os traços pertencem a outra classe. A despeito da utilização de discriminantes lineares, uma SVM não necessita, para efeitos de generalização, de classes linearmente separáveis. Tal se deve ao fato de a discriminação poder ser empregada num espaço de características, o qual já é uma transformação não-linear (*kernel*) do espaço de entradas. Assim o vetor de entrada ($\vec{x}$), no espaço de entradas, é mapeado em uma dimensão mais alta no espaço das características ($\vec{z}$). Escolhendo um mapeamento não linear a *priori*, a SVM constrói um hiperplano

[1] Um maior aprofundamento pode ser obtido em [35].

[2] Uma demonstração sobre as características desse modelo pode ser encontrada em [39].

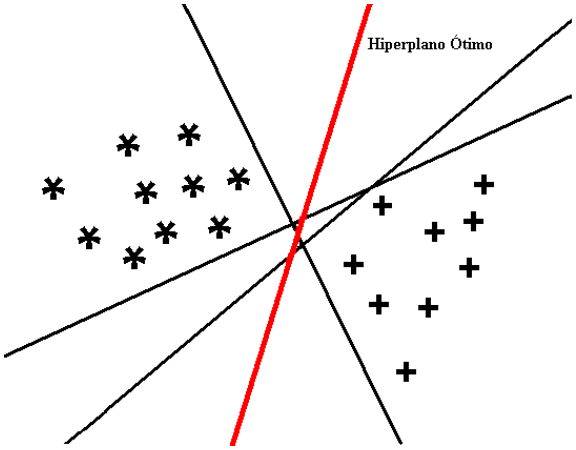

Figura 5.2: Hiperplano separador ótimo.

separador ótimo neste espaço de dimensão mais alta, conforme pode-se ver na Figura 5.1.

As restrições que o mapeamento não-linear sofre serão abordadas na Subseção 5.3.2, a qual apresentará alguns mapeamentos aceitáveis, tais como o polinomial, RBF e algumas funções sigmóides.

5.3.1 Caso linear do Modelo da SVM

Se o problema é linearmente separável por um hiperplano separador, o equacionamento para o modelo SVM é dado por:

$$\max D(\vec{\alpha}) = \sum_{k=1}^{n} \alpha_k - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \vec{x_i}, \vec{x_j} \rangle, \tag{5.11}$$

sujeito a:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{5.12}$$

$$0 \leq \alpha_i \leq C, \tag{5.13}$$

onde α_i é o multiplicador de Lagrange, $\vec{x_i}$ é o vetor de entrada e y_i é a classe associada ao i-ésimo vetor de entrada [+1,-1]. O coeficiente C tem que ser determinado. Este parâmetro introduz uma capacidade de controle adicional no classificador, podendo representar algum tipo de conhecimento a *priori* [39].

Assim, a solução ótima é fornecida por:

$$\alpha^* = arg \min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \vec{x_i}, \vec{x_j} \rangle - \sum_{k=1}^{n} \alpha_k \tag{5.14}$$

5.3.2 Transformações Não-Lineares - *Kernel*

O mapeamento do espaço de entrada num espaço de dimensão mais alta, a fim de resolver a limitação de classes que não são linearmente separáveis por hiperplanos separadores, pode ser justificada invocando-se o célebre Teorema de Cover [40], o qual afirma que padrões não-linearmente separáveis pertencentes a um dado espaço de entradas são, com alta probabilidade, linearmente separáveis num espaço de características, desde que a transformação seja não-linear e a dimensão do espaço de características seja alta o suficiente.

A teoria das funções *kernel* baseada em *Reproducing Kernel Hilbert Spaces*, RKHS [41, 42, 43, 44] afirma que um produto interno no espaço de características tem um *kernel* equivalente no espaço de entrada, ou seja,

$$K(x, x') = (\langle \Phi(x), \Phi(x') \rangle) \tag{5.15}$$

desde que garanta certas condições, como K ser uma função simétrica definida positiva, e respeite as condições de Mercer dadas por:

$$K(x, x') = \sum_{m}^{\infty} a_m \Phi_m(x) \Phi_m(x'), \quad a_m \geq 0 \tag{5.16}$$

$$\iint K(x, x') g(x) g(x') dx dx' > 0, \quad g \in L_2{}^3 \tag{5.17}$$

Assim, existem algumas restrições para que um mapeamento não-linear sob a forma de *kernel* possa ser empregado. Portanto, dado um mapeamento $\Phi : R^d \mapsto H$ (onde a dimensão de H é maior que d) a ideia é encontrar um hiperplano separador num espaço dimensional mais alto, equivalente a uma superfície não-linear em R^d.

Esta abordagem resolve um problema, já que teríamos de calcular o produto interno $\langle \Phi(x), \Phi(x') \rangle$, e isto seria complicado, visto que a dimensão H pode ser muito alta, ocasionando um problema de complexidade combinatorial, a um custo computacional impraticável. No entanto, se for conhecida a função *kernel* ($K(x, x')$), pode-se usá-la no lugar do produto interno da função Φ, o que reduziria o custo computacional, evitando que o produto interno no espaço de características fosse calculado. Esta abordagem fornece um caminho de se contornar a "maldição da dimensionalidade", porém o treinamento continuará dependente do número de

[3] Ou seja, $\int g(x)^2 dx$ é finito.

amostras, o que continua sendo uma restrição, visto que uma boa aproximação da distribuição dos dados depende de um grande número de amostras.

Assim, o *kernel* representa o produto interno no espaço de características, e, a seguir, apresentamos alguns mapeamentos que satisfazem as condições de Mercer.

5.3.2.1 Polinomial

$$K(x, x') = (\langle x, x' \rangle)^d \tag{5.18}$$

$$K(x, x') = (\langle x, x' \rangle + 1)^d \tag{5.19}$$

5.3.2.2 *Gaussian Radial Basis Function*

$$K(x, x') = \exp\left\{-\frac{\|x - x'\|^2}{2\sigma^2}\right\} \tag{5.20}$$

5.3.2.3 *Exponential Radial Basis Function*, RBF

$$K(x, x') = \exp\left\{-\frac{\|x - x'\|}{2\sigma^2}\right\} \tag{5.21}$$

5.3.2.4 *Multi-Layer Perceptron*

$$K(x, x') = \tanh(a\langle x, x' \rangle + b) \tag{5.22}$$

Dentro os mapeamentos descritos o polinomial é o método mais popular. Note-se que o *kernel* apresentado na Equação (5.19) evita possíveis problemas de singularidade quando a hessiana vai a zero.

5.3.3 Caso Não-Linear do Modelo da SVM

A SVM no espaço de características resolve um problema de programação não-linear que almeja maximizar a margem entre os vetores de entrada transformados e o hiperplano separador. A maximização se dá conforme o equacionamento na

forma dual dada por [35]:

$$\max D(\vec{\alpha}) = \sum_{k=1}^{n} \alpha_k - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\vec{x_i}, \vec{x_j}) \tag{5.23}$$

sujeito a:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{5.24}$$

$$0 \leq \alpha_i \leq C \tag{5.25}$$

onde $K(\vec{x_i}, \vec{x_j})$ é o *kernel* aplicado ao espaço de entrada e as demais variáveis estão descritas na Seção 5.3.1. Nesse caso conforme [39] a solução é dada pela Equação 5.26, e a classificação é obtida a partir da Equação 5.27.

$$\alpha^* = arg \min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\vec{x_i}, \vec{x_j}) - \sum_{k=1}^{n} \alpha_k \tag{5.26}$$

$$h(\vec{x}) = \text{sgn}(\sum_{i \in SV} \alpha_i y_i K(\vec{x_i}, \vec{x}) + b) \tag{5.27}$$

onde b representa o *bias*, e SV representa o conjunto dos vetores suporte, maiores detalhes sobre esse classificador pode ser encontrado além das referências já citadas nessa seção em [45].

Capítulo 6

Agrupamentos Hierárquicos, Abordagens Multiclasse e Estratégias

Neste capítulo analisaremos os seguintes tópicos: agrupamentos hierárquicos empregados nessa dissertação; abordagens multiclasse usadas em discriminantes lineares e máquinas de vetor suporte; e algumas estratégias associadas a essas abordagens para o sistema de reconhecimento automático de instrumentos musicais. Usaremos o conceito de família de instrumentos como sendo constituída de um conjunto particular de instrumentos musicais definida por um dos agrupamentos hierárquicos discutidos no Capítulo 2, e grupo de instrumentos como sendo um conjunto de instrumentos obtidos por um algoritmo que vise melhorar a sua diferenciação.

6.1 Agrupamentos Hierárquicos Empregados

Inicialmente no Capítulo 2 discutimos algumas formas de se agrupar os instrumentos musicais segundo taxonomias padrões da literatura, referentes ao estudo de organologia. No entanto, essa dissertação irá se restringir a basicamente 3 padrões de agrupamentos. O primeiro padrão de agrupamento consiste na coleção individual dos instrumentos. O segundo consiste na coleção das famílias de instrumentos conforme a taxonomia normalmente empregada nas orquestras sinfônicas, com um desdobramento, decorrente dos trabalhos anteriormente elaborados por [3, 4], onde as flautas constituem uma família separada. Assim, nesse agrupamento as famílias são as flautas, as palhetas, os metais, as cordas e os instrumentos de percussão. Por

fim, essas famílias de instrumentos, em algumas simulações, foram agrupadas formando outras famílias, constituída pelos instrumentos de sopro (flautas, palhetas, metais), cordas e percussão.

Assim, para cada uma das bases de dados (MIS, MUMS e RWC) foram utilizados conjuntos de instrumentos distintos, representando também taxonomias distintas quanto ao agrupamento de instrumentos.

Assim, os agrupamentos contendo 2 e 4 famílias de instrumentos, SC (sopro e cordas) e MFPC (metais, flautas, palhetas e cordas), foram utilizados para o banco de dados MIS, enquanto que os agrupamentos contendo 3, 4 e 5 famílias, SPC (sopro, percussão e cordas), MFPC e MFPPC (metais, flautas, palhetas, percussão e cordas) foram utilizados para o Banco de dados RWC e MUMS.

Além disso, foram criados 3 agrupamentos distintos de instrumentos (denominados INSTRUMENTOS) para cada base de dados e um agrupamento contendo instrumentos das 3 bases de dados.

Resumindo, foram criados dez agrupamentos para avaliar o desempenho do sistema. Seis deles representam taxonomias aplicadas à família de instrumentos e os outros quatro representam taxonomias aplicadas aos instrumentos individualmente.

Nas Tabelas 6.1, 6.2 e 6.3, apresentamos a descrição dos agrupamentos hierárquicos utilizados nesse trabalho para cada uma das 3 bases de dados, seguida da Tabela 6.4 que contém a descrição do agrupamento hierárquico reunindo as 3 bases de dados.

Família	Notas	Família	Notas	Código	Instrumento	Notas
Sopro	1622	Metais	326	2	Trombone Baixo	131
				4	Trombone Tenor	99
				11	Trompa	96
		Flautas	428	5	Flauta Contralto	99
				6	Flauta Baixo	102
				7	Flauta	227
		Palhetas	868	1	Saxofone	192
				3	Saxofone Soprano	192
				8	Fagote	122
				9	Clarinete Bb	139
				10	Clarinete Eb	119
				12	Oboé	104
Cordas	1269	Cordas	1269	13	Violoncelo	668
				14	Violino	601
Total	2891	Total	2891		Total	2891

Tabela 6.1: Tabela contendo os agrupamentos usados para a base de dados MIS.

Família	Notas	Família	Notas	Código	Instrumento	Notas
Sopro	1594	Metais	486	1	Trompa	74
				2	Trombone	148
				3	Trompete	200
				4	Tuba	64
		Flautas	412	5	Fauta Contralto	96
				6	Flauta Baixo	110
				7	Flauta	206
		Palhetas	696	8	Fagote	64
				9	Contrafagote	64
				10	Clarinete Bb	74
				11	Clarinete Baixo	50
				12	Clarinete	50
				13	Clarinete Eb	64
				14	Oboé	64
				15	Saxofone Contralto	64
				16	Saxofone Barítono	72
				17	Saxofone Soprano	64
				18	Saxofone Tenor	66
Percussão	370	Percussão	370	19	Glockenspiel	60
				20	Vibrafone	222
				21	Xilofone	88
Cordas	614	Cordas	614	22	Violoncelo	236
				23	Violino	378
Total	2578	Total	2578		Total	2578

Tabela 6.2: Tabela contendo os agrupamentos usados para a base de dados MUMS.

6.2 Abordagens Multiclasse

Em princípio existem duas abordagens normalmente empregadas para a multiclasse no uso de discriminantes lineares, um-contra-um (*one-against-one*) e um-

Família	Notas	Família	Notas	Código	Instrumento	Notas
Sopro	3191	Metais	985	1	Trompa	229
				2	Trombone	272
				3	Trompete	304
				4	Tuba	180
		Flautas	367	5	Flauta	367
		Palhetas	1839	6	Fagote	351
				7	Oboé	307
				8	Saxofone Contralto	297
				9	Saxofone Barítono	299
				10	Saxofone Soprano	297
				11	Saxofone Tenor	288
Percussão	906	Percussão	906	12	Glockenspiel	200
				13	Vibrafone	442
				14	Xilofone	264
Cordas	2448	Cordas	2448	15	Violoncelo	832
				16	Contrabaixo	791
				17	Violino	825
Total	6545	Total	6545		Total	6545

Tabela 6.3: Tabela contendo os agrupamentos usados para a base de dados RWC.

Família	Código	Instrumento	MIS	MUMS	RWC	Total
Metais	1	Trompa	96	74	229	399
	2	Trombone	230	148	272	650
	3	Trompete		200	304	504
	4	Tuba		64	180	244
Flauta	5	Flauta Contralto	99	96		195
	6	Flauta Baixo	102	110		212
	7	Flauta	227	206	367	800
Palhetas	8	Fagote	122	64	351	537
	9	Saxofone Contralto	192	64	297	553
	10	Saxofone Tenor		66	288	354
	11	Saxofone Soprano	192	64	297	553
	12	Saxofone Barítono		72	299	371
	13	Clarinete Bb	139	74		213
	14	Clarinete Eb	119	64		183
	15	Oboé	104	64	307	475
Percussão	16	Glockenspiel		60	200	260
	17	Xilofone		88	264	352
	18	Vibrafone		222	442	664
Cordas	19	Violoncelo	668	236	832	1736
	20	Violino	601	378	825	1804
			2891	2414	5754	11059

Tabela 6.4: Agrupamento usado combinando as 3 bases de dados.

contra-todos (*one-against-all*). Cada uma delas apresenta particularidades e podemos dizer que uma segue uma filosofia hierarquizada e a outra uma filosofia direta, portanto não-hierarquizada.

1. Um-contra-todos: nesse caso (representado pela Figura 6.1) o procedimento de generalização do problema de discriminação de 2 classes para o problema de discriminação multiclasse é resolvido através de um processo de decisão em cadeia, como uma árvore binária, onde cada nó representa a decisão de separar uma classe específica contra o restante. Caso a decisão seja a favor da classe específica, a amostra em teste é classificada como sendo desta classe, e o processo decisório de classificação para aquela amostra se encerra. No entanto, caso a decisão seja contrária à classe específica, o processo decisório prossegue e outra classe específica é testada contra o restante, excluindo-se para esse conjunto (dito restante) todas as classes específicas que o processo de decisão já testou. Nesse contexto, a raiz representa todos os instrumentos, e as folhas representam as classes finais a serem identificadas, ou instrumentos caso venha-se querer identificar ao nível de instrumentos. Os nós representam as decisões entre uma classe final e o conjunto restante. Esta abordagem apresenta o inconveniente de acumular o erro ocorrido na decisão do nó predecessor.

2. Um-contra-um: nesse caso, a generalização é obtida por meio de P discriminantes, onde P representa todas as duplas possíveis, a partir do total de classes que estão sendo avaliadas (vide Figura 6.2). A amostra é testada em todos os P discriminantes, e posteriormente é contabilizada a classe que foi mais "votada" para aquela amostra. A amostra é classificada como sendo a classe que recebeu mais votos. Portanto, nesse caso procura-se identificar diretamente todas as classes de instrumentos (folhas). Nesse contexto, dada uma amostra ela será identificada como correspondente à classe que apresentar a maior probabilidade. Normalmente essa abordagem costuma apresentar uma taxa média de acerto global maior que a abordagem anterior, conforme podemos constatar no trabalho de Eronem [46]. Uma das justificativas para que isso ocorra é que, ao utilizarmos a abordagem anterior, caso existam instrumentos com taxas de acerto baixas, estes irão contaminar a solução obtida para o treinamento de suas respectivas famílias (nó pai), fazendo com que o número de instrumentos

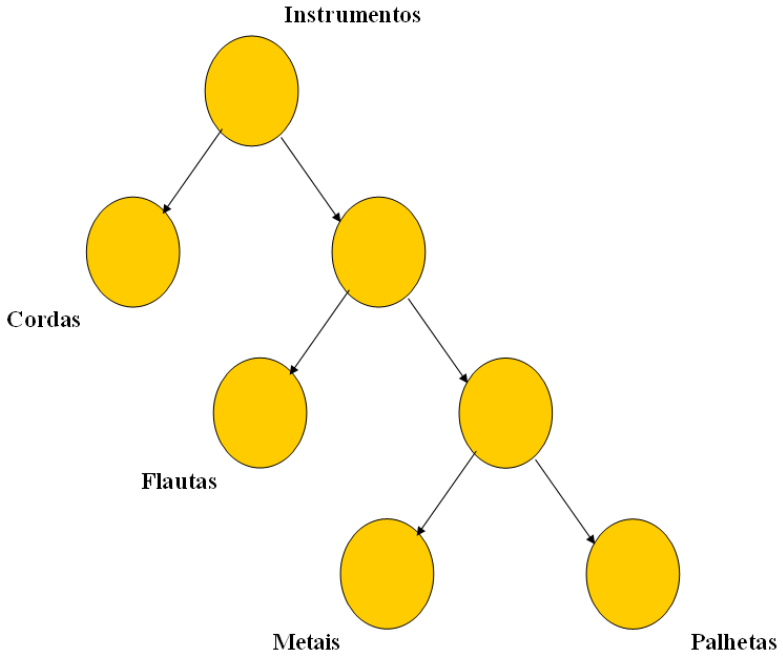

Figura 6.1: Árvore binária - representação um-contra-todos.

que venham a ser classificados para outra família (nó irmão) aumente, e assim o erro se propague. A princípio, a desvantagem do método sem hierarquização é a maior dimensionalidade que se exigirá no vetor de características para que todas as particularidades de todos instrumentos sejam igualmente apreciadas, aumentando assim a complexidade computacional e dificultando a obtenção da solução do classificador na fase de treinamento.

6.3 Estratégias

Nesta seção iremos abordar três estratégias para o sistema de reconhecimento automático com o intuito de se obter as classes pretendidas, ou seja: modelo padrão, modelo hierárquico e um modelo de reagrupamento. Cada uma das formas apresentadas nesta seção apresenta singularidades; não há nessa dissertação a pretensão de se fazer uma análise aprofundada de cada uma dessas formas a fim de se determinar qual delas é a melhor na maioria dos casos. No entanto, para algumas situações específicas, uma determinada estratégia poderá ter uma desempenho superior (taxa de

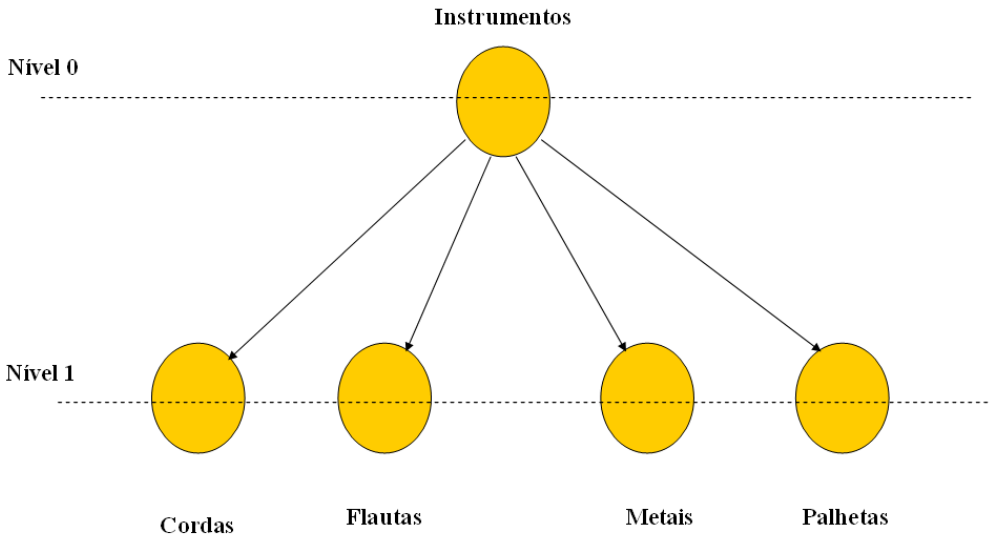

Figura 6.2: Árvore contendo as classes de interesse - representação um-contra-um.

acerto global) em relação às demais. Esse resultado é válido para um dado conjunto de treinamento e teste, e limitado a um número finito de classificadores e formas de codificação investigadas.

Em alguns casos, a taxa de acerto global pode ser afetada por suas amostras ao menos por 2 motivos:

- **problemas de contaminação**: ocasionados por existirem algumas amostras mal-posicionadas no espaço de características, ou seja, a amostra que se pretende classificar pertence a uma classe distinta da classe majoritária das amostras da sua vizinhança. Este problema pode ocorrer por medidas mal efetuadas na formação da base de dados.

- **problemas de confusão**: ocasionados por classes mal separáveis no espaço de características, ou seja, na vizinhança da amostra a ser classificada existem amostras tanto de sua classe quanto de outras classes em proporções equivalentes. Esse problema, na maioria dos casos nesse trabalho, ocorre com instrumentos de uma mesma família.

6.3.1 Estratégia 1 - Modelo Padrão

A primeira estratégia, ilustrada na Figura 6.3 e conhecida como modelo padrão, é uma abordagem direta e amplamente utilizada em diversos trabalhos de reconhecimento automático de padrões, conforme pode-se encontrar nas referências [2, 46] A estratégia 1 é composta pelos seguintes módulos:

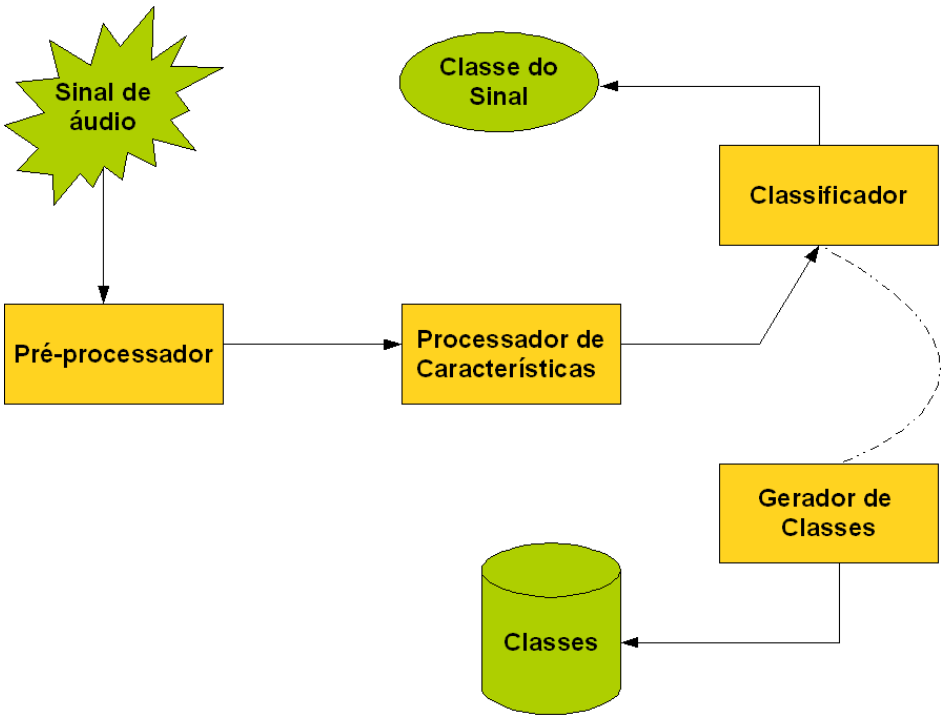

Figura 6.3: Sistema padrão de classificação.

1. Pré-processador;
2. Processador de Características;
3. Gerador de Classes;
4. Classificador de Instrumentos.

O classificador de Instrumentos pode usar qualquer uma das duas abordagens multiclasse já citadas anteriormente.

6.3.2 Estratégia 2 - Modelo Hierárquico

Uma segunda estratégia (hierárquica), proposta nesta dissertação, difere da primeira estratégia porque, ao invés de obter diretamente a taxa de acerto ao classificar as amostras em instrumentos musicais, utiliza uma abordagem indireta, ou

seja, obtém primeiro as classificações para as famílias de instrumentos, para depois obter as classificações para os instrumentos (com um classificador especializado). Essa estratégia tem sua arquitetura esboçada na Figura 6.4 e está exemplificada na Figura 6.5.

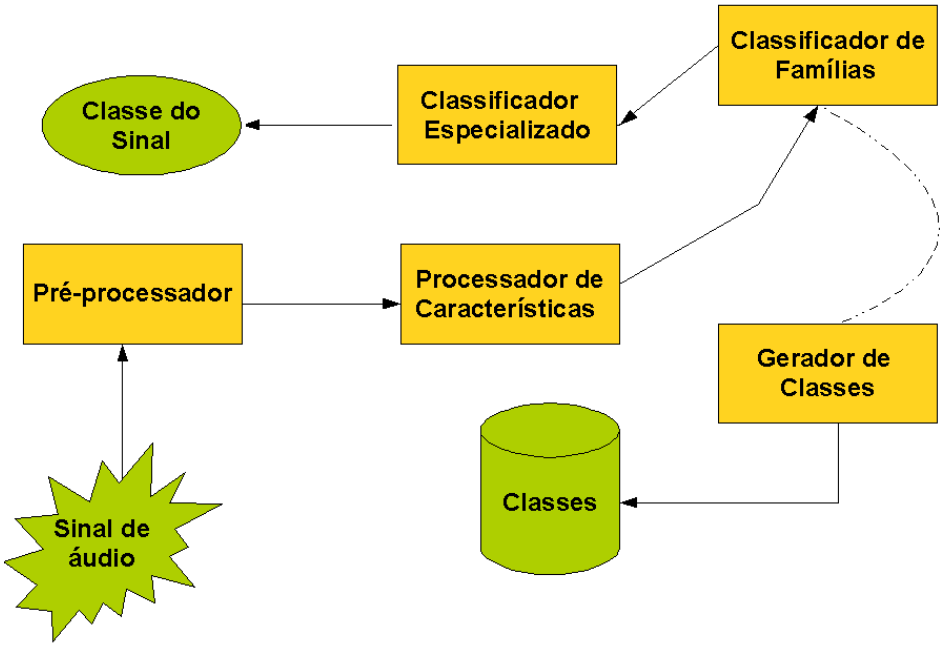

Figura 6.4: Estratégia 2.

A estratégia 2 apresenta 5 módulos independentes:

1. Pré-processador;
2. Processador de Características;
3. Gerador de Classes;
4. Classificador de Famílias;
5. Classificadores Especializados de Instrumentos por Famílias.

Assim, uma característica interessante desse método é que pode-se pensar em aplicar métodos de classificação distintos e/ou vetor de características distintos para cada nó (família, grupo, etc).

Igualmente, nessa abordagem, é possível construir esquemas onde parte das amostras do instrumento ficam num ramo e o restante das amostras fica em outro ramo, sendo que ambos conjuntos de amostras pertençam ao mesmo instrumento. Um exemplo disso se encontra em [46] em que os instrumentos com vibrato são

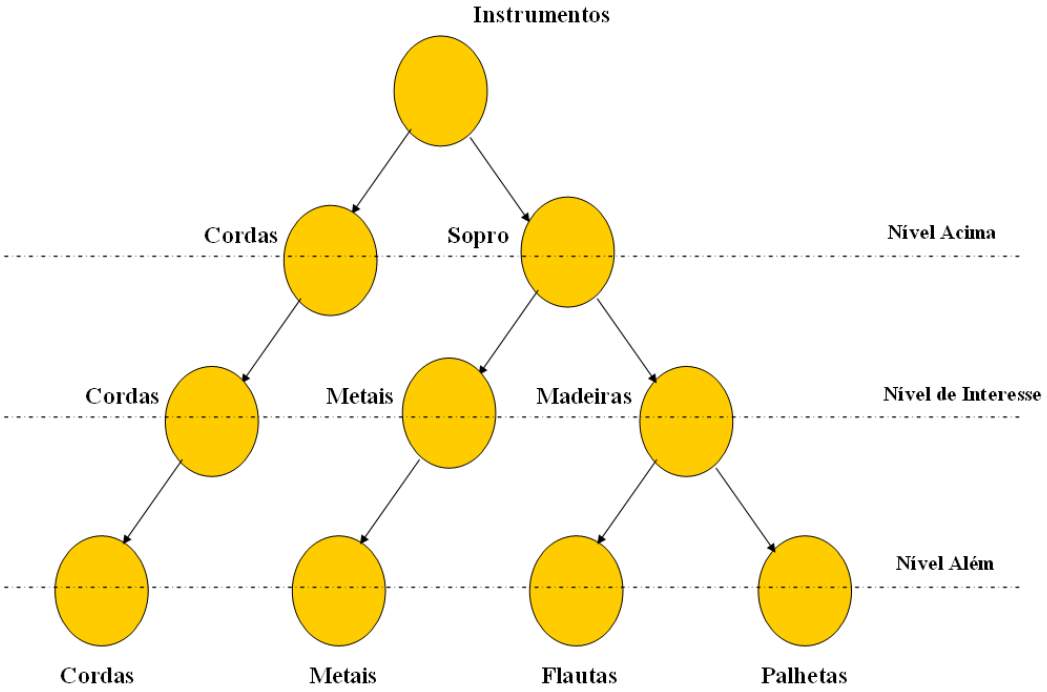

Figura 6.5: Agrupamento hierárquico de famílias de instrumentos musicais.

separados dos instrumentos sem vibrato, antes de se identificar o instrumento. Assim, em ambos os ramos após essa separação aparecerão amostras pertencendo ao mesmo instrumento. Dessa forma, não é necessário que amostras que tenham características distintas fiquem agrupadas na mesma classe. Essa possibilidade permite que se construa classes (famílias) artificiais a partir de métodos de clusterização, visando agrupar as amostras que possuam características comuns. Isso facilitará a discriminação das classes alterando a taxa de acerto na classificação.

Muitas vezes existem várias possibilidades de agrupar as amostras em famílias disjuntas, cada qual formada por um conjunto distinto de instrumentos musicais, de forma que existem diversas estruturas hierárquicas (árvores) cujas folhas consistem nos instrumentos que serão classificados. A escolha da melhor árvore (estrutura hierárquica) a ser usada pode fazer parte do problema do modelo hierárquico, uma vez que se pode agrupar os instrumentos em famílias (nós) de diversas maneiras.

Pelos motivos enumerados acima, essa abordagem pode obter excelentes taxas de acerto, podendo superar o modelo padrão que utiliza uma abordagem direta não-hierarquizada, desde que se permita nesse modelo flexibilizar o classificador e o vetor

de características empregado em cada nó.

6.3.3 Estratégia 3 - Modelo de Reagrupamento - Nível Além

Uma terceira estratégia (método 3), mais simples, consiste em classificar pelo método 1 as amostras em grupos formados em um nível além do nível de interesse (aqui denominado como subtipos de instrumentos), para depois reagrupá-los ao nível de interesse (instrumentos), conforme pode ser visto na Figura 6.6. Essa abordagem exige que as classes no nível de interesse possam ser subdivididas em classes mais atomizadas[1], o que normalmente é possível, bastando que existam amostras suficientes para isso.

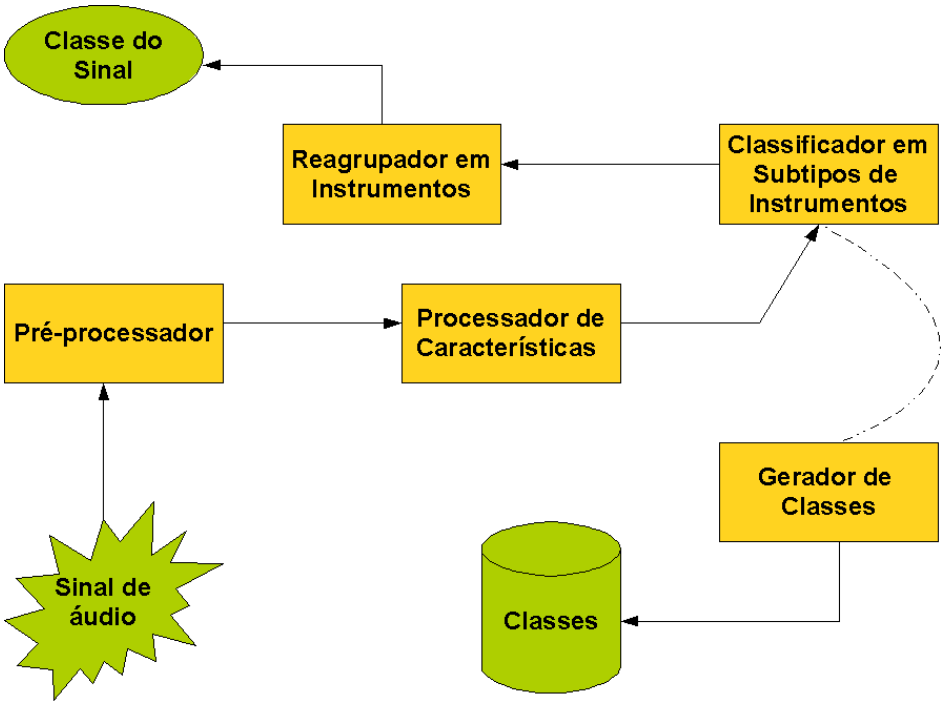

Figura 6.6: Estratégia 3.

A estratégia 3 apresenta os seguintes módulos:

1. Pré-processador;
2. Processador de Características;
3. Gerador de Classes;
4. Classificador de Subtipos de Instrumentos;
5. Reagrupador em Instrumentos.

[1]Pode se pensar em usar para cada classe um clusterizador, de forma que cada classe será atomizada em subclasses.

O reagrupamento em família difere de um treinamento direto para classificar as famílias, porque os hiperplanos separadores das famílias idealmente serão soluções ótimas para o espaço das características, o que o reagrupamento não necessariamente irá seguir. Assim, uma transformação do espaço de entrada para o espaço das características que não permita uma solução (com o uso de um hiperplano separador) capaz de obter uma separação de 100% das amostras das classes distintas (na fase de treinamento), pode em tese, com o uso do reagrupamento, obter uma curva de separação (formada por vários hiperplanos) capaz de distinguir 100% das amostras das classes distintas oferecidas na fase de treinamento, conforme pode-se observar na Figura 6.7.

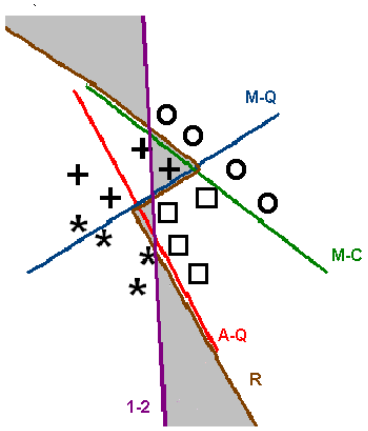

Figura 6.7: Diferença entre a classificação direta e com reagrupamento com uso de hiperplanos separadores.

Na Figura 6.7, a classe 1 é formada pelas subclasses "mais" e "asterisco", enquanto que a classe 2 é formada pelas subclasses "círculo" e "quadrado". O hiperplano separador da subclasse "asterisco" da subclasse "quadrado" é representado pela reta A-Q, enquanto que o hiperplano separador da subclasse "mais" da subclasse "círculo" e o hiperplano separador da subclasse "mais" da subclasse "quadrado" são representados respectivamente pelas retas M-C e M-Q. O hiperplano separador da subclasse "asterisco" da subclasse "círculo" não se encontra representado na Figura 6.7 porque o hiperplano representado pela reta A-Q satisfaz esta separação.

Pode-se observar na Figura 6.7 que não existe uma reta capaz de separar totalmente as amostras da classe 1 da classe 2. A reta 1-2 representa um hiperplano separador (ótimo) da classe 1 da classe 2 com erro mínimo. No entanto, a curva

obtida pelo reagrupamento dos hiperplanos separadores obtidos para as subclasses, representado pela curva "R", é capaz de separar 100% das amostras. Evidentemente o reagrupamento nem sempre representa uma solução melhor; o que poderá indicar qual separação é melhor nesse espaço de características será o erro obtido para as amostras teste nas regiões divergentes[2]. Assim, nesse exemplo, apesar da curva de separação obtida pelo reagrupamento dos hiperplanos separadores para as subclasses ser capaz de conseguir distinguir 100% das amostras das classes 1 e 2, não representa necessariamente uma separação melhor. Para que isso ocorra, basta existirem mais amostras teste da classe 2 do que da classe 1 nas regiões divergentes (cinzas).

Os erros de contaminação e confusão, nesse caso, se dão nas amostras dos subtipos de instrumentos (nível além), afetando a classificação nesse nível. Nesse caso, o erro de classificação pode ser carregado para o nível acima (o nível de interesse, instrumento), mesmo após o reagrupamento, caso o erro se dê entre subtipos de instrumentos que pertençam a instrumentos diferentes. Portanto, esses erros só serão evitados caso as amostras pertençam ao mesmo instrumento. Caso os erros ocorram entre amostras que pertençam a instrumentos distintos, pode-se pensar em redefinir os instrumentos, de forma que as amostras de subtipos de instrumentos distintos (em que ocorrem os erros) venham a pertencer ao mesmo grupo (instrumento), o que em princípio pode não ser sempre possível, face à existência de uma pré-definição das classes que se pretende classificar (instrumentos) ou o nível de confusão e contaminação que o conjunto de dados apresenta.

Assim, os erros causados entre subtipos de instrumentos distintos são eliminados quando reagrupamos os respectivos subtipos num mesmo grupo (instrumento). Da mesma forma podemos generalizar essa técnica a nível de instrumentos, com restrições, e determinar que um agrupamento hierárquico adequado pode maximizar a taxa de acerto global no nível acima (família), abrindo possibilidades de definirmos formas mais adequadas de agrupamentos de instrumentos visando a maximização da taxa de acerto (famílias), e não por convenções definidas pelas características dos instrumentos ou por modelos físicos de produção do som, conforme vimos no Capítulo 2.

Assim, essa abordagem na prática pode ter um melhor desempenho global

[2]Representada na Figura 6.7 pelas regiões preenchidas de cinza.

caso o algoritmo de posicionamento dos hiperplanos separadores (para os subtipos de instrumentos) venha a obter um posicionamento melhor para os hiperplanos separadores[3], quando comparado com o posicionamento obtido nas demais abordagens. Um dos fatores para que isso venha a ocorrer é uma parada antecipada forçada pelo algoritmo de posicionamento dos hiperplanos em função de uma convergência lenta ou outra condição satisfeita pelo critério de parada, gerando assim, uma solução ruim quando comparada àquela obtida sem essa antecipação.

Assim, os métodos 1, 2 e 3 não são equivalentes entre si, e eventualmente cada qual pode obter um desempenho global melhor que os demais, dependendo dos classificadores empregados, de como ocorre a distribuição das amostras nos espaço das características e do mapeamento do espaço de entrada para o espaço das características utilizado.

[3]Principalmente se o posicionamento do hiperplano for melhor para os instrumentos contendo mais amostras, supondo uma medida para a taxa de acerto global ponderada.

Parte III

Resultados

Capítulo 7

Construção de um Procedimento de Reconhecimento Automático

Este capítulo descreve a metodologia de busca utilizada para obter as soluções, a formação dos conjuntos de teste e treinamento, a avaliação do modelo multiclasse, a análise do desempenho da envoltória comparado ao da potência instantânea, a análise do segmento a ser utilizado, a obtenção do vetor de características, a avaliação dos classificadores e a obtenção de um método de reconhecimento automático para um dado agrupamento musical[1].

Nesse trabalho o método de reconhecimento automático para um dado agrupamento dentro de uma dada base de dados representa uma descrição dos elementos que compõem o vetor de características, combinados a um classificador limitado a um subconjunto de possibilidades pesquisadas. Assim, o conjunto descritor [vetor de características, classificador] [2] para um dado agrupamento será indistintamente chamado de solução.

7.1 Metodologia de Busca para obter as Soluções

Devido ao fato da análise ser extremamente complexa admitindo uma miríade de combinações, é de bom tom que avaliemos o custo computacional para pesquisar

[1]Podendo ser o próprio instrumento.

[2]Inseridos nesse contexto: as transformações sobre o sinal, sobre o vetor de características e sobre o classificador.

minimamente as principais combinações possíveis frente ao leque de possibilidades que essa dissertação apresenta na busca da solução ótima[3]. Tal cálculo pode ser feito de maneira aproximada da seguinte forma: temos basicamente 2 formatos (envoltória da potência instantânea ou a potência instantânea), 4 codificadores a serem testados, 4 momentos, 5 segmentos intermediários do modelo IMF[4] + ataque + subida[5], 4 quantidades de coeficientes (8, 16, 24, 32), presença ou não de características temporais, presença ou não de descritores de áudio, 3 classificadores, 3 tipos de agrupamentos e 3 bases de dados, perfazendo um total de $2 \times 4 \times 4 \times 7 \times 4 \times 2 \times 2 \times 3 \times 3 \times 3 = 96768$ possibilidades. Levando em conta que cada classificação em MATLAB leva aproximadamente 5,7 minutos[6] (DLG e SVM $>$ 8 minutos, K-NN $<$ 1 minuto) teremos 774.144 minutos, ou, 551577,6/1440 dias $\sim$ 383 dias de simulação ininterrupta, ou seja, mais que 1 ano, e sem levar em consideração as variações sobre o *kernel*, métricas e transformações do espaço de entrada, o tempo de codificação dos dados e o pré-processamento. Assim, serão adotados alguns critérios para reduzir o espaço de soluções visando garantir uma maior viabilidade na busca das melhores soluções.

Neste capítulo serão apresentados indicativos para uma dada solução como a taxa de acerto global, que representa o número de acertos divididos pelo número de amostras testadas, e a taxa de acerto média, que representa a média das taxas de acertos entre cada classe testada (como se as classes tivessem sido testadas com o mesmo número de amostras), ou seja, nesse caso cada classe entra com igual peso no cálculo da taxa de acerto média. Arbitrariamente se adotará o seguinte critério: a taxa de acerto global (mais usual) será preferida, enquanto que a taxa de acerto média virá como critério de desempate, uma vez que não se pretende avaliar todas as possíveis combinações para os agrupamentos com maior número de classes.

Assim, o procedimento será feito através de uma heurística, e seguirá uma "sequência de etapas" para a obtenção das soluções, de forma que a decisão obtida

[3]Numa busca exaustiva, ou seja, para um espaço de soluções discreto.

[4]Variando-se os limiares.

[5]Os segmentos contendo a parte final da nota foram retirados desse estudo, devido ao fato de alguns testes preliminares com esse segmento não apresentaram bons resultados.

[6]Foi usado um processador DUAL CORE 3,0 Ghz com 1Gb de memória RAM.

em cada etapa tenderá a reduzir o número de possibilidades a serem investigadas na etapa posterior. Certamente esse procedimento excluirá várias combinações, seguindo aproximadamente um algoritmo "guloso", baseado nos indicativos da taxa de acerto global e da taxa de acerto média obtidas para o agrupamento MFPC (Metais, Flautas, Palhetas e Cordas).

Portanto, não se espera que as melhores soluções encontradas para cada agrupamento pesquisado, sejam as soluções ótimas nos espaços de soluções pesquisados. No entanto, algumas (ou quiçá todas as soluções "sub-ótimas" encontradas) podem eventualmente ser as soluções ótimas para seus respectivos espaços de soluções.

Num primeiro momento, cada uma das bases de dados descritas sumariamente no apêndice A serão tratadas de forma independente, ou seja, cada qual terá um conjunto de instrumentos diferentes, mas contendo um conjunto menor de instrumentos comuns (aos pares). Esse conjunto menor de instrumentos comuns será usado no final, com o emprego de todas as bases de dados, para avaliarmos melhor a capacidade de generalização do método obtido.

Nas avaliações das primeiras etapas as estimativas serão obtidas a partir de configurações básicas. Essas configurações básicas serão denominadas sistemas de referência[7].

1. sistema de referência A - Composto por um vetor de características contendo 24 coeficientes LSF mais o desvio do segmento, conforme resultados obtidos em [33], extraídos do segmento associado intermediário proposto pelo modelo IMF com limiares de 10% e 90%, descritos no Capítulo 3. Foi usado um classificador 1-NN com uma métrica de distância euclideana, associada à estratégia definida pelo modelo padrão, com uma distribuição dos dados de 10% das amostras para teste e 90% das amostras para treinamento;

2. sistema de referência B - Igual ao sistema de referência A, exceto pelo classificador usado (SVM[8] com um *kernel* gaussiano de desvio padrão igual 1).

[7]Essa configuração básica não é tão trivial, mas foi decorrente de um estudo anterior presente em [33]. Assim, a taxa de acerto alta inicialmente se deve à presença dos codificadores.

[8]Para todos os casos que esse classificador (SVM) foi usado nessa dissertação a constante C presente na Equação 5.25 assumiu o valor infinito.

Nas etapas iniciais iremos definir qual das 3 bases de dados será empregada nas análises que determinarão a "melhor" solução. As demais bases de dados serão novamente utilizadas somente na avaliação da taxa de acerto de seus respectivos agrupamentos (após a obtenção de um conjunto de soluções mais prováveis) e ao final, na avaliação da capacidade de generalização do método, quando então misturaremos as bases de dados.

Devido ao fato de estarmos usando nas análises um dos sistemas anteriormente definidos, importa ressaltar que sempre teremos resultados parcialmente verdadeiros, embora esses resultados sejam usados como a melhor escolha no tocante à redução do espaço de soluções, o que não impede que esses resultados obtidos possam vir a ser verdadeiros para além do escopo no qual eles se mostraram verdadeiros (conforme as simulações realizadas).

7.2 Formação dos Conjuntos de Teste e Treinamento

Salvo informação expressamente contrária, todas as simulações usarão 10% das amostras de cada banco de dados para teste e 90% das amostras para treinamento. Foram usados 2 classificadores usualmente empregados em diversos trabalhos na área (K-NN e SVM), os quais também servirão de controle para a taxa de acerto obtida para a implementação elaborada nessa dissertação em DLG. Assim, pode-se avaliar se os percentuais obtidos por esse classificador (DLG) estão próximos ou não dos percentuais atingidos pelos outros classificadores.

Todas as amostras foram organizadas a partir do seguinte procedimento: os instrumentos foram subdivididos nas suas respectivas variações (*pizzicato*, *vibrato*, *tremolo*, *staccato*, etc...), inclusive para as suas variações no nível dinâmico (*Forte*, *Mezzo* e *Piano*). A partir de cada agrupamento resultante das subdivisões anteriores, as notas foram ordenadas segundo a tessitura do instrumento (em ordem crescente de *pitch*).

Para teste foram sorteadas amostras na região central da tessitura de cada instrumento, de forma que os 10% iniciais e os 10% finais da tessitura foram excluídos, restando, portanto, 80% da região central. Dessa região central sortearam-se

para cada subdivisão 10% das amostras para formarem o conjunto de teste[9]. As amostras restantes foram usadas para treinamento. Tal critério foi adotado tendo em vista que a região central da tessitura de cada instrumento é a região para a qual o instrumento musical foi melhor projetado, e portanto onde se espera encontrar seu maior uso. Assim, avaliar o desempenho com amostras sorteadas dessa região central representa avaliarmos o desempenho do classificador onde se espera ocorrer uma maior frequência das notas numa distribuição real.

7.3 Avaliação do Modelo Multiclasse

Nesta seção faremos uma breve simulação usando o sistema de referência B para avaliar qual modelo multiclasse que será empregado no resto desta dissertação[10].

Certamente o resultado obtido por algumas simulações aqui apresentadas não permite que se afirme categoricamente qual método é o melhor, e em que circunstâncias isso ocorre. No entanto, pretende-se conseguir um indicativo do modelo multiclasse que é mais adequado ao problema proposto nessa dissertação.

As avaliações feitas nessa seção contemplarão somente o agrupamento MFPC. Na Tabela 7.1 apresentamos os resultados obtidos para cada um dos métodos multiclasse descritos na Seção 6.2, considerando todas as bases de dados.

B.D.	Multiclasse	Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
MIS	1-Todos	96,7%	100,0%	59,5%	61,9%	70,56%	77,60%	79,52%
MIS	1-um	96,7%	94,1%	94,9%	84,8%	90,73%	100,00%	92,62%
MUMS	1-Todos	97,8%	84,2%	27,3%	15,5%	49,76%	55,92%	56,20%
MUMS	1-um	93,3%	89,5%	92,4%	91,4%	91,79%	100,00%	91,65%
RWC	1-Todos	98,9%	93,3%	38,9%	62,9%	63,28%	65,01%	73,50%
RWC	1-um	100,0%	100,0%	90,7%	91,4%	93,16%	100,00%	95,53%

Tabela 7.1: Tabela contendo os resultados das abordagens multiclasse nas 3 bases de dados usando o sistema de referência B.

Conforme pode-se observar, o modelo multiclasse um-contra-um obteve re-

[9]Foram feitos 3 sorteios para cada base de dados, e optou-se por aquele conjunto de amostras que apresentou uma taxa de acerto intermediária a partir do sistema de referência A. A maior variação obtida foi em torno de 3%.

[10]Não faz sentido fazer uma avaliação do modelo multiclasse para o classificador K-NN, por isso foi usado o classificador SVM.

sultados superiores para todas as bases de dados quando comparado ao desempenho do modelo um-contra-todos[11]. Assim, face aos resultados verificados, o modelo um-contra-um será adotado no restante dessa dissertação.

7.4 Análise do Desempenho da Envoltória versus Potência Instantânea

Nas avaliações feitas nessa seção foi usado o sistema de referência A associado ao classificador 1-NN para o codificador LSF, para as classes MFPC, MFPPC, SC e SPC, a depender das possibilidades em cada banco de dados.

Num primeiro momento, serão extraídos trechos do sinal a partir de sua envoltória pelo método da média e pelo algoritmo detector de envoltória, assim como diretamente da potência instantânea do sinal.

Serão avaliados 3 formatos: a potência instantânea (P_i), a envoltória da potência instantânea pelo algoritmo detector de envoltória (Ep_i), e a envoltória obtida pelo algoritmo da média RMS ($Eprms_i$), para cada uma das classes associada a cada base de dados.

Na Tabela 7.2 apresentamos as taxas de acerto para o banco de dados MIS em alguns agrupamentos num sistema de reconhecimento automático de instrumentos musicais.

Banco de dados MIS	Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Média
Detect. De Envoltória	53,33%	44,12%	55,70%	84,76%	66,13%	59,48%
Envoltória RMS	30,00%	11,76%	37,97%	64,76%	44,76%	36,12%
Potência do Sinal	96,67%	97,06%	83,54%	89,52%	**89,52%**	91,70%

Banco de dados MIS	Sopro	Cordas	Tx Global	Tx Média
Detect. De Envoltória	77,62%	84,76%	80,65%	81,19%
Envoltória RMS	64,34%	64,76%	64,52%	64,55%
Potência do Sinal	93,71%	89,52%	**91,94%**	91,62%

Tabela 7.2: Tabelas da base de dados MIS representando agrupamentos hierárquicos versus formato do sinal.

[11]Na implementação realizada para o modelo um-contra-todos verificou-se primeiro qual o desempenho de cada classe contra o restante; posteriormente ordenou-se esses resultados formando a decisão hierárquica na árvore binária.

Nas Tabelas 7.3 e 7.4 apresentamos as taxas de acerto para os bancos de dados MUMS e RWC em alguns agrupamentos hierárquicos usando o mesmo sistema de reconhecimento automático empregado nas simulações que constam na Tabela 7.2.

Banco de dados MUMS	Metais	Flautas	Palhetas	Percussão	Cordas	Tx Global	Tx Média
Detect. De Envoltória	60,00%	39,47%	69,70%	60,00%	58,62%	59,09%	57,56%
Envoltória RMS	48,89%	34,21%	51,52%	37,14%	41,38%	43,80%	42,63%
Potência do Sinal	93,33%	97,37%	86,36%	88,57%	93,10%	**91,32%**	91,75%

Banco de dados MUMS	Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Média
Detect. De Envoltória	60,00%	47,37%	69,70%	67,24%	62,80%	61,08%
Envoltória RMS	48,89%	36,84%	57,58%	60,34%	52,66%	50,91%
Potência do Sinal	93,33%	97,37%	86,36%	93,10%	**91,79%**	92,54%

Banco de dados MUMS	Sopro	Percussão	Cordas	Tx Global	Tx Média
Detect. De Envoltória	91,95%	60,00%	58,62%	79,34%	70,19%
Envoltória RMS	77,18%	37,14%	41,38%	62,81%	51,90%
Potência do Sinal	96,64%	88,57%	93,10%	**94,63%**	92,77%

Tabela 7.3: Tabelas da base de dados MUMS representando agrupamentos hierárquicos versus formato do sinal.

Banco de dados RWC	Metais	Flautas	Palhetas	Percussão	Cordas	Tx Global	Tx Média
Detect. De Envoltória	54,55%	30,00%	55,56%	58,23%	61,21%	56,68%	51,91%
Envoltória RMS	25,00%	3,33%	48,15%	27,85%	51,72%	41,12%	31,21%
Potência do Sinal	98,86%	96,67%	87,65%	86,08%	94,40%	**92,22%**	92,73%

Banco de dados RWC	Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Média
Detect. De Envoltória	54,55%	33,33%	56,17%	65,09%	58,59%	52,29%
Envoltória RMS	25,00%	3,33%	50,00%	58,19%	46,68%	34,13%
Potência do Sinal	98,86%	96,67%	87,65%	94,40%	**93,16%**	94,40%

Banco de dados RWC	Sopro	Percussão	Cordas	Tx Global	Tx Média
Detect. De Envoltória	82,50%	58,23%	61,21%	70,90%	67,31%
Envoltória RMS	61,79%	27,85%	51,72%	53,30%	47,12%
Potência do Sinal	95,36%	86,08%	94,40%	**93,74%**	91,95%

Tabela 7.4: Tabelas da base de dados RWC representando agrupamentos hierárquicos versus formato do sinal.

Ao analisarmos as Tabelas 7.2, 7.3 e 7.4 podemos concluir, caso se queira usar a codificação LSF, que a potência instantânea obteve taxas de acerto muito superiores às das envoltórias, seja pelo método DEAM, seja pelo método da Média (RMS). Isto foi verdade para todos os agrupamentos e para todas as bases de dados testados. Assim, para o restante da dissertação não se avaliará o formato das

envoltórias para se obter as características via codificadores[12].

Pode-se inferir uma possível explicação para esse comportamento[13], como oriundo da suavização da curva ao obtermos a envoltória do sinal. Dessa forma, perdemos informação útil para a codificação. Corroborando este raciocínio, temos que, quanto maior for a suavização, maior será a perda. E para constatar tal afirmação basta verificar nas Tabelas 7.2, 7.3 e 7.4, onde o método da média obteve taxas de acertos inferiores ao método DEAM, da mesma forma o DEAM obteve um desempenho inferior em relação ao método da potência instantânea, em todos os agrupamentos e para todas as bases de dados. Uma outra observação que pode ser feita é que as taxas obtidas para o formato potência instantânea nas 3 bases de dados variou menos de 4% para o agrupamento MFPC, indicando que o método usado possui capacidade de generalização para essas bases de dados.

A partir desta seção iremos privilegiar a base de dados MIS, porque ela apresentou a menor taxa de acerto para a classe MFPC. Assim, espera-se que uma configuração que angarie uma melhoria na sua taxa de acerto para um dado agrupamento implique também em melhorias nas demais bases de dados no agrupamento correspondente.

Também iremos privilegiar o agrupamento MFPC, a fim de evitarmos uma explosão combinatorial de possibilidades[14] quando usarmos os classificadores DLG ou SVM. Ao final, para as soluções que apresentarem melhores resultados, avaliaremos os demais agrupamentos em suas diversas bases de dados.

---

[12]Esse resultado foi também verificado para os codificadores LPC e MFCC em todos os conjuntos de amostras testados.

[13]Esse comportamento teve sua comprovação parcial feita no classificador 1-NN.

[14]Uma vez que a classe MFPC possui somente 4 classes, o número de máquinas classificadoras a serem executadas fica reduzido. Para cada vetor de características uma solução multiclasse (um-contra-um) contendo n classes através de hiperplanos separadores usa $\frac{n \times n-1}{2}$ classificadores para obter a solução.

7.5 Análise dos Segmentos

A ideia por trás da análise dos segmentos é determinar se existe algum segmento privilegiado[15], ou seja, que concentra mais informação com capacidade de discriminação para as famílias de instrumentos musicais ou para os próprios instrumentos musicais. Se a resposta for positiva, iremos nos concentrar em extrair as características das notas concentradas somente nesse(s) segmento(s) privilegiado(s), ao invés de tentarmos extrair características sobre toda a nota. Os segmentos pesquisados serão o ataque, a subida e o segmento estacionário (intermediário) obtido pelo modelo IMF. O segmento estacionário do modelo ADSR, ou seja, o segmento de sustentação, foi posto de lado nessas avaliações pelo fato de este segmento não estar presente em todas as notas, para todos os instrumentos. Para essas simulações usaremos o sistema de referência A[16].

Na Tabela 7.5 apresentamos os resultados obtidos para o segmento intermediário do modelo IMF usando o sistema de reconhecimento automático de instrumentos musicais descrito na seção anterior para a base de dados MIS. Variou-se o limiar superior para os valores 10%, 30%, 50%, 70% e 90% na expectativa de encontramos o limiar que define o segmento mais significativo para a discriminação das classes. Avaliaram-se os codificadores LSF, MFCC e LPC, para verificar uma possível persistência na escolha do limiar.

MIS	10%	30%	50%	70%	90%
LSF	88,7%	89,5%	89,5%	89,1%	89,5%
MFCC	96,4%	95,6%	96,4%	96,4%	96,4%
LPC	86,7%	84,3%	81,5%	82,7%	83,1%

Tabela 7.5: Tabela para a base de dados MIS contendo a taxa de acerto para o segmento intermediário para o modelo IMF.

A Tabela 7.6 apresenta as mesmas simulações para a base de dados RWC.

Conforme se pode observar ao verificarmos os percentuais obtidos para cada

[15]Esse privilégio logicamente é função da forma de extração de características e do classificador empregado.

[16]Dessa forma, o resultado dessa seção foi avaliado somente para o classificador 1-NN.

RWC	10%	30%	50%	70%	90%
LSF	90,2%	91,6%	92,4%	92,2%	93,2%
MFCC	95,7%	95,7%	96,3%	96,3%	96,3%
LPC	75,4%	75,8%	77,0%	75,2%	75,2%

Tabela 7.6: Tabela para a base de dados RWC contendo a taxa de acerto para o segmento intermediário para o modelo IMF.

codificador, não existe um segmento que se destaque, ficando as taxas de acerto sempre na mesma ordem de grandeza, independentemente do tamanho do segmento intermediário. No entanto, esse resultado é significativo, uma vez que o segmento intermediário com limiar de 90% é menor que o segmento que seria obtido caso o limiar fosse 10%. Portanto, é preferível codificar um segmento menor, obtendo taxas de acerto equivalentes. Assim, visando reduzir o número de possibilidades a serem investigadas, iremos daqui em diante sempre adotar o limiar de 90% como o limiar superior para o modelo IMF.

Além dos segmentos de ataque e de subida foi avaliado o segmento que representa os primeiros 23,2 ms[17]. Na Tabela 7.7 apresentamos os resultados obtidos na base de dados MIS para os segmentos de ataque (modelo ADSR), subida (modelo IMF)[18] e no segmento que representa os 23,2 ms iniciais.

		Subida				
MIS	(23,2ms)	10-30	10-50	10-70	10-90	Ataque
LSF s/ DCT	76,6%	77,4%	77,8%	78,6%	77,4%	82,3%
LSF c/ DCT	69,4%	67,3%	67,3%	70,6%	66,1%	59,7%
MFCC s/ DCT	89,9%	93,2%	93,2%	91,1%	90,3%	92,7%
MFCC c/ DCT	68,2%	65,7%	70,6%	66,1%	69,4%	64,1%
LPC s/ DCT	61,3%	63,3%	62,5%	66,2%	63,5%	71,1%
LPC c/ DCT	73,0%	68,6%	66,9%	66,5%	64,9%	66,1%

Tabela 7.7: Tabela para a base de dados MIS contendo a taxa de acerto nos segmentos iniciais.

[17]Para todas as notas considerou-se o limiar de 10% para a detecção da nota.

[18]Nos instrumentos em que os segmentos de ataque e subida apresentaram menos que 1024 amostras, o segmento foi redimensionado para ter 1024 amostras, que corresponde à 23,2 ms.

O mesmo procedimento feito para montar a Tabela 7.7 foi feito para a base de dados RWC apresentado na Tabela 7.8.

RWC	(23,2ms)	Subida				Ataque
		10-30	10-50	10-70	10-90	
LSF s/ DCT	73,1%	74,0%	73,6%	74,8%	75,2%	81,1%
LSF c/ DCT	64,8%	65,8%	64,3%	62,5%	62,3%	55,1%
MFCC s/ DCT	89,1%	90,2%	90,0%	88,9%	88,7%	92,0%
MFCC c/ DCT	65,8%	63,7%	63,7%	65,2%	64,5%	55,7%
LPC s/ DCT	61,3%	63,3%	62,5%	66,2%	63,5%	71,1%
LPC c/ DCT	67,8%	64,8%	70,5%	62,7%	63,5%	59,8%

Tabela 7.8: Tabela para a base de dados RWC contendo a taxa de acerto nos segmentos iniciais.

Conforme se pode observar, o uso da DCT no segmento de ataque piorou a taxa de acerto. O mesmo foi verdade para os codificadores LSF e MFCC. No entanto, para o codificador LPC, o uso da DCT foi praticamente indiferente, apresentando uma pequena melhora. Para os segmentos iniciais, os segmentos de ataque sem o uso da DCT foram os que apresentaram os melhores resultados, caso combinados com o codificador LSF ou MFCC. No entanto, até mesmo esses segmentos obtiveram taxa de acerto inferior à dos segmentos intermediários quando codificados pelos seus respectivos codificadores em suas respectivas bases de dados. De outra forma, o pior resultado obtido pelos segmentos intermediários foi melhor que o melhor resultado obtido pelos segmentos iniciais, obviamente respeitando o codificador e o banco de dados.

Portanto, nessa dissertação iremos, a partir desse ponto, trabalhar somente com o segmento intermediário proposto pelo modelo IMF. No entanto, os resultados do segmento de ataque (Modelo ADSR) para o codificador MFCC foram significativos[19], embora inferiores aos resultados obtidos no segmento intermediário. Por esse motivo esse trecho (ataque) foi colocado de lado[20].

[19] Os resultados desse trecho para o codificador LSF foram considerados razoáveis quando comparados aos resultados obtidos no segmento intermediário ou no segmento de ataque.

[20] A ideia é somente codificar um trecho que contenha capacidade de discriminar os instrumentos musicais.

7.6 Obtenção do Vetor de Características

Nesta seção serão avaliados os codificadores, os elementos estatísticos, perceptuais, e temporais a partir da potência instantânea do sinal. Os codificadores avaliados são o LSF, LPC, MFCC e CEPSTRUM, sendo que o número de coeficientes desses codificadores serão 8, 16, 24 e 32, obtidos a partir do segmento previamente escolhido da amostra na seção anterior, o qual já sofreu o escalonamento dinâmico.

Um fator que se deve destacar é que o codificador LSF é representado pelos ângulos dos números complexos que representam as raízes dos polinômios $P(z)$ e $Q(z)$ vistos nas Equações (4.9), (4.10) na Seção 4.4, portanto de natureza diversa dos coeficientes MFCC e LPC. Se por acaso estabelecermos que usaremos 16 ângulos LSFs, devemos levar em conta que toda solução complexa é um par conjugado, e sabendo que um par conjugado possui o mesmo ângulo (a menos do sinal), devemos então trabalhar com o dobro de soluções (32) para obtermos os 16 ângulos pretendidos. Logo, a ordem do polinômio para obtermos as soluções LSFs é o dobro da ordem dos polinômios para obtermos as soluções LPC e MFCC. Assim, comparações diretas entre a taxa de acerto obtida com uma solução usando LSFs e outra usando LPC ou MFCC (fixado o número de coeficientes) serão feitas nessa dissertação, apesar dessa diferença no grau do polinômio. Caso se queira compensar essa diferença, devemos avaliar as soluções com 32 coeficientes LPCs e MFCCs contra a solução contendo 16 coeficientes LSFs.

Nesta seção também serão avaliadas as características estatísticas de ordens 2, 3 e 4 do trecho da amostra, os descritores específicos de áudio tais como ZCR, RMS, SC, *flux*, BW, e as características temporais obtidas a partir da envoltória da potência instantânea do sinal, tais como largura do pulso, tempo de subida, tempo de descida, *droop*, etc.

A implementação do DLG empregada nessa dissertação utilizou dois outros critérios de parada além do erro a ser minimizado, um a partir do número máximo de iterações (6000) e outro que estima se variação do erro absoluto é menor que um número arbitrariamente pequeno. Ambos os critérios acrescentados prejudicam a otimização almejada, o que justifica em parte os resultados ligeiramente inferiores que foram obtidos. No entanto, tal procedimento evita problemas de regiões de convergência lenta, ou problemas de otimização sem solução, que ocorreram com

relativa frequência no caso da SVM.

7.6.1 Resultados dos Codificadores mais Desvio Padrão do Segmento

Uma vez definido o segmento (segmento intermediário do modelo IMF) que contém mais informação discriminante das famílias de instrumentos musicais (obtido a partir do limiar de 90%), se fará nessa seção uma análise mais detalhada do número mais adequado de coeficientes associados aos codificadores e quais codificadores que apresentam resultados mais significativos para a abordagem empregada.

Inicialmente iremos avaliar se todos os codificadores apresentam taxas de acerto similares. Caso algum fique com taxas de acerto bem abaixo dos demais, este será descartado. Da mesma forma, se algum ficar com taxas de acerto muito acima das taxas dos demais este será então preferido.

A partir dessa seção irá aparecer um outro indicativo, denominado Taxa de Treinamento ("Tx Treino"), que avalia o desempenho do treinamento. Logo, para os classificadores SVM e DLG, a solução obtida pode não ser capaz de discriminar 100% das amostras que foram usadas no treinamento. Essa informação indica se a transformação no espaço de características foi adequada ou se as amostras são facilmente separáveis. Na Tabela 7.9 apresentamos os resultados obtidos.

Coef.	Momento	Codificador	Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
8	std, std+m_3, std+m_3+m_4	LPC	83,3%	70,6%	70,9%	81,0%	76,6%	100,0%	76,4%
		CEPSTRUM	56,7%	35,3%	40,5%	61,9%	50,8%	100,0%	48,6%
		MFCC	96,7%	94,1%	92,4%	96,2%	94,8%	100,0%	94,8%
		LSF	96,7%	97,1%	82,3%	91,4%	89,9%	100,0%	91,9%
16	std, std+m_3, std+m_3+m_4	LPC	90,0%	85,3%	76,0%	81,9%	81,5%	100,0%	83,3%
		CEPSTRUM	63,3%	29,4%	62,0%	76,2%	63,7%	100,0%	57,7%
		MFCC	100,0%	100,0%	93,7%	99,1%	**97,6%**	100,0%	98,2%
		LSF	96,7%	97,1%	83,5%	93,3%	**91,1%**	100,0%	92,7%
24	std, std+m_3, std+m_3+m_4	LPC	93,3%	94,1%	74,7%	82,9%	83,1%	100,0%	86,2%
		CEPSTRUM	50,0%	35,3%	55,7%	76,2%	60,9%	100,0%	54,3%
		MFCC	100,0%	94,1%	93,7%	98,1%	96,4%	100,0%	96,5%
		LSF	96,7%	97,1%	83,5%	89,5%	89,5%	100,0%	91,7%
32	std, std+m_3, std+m_3+m_4	LPC	93,3%	94,1%	79,8%	81,0%	**83,9%**	100,0%	87,0%
		CEPSTRUM	63,3%	47,1%	67,1%	80,0%	**69,4%**	100,0%	64,4%
		MFCC	100,0%	94,1%	93,7%	99,1%	96,8%	100,0%	96,7%
		LSF	96,7%	97,1%	87,3%	85,7%	89,1%	100,0%	91,7%

Tabela 7.9: Estatística das taxas de acerto para todos os codificadores empregados nessa dissertação com o classificador 1-NN.

Pode-se observar na Tabela 7.9 que as soluções obtidas usando o codificador CEPSTRUM obtiveram desempenho bem abaixo dos demais codificadores, num patamar inferior a 69%. Portanto, o codificador CEPSTRUM não conseguiu obter um vetor de características com capacidade discriminatória (número de coeficientes igual a 8, 16, 24 e 32) usando o segmento intermediário do modelo IMF. Assim, se descartará o uso desse codificador para a pesquisa da melhor solução.

Podemos também observar que o codificador MFCC em média apresentou os melhores resultados, seguido do codificador LSF e LPC, em ordem decrescente de desempenho. Da mesma forma, optou-se por retirar das análises a codificação com 8 coeficientes, visto ter sido a que apresentou o pior desempenho para os codificadores LPC, CEPSTRUM e MFCC, e apesar de não ter sido a pior para o codificador LSF também não foi a que apresentou o melhor desempenho[21].

A seguir foram avaliadas as soluções para 16, 24 e 32 coeficientes do mesmo vetor de características para os classificadores SVM (*kernel* gaussiano) e DLG (transformação de potenciação de ordem 2) mostrados nas Tabelas 7.10 e 7.11

Coef.	Momento	Codif.	Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
16	std, std+m_3, std+m_3+m_4	LPC	93,3%	94,1%	77,2%	79,1%	82,3%	100,0%	85,9%
		MFCC	96,7%	91,2%	98,7%	98,1%	**97,2%**	100,0%	96,2%
		LSF	96,7%	94,1%	94,9%	91,4%	93,6%	100,0%	94,3%
24	std, std+m_3, std+m_3+m_4	LPC	93,3%	94,1%	79,8%	87,6%	**86,7%**	100,0%	88,7%
		MFCC	86,7%	79,4%	97,5%	99,1%	94,4%	100,0%	90,7%
		LSF	96,7%	94,1%	94,9%	84,8%	90,7%	100,0%	92,6%
32	std, std+m_3, std+m_3+m_4	LPC	96,7%	94,1%	77,2%	84,8%	85,1%	100,0%	88,2%
		MFCC	73,3%	73,5%	98,7%	99,1%	92,3%	100,0%	86,2%
		LSF	100,0%	100,0%	91,1%	91,4%	**93,6%**	100,0%	95,6%

Tabela 7.10: Estatísticas com as taxas de acerto para o classificador SVM (*kernel* gaussiano com desvio padrão unitário).

7.6.2 Resultados com as Características Temporais

Os descritores temporais usados foram os discutidos na Seção 4.1.

De todos os resultados apresentados nesta seção serão apresentados somente aqueles que apresentaram os melhores desempenhos para cada codificador[22], por-

[21] Esses resultados, onde aparece um pior desempenho no uso de 8 coeficientes, foram também observados em outras simulações envolvendo outros conjunto de amostras de teste.

[22] No entanto, para todos os casos aqui constantes foram feitas simulações que aparecem segundo

Coef.	Momento	Codif.	Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
16	std	LPC	70,0%	79,4%	73,4%	82,9%	77,8%	82,8%	76,4%
		MFCC	86,7%	91,2%	83,5%	94,3%	**89,5%**	91,0%	88,9%
		LSF	90,0%	88,2%	67,1%	87,6%	81,5%	88,2%	83,2%
	std+m_3	LPC	76,7%	82,4%	69,6%	83,8%	78,2%	83,6%	78,1%
		MFCC	86,7%	91,2%	84,8%	92,4%	89,1%	91,5%	88,8%
		LSF	90,0%	91,2%	69,6%	87,6%	82,7%	88,8%	84,6%
	std+m_3+m_4	LPC	76,7%	85,3%	68,4%	84,8%	78,6%	83,5%	78,8%
		MFCC	83,3%	94,1%	83,5%	92,4%	88,7%	91,8%	88,3%
		LSF	90,0%	91,2%	67,1%	86,7%	81,5%	89,3%	83,7%
24	std	LPC	83,3%	82,4%	70,9%	83,8%	79,4%	85,8%	80,1%
		MFCC	90,0%	88,2%	82,3%	90,5%	87,5%	93,7%	87,8%
		LSF	93,3%	94,1%	76,0%	83,8%	83,9%	90,5%	86,8%
	std+m_3	LPC	76,7%	79,4%	77,2%	83,8%	**80,2%**	86,7%	79,3%
		MFCC	90,0%	91,2%	82,3%	91,4%	88,3%	93,8%	88,7%
		LSF	90,0%	97,1%	76,0%	86,7%	85,1%	91,1%	87,4%
	std+m_3+m_4	LPC	80,0%	79,4%	70,9%	86,7%	79,8%	87,3%	79,2%
		MFCC	90,0%	91,2%	82,3%	91,4%	88,3%	93,7%	88,7%
		LSF	93,3%	97,1%	77,2%	84,8%	85,1%	90,8%	88,1%
32	std	LPC	73,3%	82,4%	65,8%	81,9%	75,8%	86,7%	75,9%
		MFCC	90,0%	94,1%	83,5%	90,5%	88,7%	**94,3%**	89,5%
		LSF	96,7%	94,1%	73,4%	87,6%	85,1%	91,7%	88,0%
	std+m_3	LPC	86,7%	85,3%	73,4%	80,0%	79,4%	87,7%	81,3%
		MFCC	86,7%	88,2%	87,3%	91,4%	89,1%	94,2%	88,4%
		LSF	93,3%	91,2%	74,7%	88,6%	85,1%	**91,8%**	86,9%
	std+m_3+m_4	LPC	90,0%	85,3%	69,6%	81,9%	79,4%	**88,1%**	81,7%
		MFCC	90,0%	88,2%	87,3%	90,5%	89,1%	94,0%	89,0%
		LSF	90,0%	91,2%	79,8%	88,6%	**86,3%**	91,5%	87,4%

Tabela 7.11: Estatísticas com as taxas de acerto para o classificador DLG (transformação de potenciação de ordem 2).

que caso apresentássemos todos os resultados, apareceriam soluções repetidas, de desempenho bem similar.

A fim de evitarmos problemas de discriminação da métrica de distância (1-NN) e de problemas de convergência com a SVM, foi feito um escalamento estatístico, usado somente para esses classificadores, ou seja, SVM e 1-NN, já que o DLG não teve problemas com a ausência do escalamento estatístico. Ao usarmos o escalamento estatístico, surge um problema que é a necessidade de determinar um escalamento para as amostras teste. A solução adotada foi escalar as amostras testes com as médias e desvios obtidos do conjunto de treinamento.

Nas Tabelas 7.12, 7.13, e 7.14 apresentamos os resultados para os classificadores DLG, 1-NN e SVM, respectivamente, acrescidos no seu vetor de características das características temporais.

---

o formato apresentado na Tabela 7.9.

Vetor de caracteristicas				Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
C.Temporais		std+m_3+m_4		30,0%	32,4%	76,0%	87,6%	69,4%	75,3%	56,5%
C.Temporais	16	LPC	std	80,0%	88,2%	83,5%	91,4%	87,1%	92,1%	85,8%
C.Temporais	24	MFCC	std+m_3+m_4	90,0%	91,2%	91,1%	96,2%	93,2%	96,3%	92,1%
		LSF	std+m_3+m_4	93,3%	91,2%	89,9%	93,3%	91,9%	96,0%	91,9%

Tabela 7.12: Melhores resultados para o classificador DLG (Potenciação de ordem 2) com características temporais.

Vetor de caracteristicas				Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
C.Temporais		std		33,3%	32,4%	60,8%	88,6%	65,3%	100,0%	53,8%
C.Temporais	16	LPC	std	90,0%	79,4%	79,8%	94,3%	87,1%	100,0%	85,9%
C.Temporais	24	MFCC	std	93,3%	94,1%	89,9%	95,2%	93,2%	100,0%	93,1%
	16	LSF	std	100,0%	94,1%	86,1%	98,1%	94,0%	100,0%	94,6%

Tabela 7.13: Tabela com os melhores resultados para o classificador 1-NN (métrica euclideana) com características temporais.

Os resultados obtidos através do classificador 1-NN (métrica euclideana, com o acréscimo dos descritores temporais sem aplicar o escalamento estatístico), se mostraram indiferentes às variações do número de coeficientes e à variação dos descritores estatísticos. Assim, a solução obtida neste caso foi a seguinte: metais 40,00%, palhetas 32,35%, flautas 59,49% e cordas 78,10%, totalizando uma taxa de acerto global de 61,29%. Portanto, podemos verificar na Tabela 7.13 que o escalamento estatístico apresentou bons resultados, melhorando o desempenho significativamente quando comparado aos resultados sem o escalamento estatístico. No entanto os classificadores SVM e 1-NN obtiveram uma taxa de acerto pior quando foi acréscido os descritores temporais aos codificadores.

Vetor de caracteristicas				Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
C.Temporais		std		46,7%	50,0%	55,7%	70,5%	60,1%	100,0%	55,7%
C.Temporais	16	LPC	std	50,0%	58,8%	65,8%	99,1%	77,0%	100,0%	68,4%
C.Temporais	16	LSF	std	70,0%	61,8%	96,2%	100,0%	89,9%	100,0%	82,0%
C.Temporais	16	MFCC	std	60,0%	20,6%	72,2%	99,1%	75,0%	100,0%	62,9%

Tabela 7.14: Melhores resultados para o classificador SVM (kernel gaussiano) com características temporais.

7.6.3 Resultados com os Descritores de Áudio

Nesta seção apresentaremos os resultados obtidos com o acréscimo dos descritores de áudio que foram vistos na Seção 4.2. Igualmente a seção anterior os dados das amostras de treinamento para os classificadores SVM e 1-NN foram escalonados estatisticamente.

Nas Tabelas 7.15, 7.16, 7.17 apresentamos as taxas de acerto para os classificadores DLG, SVM e 1-NN respectivamente.

Vetor de caracteristicas				Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
Descr. Áudio	16	LSF	std+m_3+m_4	96,7%	97,1%	91,1%	96,2%	94,8%	97,0%	95,3%
Descr. Áudio	24	LPC	std	93,3%	88,2%	91,1%	95,2%	92,7%	95,4%	92,0%
		MFCC	std+m_3	96,7%	91,2%	93,7%	96,2%	94,8%	96,3%	94,4%
Descr. Áudio		std+m_3		63,3%	52,9%	77,2%	95,2%	79,8%	77,8%	72,2%

Tabela 7.15: Resultados obtidos pelo classificador DLG.

Vetor de caracteristicas				Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
Descr. Áudio	16	LPC	std, std+m_3, std+m_3+m_4	66,7%	88,2%	82,3%	99,1%	88,3%	100,0%	84,1%
Descr. Áudio	16	MFCC	std, std+m_3, std+m_3+m_4	83,3%	73,5%	100,0%	97,1%	93,2%	100,0%	88,5%
	16	LSF	std, std+m_3, std+m_3+m_4	83,3%	97,1%	96,2%	100,0%	96,4%	100,0%	94,1%
Descr. Áudio		std+m_3, std+m_3+m_4		90,0%	70,6%	79,8%	89,5%	83,9%	98,9%	82,5%

Tabela 7.16: Resultados obtidos pelo classificador SVM.

Vetor de caracteristicas				Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
Descr. Áudio	16	MFCC	std, std+m_3, std+m_3+m_4	100,0%	100,0%	97,5%	100,0%	99,2%	100,0%	99,4%
Descr. Áudio	16	LPC	std, std+m_3, std+m_3+m_4	96,7%	97,1%	83,5%	94,3%	91,5%	100,0%	92,9%
	16	LSF	std, std+m_3, std+m_3+m_4	96,7%	100,0%	98,7%	100,0%	99,2%	100,0%	98,9%
Descr. Áudio		std, std+m_3, std+m_3+m_4		90,0%	61,8%	68,4%	92,4%	80,2%	100,0%	78,1%

Tabela 7.17: Resultados obtidos pelo classificador 1-NN.

Já o acréscimo dos descritores de áudio provocou uma melhoria na taxa de acerto para os classificadores 1-NN e DLG quando comparada aos resultados obtidos para o vetor de características contendo somente os codificadores e os descritores estatísticos (acrescidos ou não dos descritores temporais).

7.6.4 Resultados com Características Temporais e Descritores de Áudio

Nas Tabelas 7.18, 7.19, 7.20 apresentamos os quadros com todas as melhores soluções, incluindo os resultados com as características temporais e perceptuais para os classificadores DLG, 1-NN e SVM respectivamente.

Vetor de caracteristicas				Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
D. de àudio e C. Temporais	16	LPC	std+m_3	86,7%	94,1%	91,1%	97,1%	93,6%	96,3%	92,3%
		MFCC	std+m_3	100,0%	97,1%	92,4%	98,1%	96,4%	97,4%	96,9%
D. de àudio e C. Temporais	32	LSF	std	100,0%	94,1%	93,7%	95,2%	95,2%	97,6%	95,8%
D. de àudio e C. Temporais		std+m_3+m_4		70,0%	76,5%	78,5%	91,4%	82,7%	88,7%	79,1%

Tabela 7.18: Resultados obtidos pelo classificador DLG.

Vetor de caracteristicas				Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
D. de àudio e C. Temporais	24	LPC	std	93,3%	91,2%	82,3%	98,1%	91,5%	100,0%	91,2%
D. de àudio e C. Temporais	16	MFCC	std	100,0%	94,1%	84,8%	100,0%	94,4%	100,0%	94,7%
D. de àudio e C. Temporais	16	LSF	std	96,7%	94,1%	88,6%	100,0%	95,2%	100,0%	94,9%
D. de àudio e C. Temporais		std		76,7%	38,2%	73,4%	96,2%	78,6%	100,0%	71,1%

Tabela 7.19: Resultados obtidos pelo classificador 1-NN.

Vetor de caracteristicas				Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
D. de àudio e	16	LPC	std	46,7%	58,8%	59,5%	100,0%	75,0%	100,0%	66,2%
C. Temporais	16	LSF	std	70,0%	58,8%	78,5%	100,0%	83,9%	100,0%	76,8%
D. de àudio e C. Temporais	16	MFCC	std	53,3%	20,6%	45,6%	100,0%	66,1%	100,0%	54,9%
D. de àudio e C. Temporais		std		70,0%	70,6%	77,2%	98,1%	84,3%	100,0%	79,0%

Tabela 7.20: Resultados obtidos pelo classificador SVM.

Conforme pode-se observar, os resultados obtidos apresentaram para o DLG um acréscimo na taxa de acerto global, com o aumento do número de elementos no vetor de características, sendo que o vizinho mais próximo apresentou o seu ápice com o acréscimo dos descritores de áudio, assim como a SVM.

Porém pode-se observar que para a maioria dos casos a melhor solução para o codificador LSF é 16 (exceto para o classificador DLG cuja melhor solução foi com 32 coeficientes), sendo também a melhor solução para os codificadores MFCC e LPC (ou seja, o mesmo número de coeficientes).

Da mesma forma podemos dizer que a combinação para MFCC de 16 coeficientes + desvio + m_3 foi unânime, o mesmo acontecendo para LPC. A única divergência foi a presença ou não das características temporais, onde cada classificador apresentou um quadro diverso.

Pode-se verificar no quadro 7.21 uma síntese das melhores soluções.

		Vetor de caracteristicas - Soluções				Metais	Flautas	Palhetas	Cordas	Tx Global	Tx Treino	Tx Média
1	DLG	D. de áudio e C. Temporais	16	LPC	std+m_3	86,7%	94,1%	91,1%	97,1%	93,6%	96,3%	92,3%
2	DLG	D. de áudio e C. Temporais	16	MFCC	std+m_3	100,0%	97,1%	92,4%	98,1%	**96,4%**	97,4%	96,9%
3	DLG	D. de áudio e C. Temporais	32	LSF	std	100,0%	94,1%	93,7%	95,2%	95,2%	97,6%	95,8%
4	DLG	D. de áudio e C. Temporais		std+m_3+m_4		70,0%	76,5%	78,5%	91,4%	82,7%	88,7%	79,1%
5	K-NN	Descr. Áudio	16	LPC	std+m_3	96,7%	97,1%	83,5%	94,3%	91,5%	100,0%	92,4%
6	K-NN	Descr. Áudio	16	MFCC	std+m_3	100,0%	100,0%	97,5%	100,0%	**99,2%**	100,0%	99,2%
7	K-NN	Descr. Áudio	16	LSF	std	96,7%	100,0%	98,7%	100,0%	99,2%	100,0%	98,5%
8	K-NN	Descr. Áudio		std+m_3+m_4		90,0%	61,8%	68,4%	92,4%	80,2%	100,0%	78,1%
9	SVM	Descr. Áudio	16	LPC	std+m_3	66,7%	88,2%	82,3%	99,1%	88,3%	100,0%	79,1%
10	SVM		16	MFCC	std+m_3	96,7%	91,2%	98,7%	98,1%	**97,2%**	100,0%	95,5%
11	SVM	Descr. Áudio	16	LSF	std	83,3%	97,1%	96,2%	100,0%	96,4%	100,0%	92,2%
12	SVM	D. de áudio e C. Temporais		std+m_3+m_4		70,0%	70,6%	77,2%	98,1%	84,3%	100,0%	79,0%

Tabela 7.21: Melhores soluções para cada classificador/codificador.

7.6.5 Análise da Correlação e Redundância das Variáveis de Entrada

Nesta seção avaliaremos (via correlação) se as variáveis de entrada apresentam algum grau de similaridade. Caso isso seja verdade haverá a necessidade de avaliarmos a permanência dessas variáveis.

As variáveis que serão avaliadas são as características temporais, os descritores específicos de áudio, e os momentos.

Nas Tabelas 7.22, 7.23 e 7.24 que se seguem estão marcadas as correlações com mais de 40% de similaridade, sendo que as que obtiverem mais de 95% estão marcadas em tom mais escuro.

	A	B	C	D	E	F	G	H	I	ZCR	RMS	SC	BW	Flux	std	m_3	m_4
T. Rise	**1,00**	0,31	0,16	-0,07	0,04	-0,12	0,14	0,13	-0,05	-0,02	-0,09	0,35	0,33	-0,01	-0,09	-0,05	-0,05
T. Fall		**1,00**	0,18	0,01	-0,02	-0,23	0,17	0,06	-0,02	-0,01	-0,06	0,25	0,23	0,02	-0,06	-0,04	-0,04
PW			**1,00**	0,19	-0,01	-0,24	0,28	0,03	0,16	-0,12	0,12	0,49	0,50	0,20	0,08	-0,04	-0,06
T.90%				**1,00**	0,08	-0,19	-0,02	-0,14	0,27	-0,13	0,17	0,07	0,08	0,09	0,14	0,02	0,00
T.P1V1					**1,00**	0,01	-0,59	0,17	0,01	0,03	-0,01	0,00	0,01	0,02	-0,02	-0,02	-0,02
Ripple						**1,00**	-0,06	0,28	-0,11	0,13	-0,10	0,16	0,16	-0,11	-0,09	0,01	0,02
T.V1P2							**1,00**	0,22	0,01	-0,13	0,05	0,23	0,24	0,04	0,03	-0,03	-0,03
%P1V2								**1,00**	-0,46	0,04	-0,10	0,16	0,15	-0,05	-0,10	0,00	0,01
Droop									**1,00**	-0,14	0,09	0,15	0,19	0,05	0,06	0,00	-0,02
ZCR										**1,00**	-0,17	-0,05	-0,06	-0,07	-0,17	-0,04	-0,03
RMS											**1,00**	0,01	0,01	0,59	0,99	0,65	0,57
SC												**1,00**	0,97	0,04	-0,02	-0,04	-0,04
BW													**1,00**	0,06	-0,03	-0,05	-0,05
FLUX														**1,00**	0,59	0,33	0,22
std															**1,00**	0,70	0,62
m_3																**1,00**	0,98
m_4																	**1,00**

Tabela 7.22: Tabela contendo a correlação das variáveis para o banco de dados MIS.

Conforme pode-se observar nas Tabelas 7.22, 7.23 e 7.24 existem 3 correlações com alta taxa de similaridade (superior a 95%) que são comuns a todas as bases de dados, ou seja, RMS com std[23], BW com SC, e m_3 com m_4. Simulações feitas retirando-se uma das variáveis que apresentaram alto grau de correlação para SVM e DLG levaram a pequenas variações na taxa de acerto (entre 1% e 2,5%), reduzindo a taxa de acerto, o que indica que essas variáveis, apesar de possuírem entre si um alto grau de correlação, estão contribuindo positivamente no desempenho dos

[23]Essas medidas representam basicamente a mesma informação, uma apresenta o desvio polarizado e a outra o desvio não polarizado.

	A	B	C	D	E	F	G	H	I	ZCR	RMS	SC	BW	Flux	std	m_3	m_4
T. Rise	**1,00**	0,08	0,22	-0,14	-0,09	-0,22	0,18	-0,04	-0,03	0,01	0,08	0,09	0,09	0,12	0,11	0,08	0,06
T. Fall		**1,00**	-0,02	-0,01	-0,06	-0,24	0,07	-0,05	0,02	0,05	-0,03	0,13	0,15	0,02	-0,02	0,01	0,02
PW			**1,00**	-0,07	-0,22	-0,31	0,30	-0,04	-0,01	-0,07	0,27	0,19	0,21	0,39	0,29	0,19	0,17
T.90%				**1,00**	0,10	0,04	-0,09	-0,14	0,22	-0,08	-0,01	0,00	-0,01	-0,03	-0,02	-0,03	-0,03
T.P1V1					**1,00**	0,16	-0,33	-0,05	0,41	-0,11	-0,02	-0,09	-0,08	-0,10	-0,05	-0,06	-0,07
Ripple						**1,00**	-0,13	0,37	-0,17	-0,04	-0,20	0,11	0,10	-0,24	-0,26	-0,32	-0,32
T.V1P2							**1,00**	0,01	0,00	-0,03	0,13	0,17	0,17	0,12	0,14	0,07	0,05
%P1V2								**1,00**	-0,60	0,03	-0,10	0,01	0,01	-0,09	-0,11	-0,10	-0,10
Droop									**1,00**	-0,15	0,18	-0,03	-0,01	0,09	0,17	0,17	0,18
ZCR										**1,00**	-0,30	-0,14	-0,17	-0,13	-0,27	-0,14	-0,12
RMS											**1,00**	-0,05	-0,03	0,60	0,98	0,72	0,65
SC												**1,00**	0,97	-0,01	-0,04	-0,02	-0,01
BW													**1,00**	0,03	-0,03	-0,02	-0,01
FLUX														**1,00**	0,65	0,64	0,59
std															**1,00**	0,79	0,72
m_3																**1,00**	0,98
m_4																	**1,00**

Tabela 7.23: Tabela contendo a correlação das variáveis para o banco de dados MUMS.

	A	B	C	D	E	F	G	H	I	ZCR	RMS	SC	BW	Flux	std	m3	m4
T. Rise	**1,00**	0,42	0,32	-0,11	0,01	-0,19	0,22	-0,03	0,06	-0,10	0,05	0,21	0,22	0,15	0,05	-0,02	-0,03
T. Fall		**1,00**	0,22	0,04	-0,11	-0,21	0,28	-0,04	0,07	-0,05	-0,02	0,19	0,22	0,12	-0,01	-0,02	-0,02
PW			**1,00**	0,08	-0,13	-0,10	0,38	-0,08	0,20	-0,10	0,10	0,58	0,53	0,24	0,11	0,01	0,01
T.90%				**1,00**	0,25	-0,02	-0,05	-0,06	0,17	0,01	0,10	0,22	0,24	0,06	0,11	0,08	0,07
T.P1V1					**1,00**	0,11	-0,61	0,28	-0,04	-0,01	0,02	-0,05	-0,01	-0,03	0,02	0,01	0,00
Ripple						**1,00**	-0,13	0,33	-0,23	0,08	-0,23	0,09	0,08	-0,15	-0,25	-0,19	-0,18
T.V1P2							**1,00**	-0,06	0,17	-0,03	-0,01	0,35	0,34	0,10	0,00	-0,01	0,00
%P1V2								**1,00**	-0,48	-0,04	-0,12	-0,03	0,00	-0,08	-0,13	-0,09	-0,08
Droop									**1,00**	0,07	0,12	0,22	0,21	0,11	0,12	0,09	0,08
ZCR										**1,00**	-0,14	0,19	0,24	-0,04	-0,14	-0,04	-0,03
RMS											**1,00**	-0,05	-0,07	0,43	0,99	0,70	0,60
SC												**1,00**	0,95	0,08	-0,04	-0,04	-0,03
BW													**1,00**	0,09	-0,06	-0,04	-0,03
FLUX														**1,00**	0,45	0,38	0,33
std															**1,00**	0,74	0,64
m3																**1,00**	0,98
m4																	**1,00**

Tabela 7.24: Tabela contendo a correlação das variáveis para o banco de dados RWC.

classificadores empregados. Dessa forma, não se fará o descorrelacionamento das variáveis nessa dissertação, deixando tal procedimento para trabalhos futuros quando se pretende avaliar também outros classificadores, como, por exemplo, redes neurais.

A seguir aplicou-se uma ferramenta de análise de componentes principais (PCA) para verificarmos se existem variáveis linearmente dependentes das demais

variáveis. Em outras palavras, verificamos se há redundância suficiente para a redução da dimensão do espaço de entrada. Na Tabela 7.25 apresentamos as primeiras 10 observações da matriz dos dados (da base de dados MIS) contendo as observações (descritas em variáveis temporais, descritores específicos de áudio e momentos do segmento avaliado) na representação do espaço de componentes principais. Caso haja redundância em alguma variável, uma coluna aparecerá zerada na matriz no espaço das componentes principais, indicando que essa variável é linearmente dependente.

	A	B	C	D	E	F	G	H	I	ZCR	RMS	SC	BW	Flux	std	m_3	m_4
1	-4,03	0,32	-1,63	-1,37	0,39	1,70	-1,68	1,22	-1,13	-0,25	-1,23	0,31	0,44	-0,60	-0,03	0,30	-0,15
2	-3,63	1,58	-1,71	-0,57	0,14	1,39	-1,85	1,12	-0,28	0,54	-0,78	1,34	-0,57	0,25	-0,06	0,24	-0,15
3	-3,18	1,56	-1,84	0,23	-2,18	1,71	-1,61	0,59	0,94	-0,98	-0,92	0,04	0,74	-0,61	-0,04	0,16	-0,19
4	-1,54	0,77	-0,99	-0,78	0,48	0,79	-0,62	0,67	1,05	-0,56	-0,47	-0,12	-1,03	-0,49	0,00	0,13	-0,10
5	-1,87	1,81	-2,20	0,59	-2,33	0,88	0,53	-2,41	-1,38	-1,25	0,42	-0,11	0,19	-0,06	0,01	0,10	-0,12
6	-2,76	0,67	-1,30	0,23	0,05	0,90	-0,93	0,32	-0,47	-0,80	0,29	-0,24	-0,59	0,08	0,12	0,17	-0,11
7	-1,40	0,45	-0,48	-1,03	-1,30	0,62	-0,54	0,83	-0,31	-1,27	-0,36	-0,51	-0,13	-0,71	0,08	0,05	-0,13
8	-2,90	0,97	-3,35	-1,12	-0,54	0,57	-0,66	-0,37	-1,83	-0,90	-1,28	0,20	0,42	-0,09	-0,03	0,10	-0,20
9	-2,34	0,36	-1,71	-0,08	0,38	1,53	0,34	-1,74	-1,05	0,46	1,11	0,18	-0,66	-0,04	0,04	0,18	-0,10
10	-1,51	0,96	-2,26	-1,11	-0,69	0,32	0,29	-1,01	-1,07	-0,73	-0,34	-0,05	-0,45	-0,67	0,08	0,10	-0,10

Tabela 7.25: Tabela contendo as 10 primeiras observações da Matriz de dados no espaço das componentes principais.

Portanto, verificamos a inexistência de variáveis linearmente dependentes das demais no vetor de características, apesar de algumas delas apresentarem alta taxa de correlação. Logo, não se fará eliminação de nenhuma variável de entrada, uma vez que, não se verificou nenhuma redundância de variável (PCA), e a retirada das variáveis com alto grau de correlação apresentou um efeito danoso na taxa de acerto global (entre 1% e 2,5% para os classificadores SVM e DLG).

7.7 Avaliação dos Classificadores

Nesta seção iremos avaliar o desempenho dos classificadores empregados nessa dissertação variando-se o *kernel* para o classificador SVM, empregando um *kernel* polinomial de ordem 2 ou 3. Também iremos variar a transformação no espaço de entrada usando potenciação de ordem 3 ou 4, e finalmente iremos variar a métrica de distância do classificador K-NN, assim como o número de vizinhos mais próxi-

mos para 3 ou 5. Essas avaliações serão feitas somente para as melhores soluções obtidas para cada codificador a cada classificador. Ao final, cada classificador terá 4 possibilidades (3 com uso de codificadores e uma sem uso de codificador) de soluções otimizadas frente às variações a serem feitas para cada classificador, perfazendo um total de 12 candidatos.

Assim, seguindo o método proposto visando reduzir o número de possibilidades a serem analisadas, apresentam-se as melhores soluções obtidas para cada classificador (Tabelas 7.26, 7.27, 7.28) em função de algumas variações investigadas[24].

		Momento	Coef.	Codif.	Metais	Flautas	Palhetas	Cordas	Tx Gl.	Tx Tr.	Tx Méd.
city block, K=1	Descr. Áudio	std+m_3	16	MFCC	100,0%	100,0%	98,7%	99,1%	99,2%	100,0%	99,4%
city block, K=1	Descr. Áudio	std	16	LSF	96,7%	100,0%	98,7%	100,0%	**99,2%**	100,0%	98,9%
city block, K=1	Descr. Áudio	std+m_3	16	LPC	96,7%	97,1%	81,0%	98,1%	92,3%	100,0%	93,2%
city block, K=1	Descr. Áudio	std+m_3+m_4			90,0%	64,7%	64,6%	91,4%	79,0%	100,0%	77,7%
Euclideana, K=1	Descr. Áudio	std+m_3	16	MFCC	100,0%	100,0%	97,5%	100,0%	**99,2%**	100,0%	99,4%
Euclideana, K=1	Descr. Áudio	std	16	LSF	96,7%	100,0%	98,7%	100,0%	**99,2%**	100,0%	98,9%
Euclideana, K=1	Descr. Áudio	std+m_3	16	LPC	96,7%	97,1%	83,5%	94,3%	91,5%	100,0%	92,9%
Euclideana, K=1	Descr. Áudio	std+m_3+m_4			90,0%	61,8%	68,4%	92,4%	80,2%	100,0%	78,1%
city block, K=3	Descr. Áudio	std+m_3	16	MFCC	100,0%	97,1%	93,7%	99,1%	97,2%	100,0%	97,4%
city block, K=3	Descr. Áudio	std	16	LSF	96,7%	100,0%	94,9%	100,0%	98,0%	100,0%	97,9%
city block, K=3	Descr. Áudio	std+m_3	16	LPC	96,7%	97,1%	84,8%	96,2%	92,7%	100,0%	93,7%
city block, K=3	Descr. Áudio	std+m_3+m_4			93,3%	70,6%	69,6%	93,3%	82,7%	100,0%	81,7%
Euclideana, K=3	Descr. Áudio	std+m_3	16	MFCC	100,0%	97,1%	96,2%	98,1%	97,6%	100,0%	97,8%
Euclideana, K=3	Descr. Áudio	std	16	LSF	96,7%	100,0%	94,9%	100,0%	98,0%	100,0%	97,9%
Euclideana, K=3	Descr. Áudio	std+m_3	16	LPC	96,7%	97,1%	82,3%	97,1%	92,3%	100,0%	93,3%
Euclideana, K=3	Descr. Áudio	std+m_3+m_4			93,3%	73,5%	69,6%	93,3%	**83,1%**	100,0%	82,5%
city block, K=5	Descr. Áudio	std+m_3	16	MFCC	100,0%	97,1%	94,9%	99,1%	97,6%	100,0%	97,8%
city block, K=5	Descr. Áudio	std	16	LSF	96,7%	100,0%	96,2%	100,0%	98,4%	100,0%	98,2%
city block, K=5	Descr. Áudio	std+m_3	16	LPC	96,7%	97,1%	83,5%	99,1%	**93,6%**	100,0%	94,1%
city block, K=5	Descr. Áudio	std+m_3+m_4			90,0%	61,8%	69,6%	92,4%	80,7%	100,0%	78,4%
Euclideana, K=5	Descr. Áudio	std+m_3	16	MFCC	100,0%	94,1%	96,2%	99,1%	97,6%	100,0%	97,3%
Euclideana, K=5	Descr. Áudio	std	16	LSF	96,7%	100,0%	97,5%	100,0%	98,8%	100,0%	98,5%
Euclideana, K=5	Descr. Áudio	std+m_3	16	LPC	93,3%	97,1%	82,3%	97,1%	91,9%	100,0%	92,5%
Euclideana, K=5	Descr. Áudio	std+m_3+m_4			86,7%	61,8%	67,1%	91,4%	79,0%	100,0%	76,7%

Tabela 7.26: Melhores soluções obtidas para o classificador K-NN.

Ao final podemos agrupar as 12 melhores soluções (conforme a Tabela 7.29), que serão usadas para avaliar a capacidade de generalização do método e as taxas de acerto nas demais bases de dados.

A partir dessa seção todas as matrizes de confusão que aparecerão nessa dissertação farão referência ao número da solução constante na primeira coluna da Tabela 7.29.

[24] As soluções marcadas com "*" diferem do que apareceram inicialmente na Tabela 7.10 porque para essas soluções foi feito o escalonamento estatístico.

		Momento	Coef.	Codif.	Metais	Flautas	Palhetas	Cordas	Tx Gl.	Tx Tr.	Tx Méd.	
SVM - Gaussiana std=1		std+m_3	16	MFCC	83,3%	73,5%	97,5%	95,2%	91,5%	100,0%	87,4%	*
SVM - Gaussiana std=1	Descr. Áudio	std+m_3	16	LPC	66,7%	88,2%	82,3%	99,1%	88,3%	100,0%	84,1%	
SVM - Gaussiana std=1	Descr. Áudio	std	16	LSF	83,3%	97,1%	96,2%	100,0%	96,4%	100,0%	94,1%	
SVM - Gaussiana std=1	C.temporal e D. áudio	std+m_3+m_4			70,0%	70,6%	77,2%	98,1%	84,3%	100,0%	79,0%	
SVM - Gaussiana std=2		std+m_3	16	MFCC	100,0%	100,0%	98,7%	98,1%	**98,8%**	100,0%	99,2%	*
SVM - Gaussiana std=2	Descr. Áudio	std+m_3	16	LPC	86,7%	91,2%	88,6%	98,1%	92,7%	100,0%	91,1%	
SVM - Gaussiana std=2	Descr. Áudio	std	16	LSF	96,7%	100,0%	100,0%	100,0%	**99,6%**	100,0%	99,2%	
SVM - Gaussiana std=2	C.temporal e D. áudio	std+m_3+m_4			83,3%	73,5%	74,7%	97,1%	85,1%	100,0%	82,2%	
SVM - Polynomial 2		std+m_3	16	MFCC	100,0%	97,1%	94,9%	95,2%	96,0%	100,0%	96,8%	*
SVM - Polynomial 2	Descr. Áudio	std+m_3	16	LPC	96,7%	85,3%	89,9%	98,1%	**93,6%**	100,0%	92,5%	
SVM - Polynomial 2	Descr. Áudio	std	16	LSF	96,7%	100,0%	94,9%	98,1%	97,2%	100,0%	97,4%	
SVM - Polynomial 2	C.temporal e D. áudio	std+m_3+m_4			90,0%	88,2%	77,2%	95,2%	**87,9%**	98,3%	87,7%	
SVM - Polynomial 3		std+m_3	16	MFCC	100,0%	97,1%	94,9%	98,1%	97,2%	100,0%	97,5%	*
SVM - Polynomial 3	Descr. Áudio	std+m_3	16	LPC	90,0%	97,1%	87,3%	91,4%	90,7%	100,0%	91,5%	
SVM - Polynomial 3	Descr. Áudio	std	16	LSF	96,7%	100,0%	96,2%	98,1%	97,6%	100,0%	97,7%	
SVM - Polynomial 3	C.temporal e D. áudio	std+m_3+m_4			90,0%	58,8%	65,8%	95,2%	80,2%	100,0%	77,5%	

Tabela 7.27: Melhores soluções obtidas para o classificador SVM.

		Momento	Coef.	Codif.	Metais	Flautas	Palhetas	Cordas	Tx Gl.	Tx Tr.	Tx Méd.
DLG - Poten 2	C.temporal e D. áudio	std+m_3	16	MFCC	100,0%	97,1%	92,4%	98,1%	96,4%	97,4%	96,9%
DLG - Poten 2	C.temporal e D. áudio	std+m_3	16	LPC	86,7%	94,1%	91,1%	97,1%	93,6%	96,3%	92,3%
DLG - Poten 2	C.temporal e D. áudio	std	32	LSF	100,0%	94,1%	93,7%	95,2%	95,2%	97,6%	95,8%
DLG - Poten 2	C.temporal e D. áudio	std+m_3+m_4			70,0%	76,5%	78,5%	91,4%	82,7%	88,7%	79,1%
DLG - Poten 3	C.temporal e D. áudio	std+m_3	16	MFCC	100,0%	100,0%	92,4%	99,1%	**97,2%**	98,3%	97,9%
DLG - Poten 3	C.temporal e D. áudio	std+m_3	16	LPC	93,3%	94,1%	88,6%	96,2%	93,2%	97,2%	93,1%
DLG - Poten 3	C.temporal e D. áudio	std	32	LSF	100,0%	97,1%	94,9%	99,1%	**97,6%**	98,8%	97,8%
DLG - Poten 3	C.temporal e D. áudio	std+m_3+m_4			83,3%	73,5%	82,3%	92,4%	85,5%	91,6%	82,9%
DLG - Poten 4	C.temporal e D. áudio	std+m_3	16	MFCC	100,0%	97,1%	92,4%	97,1%	96,0%	98,5%	96,7%
DLG - Poten 4	C.temporal e D. áudio	std+m_3	16	LPC	96,7%	100,0%	93,7%	96,2%	**96,0%**	98,3%	96,6%
DLG - Poten 4	C.temporal e D. áudio	std	32	LSF	96,7%	94,1%	97,5%	97,1%	96,8%	98,8%	96,4%
DLG - Poten 4	C.temporal e D. áudio	std+m_3+m_4			83,3%	70,6%	84,8%	93,3%	**86,3%**	92,1%	83,0%

Tabela 7.28: Melhores soluções obtidas para o classificador DLG.

	Classificador	Vetor de caraterísticas				Metais	Flautas	Palhetas	Cordas	Tx Gl.	Tx Tr.	Tx Méd.
1	city block, K=1	Descr. Áudio	std	16	LSF	96,7%	100,0%	98,7%	100,0%	99,2%	100,0%	98,9%
2	Euclideana, K=1	Descr. Áudio	std+m_3	16	MFCC	100,0%	100,0%	97,5%	100,0%	99,2%	100,0%	99,4%
4	city block, K=5	Descr. Áudio	std+m_3	16	LPC	96,7%	97,1%	83,5%	99,1%	93,6%	100,0%	94,1%
3	Euclideana, K=3	Descr. Áudio	std+m3+m4			93,3%	73,5%	69,6%	93,3%	83,1%	100,0%	82,5%
5	SVM - Gaussiana std=2		std+m_3	16	MFCC	100,0%	100,0%	98,7%	98,1%	98,8%	100,0%	99,2%
6	SVM - Gaussiana std=2	Descr. Áudio	std	16	LSF	96,7%	100,0%	100,0%	100,0%	**99,6%**	100,0%	99,2%
7	SVM - Polynomial 2	Descr. Áudio	std+m_3	16	LPC	96,7%	85,3%	89,9%	98,1%	93,6%	100,0%	92,5%
8	SVM - Polynomial 2	C.temporal e D. áudio	std+m_3+m_4			90,0%	88,2%	77,2%	95,2%	87,9%	98,3%	87,7%
9	DLG - Poten 3	C.temporal e D. áudio	std+m_3	16	MFCC	100,0%	100,0%	92,4%	99,1%	97,2%	98,3%	97,9%
10	DLG - Poten 3	C.temporal e D. áudio	std	32	LSF	100,0%	97,1%	94,9%	99,1%	**97,6%**	98,8%	97,8%
11	DLG - Poten 4	C.temporal e D. áudio	std+m_3	16	LPC	96,7%	100,0%	93,7%	96,2%	96,0%	98,3%	96,6%
12	DLG - Poten 4	C.temporal e D. áudio	std+m_3+m_4			83,3%	70,6%	84,8%	93,3%	86,3%	92,1%	83,0%

Tabela 7.29: As melhores soluções obtidas para cada codificador em cada classificador.

Capítulo 8

Avaliação da Taxa de Acerto

Este capítulo contém a avaliação da taxa de acerto para o agrupamento MFPC, a avaliação da taxa de acerto para o agrupamento INSTRUMENTO e a estimativa da taxa de acerto do classificador proposto. Esta última tem por objetivo avaliar o desempenho do método de reconhecimento automático elaborado nos capítulos anteriores, usando amostras reservadas para teste pertencentes às 3 bases de dados (MIS, MUMS e RWC).

As amostras da base de dados MUMS só aparecerão na Seção 8.3, uma vez que essa base de dados possui poucas amostras para alguns instrumentos. As amostras da base de dados RWC serão usadas a partir da Seção 8.2. Assim, apresentaremos alguns resultados obtidos inicialmente para as amostras da base de dados MIS, variando-se os agrupamentos e as estratégias, depois contendo amostras da base de dados RWC para o agrupamento instrumento, e finalmente contendo amostras da base de dados MUMS, quando faremos uma estimativa da taxa de acerto do classificador proposto. Todas as soluções usadas foram obtidas da Tabela 7.29 presentes no capítulo anterior.

8.1 Avaliação da Taxa de Acerto para o Agrupamento MFPC

Nesta seção iremos apresentar os resultados obtidos com amostras da base de dados MIS usando o agrupamento MFPC, tanto para a estratégia padrão quanto para algumas estratégias alternativas que foram ilustradas no Capítulo 6. No en-

tanto, essas estratégias alternativas aparecem nessa dissertação somente como uma avaliação exploratória dessas estratégias (comprovando que apresentam taxas de acertos globais diferentes para a mesma solução com o mesmo agrupamento), já que não se pretende avaliar para essas estratégias (alternativas) a taxa de acerto para outros agrupamentos além do agrupamento MFPC.

O estudo feito sobre o agrupamento MFPC a partir do segmento central do modelo IMF, apresentado no capítulo anterior, determinou quais eram as melhores soluções. Para se chegar às 12 melhores soluções analisou-se 391 sistemas de reconhecimento automático, formados por 8 soluções com variações sobre o classificador DLG, 12 soluções com variações sobre o classificador SVM, 20 soluções com variações sobre o classificador K-NN, 27 soluções sem codificador, e 324 soluções resultantes da combinação dos seguintes elementos: 3 estatísticas (m_2, m_3 e m_4), 3 coeficientes (LSF, MFCC e LPC), 3 quantidades de coeficientes (16, 24, e 32), 3 classificadores (DLG, SVM e K-NN) e 4 padrões de vetor de característica (somente codificadores, codificadores e descritores temporais, codificadores e descritores de áudio, codificadores e descritores temporais e descritores de áudio).

Com intuito de sintetizar os resultados apresentados por essa busca das melhores soluções, apresentaremos somente duas tabelas contendo 351 soluções correspondentes às 324 soluções descritas anteriormente acrescidos das 27 soluções sem o uso dos codificadores.

Taxa de Acerto	Sem Cod.	LPC	LSF	MFCC	Total
[0%, 60%]	0,0%	0,9%	1,7%	3,4%	6,0%
(60%, 70%]	2,6%	2,6%	1,7%	2,6%	9,4%
(70%, 80%]	1,7%	10,0%	0,0%	2,6%	14,2%
(80%, 90%]	3,4%	8,3%	6,0%	2,6%	20,2%
(90%, 95%]	0,0%	9,1%	15,7%	14,2%	39,0%
(95%, 100%]	0,0%	0,0%	5,7%	5,4%	11,1%
Total	7,7%	30,8%	30,8%	30,8%	100,0%

Tabela 8.1: Taxa de acerto versus codificadores.

Conforme pode se observar na Tabela 8.1, todas as soluções na faixa de acerto superior a 90% usaram codificadores. E na faixa de acerto superior a 95% estão somente soluções com o uso dos codificadores LSF e MFCC. Já na Tabela 8.2 todos os classificadores apresentaram soluções em todas as faixas. Deve-se destacar que, de modo geral, o classificador K-NN teve um desempenho melhor que o DLG, o

Taxa de Acerto	DLG	SVM	1-NN	Total
[0%, 60%]	0,0%	6,0%	0,0%	6,0%
(60%, 70%]	0,9%	7,7%	0,9%	9,4%
(70%, 80%]	3,1%	8,5%	2,6%	14,2%
(80%, 90%]	9,1%	6,0%	5,1%	20,2%
(90%, 95%]	18,2%	3,4%	17,4%	39,0%
(95%, 100%]	2,0%	1,7%	7,4%	11,1%
Total	33,3%	33,3%	33,3%	100,0%

Tabela 8.2: Taxa de acerto versus classificadores.

qual, por sua vez, foi ligeiramente melhor que a SVM. No entanto, se incluirmos as 40 soluções que não constam nessas tabelas, o classificador SVM apresentou a solução com melhor desempenho. Portanto, apesar das amostras demonstrarem uma ligeira preferência para o classificador K-NN, não houve uma superioridade clara entre um classificador em relação aos demais.

As amostras usadas para se avaliar o desempenho nas matrizes de confusão que serão apresentadas nessa seção foram as mesmas em todos os casos.

8.1.1 Resultados do Agrupamento MFPC na Estratégia 1

Nessa subseção ilustraremos alguns dos resultados obtidos. As soluções 2, 6 e 10 apresentadas nas Tabelas 8.3, 8.4 e 8.5 representam, respectivamente, soluções obtidas de classificadores distintos, a saber: K-NN, SVM e DLG.

Família						Total	Tx Ac
Metais	1	30	0	0	0	30	100,0%
Flautas	2	0	34	0	0	34	100,0%
Palhetas	3	0	2	77	0	79	97,5%
Cordas	4	0	0	0	105	105	100,0%
						248	99,2%

Tabela 8.3: Matriz de confusão para o agrupamento MFPC, usando a solução 2.

Família						Total	Tx Ac
Metais	1	29	0	0	1	30	96,7%
Flautas	2	0	34	0	0	34	100,0%
Palhetas	3	0	0	79	0	79	100,0%
Cordas	4	0	0	0	105	105	100,0%
						248	99,6%

Tabela 8.4: Matriz de confusão para o agrupamento MFPC, usando a solução 6.

Família						Total	Tx Ac
Metais	1	30	0	0	0	30	100,0%
Flautas	2	0	32	2	0	34	94,1%
Palhetas	3	0	3	75	1	79	94,9%
Cordas	4	1	0	2	102	105	97,1%
						248	96,4%

Tabela 8.5: Matriz de confusão para o agrupamento MFPC, usando a solução 10.

8.1.2 Resultados do Agrupamento MFPC nas Estratégias 2 e 3

Nesta subseção ilustraremos os resultados obtidos para algumas estratégias alternativas. Deve-se ressaltar que as soluções para a estratégia 1 e estratégia 3 são as mesmas quando se usa o classificador K-NN. Assim, iremos avaliar a estratégia 3 somente para as soluções 6 e 10, que foram ilustradas na Seção 8.1.1, as quais utilizam os classificadores SVM e DLG, respectivamente.

Conforme se pode observar na matriz de confusão ilustrada pela Tabela 8.6, a taxa de acerto obtida para essa estratégia foi superior à taxa de acerto obtida usando o modelo padrão, conforme pode se ver na matriz de confusão apresentada pela Tabela 8.5, comprovando o fato de que as estratégias podem obter desempenhos distintos.

Já a solução 6 combinada com a estratégia 3 (conforme pode se ver na Tabela 8.7) apresentou um resultado ligeiramente inferior ao resultado obtido por essa

Família	Estratégia 3					Total	Tx Ac
Metais	1	29	0	0	1	30	96,7%
Flautas	2	0	34	0	0	34	100,0%
Palhetas	3	0	0	78	1	79	98,7%
Cordas	4	0	0	0	105	105	100,0%
						248	99,2%

Tabela 8.6: Matriz de confusão para o agrupamento MFPC, usando a solução 10 combinada com a estratégia 3.

solução quando combinada com o modelo padrão (estratégia 1).

Família	Estratégia 3					Total	Tx Ac
Metais	1	29	0	0	1	30	96,7%
Flautas	2	0	34	0	0	34	100,0%
Palhetas	3	0	0	78	1	79	98,7%
Cordas	4	0	0	0	105	105	100,0%
						248	99,2%

Tabela 8.7: Matriz de confusão para o agrupamento MFPC, usando a solução 6 obtida usando a estratégia 3.

Para a estratégia 2, a seguir ilustraremos para as mesmas amostras, uma escolha combinando duas das três soluções aqui investigadas, para demonstrar a potencialidade dessa abordagem.

Conforme pode-se observar, tanto para a estratégia 3 quanto para a estratégia 1, o melhor desempenho foi para a solução 6 combinada com a estratégia 1[1], demonstrado pela sua taxa de acerto de 99,6% [2]. No entanto, se combinarmos a solução 2 com a solução 6 e usarmos a estratégia 2, é possível obtermos para esse caso uma taxa de acerto de 100%

Primeiro iremos classificar as amostras cordas do total de amostras conforme

[1]Isso também foi verdade para as as demais soluções usando a estratégia 1, conforme se encontra ilustrado na Tabela 7.29.

[2]A taxa de acerto para a estratégia 3 foi de 99,2%, o mesmo valor foi obtido para a solução 2.

a Tabela 8.8.

Família	Estratégia 2			Total	Tx Ac
Cordas	1	105	0	105	100,0%
Resto 1	2	0	143	143	100,0%
	Solução 2			248	100,0%

Tabela 8.8: Matriz de confusão para o agrupamento MFPC, usando a solução 2 (cordas) combinada com a estratégia 2.

A seguir iremos classificar os metais das amostras restantes, ou seja, o total de amostras menos as amostras já classificadas como cordas, conforme a Tabela 8.9.

Família	Estratégia 2			Total	Tx Ac
Metais	1	30	0	30	100,0%
Resto 2	2	0	113	113	100,0%
	Solução 2			143	100,0%

Tabela 8.9: Matriz de confusão para o agrupamento MFPC, usando a solução 2 (metais) combinada com a estratégia 2.

Finalmente iremos classificar as amostras flautas e palhetas, conforme a Tabela 8.10.

Família	Estratégia 2			Total	Tx Ac
flautas	1	34	0	34	100,0%
palhetas	2	0	79	79	100,0%
	Solução 6			113	100,0%

Tabela 8.10: Matriz de confusão para o agrupamento MFPC, usando a solução 6 (flautas e palhetas) combinada com a estratégia 2.

Evidentemente que, devido ao fato da taxa de acerto ter sido alta para todas as estratégias, podemos dizer que elas nesse caso se equivalem. No entanto, foi

possível constatar, para esse caso específico, pequenas variações na taxa de acerto, as quais constatam diferenças entre suas abordagens.

8.2 Avaliação da Taxa de Acerto para o Agrupamento INSTRUMENTO

Nesta seção iremos analisar algumas das melhores soluções obtidas para o agrupamento INSTRUMENTO para cada codificador, e avaliaremos a capacidade de generalização do método ao verificarmos as taxas de acerto obtidas dessas mesmas soluções para a base de dados RWC, seguindo o mesmo modelo, como se fosse independente.

Conforme se pode ver na Tabela 7.29, as taxas de acerto para o agrupamento MFPC usando o classificador DLG foram ligeiramente inferiores às obtidas com o classificador SVM e K-NN. Em igual verificação pode se constatar que as soluções com o acréscimo dos codificadores LSF e MFCC se mostraram superiores em relação às demais. Portanto, só avaliaremos as soluções que representam a intersecção dessas constatações (soluções 1, 2, 5 e 6) para o agrupamento INSTRUMENTO. Por motivo de espaço, para simplificar a apresentação dos resultados, só iremos ilustrar as matrizes de confusão com a melhor e a pior taxa de acerto para essas soluções.

A fim de avaliarmos se a retirada das soluções usando o classificador DLG com o uso dos codificadores LSF e MFCC foi ruim (soluções 9 e 10), apresentamos a Tabela 8.11 contendo as soluções 1, 2, 5, 6, 9 e 10 e suas respectivas taxas de acerto para o agrupamento INSTRUMENTO, usando amostras da base de dados MIS. As demais soluções apresentaram para o agrupamento MFPC uma taxa de acerto ainda pior, e portanto não foram avaliadas para o agrupamento INSTRUMENTO[3].

Novamente pode-se constatar que as soluções para o agrupamento INSTRUMENTO usando o classificador DLG obtiveram as piores taxas de acerto. Já as taxas de acerto obtidas para as soluções 1, 2, 5 e 6 para amostras provenientes da base de dados RWC podem ser vistas na Tabela 8.12:

A seguir ilustraremos 4 matrizes de confusão nas Tabelas 8.13, 8.14, 8.15 e

[3]Esse corte contorna um procedimento guloso, portanto existe a possibilidade de uma dessas soluções apresentar resultados melhores.

Solução #	Taxa de Acerto
solução 1	95,6 %
solução 2	94,8 %
solução 5	94,8 %
solução 6	96,4 %
solução 9	92,7 %
solução 10	94,0 %

Tabela 8.11: Tabela contendo as soluções e suas taxas de acerto para o agrupamento INSTRUMENTO a partir das da base de dados MIS.

Solução #	Taxa de Acerto
solução 1	86,8 %
solução 2	94,1 %
solução 5	95,4 %
solução 6	89,8 %

Tabela 8.12: Tabela contendo as soluções e suas taxas de acerto para o agrupamento INSTRUMENTO a partir das da base de dados RWC.

8.16 do agrupamento INSTRUMENTO para as soluções 6 e 2 para as amostras da base de dados MIS e para as soluções 5 e 1 para a base de dados RWC. Cabe novamente ressaltar que os instrumentos pertencentes ao agrupamento INSTRUMENTO para base de dados RWC são ligeiramente diferentes dos instrumentos que compõem o agrupamento INSTRUMENTO para a base de dados MIS.

Uma vez que as matrizes de confusão, para as melhores soluções, do agrupamento INSTRUMENTO, em cada uma das bases de dados, apresentam resultados com erro disperso (conforme parcialmente se vê nas Tabelas 8.13, 8.14, 8.16 e 8.15), é possível definir um classificador formado por um "banco de classificadores", onde cada amostra é classificada por N classificadores, sendo que a estimativa é estabelecida pela classe mais votada nesses N classificadores. A escolha de N se dá, preferencialmente, visando a maximização da taxa de acerto[4]. No entanto, arbi-

[4]O domínio da função a ser maximizada nesse caso é discreto. Por exemplo, caso se use um espaço simplificado composto por $N = \{1, 3\}$, a partir das soluções= $\{1, 2, 5, 6\}$, deverão ser

Instrumentos		1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total	Tx Ac
Saxofone Contralto	1	18	0	0	0	0	0	0	0	0	0	0	0	0	0	18	100,0%
Trombone Baixo	2	0	12	0	0	0	0	0	0	0	0	0	0	0	0	12	100,0%
Saxofone Soprano	3	2	0	16	0	0	0	0	0	0	0	0	0	0	0	18	88,9%
Trombone Tenor	4	0	0	0	9	0	0	0	0	0	0	0	0	0	0	9	100,0%
Flauta Contralto	5	0	0	0	0	8	0	0	0	0	0	0	0	0	0	8	100,0%
Flauta Baixo	6	0	0	0	0	0	8	0	0	0	0	0	0	0	0	8	100,0%
Flauta	7	0	0	0	0	0	0	18	0	0	0	0	0	0	0	18	100,0%
Fagote	8	0	0	0	0	0	1	0	11	0	0	0	0	0	0	12	91,7%
Clarinete Bb	9	0	0	1	0	0	0	1	0	8	2	0	0	0	0	12	66,7%
Clarinete Eb	10	0	0	0	0	0	0	0	0	1	9	0	0	0	0	10	90,0%
Trompa	11	0	0	0	0	0	0	0	0	0	0	9	0	0	0	9	100,0%
Oboé	12	0	0	1	0	0	0	0	0	0	1	0	7	0	0	9	77,8%
Violoncelo	13	0	0	0	0	0	0	0	0	0	0	0	0	53	3	56	94,6%
Violino	14	0	0	0	0	0	0	0	0	0	0	0	0	0	49	49	100,0%
																248	94,8%

Tabela 8.13: Matriz de confusão para o agrupamento INSTRUMENTO, usando a solução 2.

Instrumentos		1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total	Tx Ac
Saxofone Contralto	1	18	0	0	0	0	0	0	0	0	0	0	0	0	0	18	100,0%
Trombone Baixo	2	0	12	0	0	0	0	0	0	0	0	0	0	0	0	12	100,0%
Saxofone Soprano	3	0	0	18	0	0	0	0	0	0	0	0	0	0	0	18	100,0%
Trombone Tenor	4	0	1	0	8	0	0	0	0	0	0	0	0	0	0	9	88,9%
Flauta Contralto	5	0	0	0	0	8	0	0	0	0	0	0	0	0	0	8	100,0%
Flauta Baixo	6	0	0	0	0	0	8	0	0	0	0	0	0	0	0	8	100,0%
Flauta	7	0	0	0	0	0	0	18	0	0	0	0	0	0	0	18	100,0%
Fagote	8	1	0	0	0	0	0	0	10	0	0	0	0	0	1	12	83,3%
Clarinete Bb	9	0	0	0	0	0	0	0	0	12	0	0	0	0	0	12	100,0%
Clarinete Eb	10	0	0	1	0	0	0	0	0	1	8	0	0	0	0	10	80,0%
Trompa	11	0	0	0	0	0	0	0	0	0	0	8	0	1	0	9	88,9%
Oboé	12	0	0	0	0	0	0	0	0	0	1	0	8	0	0	9	88,9%
Violoncelo	13	0	0	0	0	0	0	0	0	0	0	0	0	54	2	56	96,4%
Violino	14	0	0	0	0	0	0	0	0	0	0	0	0	0	49	49	100,0%
																248	96,4%

Tabela 8.14: Matriz de confusão para o agrupamento INSTRUMENTO, usando a solução 6.

trariamente adotou-se um critério mais simples para avaliar essa hipótese, ou seja, definiu-se o mesmo $N = 3$ para as duas bases de dados, formado pelas 3 melhores soluções para cada uma das bases de dados[5]. A seguir apresentamos os resultados

investigadas as seguintes possibilidades: $\{1, 2, 5, 6, (1, 2, 5), (1, 2, 6), (1, 5, 6), (2, 5, 6)\}$.

[5]Soluções 1, 5 e 6 para base de dados MIS; soluções 2, 4 e 5 para a base de dados RWC.

Instrumentos		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total	Tx Ac
Trompa	1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	100,0%
Trombone	2	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	100,0%
Trompete	3	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	100,0%
Tuba	4	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	18	100,0%
Flauta	5	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	30	100,0%
Fagote	6	0	0	0	0	0	23	0	0	0	0	0	0	0	0	3	1	0	27	85,2%
Oboé	7	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	1	27	96,3%
Saxofone Contralto	8	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	1	27	96,3%
Saxofone Barítono	9	0	0	0	0	0	0	0	0	25	0	0	0	0	0	1	0	1	27	92,6%
Saxofone Soprano	10	0	0	0	0	0	2	0	2	0	22	1	0	0	0	0	0	0	27	81,5%
Saxofone Tenor	11	0	0	0	0	1	0	0	1	0	2	22	0	0	0	0	0	1	27	81,5%
Glockenspiel	12	1	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	19	94,7%
Vibrafone	13	0	0	0	0	0	0	0	0	0	0	0	0	35	1	0	0	0	36	97,2%
Xilofone	14	0	0	0	0	0	0	0	0	0	0	0	0	1	23	0	0	0	24	95,8%
Violoncelo	15	0	0	0	0	0	0	0	0	1	0	0	0	0	0	76	1	0	78	97,4%
Contrabaixo	16	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	76	0	77	98,7%
Violino	17	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	74	77	96,1%
																			591	95,4%

Tabela 8.15: Matriz de confusão para o agrupamento INSTRUMENTO, usando a solução 5.

Instrumentos		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total	Tx Ac
Trompa	1	17	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	89,5%
Trombone	2	1	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	95,8%
Trompete	3	3	0	23	0	0	0	1	0	0	0	0	0	0	0	0	0	0	27	85,2%
Tuba	4	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	18	100,0%
Flauta	5	0	0	1	0	25	2	0	0	0	1	0	0	0	0	1	0	0	30	83,3%
Fagote	6	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	27	100,0%
Oboé	7	0	0	1	0	1	0	18	0	0	1	0	0	0	0	3	0	3	27	66,7%
Saxofone Contralto	8	0	0	0	0	0	0	0	24	1	0	0	0	0	0	0	0	2	27	88,9%
Saxofone Barítono	9	0	0	0	0	0	0	0	1	25	0	0	0	0	0	0	0	1	27	92,6%
Saxofone Soprano	10	1	0	1	0	1	0	0	0	0	21	3	0	0	0	0	0	0	27	77,8%
Saxofone Tenor	11	0	0	0	0	0	0	0	1	0	0	25	0	0	0	1	0	0	27	92,6%
Glockenspiel	12	0	0	0	0	1	0	0	0	0	0	0	18	0	0	0	0	0	19	94,7%
Vibrafone	13	0	0	0	0	1	0	0	0	0	1	0	0	29	1	1	0	3	36	80,6%
Xilofone	14	0	0	0	0	0	0	0	0	0	0	0	0	5	19	0	0	0	24	79,2%
Violoncelo	15	0	0	0	0	1	0	1	0	0	0	0	0	0	0	72	0	4	78	92,3%
Contrabaixo	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	74	0	77	96,1%
Violino	17	0	0	0	0	2	1	1	3	6	1	3	0	1	1	3	0	55	77	71,4%
																			591	86,8%

Tabela 8.16: Matriz de confusão para o agrupamento INSTRUMENTO, usando a solução 1.

obtidos para as bases de dados MIS e RWC, respectivamente, presentes nas matrizes de confusão das Tabelas 8.17 e 8.18:

Somente para ilustrarmos, as taxas de acerto obtidas para o agrupamento

Instrumentos		1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total	Tx Ac
Saxofone Contralto	1	18	0	0	0	0	0	0	0	0	0	0	0	0	0	18	100,0%
Trombone Baixo	2	0	12	0	0	0	0	0	0	0	0	0	0	0	0	12	100,0%
Saxofone Soprano	3	0	0	18	0	0	0	0	0	0	0	0	0	0	0	18	100,0%
Trombone Tenor	4	0	1	0	8	0	0	0	0	0	0	0	0	0	0	9	88,9%
Flauta Contralto	5	0	0	0	0	8	0	0	0	0	0	0	0	0	0	8	100,0%
Flauta Baixo	6	0	0	0	0	0	8	0	0	0	0	0	0	0	0	8	100,0%
Flauta	7	0	0	0	0	0	0	18	0	0	0	0	0	0	0	18	100,0%
Fagote	8	1	0	0	0	0	0	0	11	0	0	0	0	0	0	12	91,7%
Clarinete Bb	9	0	0	0	0	0	0	0	0	12	0	0	0	0	0	12	100,0%
Clarinete Eb	10	0	0	0	0	0	0	0	0	1	9	0	0	0	0	10	90,0%
Trompa	11	0	0	0	0	0	0	0	0	0	0	8	0	1	0	9	88,9%
Oboé	12	0	0	0	0	0	0	0	0	0	1	0	8	0	0	9	88,9%
Violoncelo	13	0	0	0	0	0	0	0	0	0	0	0	1	54	1	56	96,4%
Violino	14	0	0	0	0	0	0	0	0	0	0	0	0	0	49	49	100,0%
																248	97,2%

Tabela 8.17: Matriz de confusão com o banco de classificadores 1, 5 e 6 - MIS.

Instrumentos		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total	Tx Ac
Trompa	1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	100,0%
Trombone	2	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	100,0%
Trompete	3	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	100,0%
Tuba	4	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	18	100,0%
Flauta	5	1	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	30	96,7%
Fagote	6	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	27	100,0%
Oboé	7	0	0	1	0	0	0	25	0	0	1	0	0	0	0	0	0	0	27	92,6%
Saxofone Contralto	8	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	1	27	96,3%
Saxofone Barítono	9	0	0	0	0	0	0	0	0	26	0	0	0	0	0	1	0	0	27	96,3%
Saxofone Soprano	10	0	0	0	0	0	1	0	2	0	23	1	0	0	0	0	0	0	27	85,2%
Saxofone Tenor	11	0	0	0	0	0	1	0	2	0	1	23	0	0	0	0	0	0	27	85,2%
Glockenspiel	12	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	19	100,0%
Vibrafone	13	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0	0	0	36	100,0%
Xilofone	14	0	0	0	0	0	0	0	0	0	0	0	0	1	23	0	0	0	24	95,8%
Violoncelo	15	0	0	0	0	0	0	0	0	1	0	0	0	0	0	77	0	0	78	98,7%
Contrabaixo	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	77	0	77	100,0%
Violino	17	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	75	77	97,4%
																			591	97,1%

Tabela 8.18: Matriz de confusão com o banco de classificadores 2,4 e 5 - RWC.

MFPC e MFPPC (base de dados RWC) usando os banco de classificador propostos pelas 3 melhores soluções para as bases de dados RWC e MIS, respectivamente, são apresentadas nas matrizes de confusão das Tabelas 8.19, 8.20 e 8.21.

Assim, podemos constatar que para ambas bases de dados o banco de classificadores proporcionou um ganho na taxa de acerto para o agrupamento INSTRUMENTO, ficando próxima em ambas, ou seja, em torno de 97%. Assim, passaremos

Família	Estratégia 3					Total	Tx Ac
Metais	1	29	0	0	1	30	96,7%
Flautas	2	0	34	0	0	34	100,0%
Palhetas	3	0	0	79	0	79	100,0%
Cordas	4	0	0	1	104	105	99,0%
						248	99,2%

Tabela 8.19: Matriz de confusão com o banco de classificadores 1,5 e 6 - MIS.

Família	Estratégia 3					Total	Tx Ac
Metais	1	88	0	0	0	88	100,0%
Flautas	2	1	29	0	0	30	96,7%
Palhetas	3	1	0	159	2	162	98,1%
Cordas	4	0	0	2	230	232	99,1%
						512	98,8%

Tabela 8.20: Matriz de confusão com o banco de classificadores 2,4 e 5 - RWC.

Família	Estratégia 3						Total	Tx Ac
Metais	1	88	0	0	0	0	88	100,0%
Flautas	2	1	29	0	0	0	30	96,7%
Palhetas	3	1	0	159	0	2	162	98,1%
Percussão	4	0	0	0	79	0	79	100,0%
Cordas	5	0	0	2	0	230	232	99,1%
							591	99,0%

Tabela 8.21: Matriz de confusão com o banco de classificadores 2, 4 e 5 - RWC.

a partir desta seção a definir o classificador proposto como sendo formado por um banco de classificadores composto pelas 4 melhores soluções (1, 2, 5 e 6) obtidas para o agrupamento MFPC, a fim de possibilitar que a mesma solução seja empregada independentemente da base de dados escolhida. Tal abordagem não foi encontrada na literatura pesquisada nessa dissertação. Na Figura 8.1 apresentamos a arquitetura do classificador proposto nessa dissertação, onde os processos em amarelo representam o pré-processamento, em branco representam a codificação (e obtenção do vetor de características) e em verde representam a classificação. As abreviações "E.D." e

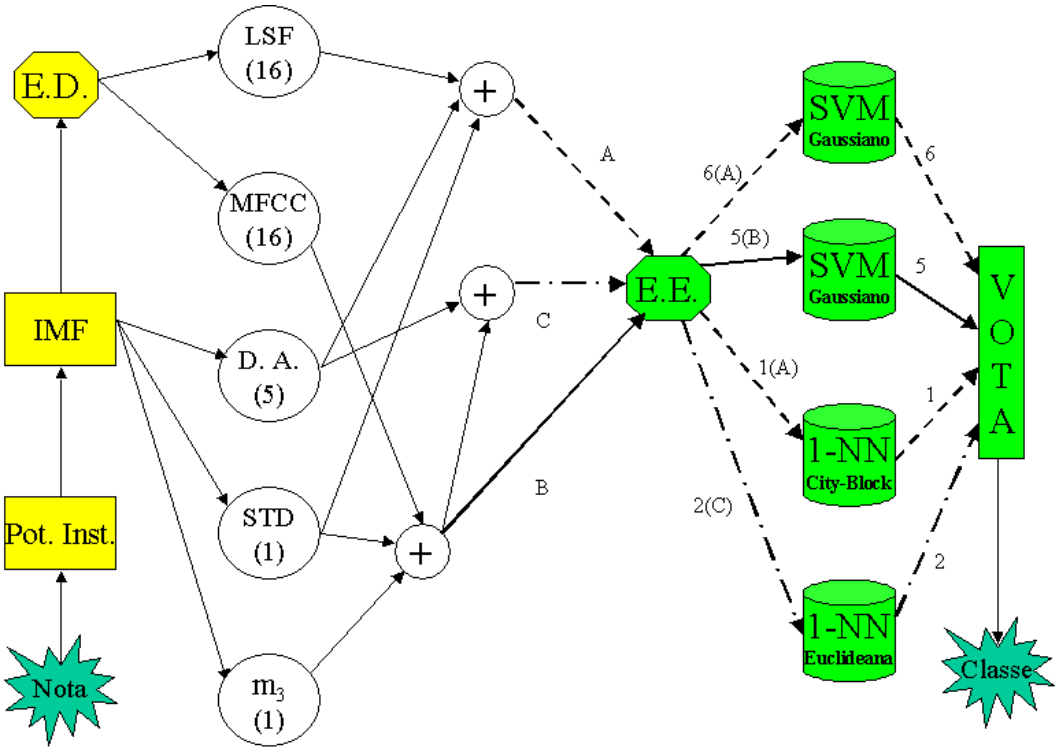

Figura 8.1: Arquitetura do classificador proposto.

"E.E." significam escalamento dinâmico e escalamento estatístico. Já as abreviações "D.A." e "Pot. Inst." significam respectivamente descritores específicos de áudio e potência instantânea da nota. As letras "A", "B" e "C" representam os três vetores de características extraídos pelo classificador proposto, cujas dimensões são 22, 18 e 23 respectivamente. E finalmente, os números 1, 2, 5 e 6 representam as predições das classes provenientes das soluções 1, 2, 5 e 6, respectivamente, conforme a Tabela 7.29

8.3 Estimativa da Taxa de Acerto do Classificador Proposto

Nesta seção variaremos o conjunto de treinamento, em relação às amostras e ao percentual, a fim de termos uma estimativa da taxa de acerto do classificador proposto.

Para avaliarmos a taxa de acerto do classificador proposto usaremos um novo agrupamento, de forma a combinar os instrumentos comuns das bases de dados

distintas. Esse agrupamento será formado por 20 instrumentos, conforme a Tabela 6.4. Poderia-se contra-argumentar que o melhor seria empregar a solução obtida numa base de dados em outra (sem retreinar). O problema dessa abordagem é que, em nosso caso, existem diferenças significativas entre as bases de dados, ou seja, na forma de detecção, na presença ou não de um ruído de fundo, além do fato de apresentarem distribuições das amostras por instrumento diversas. Todas essas diferenças quando combinadas poderão ocasionar variações significativas na taxa de acerto global.

Dependendo da escolha que se faz das amostras (contendo ou não os problemas confusão e contaminação anteriormente referidos), estas irão contribuir ou para o treinamento do classificador ou para serem usadas como teste. Portanto, para cada escolha feita a taxa de acerto obtida pode variar, em função das contribuições que cada amostra dará para o classificador, seja para definir a classe, seja para estimar a taxa de acerto. Assim, independentemente de se manter constante o percentual escolhido para o treinamento, a taxa de acerto poderá variar.

Uma forma de contornar esse problema e se obter uma estimativa da taxa de acerto mais adequada é escolhermos um conjunto de amostras (teste e treinamento) capaz de obter uma medida razoável da capacidade que o classificador tem para discriminar as classes que ele se propõe a classificar. Outra forma é estimarmos através da média de todas ou de diversas combinações possíveis entre as amostras existentes no banco de dados para o percentual usado no treinamento[6]. Dessa forma, se levarão em conta as diversas possibilidades para a formação da estimativa, o que impede parcialmente que uma escolha mal feita do conjunto de amostras para teste e treinamento possa servir para obter uma medida ruim da taxa de acerto.

Adotaremos uma abordagem intermediária entre os dois métodos normalmente usados para se ter uma estimativa da taxa de acerto. Portanto, para avaliarmos a taxa de acerto a partir desse agrupamento fez-se dois tipos de simulações, variando tanto o conjunto de teste quanto a quantidade de amostras empregadas. Além disso, foi elaborado um procedimento para sortear as amostras de testes, restringindo essas amostras à região central da tessitura de cada instrumento musical

[6]Caso haja amostras em quantidade suficientes no conjunto de treinamento para que essa medida não represente uma distorção em relação à distribuição real.

(80%). Tal procedimento está melhor detalhado na Seção 7.2, e tem como conceito central a expectativa de uso das notas de cada instrumento ocorrer de forma desigual (na prática), ou seja, espera-se que as notas da região central de cada instrumento musical tenham uma maior probabilidade de surgimento. Portanto, o conjunto de teste sorteado, levando em conta essa preferência, terá uma maior representatividade. A seguir descrevemos os dois modos nos quais foram avaliadas as taxas de acerto.

1. MODO 1: nesse modo fixou-se o conjunto de treinamento com a seguinte composição: 90% da base de dados RWC, 90% da base de dados MIS, e X da base de dados MUMS. O valor de X variou de 50% a 80%, no passo de 10%. Nesse caso as amostras para teste foram somente as amostras que não participaram do treinamento pertencentes à base de dados MUMS;

2. MODO 2: nesse modo variou-se igualmente o conjunto de treinamento para todas as bases de dados na proporção de X. O valor de X variou entre 50% e 90% no passo de 10%. O conjunto de teste foi formado por todas as demais amostras das 3 bases de dados que não participaram do conjunto de treinamento.

Os resultados obtidos aparecem nas Tabelas 8.22 e 8.23, onde cada percentual apresentado representa um conjunto de amostras distinto, uma vez que, para cada percentual, foi feito um novo sorteio.

MODO 1	50%	60%	70%	80%
solução 5	78,71%	80,82%	82,73%	87,26%
solução 6	83,35%	83,14%	86,80%	88,99%
solução 1	81,94%	81,88%	84,13%	87,69%
solução 2	77,38%	78,19%	79,78%	82,72%
Taxa Global	86,16%	86,72%	88,48%	92,01%

Tabela 8.22: Tabela contendo a taxa de acerto para o MODO 1.

Independentemente do percentual de treinamento, em todos os casos a solução obtida pelo banco de classificadores foi sempre melhor que a melhor solução obtida individualmente (por cada classificador), o que valida essa abordagem e a escolha das soluções 1, 2, 5 e 6 para formar o classificador proposto, considerando

MODO 2	50%	60%	70%	80%	90%
solução 5	86,75%	89,08%	89,11%	91,11%	91,62%
solução 6	86,64%	87,08%	88,79%	89,94%	91,21%
solução 1	82,82%	84,25%	85,96%	86,99%	87,90%
solução 2	83,40%	85,56%	87,77%	88,27%	92,04%
Taxa Global	90,42%	92,22%	92,83%	94,50%	95,76%

Tabela 8.23: Tabela contendo a taxa de acerto para o MODO 2.

os bancos de dados e o sistema de reconhecimento automático presentes nessa dissertação.

A solução 5 para o MODO 2, apresentada na Tabela 8.23, sempre foi a que obteve melhor resultado. Isso já era esperado, uma vez que a maioria dos dados, quando se misturam as 3 bases de dados, provém da RWC. A Tabela 8.12 já indicava essa tendência.

Já a solução 6 pela Tabela 8.11 indicava ser a melhor solução para as amostras provenientes do banco de dados MIS. É ela também que apresenta a melhor solução para os dados provenientes do banco de dados MUMS[7], o que se pode constatar ao verificarmos a Tabela 8.22 para o MODO 1.

Conforme pode ser observado nas Tabelas 8.22 e 8.23, caso o treinamento ocorra com um número superior a 80% das amostras, independentemente da origem da amostra[8] podemos afirmar que a taxa de acerto será superior a 92 %.

Foram feitas 5 estimativas da taxa de acerto com conjuntos distintos (tanto na composição quanto na quantidade) contendo instrumentos das 3 bases de dados. O percentual da taxa de acerto variou pouco, de cerca de 90,42% a 95,76% para uma variação de 50% a 90% no conjunto de treinamento. Todas as estimativas com percentuais entre as duas medidas apresentaram taxas de acerto em sequência, indicando que a estimativa correta se encontra num valor entre as estimativas dos extremos. Assim, podemos constatar que um aumento de 40% na quantidade de amostras a serem testadas concomitantemente com uma redução de 40% do conjunto de treinamento reduziu menos de 6% na taxa de acerto.

Por outro lado, a base de dados MUMS, que contém o menor número de

[7]Essas afirmações obviamente estão restritas às soluções pesquisadas nessa dissertação, por isso as afirmações tem caráter relativo.

[8]Restringindo a somente uma das 3 bases de dados.

amostras, isoladamente apresentou taxas inferiores. Esse resultado se justifica em parte pelo fato de suas gravações não terem sido feitas em câmaras anecóicas, portanto contendo reverberações[9]. Assim, preferiu-se usar essa estimativa mais rigorosa para representar a capacidade de classificação do conjunto contendo as 3 bases de dados caso não se conheça as características das amostras (com reverberação ou não). Optou-se pelo ponto de 80% das amostras porque atende ao primeiro critério de estar entre os percentuais de 90,42% e 95,76%.

Logo, uma estimativa razoável para a taxa de acerto do classificador, caso ele venha ser treinado com 80% das amostras, é cerca de 92% se o conjunto de amostras a serem testadas contém reverberação; caso contrário a estimativa da taxa de acerto de 94% é mais consistente para classificar os 20 instrumentos.

Como mostrado no Apêndice B, estes índices são coerentes, superando o estado da arte em reconhecimento de instrumentos para o agrupamento contido nesse apêndice.

[9]Inclusive a própria base de dados indica o tempo de reverberação presente nas gravações.

Capítulo 9

Resultados frente às Misturas Instantâneas

Os métodos propostos nesse capítulo visam avaliar o desempenho do classificador quando há superposição temporal, seja entre instrumentos, seja entre estes e ruídos. Para isso, iremos avaliar o impacto da superposição temporal frente à inserção de ruído ou à inserção de sinal interferente.

Um dos problemas que se irá contornar é uma restrição do classificador proposto, qual seja a de ter sido treinado para reconhecer notas isoladas, já que seu aprendizado é baseado em condições idealizadas (equivalentes a câmaras anecóicas). Assim, iremos introduzir alguns blocos que almejam generalizar a aplicação do classificador, contornando em parte tal restrição. De antemão, fica esclarecido que a proposta aqui apresentada é somente uma das possibilidades, sendo possível adotar-se outras arquiteturas.

Primeiramente, deve-se levar em conta que, normalmente, um sinal de áudio gravado é na maior parte produto de configurações polifônicas (gravações contendo diversas fontes). São essas gravações que normalmente apresentam maior interesse para a comunidade científica. O procedimento elaborado nessa dissertação para classificar notas isoladas pode ser facilmente generalizado para músicas monofônicas (gravações com uma única fonte), bastando para isso treinar os classificadores com pequenos trechos obtidos a partir de músicas previamente eleitas (logicamente o vetor de características deverá ser refeito para se ter um melhor desempenho). Assim, o banco de amostras (tanto para teste quanto para treinamento) nessa abor-

dagem deverá ser construído a partir de pequenos trechos retirados de músicas que compõem a base de dados empregada. No entanto, essa abordagem apresenta o inconveniente de exigir um re-treinamento do classificador para esse novo conjunto de amostras, o que acaba por dificultar uma comparação direta entre o resultado obtido nos capítulos anteriores (taxa de acerto) e a robustez do classificador que se pretende avaliar nesse capítulo. Por causa desse problema, essa abordagem mais natural será descartada. Adotaremos uma abordagem diferente, na qual o classificador continuará classificando notas, sendo necessário para isso recorrer a uma etapa de pré-processamento em que se extrairão as notas contidas nas gravações polifônicas ou monofônicas.

Uma possibilidade interessante (embora ressaltemos que não é a única), de se contornar este problema consiste na extração, a partir da gravação polifônica, dos instrumentos em separado (sequências monofônicas). Supondo ausência de conhecimento das amostras de todos os instrumentos, bem como do sistema que efetua a mistura, as técnicas que extraem os componentes são chamadas de técnicas de separação cega de fontes (SCF). Na nomenclatura das técnicas de SCF, "fonte" significa cada componente da mistura[1]. Assim, nesta dissertação, cada fonte identifica-se com um instrumento de uma base de dados.

Desta forma, o problema do reconhecimento de instrumentos em misturas polifônicas pode ser dividido em três etapas:

1. Separação dos instrumentos (sinal polifônico) em sequências monofônicas;

2. Extração das notas de cada sequência monofônica;

3. Classificação das notas de cada sequência monofônica.

O instrumento será identificado por um critério de votação que verifica o instrumento majoritário obtido na classificação de cada nota de uma sequência monofônica. A razão desta abordagem se baseia no fato de que o algoritmo usado na separação das sequências monofônicas é um separador de fontes, portanto espera-se que as sequências de notas sejam formadas majoritariamente por notas provenien-

[1]Um componente da mistura nessa dissertação é uma sequência de notas proveniente de um único instrumento (sequência monofônica).

tes de um determinado instrumento musical (com reduzida interferência de outras fontes).

Conforme se pode observar na descrição das três etapas feitas anteriormente, será necessário empregar sinais polifônicos para simularmos o nosso problema. Uma forma de resolvermos essa questão foi gerarmos artificialmente para cada instrumento musical, a partir das notas do conjunto de teste, sequências monofônicas. Posteriormente, iremos gerar misturas instantâneas dessas sequências monofônicas, de forma que ocorra superposição temporal, construindo assim um sinal polifônico a partir das notas usadas para teste.

A partir deste instante, avaliaremos os problemas inerentes a cada um dos algoritmos referenciados nos itens 1 e 2 e na construção dos sinais polifônicos, sendo que o algoritmo previsto no item 3 foi coberto pelos capítulos precedentes.

Assim, esse capítulo descreverá a construção artificial de um sinal polifônico, o método de identificação de fontes com separador de fontes, o método de identificação de fontes sem separador de fontes, a extração das notas isoladas de uma sequência monofônica e os resultados.

9.1 Construção Artificial de um Sinal Polifônico

Conforme comentado anteriormente, a construção artificial de sinais polifônicos facilita a avaliação do classificador elaborado nos capítulos precedentes. Portanto, não se pretende simular uma música polifônica com todas as suas características, sendo que a motivação para a construção desse sinal (polifônico) nessa dissertação é somente de servir como um mecanismo útil para a introdução de outros sinais simultâneos no tempo. Assim, é possível avaliar a robustez do classificador com relação à superposição temporal. De outra forma, teríamos que enfrentar o problema de detecção das notas de um sinal real polifônico (música polifônica) de maneira bem mais aprofundada, o que foge ao escopo dessa dissertação. Assim, foi utilizado um algoritmo simples para a construção dos sinais polifônicos.

O procedimento usado para se construir um sinal polifônico primeiramente cria sequências monofônicas e depois efetua misturas instantâneas (somando-se diretamente os sinais). O tamanho dessas sequências monofônicas (comprimento) é

proporcional à quantidade de notas que a fonte tem associada no conjunto de teste.

Antes de se somar os sinais monofônicos para obtermos uma sequência polifônica, é necessário verificar se os sinais monofônicos apresentam tamanhos diferentes. Se isso acontecer, deve-se igualar o tamanho deles para que não haja uma facilitação no processo de separação das fontes, porque em determinado instante teríamos uma fonte contribuindo com o sinal (polifônico) enquanto que a outra estaria em silêncio. Tal situação em determinados casos é um facilitador para os algoritmos de separação. Preferiu-se evitar esses casos, uma vez que estamos tentando avaliar a condição em que a superposição temporal afeta o classificador. Assim, optou-se por truncar a sequência monofônica de maior comprimento, de forma que seu tamanho ficasse igual à sequência de menor comprimento. Assim, nas misturas polifônicas, sequências de notas contendo várias fontes terão seu comprimento definido pela sequência monofônica de menor comprimento.

Cada sequência monofônica foi construída usando as notas do conjunto de teste, conforme o procedimento anteriormente descrito de sortear um percentual da região central da tessitura do instrumento musical. Essas notas, pertencentes ao conjunto de teste, foram sorteadas e separadas por intervalos aleatórios positivos escolhidos arbitrariamente variando entre 0,045 ms e 0,3 ms. Procurou-se usar intervalos pequenos para que a sequência das notas tivesse uma maior continuidade auditiva, sem que chegassem a ponto de interromper a nota precedente. Em alguns sinais polifônicos reais foi observada a ocorrência de uma nota posterior interromper a nota anterior, como se o intervalo fosse negativo. Esses problemas não foram tratados nessa dissertação, visto que teriam influência maior no algoritmo de extração das notas do que no classificador. A princípio, a influência da perda da parte final da nota não afeta o classificador, uma vez que o mesmo somente utiliza o segmento intermediário da nota (modelo IMF). No entanto, se a interrupção ocorrer a ponto de se perder parte desse segmento intermediário, poderemos ter erros de classificação para essa nota corrompida.

Apresentamos na Figura 9.1 um sinal polifônico real e um sinal polifônico construído pelo algoritmo acima usando notas de instrumento de sopro para as bases de dados RWC e MIS.

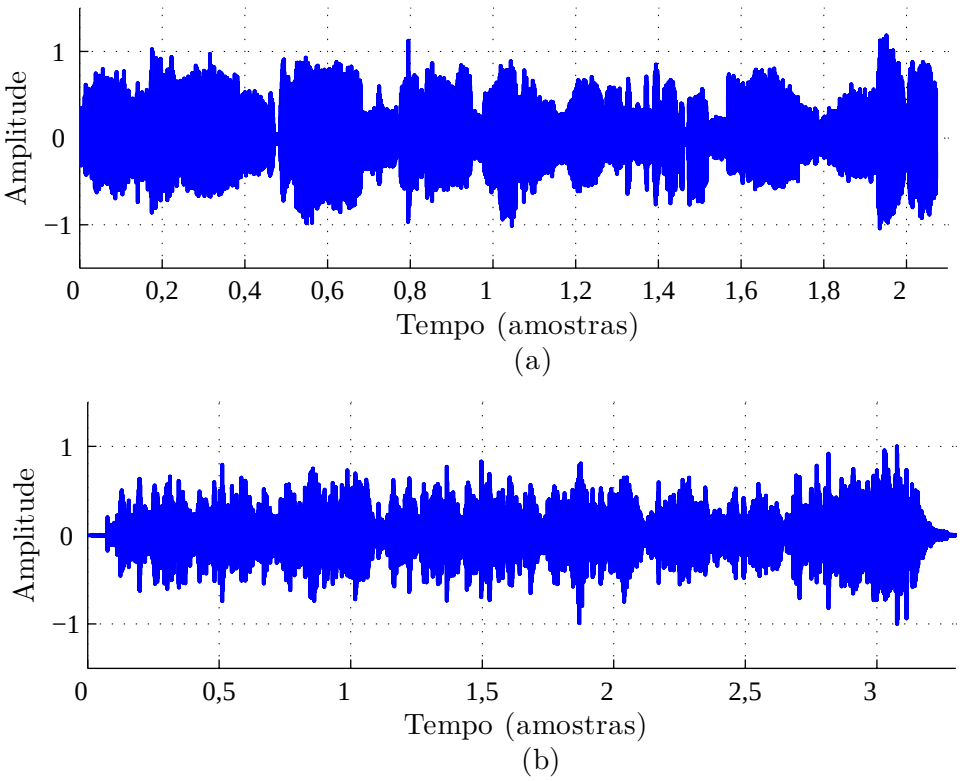

Figura 9.1: Exemplo de um sinal polifônico (a) artificial; (b) real.

9.2 Método de Identificação de Instrumentos com Separador de Fontes

Dada a sequência polifônica artificialmente construída e exemplificada na Figura 9.1, o objetivo desse algoritmo é obter as sequências (monofônicas) originais que serviram para a construção do sinal polifônico.

O caso mais difícil de separação de fontes ocorre quando temos apenas uma mistura. Uma forma de tentar resolver este problema consiste em converter a mistura para o domínio da frequência (via janelamento/STFT) e associar cada raia de cada quadro a uma das estimativas. Esta associação em geral necessita de um conhecimento estatístico acerca das fontes, o qual permite-nos efetuar uma inferência estatística à hora da associação. Duas formas muito difundidas de associação são o mascaramento *hard* (binário) e o *soft* (suave).

Num modelo instantâneo e monocanal, uma raia de um certo quadro da mistura é resultante da soma das fontes escaladas. Supondo a presença de apenas

duas fontes e que apenas uma delas seja a dominante, podemos associar esta raia a esta fonte, impondo que a da outra seja zero (isto deve ser feito quadro-a-quadro). A proposta nessa seção é verificar a viabilidade deste tipo de técnica, supondo que temos acesso às fontes durante a separação.

Este acesso às fontes permite-nos escolher a fonte à qual associaremos uma dada amostra (uma raia num certo quadro) de forma ótima. Isto significa que estamos trabalhando no universo do melhor caso possível que esta técnica de separação de fontes permite. Em outras palavras, analisaremos o melhor desempenho (limitante superior) que o reconhecimento de instrumentos pode ter nesta configuração.

Seja uma amostra $X(f,t)$ da mistura. Conhecendo $S_1(f,t)$ da fonte 1, e $S_2(f,t)$ da fonte 2, é possível estimarmos $\hat{S}_1(f,t)$ e $\hat{S}_2(f,t)$. No mascaramento *Hard* (binário) a estimativa será:

caso $|X(f,t) - S_1(f,t)| < |X(f,t) - S_2(f,t)|$, então $\hat{S}_1(f,t) = X(f,t)$ e $\hat{S}_2(f,t) = 0$;

caso $|X(f,t) - S_2(f,t)| < |X(f,t) - S_1(f,t)|$, então $\hat{S}_1(f,t) = 0$ e $\hat{S}_2(f,t) = X(f,t)$;

No mascaramento *Soft* (suave) é feita uma ponderação pela soma, ou seja, $S(f,t) = |S_1(f,t)| + |S_2(f,t)|$. Nesse caso a estimativa ótima será: $\hat{S}_1(f,t) = [|S_1(f,t)|/S(f,t)] \times X(f,t)$ e $\hat{S}_2(f,t) = [|S_2(f,t)|/S(f,t)] \times X(f,t)$.

Em geral, o mascaramento suave gera resultados melhores que o binário. Para maiores detalhes, vide o penúltimo capítulo constante em [47] e as respectivas referências.

Obviamente, os mascaramentos anteriormente apresentados prestam-se também a casos onde temos mais do que duas fontes. O principal fator que esse algoritmo trará de forma a afetar a taxa de acerto será a distorção que ele provocará nas notas, decorrente de um desembaralhamento das sequências mal efetuado. Assim, a nota quando extraída carregará invariavelmente informação proveniente de outra fonte ou de algum sinal contaminante (ruído ou sinal interferente) na maioria dos casos. O fato de estarmos usando um limitante superior nesse algoritmo significa que não estamos interessados em avaliar o real impacto desse algoritmo no processo de classificação como um todo, mas somente o impacto (para o melhor caso) dele nas notas, a fim de avaliarmos se o classificador se mantém consistente apesar da perturbação

residual que esse algoritmo causará nas notas a serem classificadas.

Na Figura 9.2 apresentamos uma mistura contendo notas provenientes da base de dados RWC.

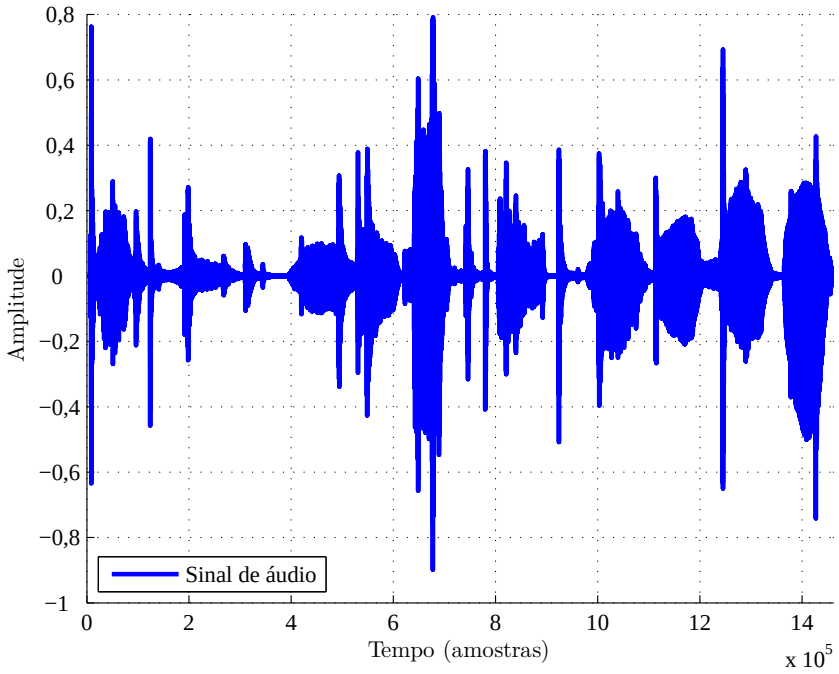

Figura 9.2: Mistura com notas de cordas e percussão.

9.3 Método de Identificação de Instrumentos sem Separador de Fontes

Na prática, o acesso às fontes revela-se uma hipótese demasiado restritiva. No entanto, podemos pensar numa abordagem alternativa sem o uso de separadores de fontes, que consiste em obtermos uma estimativa do ruído do ambiente e usarmos essa mesma técnica para separarmos o ruído (ou interferência) do restante do sinal $X(f,t) = R(f,t) + S(f,t)$. Posteriormente, extrair-se-ão as notas do sinal filtrado $S(f,t)$, classificando o instrumento (fonte) originário correspondente à cada notas. Caso se tenha de antemão a informação do número N de fontes presentes no cenário, pode-se optar por identificar as N fontes como sendo as N mais votadas no processo de classificação.

Essa técnica necessariamente precisará de algoritmos de extração de notas

mais elaborados, visto que a extração se dará diretamente sobre o sinal $X(f,t)$ contendo as superposições temporais ocorridas entre as fontes distintas. Portanto, nessa dissertação não avaliaremos essa técnica da mesma forma que faremos com a abordagem anterior. No entanto, a fim de ilustrarmos o conceito iremos mostrar a viabilidade dessa técnica para misturas com poucas fontes (2, 3 e 5 para a base de dados RWC), já que nesse caso ocorrerá uma menor superposição temporal em função do menor número de fontes.

9.4 Extração das Notas Isoladas de uma sequência Monofônica

O uso de um algoritmo de extração das notas a partir de um sinal polifônico real pode introduzir uma perda significativa na taxa de acerto do classificador, visto que num sinal polifônico real as notas podem não estar espaçadas por intervalos, dificultando a extração. Essa má extração pode dificultar uma avaliação da robustez do classificador. Uma maneira de se contornar o problema seria fazer uma extração fina (supervisionada) de forma a minimizar ao máximo possível o impacto de uma possível má extração da nota na taxa de acerto do classificador. Dessa forma, o classificador não teria sua medida de robustez (em relação à superposição temporal com outros sinais) afetada pelo algoritmo de extração, facilitando assim uma medida posterior que caracterizaria uma perda na taxa de acerto decorrente do uso desses algoritmos de extração de notas em sinais polifônicos reais. Devido à grande quantidade de dados disponíveis para essa dissertação, uma extração fina irá requerer uma grande quantidade de tempo para que se obtenha cada nota. Por outro lado, poderíamos pensar em usar poucas amostras juntamente com uma extração supervisionada, mas isto também não é interessante, pois a medida na taxa de acerto do classificador obtida poderá permanecer viciada pelas características não representativas que essas poucas amostras poderiam reter. Assim, optou-se por uma solução intermediária, ou seja, o uso de uma quantidade significativa de amostras (notas) representadas pelos sinais polifônicos artificiais montados a partir de sequências monofônicas. Estas sequências foram formadas por notas espaçadas, de forma que a extração das notas fosse mais simples, reduzindo a perda na taxa de

acerto do classificador ocasionada pela extração. Portanto, na simulação, ainda se espera que ocorram algumas extrações de notas ruins, de sorte que, caso ocorra uma perda na subida da nota ou na descida da nota, a mesma não afetará a classificação.

Para a obtenção das notas a partir dos arquivos que foram fornecidos (base de dados MIS e RWC) já havia sido elaborado um algoritmo para retirar as notas desses arquivos (trens de notas). Esse algoritmo, derivado do método do desvio, se baseia na análise da envoltória da potência instantânea do sinal, onde o início e o fim de cada nota são detectados ao se passear uma janela previamente definida sobre o sinal e verificar variações do desvio padrão e da média da janela em relação a um valor de referência (1/10 do desvio padrão e da média do sinal, incluindo o ruído de fundo e/ou intervalos entre as notas) medido sobre o sinal inteiro. Esse algoritmo funciona bem na maioria dos casos, requerendo somente algumas adequações para os casos em que o cenário gravado possua uma baixa SNR[2]. Assim, neste caso também foram verificados alguns algoritmos que comumente são sugeridos na literatura. Foram implementados, além do método anteriormente descrito, outros dois métodos (usando a envoltória RMS, e visando a detecção da f_0 ou *pitch*) presentes em [48]. Em ambos os casos, a indicação dos instantes que determinam o início e o fim da nota é feita acompanhando os máximos e mínimos da função derivada do sinal correspondente, a depender do método escolhido. O método final usado nessa dissertação foi uma combinação de dois deles (envoltória RMS e desvio).

O algoritmo de detecção pelo *pitch* se mostrou mais sensível que os anteriores (conforme verificado na Figura 9.3) e por este motivo não foi usado na composição final do extrator da notas. Isto ocorre porque, além do problema da detecção do início e final da nota, ele também mostra uma sensibilidade maior em relação ao algoritmo usado para a estimação do *pitch*, conforme pode ser observado na Figura 9.3. Nesse gráfico foram usados três métodos para estimação do *pitch*: coeficientes cepstrais [49], correlação [22] e acompanhamento da f_0. Esse último não faz a estimação do *pitch*, mas pressupõe que o *pitch* acompanha a frequência fundamental, uma vez que esse componente espectral é aquele que dará a maior contribuição na composição do *pitch*.

[2]É definida pela seguinte equação: $SNR = 10 \times \log_{10}\left(\frac{P_s}{P_r}\right)$, onde P_s é potência do sinal e P_r é a potência do ruído.

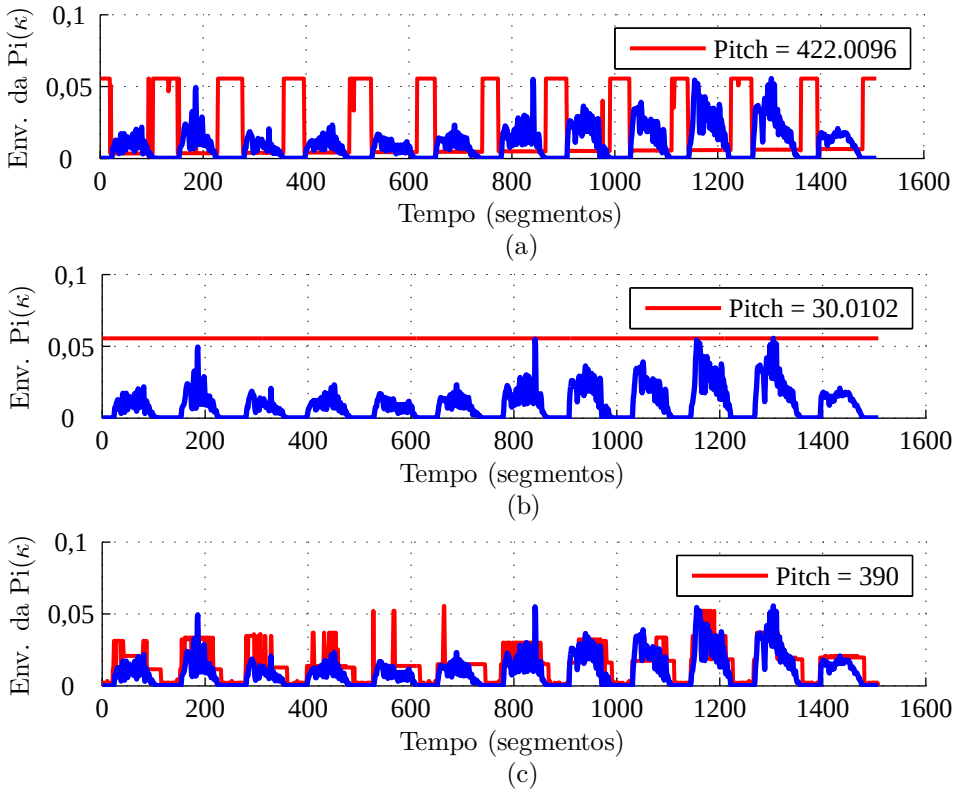

Figura 9.3: Identificação da nota pelo *Pitch*: (a) Método da Correlação; (b) Método dos coeficientes cepstrais; (c) Método de acompanhamento da f_0.

Foi avaliado que a combinação escolhida conseguiu detectar corretamente as notas em cerca de 75% das notas para alguns dos instrumentos que compõem a base de dados MIS (esse índice foi obtido de algumas sequências monofônicas sem contaminações de ruído ou sinal interferente). Não se procurou melhorar o desempenho desse algoritmo, uma vez que não se espera que seja necessária uma detecção perfeita de todas as notas que compõem a música, pois além dos fatores já mencionados garantirem uma certa robustez do classificador, será empregado um método de votação, bastando a princípio que a maioria das notas esteja corretamente identificada para que o classificador consiga determinar corretamente qual foi o instrumento que gerou a música (no caso de estarmos usando um separador de fontes), já que o classificador possui uma boa taxa de acerto (superior a 90%). Portanto, espera-se que os algoritmos que extraem notas corretamente (numa taxa em torno de 60%) sejam suficientemente bons para serem empregados conjuntamente com o classificador desenvolvido nessa dissertação.

9.5 Resultados

Foram escolhidos 11 instrumentos das músicas monofônicas, que serão misturadas para formarem as músicas polifônicas que pretendemos investigar, conforme o padrão abaixo:

1. Instrumentos de Sopro (6):

   Saxofone Contralto, Saxofone Soprano, Trompa, Oboé, Flauta e Fagote.

2. Instrumentos de Cordas (2):

   Violoncelo e Violino.

3. Instrumentos de Percussão (3):

   Glockenspiel, Vibrafone e Xilofone.

Foram usadas notas contidas em duas bases de dados, RWC e MIS. A principal diferença entre as duas bases de dados para essa seção, além do próprio gravador, é o ambiente de gravação. Na base de dados MIS as notas estão espaçadas por um ruído de fundo variável com o instrumento. Já as notas da base de dados RWC estão espaçadas com silêncio (sequência de zeros).

Foram gravadas notas dos 11 instrumentos que estão presentes em ambas bases de dados, exceto os instrumentos de percussão que só existem na base de dados RWC.

A nota de menor *pitch* para as bases de dados empregadas é o Dó de primeira oitava (cerca de 32 Hz); logo, a frequência mínima usada foi de 30 Hz. De posse dessa informação e a de que o algoritmo usado exige que a janela seja divisível por 4, ajustou-se uma janela de 1472 amostras no algoritmo de separação de fontes, já que a taxa de amostragem para ambas as bases de dados é de 44.100 Hz.

Foram gerados arquivos de músicas monofônicas para todos os instrumentos de cada base de dados (14-MIS, e 17-RWC), ou seja, um total de 31 arquivos de músicas monofônicas formadas a partir das notas pertencentes ao conjunto de teste, definido na proporção de 10% para teste e 90% para treinamento.

De antemão, deve-se ressaltar também que se esperam variações nas taxas de acerto em relação às taxas obtidas pelo classificador quando comparamos um instrumento específico, mesmo que a contaminação não tenha afetado o classificador,

porque o conjunto de notas usado, nas músicas monofônicas, é um subconjunto (aleatório) das notas usadas para testar o classificador, além do fato de o algoritmo extrator de notas poder introduzir perdas de algumas notas quando não consegue extraí-las, ou introduzir notas falsas quando faz uma má extração. Essa variação na taxa de acerto tenderá a ficar menor com o aumento de notas usadas. Assim, espera-se também que taxas de acerto que representam totalizações fiquem mais próximas das taxas de acertos globais obtidas pelo classificador quando a contaminação (efeito da superposição temporal) tenha tido pouco efeito.

9.5.1 Análise dos Resultados para Misturas contendo várias Fontes

Nesta seção mostraremos como foram feitas as simulações para avaliar o desempenho do classificador frente às misturas contendo mais de uma fonte. A ideia foi determinar a capacidade do classificador em reconhecer as diversas fontes.

Para cada uma das 19 músicas monofônicas que se pretendia usar (11 da RWC e 8 da MIS) foram elaborados 16 arquivos de músicas polifônicas conforme o padrão a seguir:

Misturas com notas da base de dados MIS:

1. Instrumentos de Cordas (2 fontes): Violino e Violoncelo;
2. Instrumentos de Sopro (6 fontes): Saxofone Contralto, Saxofone Soprano, Trompa, Oboé, Flauta e Fagote;
3. Instrumentos de Cordas e Sopro (contendo todas as fontes acima relacionadas).

Misturas com notas da base de dados RWC:

1. Instrumentos de Cordas (2 fontes): Violino e Violoncelo;
2. Instrumentos de Sopro (6 fontes): Saxofone Contralto, Saxofone Soprano, Trompa, Oboé, Flauta e Fagote;
3. Instrumentos de Percussão (3 fontes): Glockenspiel, Xilofone e Vibrafone;
4. Instrumentos de Sopro e Percussão (9 fontes);

5. Instrumentos de Cordas e Percussão (5 fontes);

6. Instrumentos de Sopro e Cordas (8 fontes);

7. Instrumentos de Sopro, Percussão e Cordas (11 fontes).

Misturas com notas de ambas as bases de dados (MIS e RWC)

1. Instrumentos de Cordas (4 fontes): Violino(2) e Violoncelo(2);

2. Instrumentos de Sopro (12 fontes): Saxofone Contralto(2), Saxofone Soprano(2), Trompa(2), Oboé(2), Flauta(2) e Fagote(2);

3. Instrumentos de Sopro e Percussão (15 fontes);

4. Instrumentos de Cordas e Percussão (7 fontes);

5. Instrumentos de Sopro e Cordas (16 fontes);

6. Instrumentos de Sopro, Percussão e Cordas (19 fontes).

Assim, o número de fontes varia entre 2 e 19. A partir desse ponto foi usado o algoritmo de extração de notas, combinado com o algoritmo separador de fontes (SF). Assim, se espera que após o SF tenhamos sequências monofônicas, pertencentes a fontes distintas. Para cada sequência o algoritmo extrator de notas identificará as notas que compõem a sequência.

Após a obtenção das notas, essas foram codificadas em 3 formas distintas, que comporão os padrões de extração de características de melhor desempenho obtidos no estudo do classificador, conforme mostrado abaixo:

1. Descritores de áudio + 16 coeficientes LSF + desvio padrão (soluções 1 e 6);

2. Descritores de áudio + 16 coeficientes MFCC + desvio padrão + m_3 (solução 2);

3. 16 coeficientes MFCC + desvio padrão + m_3 (solução 5).

O conjunto de notas codificadas foi avaliado pelo classificador após o seu treinamento (90% restante). As Tabelas 9.1 a 9.5 contém os resultados obtidos para ambas as bases de dados usando o método de identificação de instrumento

numa sequência polifônica com o uso do separador de fontes. Os números que aparecem entre parêntesis em algumas dessas tabelas representam os números de notas extraídas.

			MIS		
	Fontes		Tx Acerto	Estimativa	Tx Acerto
1	Violoncelo	Mistura com Cordas	90,2%	Violoncelo	93,0%
2	Violino			Violino	83,3%
1	Saxofone Contralto	Mistura com Sopro	65,0%	Saxofone Contralto/ Fagote	33,3%
2	Saxofone Soprano			Saxofone Soprano	75,0%
3	Trompa			Trompa	50,0%
4	Oboé			Oboé	100,0%
5	Flauta			Flauta	100,0%
6	Fagote			Fagote	50,0%

Tabela 9.1: Taxa de acerto (Tx Acerto) das fontes para misturas polifônicas.

			RWC		
	Fontes		Tx Acerto	Estimativa	Tx Acerto
1	Violoncelo	Mistura com Cordas	92,8%	Violoncelo	97,4%
2	Violino			Violino	86,7%
1	Saxofone Contralto	Mistura com Sopro	52,5%	Saxofone Contralto	57,9%
2	Saxofone Soprano			Saxofone Soprano	37,0%
3	Trompa			Trompa	56,0%
4	Oboé			Oboé	55,0%
5	Flauta			Flauta	47,6%
6	Fagote			Fagote	80,0%
1	Glockenspiel	Mistura com Percussão	94,4%	Glockenspiel	100,0%
2	Vibrafone			Vibrafone	100,0%
3	Xilofone			Xilofone	91,7%

Tabela 9.2: Taxa de acerto (Tx Acerto) das fontes para misturas polifônicas.

			RWC		
	Fontes		Tx Acerto	Estimativa	Tx Acerto
1	Violoncelo	Mistura com Cordas, Sopro e Percussão	63,2%	Violoncelo(12)	100,0%
2	Violino			Violino(12)	75,0%
3	Saxofone Contralto			Saxofone Contralto(14)	50,0%
4	Saxofone Soprano			Fagote(17)	17,7%
5	Trompa			Trompa(18)	61,1%
6	Oboé			Oboé(14)	50,0%
7	Flauta			Flauta(22)	68,2%
8	Fagote			Fagote(8)	50,0%
9	Glockenspiel			Glockenspiel(4)	75,0%
10	Vibrafone			Vibrafone(8)	62,5%
11	Xilofone			Xilofone(23)	87,0%

Tabela 9.3: Taxa de acerto (Tx Acerto) das fontes para misturas polifônicas.

			MIS+RWC		
	Fontes		Tx Acerto	Estimativa	Tx Acerto
1	Violoncelo	Mistura com Cordas	90,4%	Violoncelo(43)	88,4%
2	Violino			Violino(18)	83,3%
3	Violoncelo			Violoncelo(36)	94,4%
4	Violino			Violino(28)	92,9%

Tabela 9.4: Identificação das fontes para misturas polifônicas.

			MIS+RWC		
	Fontes		Tx Acerto	Estimativa	Tx Acerto
1	Violoncelo	Mistura com Cordas e Percussão	87,2%	Violoncelo(14)	85,7%
2	Violino			Violino(15)	73,3%
3	Violoncelo			Violoncelo(10)	90,0%
4	Violino			Violino(11)	81,8%
5	Glockenspiel			Glockenspiel(4)	100,0%
6	Vibrafone			Vibrafone(8)	100,0%
7	Xilofone			Xilofone(24)	91,7%

Tabela 9.5: Identificação das fontes para misturas polifônicas.

As estimativas sombreadas indicam que houve erro na estimação. Observando as Tabelas 9.1 a 9.5 podemos verificar que para até 6 fontes houve estimação correta de todos os instrumentos presentes na mistura[3]. Acima de 6 fontes, dependendo da origem das fontes na composição da mistura, ocorre erro de ao menos uma

[3]Na Tabela 9.1 o sombreamento não identifica propriamente um erro, mas indica que houve uma segunda estimativa para outro instrumento com igual probabilidade que a estimativa correta.

estimativa.

As Tabelas 9.6 a 9.11 contém os resultados obtidos para ambas as bases de dados usando o método de identificação da fonte numa sequência polifônica sem o uso de SF.

			MIS		
	Fontes		Tx Acerto	Estimativa	Ocorrência
1	Violoncelo	Mistura com Cordas	91,3%	Violoncelo(27)	58,7%
2	Violino			Violino(15)	32,6%
3		Outros	8,7%	Outros(4)	8,7%

Tabela 9.6: Identificação das fontes sem separador de fontes.

			RWC		
	Fontes		Tx Acerto	Estimativa	Ocorrência
1	Violoncelo	Mistura com Cordas	79,2%	Violoncelo(49)	48,5%
2	Violino			Violino(31)	30,7%
3		Outros	20,8%	Outros(21)	20,8%

Tabela 9.7: Identificação das fontes sem separador de fontes.

			MIS+RWC		
	Fontes		Tx Acerto	Estimativa	Ocorrência
1	Violoncelo	Mistura com Cordas	89,3%	Violoncelo(26)	46,4%
2	Violino			Violino(24)	42,9%
3		Outros	10,7%	Outros(6)	10,7%

Tabela 9.8: Identificação das fontes sem separador de fontes.

			RWC		
	Fontes		Tx Acerto	Estimativa	Ocorrência
1	Glockenspiel	Mistura com Percussão	92,6%	Glockenspiel(18)	66,7%
2	Vibrafone			Vibrafone(6)	22,2%
3	Xilofone			Xilofone(1)	3,7%
4		Outros	7,4%	Violino(2)	7,4%

Tabela 9.9: Identificação das fontes sem separador de fontes.

Conforme pode ser observado nas Tabelas 9.6 a 9.11, as duas estimativas com maior frequência se mostraram corretas para misturas com até 5 fontes. Eventualmente aparece na coluna "Fontes" o instrumento denominado "Outros" que consiste de vários insturmentos. Em todos esses casos ("Outros"), o instrumento mais votado

			RWC		
	Fontes		Tx Acerto	Estimativa	Ocorrência
1	violoncelo	Mistura com Cordas e Percussão	95,0%	Violoncelo(7)	35,0%
2	violino			Violino(5)	25,0%
3	Glockenspiel			Glockenspiel(1)	5,0%
4	Vibrafone			Vibrafone(1)	5,0%
5	Xilofone			Xilofone(5)	25,0%
6		Outros	5,0%	Saxofone Baritono(1)	5,0%

Tabela 9.10: Identificação das fontes sem separador de fontes.

			MIS+RWC		
	Fontes		Tx Acerto	Estimativa	Ocorrência
1	violoncelo	Mistura com Cordas e Percussão	79,2%	Violoncelo(7)	29,2%
2	violino			Violino(8)	33,3%
3	Glockenspiel			Glockenspiel(0)	0,0%
4	Vibrafone			Vibrafone(1)	4,2%
5	Xilofone			Xilofone(3)	12,5%
6		Outros	20,8%	Outros(5)	20,8%

Tabela 9.11: Identificação das fontes sem separador de fontes.

ficou abaixo de 8,33%. Assim, mesmo que não se tenha um bom algoritmo extrator de notas, é possível usar esse modelo para identificar as fontes de misturas com sequências polifônicas (formadas a partir das bases de dados MIS e RWC) quando as misturas são formadas por apenas duas fontes. Este proceder pode reduzir a complexidade para esses casos, uma vez que dispensa o uso dos algoritmos SF.

9.5.2 Análise dos Resultados para Misturas contendo Sinal Interferente ou Ruído Branco

Nesta seção pretende-se avaliar o desempenho do classificador quando as notas que deverão ser classificadas são contaminadas, seja com sinal interferente ou com ruído branco. Foram gerados os sinais na qual será feita a análise, para todas as músicas monofônicas anteriormente relatadas, com contaminação de ruído branco gaussiano, variando-se a SNR (entre 10 e 26 dB). O mesmo foi feito com sequências monofônicas contaminadas por um sinal interferente, para os valores de SIR (relação

sinal interferência)[4] entre 10 e 20 dB.

Foi escolhida arbitrariamente uma mistura contendo notas de todos os instrumentos musicais (19 fontes distintas) das duas bases de dados para servir de sinal interferente. Assim, o sinal interferente não é um sinal específico de um instrumento particular, o que poderia suscitar dúvidas em relação a um possível favorecimento ou não da interferência sobre o sinal.

A seguir apresentamos o padrão do sinal interferente e do ruído branco usados para uma SIR e uma SNR iguais a 10 dB para um sinal polifônico.

Nas Figuras 9.4 e 9.5, o primeiro gráfico representa o sinal contaminante, o segundo gráfico o sinal original e o terceiro gráfico ilustra o sinal contaminado.

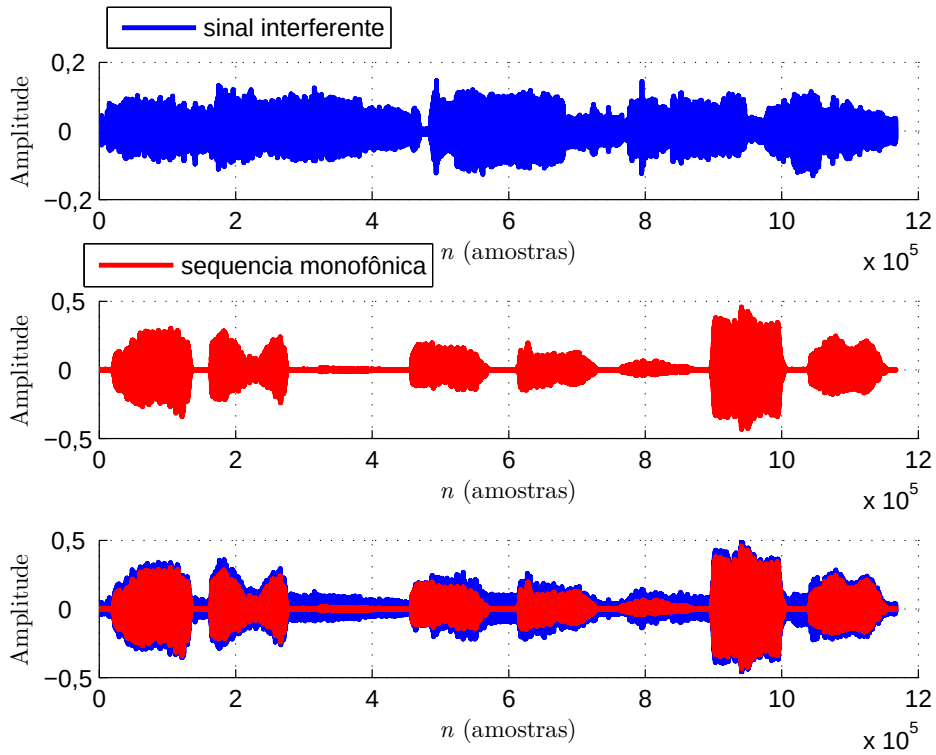

Figura 9.4: Efeito do sinal interferente sobre uma sequência monofônica.

Nas Figuras 9.6 a 9.13 apresentamos alguns resultados, sendo que cada instrumento possui dois gráficos. Um contém o erro sem o separador (curva azul), em que o ruído ou o sinal interferente não foi retirado e o outro contém o erro com separador (curva vermelha), em que o ruído ou o sinal interferente foi retirado.

[4]É definida pela seguinte equação: $10 \times \log_{10}\left(\frac{P_s}{P_i}\right)$, onde P_i a potência do sinal interferente e P_s a potência do sinal.

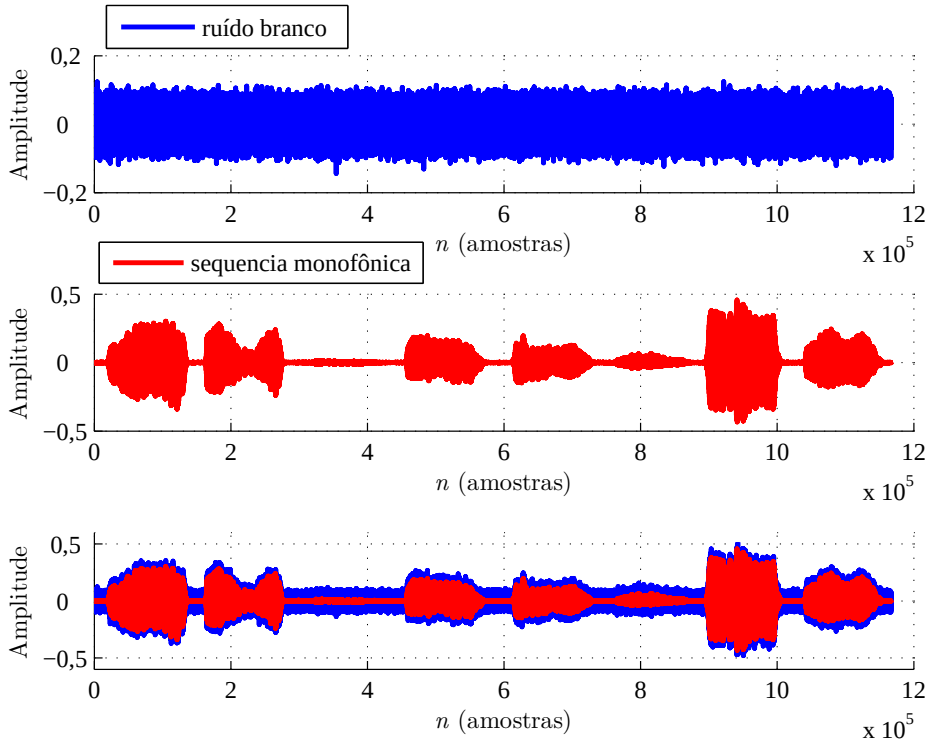

Figura 9.5: Efeito do ruído branco sobre uma sequência monofônica.

Em ambos os gráficos aparecem curvas em verde que representam estimativas do instrumento majoritário para uma dada sequência monofônica, ou seja, se o classificador conseguiu acertar o instrumento (erro=0) ou não (erro=1).

Os valores intermediários, que porventura aparecem no esboço gráfico para essa curva (entre zero e um), representam mudanças na estimativa do classificador em algum ponto do intervalo.

Primeiramente apresentamos os resultados obtidos para a base de dados MIS com sinal interferente polifônico nas Figuras 9.6 e 9.7.

Nas Figuras 9.8 e 9.9, apresentaremos os resultados com a contaminação por ruído branco, para a base de dados MIS.

Conforme pode ser observado nas Figuras 9.6 a 9.9, o classificador se mostrou mais sensível ao ruído branco do que ao sinal interferente. Uma das possíveis explicações para isso é o fato do ruído branco afetar o espectro inteiro, tornando mais difícil a sua separação com o algoritmo SF utilizado. Assim, para contaminações com SIR, mesmo quando não se usou o separador para extrair o sinal interferente da sequência monofônica, o classificador apresentou bons resultados, o que não ocorreu com

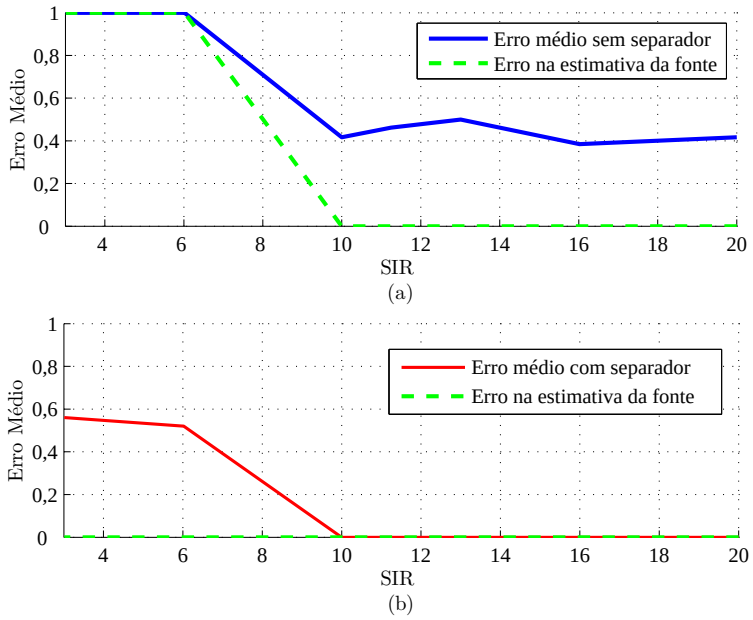

Figura 9.6: Erros estimados na contaminação com sinal interferente: (a) Clarinete Bb sem separador e (b) Clarinete Bb com separador.

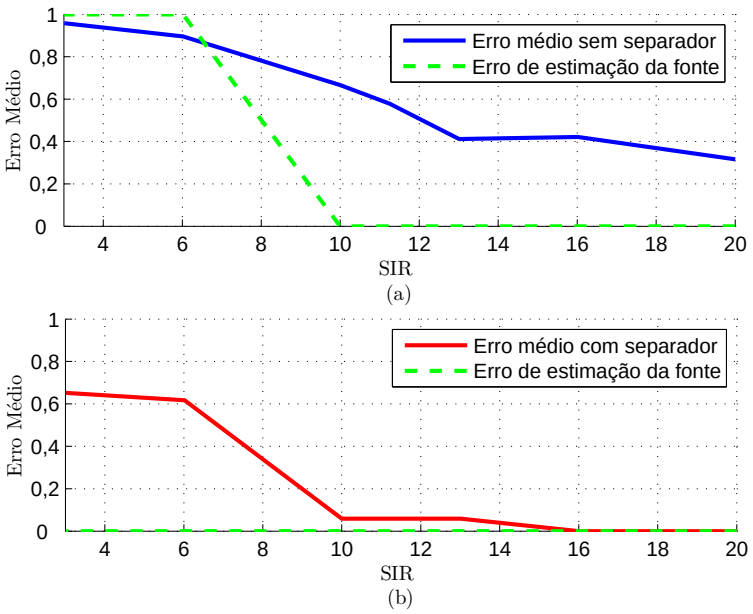

Figura 9.7: Erros estimados na contaminação com sinal interferente: (a) Saxofone Soprano sem separador e (b) Saxofone Soprano com separador.

contaminações de ruído branco, conforme pode-se observar na Figura 9.9. Portanto, o uso do separador é fundamental para melhorar o desempenho do classificador.

Podemos observar que a taxa de acerto varia muito em função do tipo de

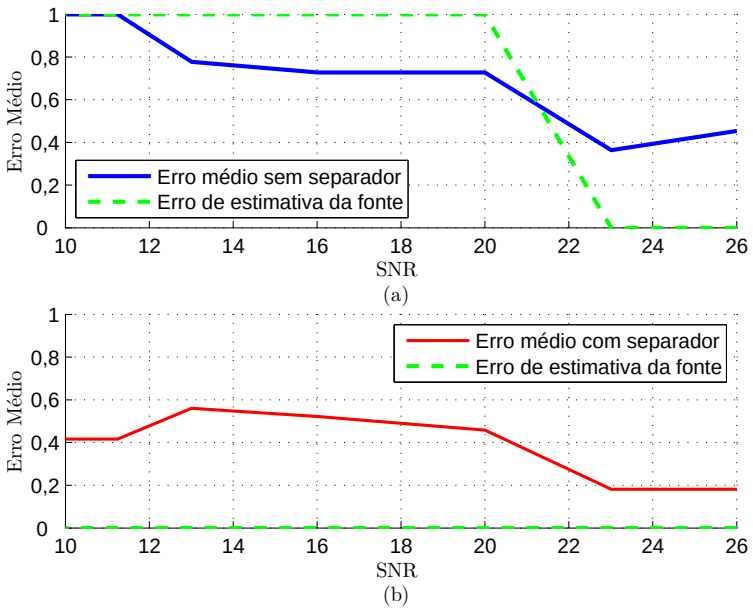

Figura 9.8: Erros estimados na contaminação com ruído branco: (a) Clarinete Bb sem separador e (b) Clarinete Bb com separador.

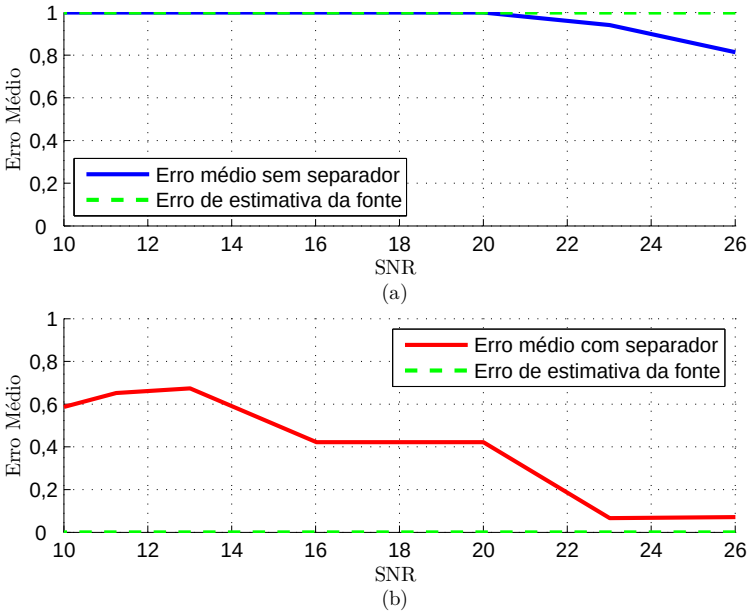

Figura 9.9: Erros estimados na contaminação com ruído branco: (a) Saxofone Soprano sem separador e (b) Saxofone Soprano com separador.

fonte. A seguir apresentamos nas Figuras 9.10 a 9.13, para ambas as bases de dados, o erro total em função da SNR e SIR para todos os instrumentos usados na base de dados MIS e RWC.

O fato de o erro estar acima de 50% não significa que a estimativa estará errada, porque o classificador usa um critério de votação para determinar qual é a fonte daquela sequência monofônica para um dado conjunto de notas classificadas. Assim, em 10 notas, uma votação de três pode representar o mais votado. Isso pode ser atestado na Figura 9.9, que ilustra a contaminação com ruído branco para o Saxofone Soprano com separador.

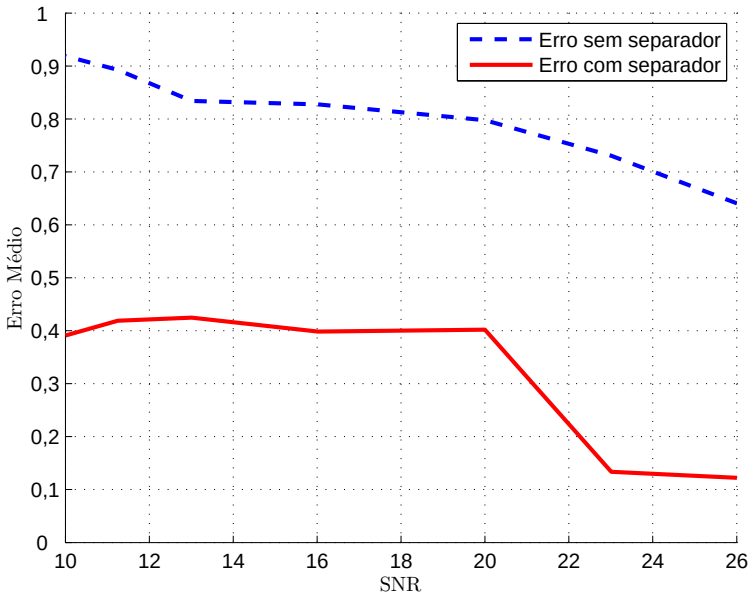

Figura 9.10: Erro do classificador para misturas contaminadas com ruído branco, para amostras provenientes da base de dados MIS.

Sabendo de antemão que a taxa de acerto do classificador é superior a 90%, espera-se um erro residual inferior a 10% para as relações em que a potência do sinal é muito superior a do sinal contaminante.

Independentemente do tipo de sinal contaminante utilizado na sequência monofônica, podemos observar a relação esperada, ou seja, uma dependência proporcional da taxa de acerto com a SIR ou com a SNR.

Flutuações (a princípio inversamente proporcionais à taxa de acerto como ocorre na Figura 9.13 para SNR entre 10 e 12 dB) são decorrentes de variações da quantidade de notas que é testada quando a SNR varia. Assim foi observado que quanto menor é a SNR, maior a dificuldade tanto no algoritmo separador (SF) quanto no algoritmo extrator de notas, ocasionando além da má extração uma grande variação na quantidade de notas que foram identificadas.

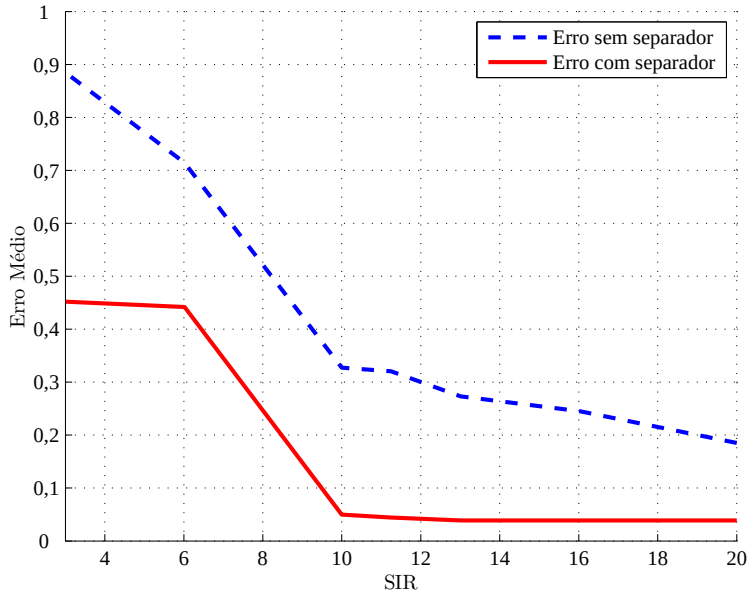

Figura 9.11: Erro do classificador para misturas contaminadas com sinal interferente, para amostras provenientes da base de dados MIS.

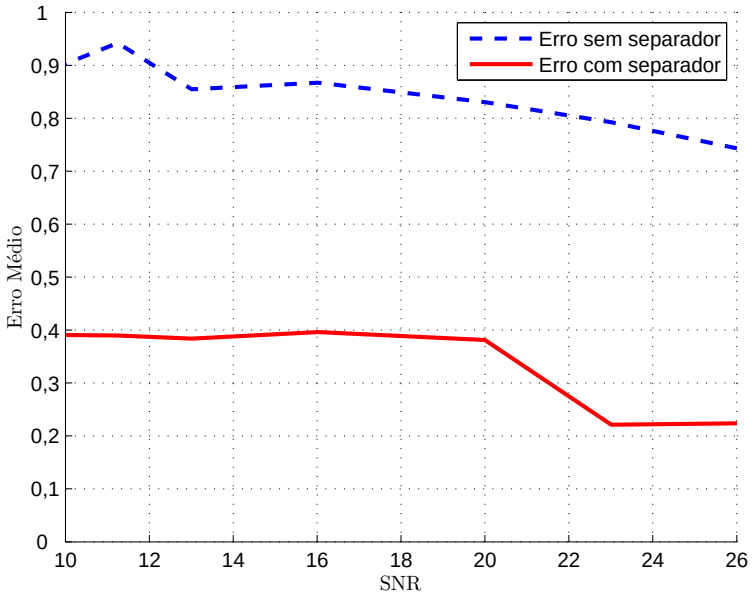

Figura 9.12: Erro do classificador para misturas contaminadas com ruído branco, para amostras provenientes da base de dados RWC.

Além disso, podemos também observar que para as contaminações de ruído branco e sinal interferente, os gráficos apresentam pontos de quebra[5]. Isso indica

[5]Em torno dos limiares de 20 dB para SNR e 6 dB para SIR.

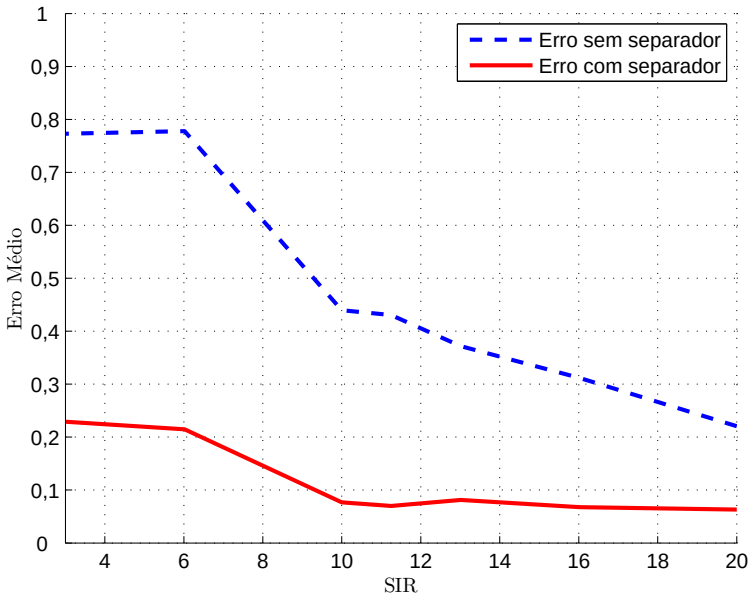

Figura 9.13: Erro do classificador para misturas contaminadas com sinal interferente, para amostras provenientes da base de dados RWC.

que, independentemente dos bons resultados obtidos com o uso do separador para SNR maiores que 20 dB ou SIR maiores que 6 dB, a taxa de acerto passa a ter um salto qualitativo a partir desse ponto.

Já o uso do separador, para todas as SNRs simuladas, não resultou em erro com valor inferior a 10%, que seria o valor esperado caso não houvesse contaminação, indicando que o algoritmo SF teve dificuldades em separar o sinal do ruído branco.

Já em relação à contaminação com sinal interferente, o uso do separador de fontes garantiu uma taxa de erro inferior a 10% desde o início, que é aproximadamente o erro do classificador, explicando porque o erro (com o uso do SF) praticamente não variou com a SIR.

Portanto, o uso do separador mostrou-se eficiente para sinais interferentes, e garantiu uma redução na taxa de erro para em torno de 50 % para ambas as bases de dados quando os sinais são submetidos a ruído branco.

Capítulo 10

Resultados frente às Misturas Convolutivas

Na prática, as misturas do sinal com ruído (ou outros sinais) ocorrem de forma convolutiva sendo que o conjunto dos sistemas de mistura e separação pode ser classificado em 4 topologias básicas: SISO (*single input and single output*), SIMO (*single input and multiple output*), MISO (*multiple input and single output*) e MIMO (*multiple input and multiple output*). Nesse capítulo avaliaremos o classificador quando se depara com misturas convolutivas, segundo o padrão MIMO.

Esse capítulo contém o modelo teórico para misturas convolutivas com duas fontes e dois gravadores, a compensação da distorção causada pela mistura convolutiva, e os resultados obtidos para misturas convolutivas.

10.1 Modelo de Misturas Convolutivas - Duas Fontes e Dois Microfones

Consideraremos em todo o desenvolvimento a seguir, que tanto o número de fontes quanto o de misturas (gravações) é igual a 2. Apesar dessa restrição, esse caso pode servir para separar duas fontes pontuais de sinais (podendo ser uma delas um sinal interferente). Um ambiente acústico altera, por meio de uma filtragem, cada uma das fontes. Um microfone capta os sinais das fontes somados, cada qual distorcido de forma diferente, já que as posições das fontes são distintas. Assim, podemos modelar a distorção de cada fonte por um filtro FIR com comprimento

da ordem de centenas ou mesmo milhares de coeficientes. Assim, o modelo que representa essa situação é ilustrado na Figura 10.1.

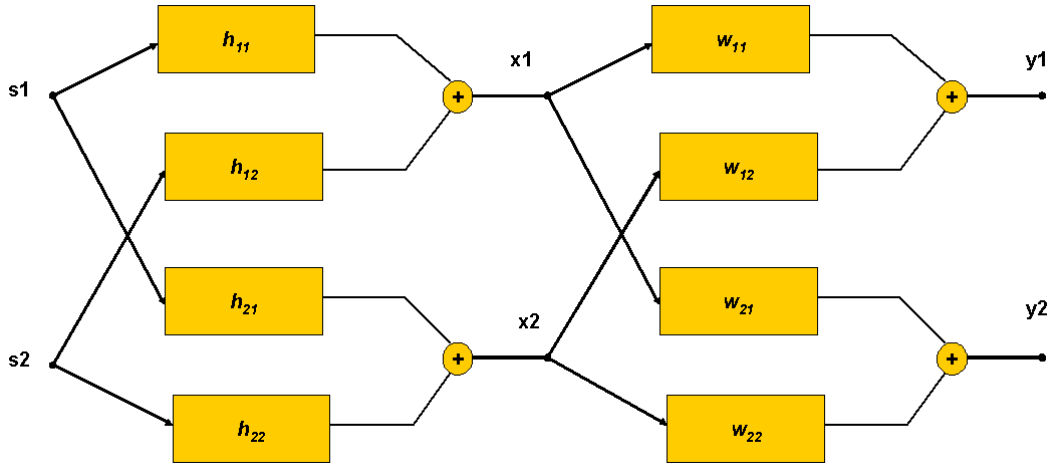

Figura 10.1: Modelo de Separação de fontes.

Na Figura 10.1, $s_1(n)$ e $s_2(n)$ representam as fontes (em nosso caso, sequências monofônicas originadas de instrumentos distintos), h_{ij} é a função de transferência (filtro FIR) entre a j-ésima fonte e a i-ésima mistura (ou gravação), $x_1(n)$ e $x_2(n)$ representam os sinais captados pelos microfones, os quais podem ser expressos como:

$$x_1(n) = h_{11}(n) * s_1(n) + h_{12}(n) * s_2(n) \tag{10.1}$$

$$x_2(n) = h_{21}(n) * s_1(n) + h_{22}(n) * s_2(n) \tag{10.2}$$

onde "*" significa convolução. Num formato matricial, podemos reescrever as Equações (10.1) e (10.2) na seguinte forma:

$$\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} = \begin{bmatrix} h_{11}(n) & h_{12}(n) \\ h_{21}(n) & h_{22}(n) \end{bmatrix} * \begin{bmatrix} s_1(n) \\ s_2(n) \end{bmatrix} \tag{10.3}$$

onde w_{ij} é a função de transferência (filtro FIR) que representa a separação, e $y_1(n)$ e $y_2(n)$ representam as estimativas das fontes obtidas após a separação, podendo ser expressas como:

$$y_1(n) = w_{11}(n) * x_1(n) + w_{12}(n) * x_2(n) \tag{10.4}$$

$$y_2(n) = w_{21}(n) * x_1(n) + w_{22}(n) * x_2(n) \tag{10.5}$$

ou na forma matricial:

$$\begin{bmatrix} y_1(n) \\ y_2(n) \end{bmatrix} = \begin{bmatrix} w_{11}(n) & w_{12}(n) \\ w_{21}(n) & w_{22}(n) \end{bmatrix} * \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} \tag{10.6}$$

Conforme [50, 51], a escolha ideal dos filtros de separação (a menos de eventuais constantes de escalamento) é dada por:

$$\begin{bmatrix} w_{11}(n) & w_{12}(n) \\ w_{21}(n) & w_{22}(n) \end{bmatrix} = \begin{bmatrix} h_{22}(n) & -h_{12}(n) \\ -h_{21}(n) & h_{11}(n) \end{bmatrix} \tag{10.7}$$

10.2 Compensação da Distorção Causada pela Mistura Convolutiva

Nesta seção daremos prosseguimento ao desenvolvimento teórico para misturas convolutivas a partir de duas fontes e dois gravadores, e mostraremos que a solução obtida pela Equação (10.7) insere uma distorção, que deverá ser compensada a fim de que o classificador consiga reconhecer o padrão correto da nota que se pretende classificar.

A solução da Equação (10.7) é de fácil obtenção, bastando para isso substituir $x_1(n)$ e $x_2(n)$ conforme as Equações (10.1) e (10.2) nas Equações (10.4) e (10.5) para termos:

$$\begin{aligned} y_1(n) &= [w_{11}(n) * h_{11}(n) + w_{12}(n) * h_{21}(n)] * s_1(n) \\ &\quad + [w_{11}(n) * h_{12}(n) + w_{12}(n) * h_{22}(n)] * s_2(n) \end{aligned} \tag{10.8}$$

$$\begin{aligned} y_2(n) &= [w_{21}(n) * h_{12}(n) + w_{22}(n) * h_{22}(n)] * s_1(n) \\ &\quad + [w_{21}(n) * h_{11}(n) + w_{22}(n) * h_{21}(n)] * s_2(n) \end{aligned} \tag{10.9}$$

Para que ocorra a separação queremos que $y_1(n)$ seja somente dependente de $s_1(n)$, e $y_2(n)$ seja somente dependente de $s_2(n)$. Portanto, obrigaremos que os termos cruzados sejam zero, ou seja:

$$w_{11}(n) * h_{12}(n) + w_{12}(n) * h_{22}(n) = 0 \tag{10.10}$$

$$w_{21}(n) * h_{11}(n) + w_{22}(n) * h_{21}(n) = 0 \tag{10.11}$$

Podemos observar que as soluções triviais da Equação (10.10) são:
$w_{11}(n) = h_{22}(n)$ e $w_{12}(n) = -h_{12}(n)$ ou $w_{11}(n) = -h_{22}(n)$ e $w_{12}(n) = h_{12}(n)$, e para a Equação (10.11) são:
$w_{21}(n) = -h_{21}(n)$ e $w_{22}(n) = h_{11}(n)$ ou $w_{21}(n) = h_{21}(n)$ e $w_{22}(n) = -h_{11}(n)$,

Assim, combinando as soluções anteriores, temos as seguintes matrizes de soluções triviais:

$$\begin{bmatrix} w_{11}(n) & w_{12}(n) \\ w_{21}(n) & w_{22}(n) \end{bmatrix} = \pm \begin{bmatrix} h_{22}(n) & -h_{12}(n) \\ -h_{21}(n) & h_{11}(n) \end{bmatrix} \tag{10.12}$$

ou

$$\begin{bmatrix} w_{11}(n) & w_{12}(n) \\ w_{21}(n) & w_{22}(n) \end{bmatrix} = \pm \begin{bmatrix} h_{22}(n) & -h_{12}(n) \\ h_{21}(n) & -h_{11}(n) \end{bmatrix} \tag{10.13}$$

Supondo o conhecimento da função de transferência dos canais, podemos implementar a escolha ótima dos filtros. Porém, cabe ressaltar que as estimativas das fontes não equivalem às fontes, mas a versões filtradas das mesmas.

Ao final de um processo de separação bem sucedido, as estimativas das fontes $y_1(n)$ e $y_2(n)$ não apresentam interferência das outras fontes, mas são distorcidas. Para verificarmos isso, basta efetuar a seguinte operação:

$$\begin{bmatrix} w_{11}(n) & w_{12}(n) \\ w_{21}(n) & w_{22}(n) \end{bmatrix} * \begin{bmatrix} h_{11}(n) & h_{12}(n) \\ h_{21}(n) & h_{22}(n) \end{bmatrix} = \begin{bmatrix} h_{22}(n) & -h_{12}(n) \\ -h_{21}(n) & h_{11}(n) \end{bmatrix} * \begin{bmatrix} h_{11}(n) & h_{12}(n) \\ h_{21}(n) & h_{22}(n) \end{bmatrix} \tag{10.14}$$

para constatarmos que:

$$\begin{bmatrix} y_1(n) \\ y_2(n) \end{bmatrix} = \begin{bmatrix} h_{22}(n) * h_{11}(n) - h_{12}(n) * h_{21}(n) \\ h_{22}(n) * h_{11}(n) - h_{12}(n) * h_{21}(n) \end{bmatrix} * \begin{bmatrix} s_1(n) \\ s_2(n) \end{bmatrix} \tag{10.15}$$

Assim, ocorre uma distorção do sinal original nas estimativas obtidas por um fator de distorção $t(n)$ dado por:

$$t(n) = h_{11}(n) * h_{22}(n) - h_{12}(n) * h_{21}(n) \tag{10.16}$$

que é o mesmo em ambas estimativas $y_1(n)$ e $y_2(n)$. Portanto, antes de usarmos o classificador para estimarmos o instrumento a partir de cada uma das sequências monofônicas obtidas após a separação, faremos uma correção nessa distorção causada pelo método de separação. Essa proposta de correção tem o intuito de aumentar a taxa de acerto do classificador. Usualmente tal procedimento é denominado derreverberação ou deconvolução. Após a obtenção das estimativas $y_1(n)$ e $y_2(n)$ determinaremos o fator de distorção $t(n)$.

Assim, combinando as Equações (10.4) e (10.5) com a Equação (10.16) e após aplicarmos a DFT teremos:

$$Y_1(k) = T(k)S_1(k) \tag{10.17}$$

e

$$Y_2(k) = T(k)S_2(k) \tag{10.18}$$

Sabendo que o comprimento de $t(n)$ é igual a M e o comprimento de $Y_1(k) = Y_2(k)$ é igual a N podemos afirmar que o comprimento de $S_1(k)$ e $S_2(k)$ é $N - M + 1$. No entanto, devemos ter o cuidado de observar que as equações acima podem ser reescritas conforme as Equações (10.19) e (10.20):

$$Y_1(k) = DFT(t(n), N).DFT(s_1(n), N), \tag{10.19}$$

e

$$Y_2(k) = DFT(t(n), N).DFT(s_2(n), N), \tag{10.20}$$

onde cada termo tem comprimento N. Definindo o fator de correção $\Gamma(k)$ como:

$$\Gamma(k) = \frac{1}{DFT(t(n), N)}, \tag{10.21}$$

onde a inversão ocorrerá sobre cada elemento do vetor $T(k)$, e, combinando com as Equações (10.19), (10.20) e (10.21), teremos:

$$s_1(n) = FFT^{-1}(\Gamma(k).Y_1(k)), \tag{10.22}$$

e

$$s_2(n) = FFT^{-1}(\Gamma(k).Y_2(k)), \tag{10.23}$$

Pode-se observar que o comprimento obtido para $s_1(n)$ e $s_2(n)$ é N e não $N - M + 1$ como deveria ser. Portanto, reduz-se esse comprimento extraindo as primeiras $N - M + 1$ amostras de $s_1(n)$ e de $s_2(n)$.

Para as simulações foram usados somente instrumentos da base de dados MIS. Foram feitas duas misturas convolutivas, sendo a primeira mistura composta por dois instrumentos de sopro (saxofone contralto e saxofone soprano) e a segunda mistura composta por dois instrumentos de cordas (violoncelo e violino). A disposição dos componentes na sala foi estabelecida arbitrariamente para uma distância de 1 m entre os microfones, com as fontes fixadas nas posições 1 e 2 conforme a Figura 10.2.

A seguir apresentaremos gráficos a partir de um índice normalmente utilizado para comparar sinais de áudio, a razão sinal distorção, dada por:

$$SDR = 10 \times \log_{10} \left(\frac{\sum_{i=1}^{N} |s(i)|}{\sum_{i=1}^{N} |s(i) - y(i)|} \right) \tag{10.24}$$

onde $s(i)$ representa o sinal e $y(i)$ representa a estimativa do sinal. Esta medida é equivalente ao erro RMS numa escala logarítmica, conforme pode ser observado na Figura 10.3.

10.3 Resultados Obtidos para Misturas Convolutivas

Foram usadas funções de transferências reais [50] nas simulações, medidas na sala representada na Figura 10.2.

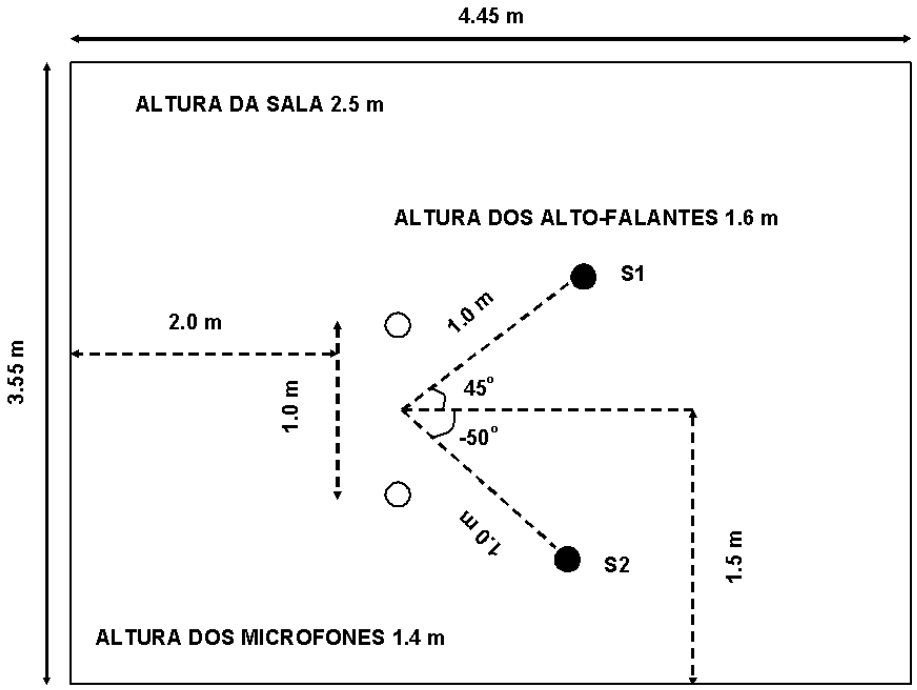

Figura 10.2: Configuração da sala.

O espaçamento entre os microfones foi ora de 5 cm e ora de 1 m. Logo, existem um total de 2 posições de microfone (com espaçamentos de 5 cm e 1 m) e 4 posições das fontes, sendo que um par de fontes ocupa duas posições necessariamente distintas. Portanto, teremos: (2 posições de microfones)×(2×6 combinações de

posições das fontes), perfazendo um total de 24 funções de transferência, uma vez que não ocorre simetria entre as posições (i, j) e (j, i) das fontes em relação aos microfones e das paredes da sala.

Foram simulados 49 pontos, onde cada ponto representa uma totalização de 4 estimativas de fontes (sequência monofônica). Portanto, cada ponto é o resultado da separação de duas misturas convolutivas distintas, uma mistura contendo instrumentos de cordas (violino e violoncelo) e outra contendo instrumentos de sopro (Saxofone Contralto e Saxofone Soprano). Cada sequência monofônica foi prejudicada no processo de separação, devido a uma contaminação com ruído branco feita diretamente sobre os coeficientes da matriz de mistura H[1]. Essa contaminação ocorreu de forma solidária em relação a SNR, ou seja, se um coeficiente foi contaminado com uma dada SNR, todos os demais coeficientes também sofreram contaminações de ruído branco (aleatório) com a mesma SNR. Essa perturbação nos coeficientes da matriz de mistura visa provocar um erro na estimativa dessa matriz. A idéia é identificar a capacidade do classificador de estimar corretamente as fontes, caso ocorra erros na estimação da matriz de mistura.

As contaminações do sinal com ruído branco aleatório foram feitas a partir de 48 níveis de SNR.

Os gráficos apresentados nas Figuras 10.3 a 10.4 foram feitos a partir das duas misturas anteriormente descritas.

Na Figura 10.3 o primeiro gráfico representa o erro de separação estimado com a SDR e o outro o gráfico com a estimação do erro RMS[2](em escala logarítmica).

O algoritmo de derreverberação se baseou na suposição que os coeficientes h_{ij} estivessem corretamente estimados. Erros nessas estimativas propagam os erros na cadeia de manipulações algébricas. Isso ocorre porque essas variações (nos coeficientes) afetam a separação obtida, pois a matriz de separação empregada é função dos h_{ij}. Portanto, a derreverberação ficará também prejudicada, já que tentará compensar a partir desses mesmos h_{ij} os sinais já erradamente separados pela matriz de separação. Uma possibilidade alternativa seria estimarmos diretamente os coeficientes de uma Matriz Inversa (correspondente a Separação + Derreverberação)

[1] Cada elemento da matriz de mistura possui 4.000 amostras.

[2] $E_{rms} = \sqrt{\frac{1}{N} \times \sum_{i=1}^{N} [x(i) - \hat{x}(i)]^2}$

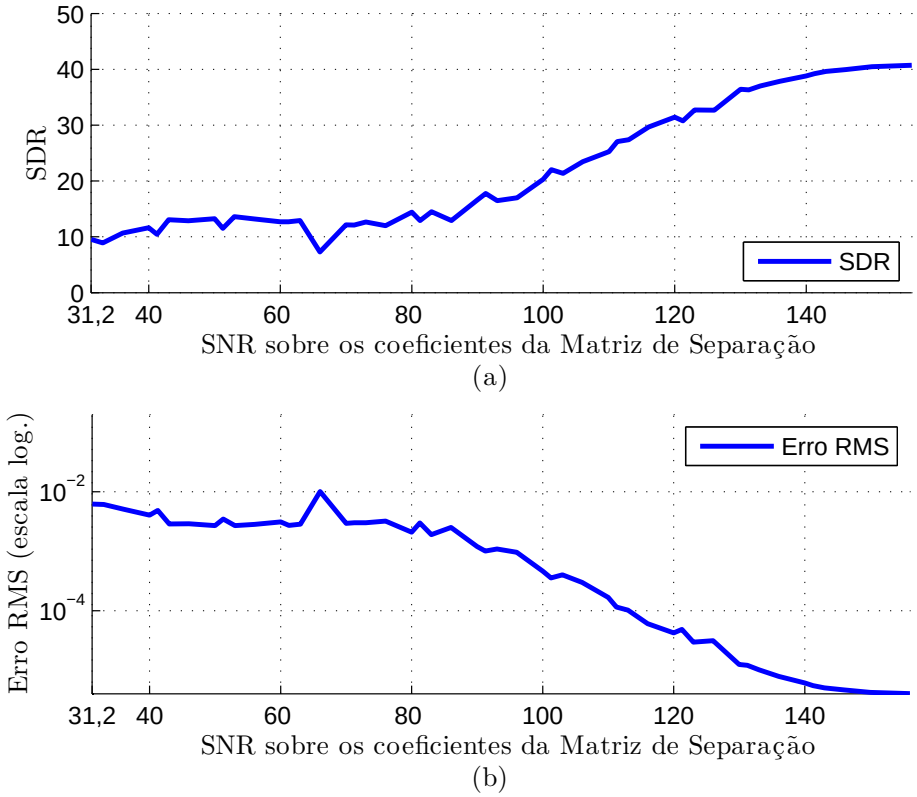

Figura 10.3: (a) Erro RMS da separação (MIS); (b) SDR (MIS).

conforme a proposta presente em diversos trabalhos tal como [52]. No entanto, a tendência foi melhorar a SDR com o aumento da SNR, sendo que, a partir de uma SNR maior que 50 dB a tendência foi estimar corretamente todas as classes dos instrumentos testados.

Houve em alguns pontos da curva, pequenas variações contra a tendência. Isso certamente tem como um dos fatores o fato de terem sido feitas poucas simulações. Conforme já foi dito, cada ponto é representado por duas misturas, onde cada mistura foi equivocadamente separada já que a matriz de separação foi contaminada com ruído branco (um padrão de ruído branco para cada coeficiente da matriz), quando o mais correto seria termos para cada mistura a sua separação perturbada por várias matrizes de contaminação com ruído branco. No entanto, em cada ponto, cada mistura teve sua separação contaminada com uma matriz de contaminação de ruído branco diferente.

Pode-se observar na Figura 10.4 que, até uma SNR em torno de 90 dB (SDR próxima a 20 dB), a separação das fontes evolui pouco com a SNR, e é exatamente

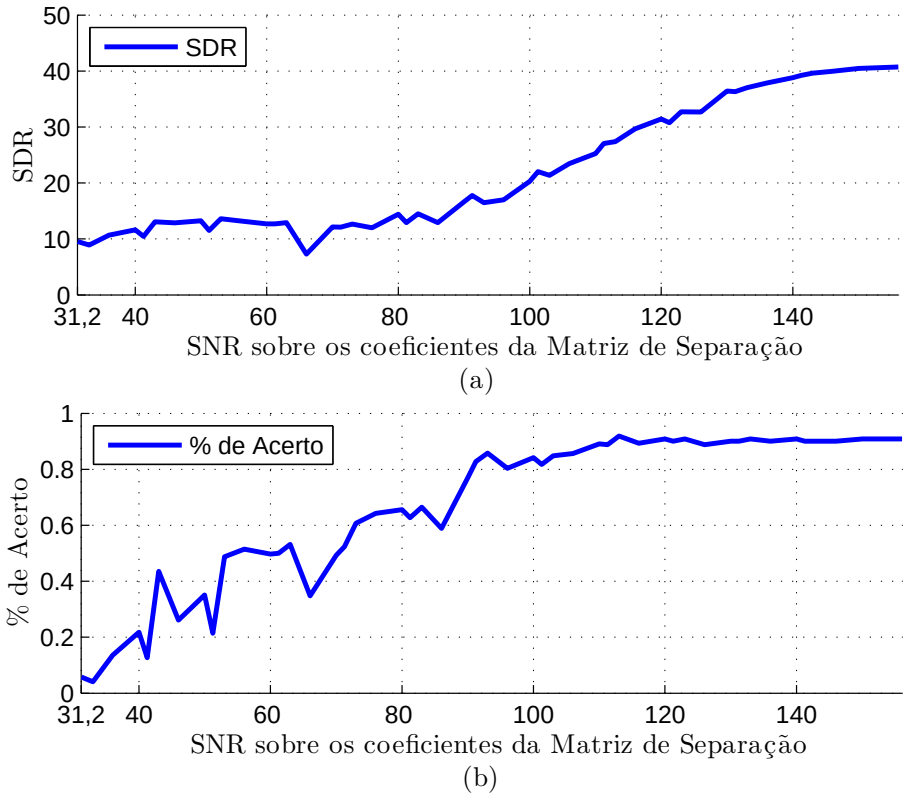

Figura 10.4: (a) Erro de separação (MIS); (b) Taxa de acerto classificação (MIS).

nesse trecho que a taxa de acerto do classificador mais progride com a SNR. A partir desse ponto, o classificador estabiliza a sua taxa de acerto média e passa a ocorrer uma melhora significativa na separação.

A dependência com a separação do classificador ocorre na faixa da SNR em que o classificador não apresenta suas melhores taxas, indicando que nessa região uma melhora ou piora na separação afeta a capacidade do classificador em acertar. A partir desse ponto (SNR próximo a 90 dB, onde o ruído já não afeta tanto o sinal), a melhora na separação não afeta a capacidade do classificador em acertar. Assim, existem duas regiões em relação à SNR quanto a taxa de acerto do classificador: uma em que a taxa de acerto do classificador é dependente da separação, e outra em que a taxa de acerto do classificador não é "dependente" da separação. Essa segunda região representa a região em que as perturbações nos coeficientes h_{ij} não causam perturbações nas notas suficientes para afetar a capacidade do classificador de identificar o instrumento.

Parte IV

Conclusão

Capítulo 11

Conclusão

Essa dissertação apresentou vários estudos, assim como alguns comparativos com trabalhos internacionais, em classificação de instrumentos. Importa ressaltar que os resultados que aqui serão relatados estão limitados às bases de dados que foram empregadas. Portanto, a despeito de os resultados sugerirem um alcance amplo, cumpre ser cuidadoso para não estender em demasia o campo de aplicabilidade das conclusões.

No Capítulo 3 diversos métodos de extração de segmentos de uma nota musical foram descritos. A ideia por trás dessa abordagem é a hipótese de se poder identificar um instrumento musical através de uma codificação sobre uma região particular da nota musical; ou seja, não é imperioso codificarmos aspectos relativos à nota inteira. Esse princípio é comumente encontrado em diversos problemas presentes em nosso cotidiano, como por exemplo, na identificação de pessoas através das digitais, onde não é necessário codificarmos informações sobre o corpo inteiro ao discriminarmos os indivíduos. Assim, em alguns casos, concentrar a extração de informações numa região particular (redução do espaço de codificação) é mais efetivo que obter a mesma quantidade de informações de uma região muito maior, ou aumentar a quantidade de informações de forma a cobrir o corpo inteiro com a mesma densidade[1].

Essa abordagem não foi observada em outros trabalhos encontrados na literatura em identificação de instrumentos musicais possuindo dois fatores que a justificam. Em primeiro lugar, em muitos casos o excesso de informação atua mais como

[1] O que implica um maior custo computacional oriundo do aumento da dimensionalidade.

um agente incrementador da complexidade na busca da solução do que propriamente como um agente viabilizador da solução. Isso ocorre em nosso problema porque de fato diversos algoritmos de reconhecimento são algoritmos de otimização e, portanto, sensíveis aos problemas de dimensionalidade, que podem afetar o seu processo de convergência na busca dos mínimos ou máximos globais. Dessa forma, podemos perceber que o excesso de informação pode ser prejudicial na busca do ótimo. Em segundo lugar, resolver problemas buscando informações numa região menor para a maioria dos casos levará menos tempo do que resolver o mesmo problema buscando informações numa região muito maior. Esses são alguns dos aspectos que diferem o que é necessário descrever (vetor descritor) para uma classificação do que é necessário descrever para uma reprodução.

Assim, a ideia (diferente da usual, que consiste em passear uma janela ao longo da nota e ir extraindo características da nota) é determinarmos somente um segmento da nota que a princípio contenha informação suficiente para discriminar os instrumentos musicais, a partir do procedimento de extração de características previamente estabelecido.

Dessa forma, avaliaram-se algumas formas distintas de se obter segmentos de interesse das notas musicais. Os modelos de segmentação avaliados nessa dissertação podem ser subdivididos em classes: segmentadores por limiar, segmentadores por *pitch* e o modelo ADSR [2]. Nessa dissertação avaliaram-se os seguintes modelos: o modelo ADSR, que segmenta a nota musical nos trechos de ataque, decaimento, sustentação e relaxação; um modelo elaborado para essa dissertação, o modelo IMF, baseado em parte nos modelos TP [19] e no modelo de detecção da nota por limiar; e finalmente um modelo de segmentação baseado na estimação do *pitch*. Todos esses modelos podem ser adaptados para uso com outros sinais, tais como pulso RADAR.

Os segmentadores por *pitch* se mostraram particularmente pouco eficientes para discriminar segmentos dentro da nota, sendo capazes somente de detectar a presença da nota em meio ao ruído de fundo. Portanto, não foram utilizados nos testes do classificador para avaliação de desempenho.

Nem todos os segmentos obtidos foram investigados. O segmento de sustentação do modelo ADSR não foi avaliado visto, que em diversos casos havia notas que não apresentavam esse segmento. Isto ocorreu nas notas dos instrumentos de

cordas com *pizzicato*, em que o padrão temporal é basicamente formado pelo ataque seguido de relaxação, a qual sua envoltória segue um padrão aproximado de uma exponencial decrescente. Assim, foram avaliados os segmentos de ataque do modelo ADSR, de subida do modelo IMF e o intermediário do modelo IMF. Desses segmentos, destacaram-se como bons discriminadores o segmento intermediário com codificação MFCC, LSF e LPC, e o segmento de ataque com codificação MFCC e LSF. Os resultados obtidos para o segmento intermediário com MFCC e LSF sempre apresentaram taxas de acerto superiores às taxas de acerto obtidas pelos segmentos de ataque para seus respectivos codificadores, independentemente dos tamanhos testados. Assim, preferiu-se somente classificar as notas usando segmentos intermediários do modelo IMF.

Um resultado interessante foi a constatação de que variando-se dentro de certos limites o tamanho do segmento intermediário, não ocorre uma tendência na redução da taxa de acerto. Pelo contrário, em alguns casos propicia-se um aumento na taxa de acerto, indicando que a redução do tamanho desse segmento central pode continuar. Logo, muito provavelmente uma tendência na redução da taxa de acerto ocorrerá para algum valor de limiar acima de 90% do valor médio segundo o modelo IMF. Nessa dissertação usamos limiares variando de 10% a 90% do valor médio da potência do sinal. Provavelmente, o valor ótimo desse limiar será variável em função do padrão da nota. Pelo que sabemos esse fenômeno não figura na literatura da área, sendo necessário recorrer a mais simulações em outras bases de dados para ser generalizado.

Alguns trabalhos de reconhecimento citam o trecho de ataque como importante para se extrair características capazes de discriminar fontes distintas [18], dessa forma, vários trabalhos codificam o segmento central juntamente com outros segmentos (parte do ataque e parte da relaxação) [3, 6, 25, 46]. No entanto, os resultados obtidos nessa dissertação propõe que haja uma redução da região que se pretende codificar (segmento central pelo modelo IMF). Assim, para o uso exclusivo do segmento de ataque (ou segmentos iniciais) num sistema de reconhecimento automático de instrumentos musicais, deve-se tomar cuidado com pelo menos três fatores que podem atuar como agentes perturbadores da taxa de acerto:

1. a imprecisão do instante de encerramento do trecho de ataque, o que pode

determinar, para amostras do mesmo instrumento, perdas de características importantes ou acréscimos não pertinentes, dificultando a identificação de padrões;

2. o reduzido número de amostras, em alguns instrumentos, obtidas para caracterizar esse segmento. Tal restrição poderia ser contornada aumentando-se a taxa de amostragem. Isto foi percebido pelo fato de várias notas apresentarem poucas amostras para esse segmento, obrigando a inserção de uma regra de tamanho mínimo (1024 amostras)[2];

3. a escolha do conjunto de características que melhor se presta para diferenciar os instrumentos a partir desse segmento. Nesse último caso, o codificador MFCC se mostrou promissor, abrindo possibilidades de se investigar o uso de mais de um segmento com métodos de extração de características distintas para a formação do vetor de características.

Existem outros métodos de segmentação além dos elaborados ou reportados nesta dissertação. Como exemplo, podemos citar modelos preditivos usando SVM [53]. Cabe enfatizar que não avaliamos todos os tipos de codificadores, tais como: LPCC [3], *Wavelet* [54, 55, 56] e RASTA-PLP [54, 57]. Uma avaliação exaustiva do impacto do uso dos parametrizadores acima sobre o desempenho do sistema classificador seria extremamente laboriosa, haja vista o alto número de combinações envolvidas e as complexas dependências entre os vários estágios do sistema de classificação. Sendo assim, optou-se por utilizar um conjunto limitado de parâmetros. Contudo, foi possível constatar que, para o emprego usado nessa dissertação, os classificadores empregados mostraram uma preferência em ordem decrescente da taxa de acerto para as codificações LSF, MFCC, LPC e CEPSTRUM. Esse resultado também confirma as conclusões de um dos primeiros trabalhos na área [6], acrescentando somente a presença do codificador LSF descrita em [3].

O resultado de Krishna e Sreenivas [3], sugere o uso preferencial do LSF no lugar do MFCC conjuntamente para os classificadores GMM e K-NN, e tem seu

[2]Pode-se argumentar que um aumento na taxa de amostragem não necessariamente incrementará a quantidade de informação que extrairemos do sinal, já que o sinal pode vir a ser limitado em banda; por outro lado, esses segmentos costumam ser de banda larga, e eventualmente essas informações de alta frequência podem ser determinantes para caracterizar o instrumento.

escopo parcialmente confirmado nas soluções finais observadas para o agrupamento MFPC, apesar da forma de extração de características desenvolvida nessa dissertação ser diferente da desenvolvida em [3] (ou seja, para os classificadores K-NN, DLG e SVM, a LSF apresentou resultados ligeiramente superiores aos da MFCC, com as diferenças observadas não sendo tão significativas quanto as constatadas em [3]). Esse resultado final não foi observado em várias outras configurações, quando a MFCC apresentou desempenho um pouco melhor que a LSF. Portanto, pode-se afirmar que ambos os codificadores apresentaram desempenhos equivalentes para a forma de extração e classificação utilizada nessa dissertação.

Essa dissertação não teve a pretensão de enumerar todas as abordagens de classificadores, tampouco todas formas de se obter uma classificação multiclasse. No entanto, a abordagem multiclasse "um-contra-um" obteve resultados significativamente superiores aos da abordagem "um-contra-todos", sendo tal resultado consoante com [6].

Foi também proposta nessa dissertação uma implementação de um classificador DLG, o qual não foi encontrado na literatura em identificação de instrumentos musicais. Este classificador foi idealizado como um misto de alguns conceitos do modelo usado pela SVM (uso de hiperplano separador para discriminar duas classes no espaço de características) e de alguns conceitos presentes em Redes Neurais (similaridade na função custo a ser minimizada). Tal classificador aceita transformações no espaço de entrada, sendo que para se obter a solução multiclasse foram usadas técnicas descritas no Capítulo 6, e que normalmente são usadas para a SVM. Já as transformações no espaço de entrada que o classificador DLG utilizou nessa dissertação foram baseadas em [33].

Ao compararmos os melhores resultados obtidos por cada classificador não se observou uma preferência clara entre os classificadores, sendo que o classificador DLG obteve taxas ligeiramente inferiores que as dos demais (cerca de 2 a 3 pontos percentuais para menos). No entanto, a melhor solução com o codificador LPC foi obtida com o uso do classificador DLG (entre 2 e 3 pontos percentuais para mais). Essa pequena diferença na taxa de acerto entre os demais classificadores e o DLG persistiu também para o agrupamento INSTRUMENTO da base de dados MIS. Essa diferença pode em parte ser explicada pela taxa de acerto obtida para

o conjunto de treinamento. Observamos também que para nenhum caso a DLG conseguiu 100% de separação para o conjunto de treinamento, o que não ocorreu com a SVM. Esse resultado indica que ou a transformação polinomial não foi adequada para conseguir uma separação total dos conjuntos das amostras no espaço das características ou a condição de parada da DLG interrompeu antecipadamente o processo de convergência (por estar lento), ou ambos. Independente de qual foi o principal responsável, a consequência é um hiperplano não posicionado no ótimo para o conjunto de treinamento. Em alguns casos identificou-se que houve uma parada antecipada. Nesses casos, aumentou-se um pouco mais o limite de iterações que determina a parada. Tal procedimento afeta de forma combinatorial o tempo de execução do algoritmo como um todo, já que o problema multiclasse foi tratado usando a abordagem "um-contra-um".

Nas simulações com a envoltória da potência do sinal, verificou-se para todas as bases de dados que quanto maior é a suavização menor é a taxa de acerto. Assim, a envoltória pelo método da média RMS obteve taxas de acerto inferiores às da envoltória pelo método DEAM, o qual por sua vez obteve desempenho inferior ao método da Potência Instantânea do sinal. Isso ocorreu nos três codificadores que foram testados e em todos os agrupamentos para as três bases de dados utilizadas. Entretanto, só se avaliou esse resultado com um classificador K-NN, o que restringe o escopo dessa afirmação. Algumas dessas diferenças (em determinadas famílias, para todas as bases de dados) foram superiores a 50% na taxa de acerto. Esses resultados sugerem que houve perda de informação.

Apresentou-se nesse trabalho abordagens para se obter a classificação de sinais alternativas ao modelo padrão, mostrando que o reagrupamento pode eventualmente apresentar resultados melhores. Da mesma forma, essa comparação entre o modelo padrão e o modelo hierárquico com o modelo de reagrupamento não foi observada em outros trabalhos em indentificação de instrumentos musicais.

Famílias de instrumentos distintas formando agrupamentos hierárquicos com o mesmo número de classes (MFPC, FRBS) foram utilizadas apresentando resultados ligeiramente distintos para o mesmo conjunto de teste e treinamento.

Ao final do procedimento de elaboração dos melhores classificadores, obtivemos quatro soluções (1, 2, 5 e 6) que apresentaram os melhores desempenhos para

a base de dados MIS, usando o agrupamento MFPC, contendo quatro classes representadas por famílias de instrumentos, de um total de 14 instrumentos. Todas obtiveram taxas de acerto iguais ou superiores a 98,8%. Essas soluções foram avaliadas para um novo agrupamento contendo 14 classes (instrumentos), mantendo o excelente desempenho, com as seguintes taxas de acerto: 95,6%, 94,8%, 94,8% e 96,4% para as soluções 1, 2, 5 e 6, respectivamente. Verificou-se que essas soluções mantem o mesmo desempenho para um novo agrupamento contendo 17 classes formadas por amostras de outra base de dados (RWC), obtendo as seguintes taxas de acerto: 86,8%, 94,1%, 95,4% e 89,8% para as soluções 1, 2, 5 e 6, respectivamente. Isso demonstra a capacidade de migração da solução obtida de um agrupamento contendo 4 classes, usando uma determinada base de dados (MIS), para um novo agrupamento contendo 17 classes (instrumentos) com uma outra base de dados (RWC). Cabe observar que a solução 2 e a solução 5 praticamente não apresentaram degradação no desempenho com a mudança de base de dados.

Foi observado que o uso de um banco de classificadores para determinar a classe da amostra apresentou ganhos em relação às soluções individuais dos classificadores para os dois agrupamentos testados (MFPC e INSTRUMENTO). Assim, o classificador proposto nessa dissertação é composto por 4 classificadores, 2 SVM e 2 K-NN, configurados para as soluções 1, 2, 5 e 6 descritas na Tabela 7.29. Essa abordagem de uso de banco de classificadores para identificação de instrumentos musicais também não foi observada em outros trabalhos para esse mesmo propósito.

As estimativas da taxa de acerto para reconhecer os 14 instrumentos da base de dados MIS e os 17 instrumentos da base de dados RWC usados nessa dissertação são de 97,2% e 97,1%, respectivamente, para o classificador proposto em cada uma das bases de dados, e para suas famílias de instrumentos de 99,2% e 99,0%, respectivamente.

A estimativa da taxa de acerto para reconhecer os 20 instrumentos das 3 bases de dados proposta nessa dissertação na forma mais rigorosa (com 100% das amostras para teste com reverberações) é de 92% e na forma menos rigorosa (com 22% das amostras para teste com reverberações) é de 94%. Esse resultado está conforidade com os resultados obtidos na literatura, sendo a taxa de acerto global ligeiramente superior às taxas de acertos apresentadas em trabalhos anteriores. Contudo, tais

comparações com esses trabalhos devem ser atenuadas devido às eventuais diferenças nos testes, como as bases de dados utilizadas e os agrupamentos hierárquicos avaliados.

Nessa dissertação foi também abordado o problema de identificação das fontes em um sinal polifônico a partir de misturas instantâneas. Foram usadas duas abordagens, uma contendo um separador de fontes e outra sem o separador de fontes. O separador de fontes empregado foi otimizado de forma a interferir minimamente na taxa de acerto do classificador. De fato, as simulações avaliaram a capacidade do classificador conseguir classificar corretamente as notas distorcidas pela superposição temporal, a qual não foi possível resolver com separador de fontes otimizado. Para esse caso, verificou-se que misturas contendo até 6 fontes foram estimadas corretamente. Já para misturas sem o separador de fontes verificou-se que misturas com até 5 fontes tiveram as duas estimativas mais votadas corretamente identificadas.

Também se avaliou a robustez do classificador ao contaminar uma sequência monofônica com ruído branco e sinal interferente. Verificou-se que o classificador apresentou uma maior sensibilidade à contaminação com ruído branco do que à contaminação com sinal interferente. O motivo dessa sensibilidade se deve ao fato do ruído branco perturbar significativamente todas as raias no domínio da frequência, para todos os quadros, o que não ocorre com o sinal musical interferente. Essa característica (do ruído branco não possuir uma estrutura harmônica) dificulta a associação das amostras às fontes originárias. Foram avaliadas as curvas da taxa de acerto do classificador em função da SNR e da SIR. Para valores de SNR superiores a 20 dB e SIR superiores a 6 dB, o classificador apresentou excelente desempenho. Esse resultado evidencia a maior dificuldade que o classificador obteve para identificar sinais contaminados com ruído branco.

Posteriormente verificou-se a capacidade do classificador de identificar misturas convolutivas. Percebeu-se que a distorção causada por um algoritmo separador é suficiente para fazer com que o classificador errasse as estimativas. Assim, foi necessário usar um algoritmo de derreverberação para corrigir essas distorções causadas pelo algoritmo separador de misturas convolutivas. Avaliou-se nesse caso uma perturbação com ruído branco nos coeficientes da matriz de separação que serão determinantes para a deconvolução da distorção causada pelo algoritmo separador.

Os resultados indicam que a SNR necessária para que o classificador consiga estimar corretamente as fontes é alta, superior a 90 dB. Esse resultado é dependente do método que foi avaliado nessa dissertação.

No apêndice B foi feita uma comparação direta entre os resultados obtidos em [3] e os resultados obtidos nessa dissertação. A base de dados empregada foi a MIS, e os instrumentos avaliados em ambos os trabalhos foram os mesmos (14 instrumentos). As diferenças residem somente na formação do conjunto de treinamento (que em [3] não foi informado) e nos agrupamentos familiares, que foram alterados no apêndice B de forma a se poder fazer a comparação. As taxas de acerto obtidas para o agrupamento familiar foram de 95% em [3] contra 98,8% nessa dissertação, e para os instrumentos foram de 90% em [3] contra 96,4% obtida pela solução 6 ou 97,2% obtida pelo banco de classificadores formado pelos sistemas de classificação definidos nas soluções 1, 5 e 6, de forma que os resultados obtidos nessa dissertação para esse agrupamento com essa base de dados superaram os melhores resultados observados na literatura para esses mesmos agrupamento e base de dados.

Essa dissertação levanta algumas questões que possivelmente podem resultar em futuros trabalhos nas áreas de segmentação, codificação, classificação e agrupamentos hierárquicos visando um sistema de reconhecimento automático para instrumentos musicais.

Na área de segmentação, não foi observada uma redução significativa da taxa de acerto com a redução do segmento central. Com isso abre-se a possibilidade de um trabalho que consiga caracterizar o tamanho mínimo desse segmento para uma dada nota. Uma característica que se deve levar em conta no modelo IMF é que o segmento central possui parte do segmento de ataque mais o decaimento acrescido de parte do segmento de sustentação (ou, quando esses não ocorrem, parte da relaxação). Também se deve destacar que grande parte das notas com *pizzicato* dos instrumentos de cordas, que não possuem sustentação conforme pode ser observado na Figura 3.7 (envoltória da nota C4 de um violino), tiveram uma alta taxa de acerto. Assim, para todos os casos sempre esteve presente parte do segmento de ataque mais o segmento de decaimento ou parte do segmento da relaxação. Se assumirmos que para os instrumentos que não possuem sustentação a parte inicial da relaxação se confunde em parte com o decaimento, podemos supor que bastaria

codificar parte do ataque mais o decaimento (ou parte da relaxação, caso não exista sustentação) para caracterizar o instrumento musical, uma vez que esses trechos da nota representam o intervalo entre o começo do fim do ataque da nota e o início da sustentação ou relaxação da nota, ou seja, onde o instrumentista teria menor controle sobre a nota. Possivelmente por esse motivo esses trechos carregariam menor contaminação volitiva, e portanto, uma informação mais "limpa" do instrumento musical. Além dessa questão, existe a possibilidade de avaliar-se o desempenho do trecho de relaxação, que não foi abordado nesse trabalho, como também de qualquer outro segmento decorrente de métodos de segmentação que não foram investigados nesse trabalho.

Na área de codificação foi observado que o codificador MFCC apresentou o melhor resultado para o trecho de ataque, enquanto que o codificador LSF apresentou um desempenho ligeiramente superior para o segmento central do modelo IMF. Assim, pode-se investigar em trabalhos futuros a associação desses 2 trechos com essas codificações distintas combinados ou não com a transformada *Wavelet* para a formação de um vetor de características com potencial de discriminação maior. Além dessas questões, pode-se avaliar o uso de outros codificadores tais como LPCC e RASTA-PLP.

Na área de classificação o classificador DLG com uma transformação polinomial para o espaço de características apresentou taxas de acerto superiores a 97% (para o agrupamento MFPC), provando que o desempenho da sua implementação na forma generalizada proposta nessa dissertação para essas bases de dados é comparável à dos demais classificadores empregados nessa dissertação, de forma que o uso de discriminantes lineares na área de classificação de instrumentos musicais não pode ser desprezado. Assim, pode-se pensar em trabalhos futuros usando uma implementação otimizada do DLG com uso de variados *kernels* a fim de reduzir o seu tempo de resposta e de melhorar seu desempenho, uma vez que a transformação para o espaço de característica que a SVM (gaussiana) empregou pode ter favorecido esse último na conquista das melhores soluções. Outra melhoria possível é que essa implementação de DLG permite variações nas funções objetivos, o que pode angariar melhorias no seu desempenho. O classificador DLG elaborado nessa dissertação usou um método de busca de mínimo local baseado no algoritmo LMS normalizado.

No entanto, nada impede que se usem outros métodos de busca de mínimos locais ou globais mais eficientes ou mais rápidos, dependendo da situação. Além dessa questão, observou-se ainda que a maioria dos trabalhos nessa área utilizam GMM ou SVM. Existem alguns casos específicos com Redes Neurais e HMM. Portanto, a abordagem apresentada nessa dissertação pode ser avaliada juntamente com esses demais classificadores.

Com relação aos agrupamentos foi mostrado nessa dissertação que o modelo de reagrupamento difere do modelo padrão. Portanto, em trabalhos futuros pode-se desenvolver novos agrupamentos das amostras dos instrumentos a partir das semelhanças entre elas (medidas por métrica) ou estimadas pela taxa de acerto (por um sistema de reconhecimento previamente definido) para uma dada estratégia, visando reduzir erros de confusão entre instrumentos, e com isso atingir melhores resultados. Nessa busca possivelmente será necessário usarmos clusterizadores.

Com relação ao modelo hierárquico obteve-se no único caso estudado uma configuração capaz de discriminar 100% das amostras da família MFPC tanto em teste quanto em treinamento. Isso foi possível alterando-se para cada nó a solução (vetor de característica mais classificador) o que levanta a questão de se estudar qual a melhor estrutura hierárquica (árvore) para se classificar um agrupamento (conjunto de classes, folhas).

Verificou-se que a presença da reverberação afeta a taxa de acerto. Isso foi verificado tanto na base de dados MUMS (cujas amostras estão contaminadas com reverberação) quanto nas misturas convolutivas em uma sala (com reverberação). Assim, pode-se pensar em usar um conjunto de descritores que sejam menos sensíveis à reverberação, ou em alguma transformação sobre o vetor de caracteristicas já formado, uma vez que contornar a distorção causada pela reverberação é algo extremamente complicado visto que a reverberação pode alterar drasticamente o espectro de forma desconhecida.

Referências Bibliográficas

[1] Anssi Klapuri e Manuel Davy, *Signal Processing Methods for Music Transcription,* Springer, Science+Business Media LLC, 2006.

[2] Hyoung-Gook Kim, Nicolas Moreau e Thomas Sikora, *Introduction to MPEG-7 audio,* John Wiley & Sons Ltd, 2005.

[3] A. G. Krishna e T. V. Sreenivas, *Music instrument recognition: from isolated notes to solo phrases, Proc. of ICASSP,* pp. 265-268, 2004.

[4] Keith D. Martin e Youngmoo E. Kim, *Musical instrument identification: A pattern-recognition approach, 136th meeting of the Acoustical Society of America,* 1998.

[5] Frank Opolko e Joel Wapnick, *McGill University Master Samples,* conjunto com 3 DVDs, disponibilizada pela *McGill University*, Montreal, via http://www.music.mcgill.ca/resources/mums/html/mums.html, 1987.

[6] Janet Marques e Pedro J. Moreno, *A Study of Musical Instrument Classification Using Gaussian Mixture Models and Support Vector Machines, Cambridge Research Labs Technical Report Series CRL/4,* 1999.

[7] Antii Eronen e Anssi Klapuri, *Music instrument recognition using cepstral coefficients and temporal features, Proc. of ICASSP,* pp. 753-756, 2000.

[8] G. Agostini, M. Longari e E. Pollastri, *Music instrument timbres classification with spectral features, Proc. of ICME,* pp. 97-102, 2001.

[9] Teisuro Kitahara, M. Goto e H. G. Okuno, *Music instrument identification based on F0-dependent multivariate normal distribution, Proc. of ICASSP,* pp. 421-424, 2003.

[10] Lawrence Fritts, *Musical Instruments Samples of IOWA University,* MIS, disponibilizada pela *University of IOWA*
via http://theremin.music.uiowa.edu/MIS.html, 1997.

[11] Masataka Goto e Takuichi Nishimura, *RWC Music Database: Music Genre Database and Musical Instrument Sound Database,* disponibilizada pela *National Institute of Advanced Industrial Science and Technology* (AIST), via http://staff.aist.go.jp/m.goto/RWC-MDB/m, ISMIR, pp. 229-230, 2003.

[12] Bohumil Med, *Teoria da Música*, 4.ed. rev. e ampl., Brasília, DF, Musimed, 1996.

[13] Margaret J. Kartomi, *On Concepts and Classifications of Musical Instruments, Chicago: University of Chicago Press,* 1990.

[14] Margaret J. Kartomi, Wikipedia, *Instrumento Musical,* enciclopedia livre, licenciado sob CC-BY-SA , http://pt.wikipedia.org/wiki/Instrumento_musical, maio de 2008.

[15] Erich M. von Hornbostel e Curt Sachs, *Classification of Musical Instruments: Translated from the Original German by Anthony Baines and Klaus P. Wachsmann, The Galpin Society Journal,* vol. 14, pp. 3-29, 1961.

[16] Tiago de Oliveira Pinto, *Som e música. Questões de uma Antropologia Sonora,* Rev. Antropol., vol.44, no.1, 2001.

[17] Hwei P. Hsu, *Teoria e Problemas de Comunicação Analógica e Digital,* 2.ed., Porto Alegre, Bookman, pp. 56, 2006.

[18] Keith Dana Martin, *Sound-Source Recognition: A Theory and Computational Model,* Tese de Ph.D. submetida ao departamento de Engenharia Elétrica e Ciência da Computação, *Massachusetts Institute of Technology*, MIT, 1999.

[19] RAS, Electronic Inteligence, *Introduction to Radar, Signal Interception and EW Databases*, Notas, 1995.

[20] Judith C. Brown, *Calculation of a constant Q spectral transform, J. Acousti. Soc. Am.,* 89, pp. 425-434, 1991.

[21] Judith C. Brown e M. S. Puckette, *An efficient algorithm for the calculation of a constant Q transform, J. Acoust. Soc. Am.,* 92, pp. 2698-2701, 1992.

[22] *UCL Department of Phonetics and Linguistics*, *Introduction to Computer Programming with MATLAB, Lecture 10: Speech Signal Analysis*, disponibilizada via http://www.phon.ucl.ac.uk/courses/spsci/matlab/lect10.html, setembro de 2008.

[23] Peyton Z. Peebles, Jr, *Probability, Random Variables, and Random Signal Principles*, 4.ed., McGraw Hill, New York, 2001.

[24] Adam Kawalec e Robert Owczarek, *Radar Emitter Recognition Using Intrapulse Data, Microwaves, Radar and Wireless Communications*, MIKON-2004, *15th International Conference,* vol. 2, pp. 435-438, 2004.

[25] Jeremiah D. Deng, Christian Simmermacher e Stephen Cranefield, *A Study on Feature Analysis for Musical Instrument Classification, IEEE Transactions On Systems, Man, And Cybernetics-Part B: Cybernetics,* vol. 38, no. 2, 2008.

[26] Lawrence Rabiner e Biing-Hwang Juang, *Fundamentals of Speech Recognition,* Englewood Cliffs, Prentice-Hall, New Jersey, 1993.

[27] Alan V. Oppenheim e Ronald W. Schafer, *Discrete-Time Signal Processing,* Prentice Hall, Englewood Cliffs, New Jersey, 1989.

[28] Marina Bosi e Richard E. Goldberg, *Introduction to Digital Audio Coding and Standards, Kluwer Academic Publishers,* 2.ed., Norewll, Massachusetts, USA, 2003.

[29] Alexandre Leizor Szczupak *Identificação de Notas Musicais em Registros Solo de Violão e Piano,* Dissertação de Mestrado, COPPE/UFRJ, 2008.

[30] Jorge C. Pires Filho, Diego B. Haddad e Luiz P. Caloba, *Classificação de Padrões de Varredura de Radares,* Anais do VIII Congresso de Redes Neurais, vol. 1, 2007.

[31] Antii Eronen, *Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs, Signal Processing and Its Appli-*

cations, In Proceedings. Seventh International, Symposium, vol. 2, pp. 133-136, 2003.

[32] Jorge C. Pires Filho, Diego B. Haddad e Luiz P. Calôba, *Técnicas de Reconhecimento de Padrões aplicadas na Classificação de Varreduras de Radar,* Anais do IX Simpósio Internacional de Guerra Eletrônica, vol. 1, 2007.

[33] Jorge C. Pires Filho, Paulo Antonio Andrade Esquef e Luiz Wagner Pereira Biscainho, *Classificação Automática de Sons de Instrumentos Musicais usando Discriminantes Lineares,* 6o Congresso da AES Brasil, 12a Convenção Nacional da AES Brasil, pp. 112-118, 2008.

[34] Ian Kaminskyj e Tadeusz Czaszejko, *Automatic Recognition of Isolated Monophonic Musical Instrument Sounds using kNNC, Journal of Intelligent Information Systems,* vol. 24, no. 2-3, pp. 199-221(23), 2005.

[35] Richard O. Duda, Peter E. Hart e David G. Stork, *Pattern classification,* John Wiley & Sons, Inc, New York, 2000.

[36] John R. Deller, John H. L. Hansen e John G. Proakis, *Discrete-Time Processing of Speech Signals,* IEEE Press, 2000.

[37] Simon Haykin, *Neural Networks a Comprehensive Foundation,* Prentice Hall, 2a. Ed., 1999.

[38] Vladimir Vapnik, *The Nature of Statistical Learning Theory,* Springer, New York, 1995. ISBN 0-387-94559-8.

[39] Steve R. Gunn, *Support Vector Machines for Classification and Regression, Technical Report - Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science,* Southtampton University, 1998.

[40] Simon Haykin, *Adaptive Filter Theory,* Prentice Hall, New Jersey, 1996.

[41] Nachman Aronszajn, *Theory of Reproducing Kernels, Transactions of the American Mathematical Society,* vol. 68, no. 3, pp. 337-404, 1950.

[42] Grace Wahba, *Spline Models for Observational Data,* SIAM, 1990.

[43] Nancy Heckman, *The theory and application of penalized least squares methods or reproducing kernel Hilbert spaces made easy, UBC Statistics Department Technical Report,* no. 216, 1997.

[44] E. Osuna, R. Freund e Fredrico Girosi, *Training support vector machines: An application to face detection, In Proc. IEEE Conference on Computer Vision and Pattern Recognition,* pp. 130-136, IEEE Press, 1997.

[45] Christopher J.C. Burges, *A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2*, pp. 121-167, 1998.

[46] Antti Eronen, *Automatic Musical Instrument Recognition,* Dissertação de Mestrado, Departamento de Tecnologia da Informação, Tampere University of Technology, 2001.

[47] Diego Barreto Haddad, *Propostas para Separação Cega e Supervisionada de Fontes,* Dissertação de Mestrado, COPPE/UFRJ, 2008.

[48] Tairone Magalhães et al, *Segmentação automática de sinais musicais monofônicos para análise da expressividade,* XVIII Congresso da Associação Nacional de Pesquisa e Pós-Graduação (ANPPOM), 2008.

[49] Michael Noll, *Short-time Spectrum and Cepstrum Techniques for Vocal-Pitch Detection, The Journal of American Society of Acoustic.,* vol. 36, no. 2, pp. 296-302, 1964.

[50] Herbert Buchner e Walter Kellerman, *A Fundamental Relation Between Blind and Supervised Adaptive Filtering Illustrated For Blind Source Separation and Acoustic Echo Cancellation,* HSCMA, pp. 17-20, 2008.

[51] Shoji Makino e Te-Won Lee, *Blind Speech Separation,* Series: Signals and Communication Technology, Sawada, Hiroshi (Eds.), XV, p. 432, 2007.

[52] Isaac Osunkunle e Sayed ali Shekarchi, *A survey on methods for blind acoustic dereverberation,* Blekinge Institute of Technology/(TEK), 2007.

[53] Manuel Davy e Simon Godsill, *Detection of Abrupt Spectral Changes Using Support Vector Machines an Application to Audio Signal Segmentation, European research project MOUMIR,* http://www.moumir.org.

[54] H. Hermansky, *Perceptual linear predictive (PLP) analysis of speech, J. Acoust. Soc. Am.,* vol. 87, no. 4, pp. 1738-1752, 1990.

[55] C. Pruysers, J. Schnapp e Ian Kaminskyj, *Wavelet Analysis in Musical Instrument Sound Classification, 8th Int. Symp Signal Proc. & Applns, University of Wollongong,* pp. 1-4, 2005.

[56] Nishan Canagarajah, *Instrument Recognition Based Wavelet Packet Tree in Audio Feature Extraction, University of Bristol, Digital Music Research Group, in the Proceedings of International Symposium on Musical Acoustics,* (ISMA'2001), pp. 465-468, 2001

[57] H. Hermansky e N. Morgan, *RASTA processing of speech, IEEE Trans. on Speech and Audio Proc.,* vol. 2, no. 4, pp. 578-589, 1994.

Parte V

Apêndices

Apêndice A

Banco de Dados de Instrumentos

Foram utilizados no desenvolvimento deste trabalho de reconhecimento de instrumentos musicais três bancos de dados de instrumentos: um fornecido pela Universidade de Iowa [10], outro fornecido pela Universidade McGill [5] e o último uma base de dados japonesa denominada *RWC Music Database* [11]. Nem todas as amostras contidas em cada banco de dados foram utilizadas, ou seja, foram utilizados nessa dissertação somente alguns dos instrumentos, tendo havido preferência para os instrumentos que aparecem em todas as bases de dados. No entanto, especificamente em relação à base de dados MIS, a escolha do subconjunto de instrumentos musicais foi feita com o intuito de avaliar o desempenho do algoritmo proposto em relação a um outro trabalho internacional (desenvolvido por Krishna [3]). Somente a partir desse subconjunto é que buscou-se os instrumentos equivalentes nas demais bases de dados. Posteriormente, acrescentamos mais alguns instrumentos das outras bases de dados, principalmente os instrumentos de percussão, visto que os mesmos não existem na base de dados MIS.

A.1 Banco de Dados de Instrumentos MIS-IOWA

As amostras de Instrumentos Musicais da Universidade de Iowa (UIowa)[10] foram criadas em 1997 por Lawrence Fritts, Diretor da *Electronic Music Studios and Associate* e professor de composição da mesma Universidade. Os sons dos instrumentos musicais foram gravados em uma câmara anecóica no *Johnson Speech and Hearing Center* na Universidade de Iowa com os seguintes equipamentos: Microfone

Neumann KM 84; *Mixer Mackie 1402-VLZ*; Gravador *DAT anasonic SV-3800.*

As gravações foram transferidas digitalmente para um *Macintosh Power PC 8500* através de uma interface *Digidesign Audiomedia III* (1997-1999) e para um *Macintosh G4* por uma interface digital *Digidesign Digi-001* (2000-). Posteriormente foram editadas para arquivos de áudio utilizando a ferramenta *Digidesign Sound Designer II* (1997-1999) ou a *Bias Peak* (2000-). Para cada instrumento musical selecionado foram gravados (cobrindo toda a tessitura do instrumento) escalas cromáticas com três níveis dinâmicos não normalizados (pp, mf, ff, ou seja, pianíssimo, mezzo forte, e fortíssimo), feitas em mono, 16 bit, 44,1 kHz, e formato AIFF. A única exceção é o piano cujo som foi gravado em um pequeno estúdio (ambiente não anecóico) na forma *stereo.*

Cada nota tem aproximadamente 2 segundos de duração e é imediatamente precedida e seguida de um intervalo. Quando pertinentes, diferentes estilos de execução e recursos expressivos foram gravados (por exemplo: amostras de sons de violino, viola e violoncelo com ou sem vibrato além de execução com arco em *pizzicato*).

Na Tabela A.1 são apresentados todos os instrumentos contemplados por este banco de dados.

Instrumento	Arquivo	Notas
Alto Flute	11	99
Alto Saxophone	18	192
Bass Clarinet	12	139
Bass Flute	10	102
Bass Trombone	12	131
Bassoon	15	122
Bb Clarinet	13	139
Bb Trumpet	24	212
Cello	77	668
Double Bass	69	571
Eb Clarinet	13	119
Flute	22	227
French Horn	12	96
Oboe	12	104
Piano	259	259
Soprano Saxophone	24	192
Tenor Trombone	12	99
Tuba	9	111
Viola	27	257
Violin	71	601

Tabela A.1: Tabela de instrumentos da base de dados MIS.

As amostras estão organizadas em arquivos separados por nível dinâmico, que armazenam uma nota ou um conjunto de notas de um dado instrumento em ordem crescente de *pitch*.

As Figuras A.1 e A.2 ilustram a sequência de notas do instrumento Flauta Contralto (do arquivo AltoFlute.ff.C4B4) e Trompa (do arquivo Horn.ff.C4B4).

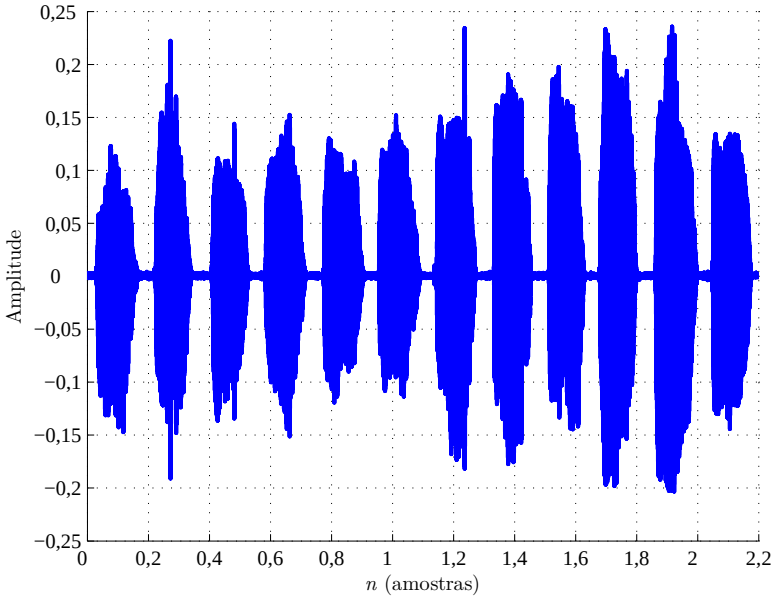

Figura A.1: Flauta Contralto C4-B4.

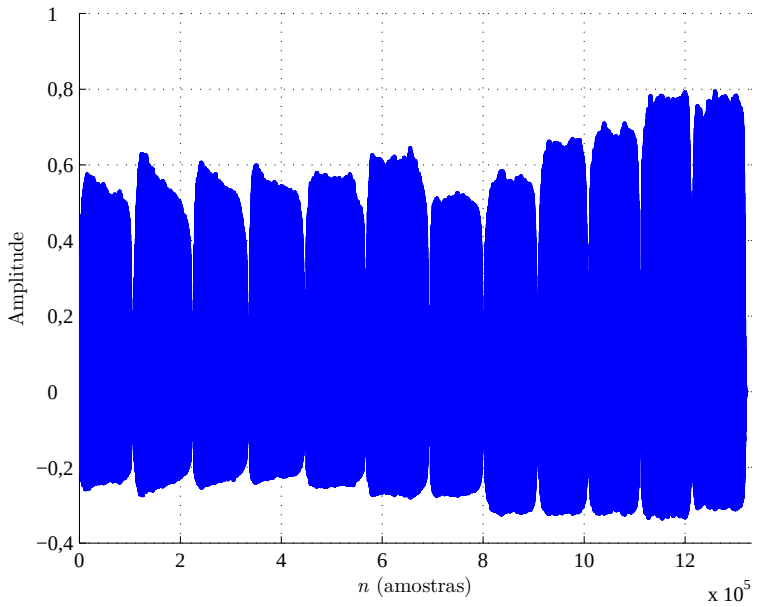

Figura A.2: Trompa C4-B4.

A.2 Banco de Dados de Instrumentos MUMS

A maioria dos sons da biblioteca MUMS (*McGill University Master Samples*) [5] foram gravados diretamente de um gravador DASH Sony PCM 3202, em um estúdio de gravação preparado para ser acusticamente neutro e tendo um tempo de reverberação de aproximadamente 0,4 segundos. Já os instrumentos de cordas e o

piano foram gravados em um *concert hall* tendo o tempo de reverberação variando de 2,5 a 5 segundos. As amostras foram gravadas com uma taxa de amostragem de 44,1 kHz com 24 bits e possuem um tempo de duração variando entre 2 s e 10 s.

Todas as amostras desta biblioteca são *stereo*, e diferentemente das bibliotecas MIS e RWC (onde cada arquivo apresenta um trem de nota), os arquivos desta base de dados apresentam somente uma nota e se encontram no padrão ".wav", sendo armazenados em 3 DVDs.

Existem 6546 amostras de som nessa biblioteca, sendo divididos em 2204 para cordas, 1595 para teclado, 1197 para madeira, 1087 para percussão e 463 para metais.

Esta biblioteca apresenta uma quantidade numerosa de instrumentos musicais, perfazendo um total de 211 instrumentos. Assim, uma lista completa de seus instrumentos musicais pode ser encontrada em [5]. No entanto, o número de amostras por instrumento é inferior ao das demais bases de dados usadas nessa dissertação. Devido a esse fato, para cada gravação, as notas dos dois canais foram usadas nas simulações.

A seguir na Figura A.3 apresentamos a nota A4 de um Saxofone Contralto nos dois canais.

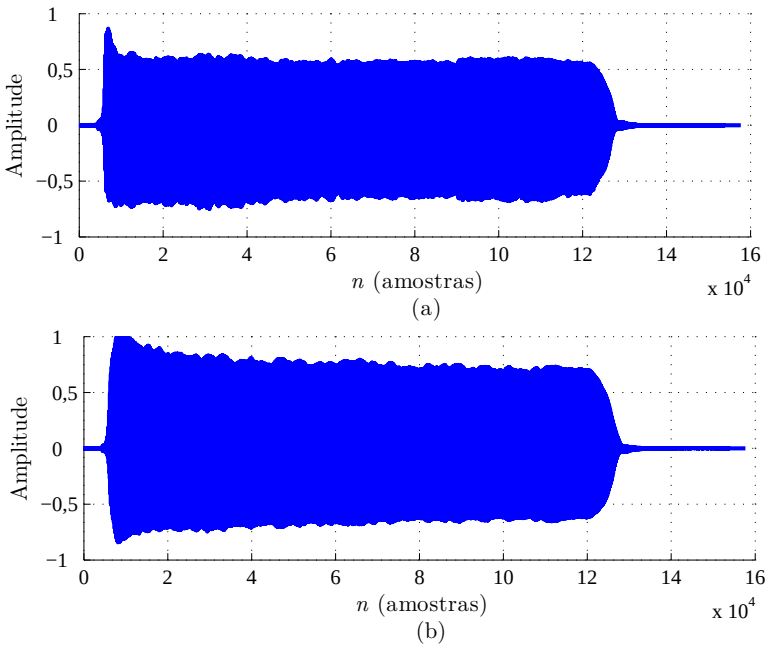

Figura A.3: Saxofone Contralto - A4. (a) canal 1; (b) canal 2.

A.3 Banco de Dados de Instrumentos RWC

A base de dados da RWC *Real Word Computing*, *RWC Music Database* [11] é um projeto japonês, formada por 6 bases de dados, a saber: *Popular Music Database*, *Royalty-free Music Database*, *Classical Music Database*, *Jazz Music Database*, *Music Genre Database* e *Musical Instrument Sound Database*. O pacote usado nessa dissertação é o *Musical Instrument Sound Database*, composto por 50 instrumentos distintos e armazenados em 12 DVDs. As amostras foram gravadas em 44,1 kHz, com 16 bits em formato monoaural.

Ela fornece, a princípio, 3 variações para cada instrumento, totalizando cerca de 150 performances de instrumentos de música, sendo também caracterizada por 4 tópicos, a saber:

1. **As Variações** são decorrentes de gravações oriundas de 3 fabricantes de instrumentos, sendo estes instrumentados por até 3 músicos distintos. Assim, cada variação é caracterizada, em princípio, por um instrumento de um diferente fabricante tocado por um diferente músico. Cada profissional empregado teve em média 17 anos de experiência assegurados para cada instrumento musical. Entretanto, para alguns instrumentos musicais, foi incluída uma variação a partir de um outro tipo de instrumento musical;

2. **Estilos de execução** (dependentes do instrumento). Muitos estilos de execução foram registrados, dentro da gama de possibilidades para cada instrumento. No entanto, para instrumentos de percussão (RWC-MDB-I-2001 No 40-44), cada tipo foi quebrado dentro dos específicos instrumentos e contabilizado como estilo de execução por conveniência (e gravado para cada um destes múltiplos estilos de execução);

3. ***Pitch***. Para cada estilo de execução de um instrumento, o músico geralmente tocou sons individuais com intervalos de meio tom sobre a faixa inteira de tons que poderiam ser produzidas pelo instrumento. Para instrumentos de corda, a faixa total de sons foi obtida para cada corda.

4. **Nível Dinâmico** (Três níveis dinâmicos). Cada estilo de execução de um instrumento foi também gravado com 3 (três) níveis dinâmicos (forte, mezzo,

piano) sobre a faixa total do instrumento em questão.

As gravações dos arquivos seguiram o procedimento de agrupar os sons individuais em ordem crescente de *pitch* sobre a faixa total do instrumento (tessitura), inserindo um intervalo de silêncio (*gap*) entre sons individuais e adjacentes. Dessa forma é facilitado o procedimento de segmentação das notas, bastando para isso um simples detector de silêncio. O nome de cada arquivo é formado por oito caracteres com uma extensão ".wav". Estes oito caracteres consistem em:

1. dois dígitos para o identificador do instrumento musical;
2. um dígito para a variação;
3. dois caracteres para uma abreviação do instrumento;
4. dois caracteres para identificação do estilo de execução;
5. um caractere indicando o nível dinâmico.

Um total de 3544 arquivos compõe a totalidade dos instrumentos dessa base de dados, ocupando um espaço de cerca de 29,1 Gbytes e um tempo total de gravação de cerca de 91 horas 37 minutos e 38 segundos (incluído os intervalos de silêncio).

Na Tabela A.2 apresentamos uma lista com os 50 instrumentos contemplados por essa base de dados conforme sua denominação em inglês:

Apresentamos na Figura A.4 o arquivo 261ASNOF.WAV, o qual contém as notas em toda a faixa do instrumento Saxofone Contralto sem vibrato no nível dinâmico forte.

Inst. No.	Instrument symbol	Instrument name	Inst. No.	Instrument symbol	Instrument name
1	PF	Pianoforte	25	SS	Soprano Sax
2	EP	Electric Piano	26	AS	Alto Sax
2	CV	Clavinet	27	TS	Tenor Sax
3	HC	Harpsichord (Cembalo)	28	BA	Baritone Sax
4	VI	Vibraphone	29	OB	Oboe
4	GL	Glockenspiel	29	EH	English Horn
5	MB	Marimba	30	FG	Bassoon (Fagotto)
5	XY	Xylophone	31	CL	Clarinet
6	OR	Pipe Organ	32	PC	Piccolo
6	HO	Hammond Organ	33	FL	Flute
7	AC	Accordion	33	PN	Pan Flute
8	HM	Harmonica	34	RC	Recorder
9	CG	Classic Guitar (Nylon String)	35	SH	Shakuhachi
10	UK	Ukulele	36	BJ	Banjo
11	AG	Acoustic Guitar (Steel String)	37	SA	Shamisen
12	MD	Mandolin	38	KT	Koto
13	EG	Electric Guitar	39	SO	Sho
14	EB	Electric Bass	40	ND etc.	Japanese Percussion
15	VN	Violin	41	BD etc.	Concert Drums 3
16	VL	Viola	42	BD etc.	Rock Drums 3
17	VC	Cello	43	BD etc.	Jazz Drums 3
18	CB	Contrabass (Wood Bass)	44	diversos	Percussion
19	HP	Harp	45	SP	Soprano (Female)
20	T1/T2/T3/T4	Timpani	46	AL	Alto (Female)
21	TR	Trumpet	47	TN	Tenor (Male)
21	CR	Cornet	48	BT	Baritone (Male)
22	TB	Trombone	49	BS	Bass (Male)
23	TU	Tuba	50	VM	R&B Vocal (Male)
24	HR	Horn	50	VF	R&B Vocal (Female)

Tabela A.2: Tabela de instrumentos da base de dados RWC.

A.4 Segmentador Elaborado usando Média e Desvio.

Esse segmentador foi inicialmente elaborado para extrair as notas dos arquivos fornecidos pela base de dados MIS. Conforme pode se observar nas Figuras A.2 e A.1 é necessário um algoritmo de segmentação que seja capaz de extrair cada nota do trem de notas do qual é composta a amostra do sinal. Infelizmente, os arquivos da base de dados MIS possuem as suas notas espaçadas ora por silêncio ora por ruído de fundo, o que descarta o emprego de um detector de silêncio.

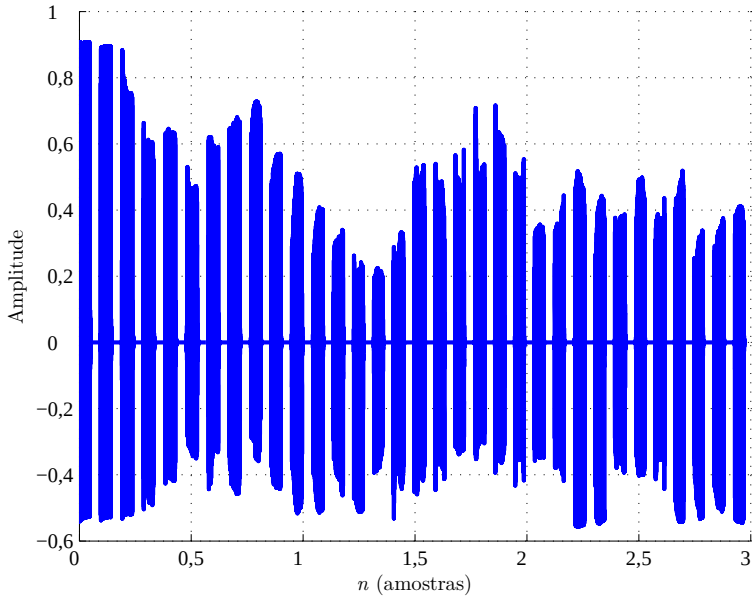

Figura A.4: Saxofone Contralto.

Posteriormente esse segmentador mostrou-se útil quando foram feitas as análises presentes nos capítulos 9 e 8, onde foi empregado conjuntamente com outro extrator de notas, basicamente devido ao fato de ter se mostrado bastante robusto.

O segmentador implementado usou o fato que quando uma janela passeia sobre a energia do sinal, o desvio padrão do sinal na janela que contém o início ou final da nota sofre um acréscimo, visto que a variação do nível de energia quando a nota se inicia ou se encerra ser maior que a variação do nível de energia que contém somente amostras dos instantes de silêncio ou de presença da nota. Assim, ao variarmos as entradas (tamanho da janela, desvio padrão da janela) é possível gerar uma segmentação correta das notas contidas na amostra do sinal.

Após a segmentação persiste um segundo problema que é saber se o número de notas obtidas pelo segmentador representa o número real de notas distintas contidas no arquivo. Para contornar esse problema fez-se uma inspeção visual e às vezes auditiva em cada arquivo da base de dados, anotando o número de notas contidas no arquivo e armazenando este resultado num arquivo gabarito. De posse dessa informação foi possível ajustar o par de características do segmentador de forma a se conseguir uma segmentação correta. Nas Figuras A.5 e A.6 vemos duas notas já segmentadas referentes a cada um dos arquivos mencionados anteriormente:

Posteriormente, para emprego nos Capítulos 9 e 10, foi necessário automatizar

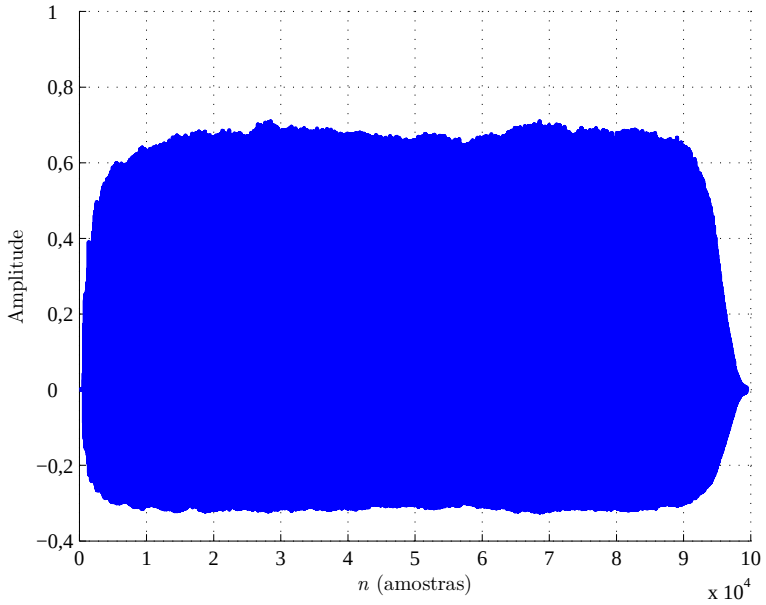

Figura A.5: Trompa - A4.

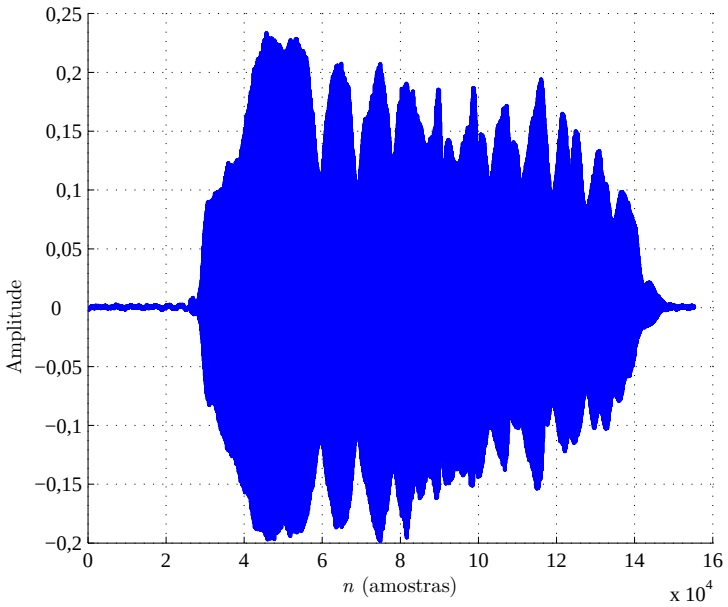

Figura A.6: Flauta Contralto A4.

o processo. Para isso, mais uma entrada foi acrescentada: a média da janela, a qual passou a ser comparada juntamente com o desvio da janela com a média e o desvio do sinal inteiro. Essa modificação não garantiu o sucesso de 100% dos casos, mas tornou o processo robusto o suficiente para que fosse empregado nas avaliações dos capítulos citados anteriormente.

Apêndice B

Comparação com outros Trabalhos

Esse apêndice tem como objetivo efetuar uma comparação entre os resultados obtidos pelo classificador proposto por essa dissertação e os resultados obtidos em [3].

Existem algumas poucas limitações para uma comparação direta dos resultados obtidos pelos dois trabalhos, as quais relacionam-se com o conjunto de treinamento escolhido para treinar os classificadores, porque traduz uma alteração do conjunto de teste. Portanto, pequenas variações na taxa de acerto para mais ou para menos nos instrumentos ou nas famílias podem ocorrer, sem que haja uma preferência clara. Além da limitação anterior, existe uma segunda restrição que é o percentual das notas que foram usadas para o treinamento. Esse percentual não aparece claramente no artigo, dizendo somente os percentuais globais atingidos pelos dois agrupamentos usados e os percentuais obtidos por cada instrumento. Assim, iremos comparar o resultado do artigo supondo uma formação com 90% das amostras para treinamento, que foi a mais frequentemente simulada nesse trabalho.

A taxa de acerto global relatada em [3] foi de 90% para os instrumentos e 95% para a família dos instrumentos.

A fim de podermos fazer uma comparação direta entre as taxas de acerto é necessário verificarmos também se a taxa de acerto global foi obtida de forma ponderada com a distribuição das amostras nos instrumentos, conforme foi feita nessa dissertação. O conjunto de amostras usados em [3] aparentemente foi proporcional à quantidade de amostras por instrumento, pois a taxa global indicada foi de 90%, sendo que, se ponderarmos pelos instrumentos as taxas relatadas para cada

instrumento com o total de amostras associada a cada instrumento, obtemos uma taxa de acerto global de 90,1%. Essa diferença de 0,1% entre o valor relatado e o estimado foi também notada quando se ponderaram pelo mesmo critério as taxas obtidas pelo classificador elaborado nessa dissertação. Portanto, ocorreu um erro de 0,1% para menos na estimativa, ou seja, de 97,2% avaliado contra 97,3% estimado. Logo, atribuiu-se esse erro a problemas de arredondamento.

O artigo [3] usou 2 agrupamentos definidos de forma quase idêntica ao que foi feito nessa dissertação. Essa semelhança foi proposital para a base de dados MIS, pois já visava essa comparação.

A princípio o trabalho elaborado no artigo [3] usou todas as variações existentes para os instrumentos presentes na base de dados. Contudo, existe uma diferença, que é uma inconsistência organológica com os agrupamentos apresentados no Capítulo 2, uma vez que Krishna e Sreenivas inseriram o instrumento Saxofone na família metais, onde normalmente ele aparece na família palhetas. Denominaremos esse novo agrupamento, com a presença do saxofone na família dos metais, de FRBS. Nessa dissertação o instrumento saxofone foi inserido na família palhetas. Contudo, para podermos comparar a taxa de acerto do agrupamento de família de instrumentos, iremos alterar a classificação hierárquica usada nessa dissertação visando padronizá-la com a classificação hierárquica usada no artigo.

Nas tabelas apresentadas a seguir as duas linhas finais (de cima para baixo) representam as taxas de acerto globais e as taxas de acerto média respectivamente. Na Tabela B.1 apresentamos as taxas de acerto obtidas por Krishna e Sreenivas para os instrumentos que foram usados.

Na Tabela B.2 apresentamos os resultados obtidos para os três melhores classificadores elaborados nessa dissertação para a base de dados MIS, com 90% da base de dados usada para treinamento.

Conforme se pode ver, para esse caso, as 3 melhores soluções obtiveram individualmente um desempenho superior à melhor solução obtida por Krishna usando GMM com 46 misturas e codificação LSF.

Se compararmos instrumento a instrumento o desempenho entre o melhor classificador de Krishna e Sreenivas (LSF + GMM) e o melhor classificador individual (LSF, solução 6 + SVM) do banco de classificadores elaborado nesse trabalho,

Instrumentos	Notas	%	LSF	LPCC	MFCC
Saxofone Contralto	192	6,64%	97,10%	88,13%	96,00%
Trombone Baixo	131	4,53%	72,28%	66,42%	95,40%
Saxofone Soprano	192	6,64%	85,68%	83,64%	79,90%
Trombone Tenor	99	3,42%	92,41%	87,16%	87,60%
Flauta Contralto	99	3,42%	85,66%	87,64%	56,53%
Flauta Baixa	102	3,53%	100,00%	94,07%	82,70%
Flauta	227	7,85%	85,86%	80,67%	83,15%
Fagote	122	4,22%	91,04%	90,10%	75,00%
Clarinete Bb	139	4,81%	73,43%	77,62%	26,50%
Clarinete Eb	119	4,12%	62,81%	55,78%	44,80%
Trompa	96	3,32%	55,71%	62,50%	75,30%
Oboé	104	3,60%	98,60%	93,33%	66,00%
Violoncelo	668	23,11%	99,00%	100,00%	95,75%
Violino	601	20,79%	96,44%	86,00%	91,75%
Total	2891	100%	90,1%	86,3%	82,7%
			85,4%	82,4%	75,5%

Tabela B.1: Taxas de acerto obtidas por Krishna e Sreenivas usando GMM com 46 misturas.

Instrumentos	Notas	%	LSF-1	LSF-6	MFCC-5	Prop
Saxofone Contralto	192	6,64%	100,00%	100,00%	94,44%	100,00%
Trombone Baixo	131	4,53%	100,00%	100,00%	100,00%	100,00%
Saxofone Soprano	192	6,64%	100,00%	100,00%	94,44%	100,00%
Trombone Tenor	99	3,42%	77,78%	88,89%	100,00%	88,89%
Flauta Contralto	99	3,42%	87,50%	100,00%	100,00%	100,00%
Flauta Baixa	102	3,53%	100,00%	100,00%	100,00%	100,00%
Flauta	227	7,85%	100,00%	100,00%	100,00%	100,00%
Fagote	122	4,22%	83,33%	83,33%	91,67%	91,67%
Clarinete Bb	139	4,81%	91,67%	100,00%	75,00%	100,00%
Clarinete Eb	119	4,12%	90,00%	80,00%	90,00%	90,00%
Trompa	96	3,32%	88,89%	88,89%	100,00%	88,89%
Oboé	104	3,60%	88,89%	88,89%	77,78%	88,89%
Violoncelo	668	23,11%	98,21%	96,43%	94,64%	96,43%
Violino	601	20,79%	97,96%	100,00%	97,96%	100,00%
Total	2891	100%	95,6%	96,4%	94,8%	97,2%
			93,2%	94,7%	94,0%	96,1%

Tabela B.2: Taxas de acerto obtidas pelos melhores classificadores para a base de dados MIS.

podemos ver que alguns instrumentos foram melhores classificados com o classificador elaborado nessa dissertação enquanto que outros instrumentos foram melhores classificados com o classificador elaborado por Krishna e Sreenivas. No entanto, convém destacar a diferença de desempenho entre os dois classificadores para os

instrumentos trompa, clarinetes (Bb e Eb) e oboé, os três primeiros a favor do classificador elaborado nessa dissertação e o último a favor de Krishna e Sreenivas.

Para avaliarmos o resultado para o agrupamento FRBS com o classificador elaborado nessa dissertação usaremos somente a estratégia 3, pelo simples motivo dela servir tanto para o K-NN quanto para a SVM. Infelizmente, nesse caso não é possível fazer uma comparação direta com os resultados de cada família, visto que Krishna e Sreenivas não apresentaram no seu artigo a taxa de acerto de cada família. Apesar disso, apresentaremos na Tabela B.3 os resultados para o agrupamento FRBS obtidos pelos principais classificadores que compõem a solução proposta nessa dissertação.

Família	LSF-1	LSF-6	MFCC-5	Prop
Metais	98,48%	98,48%	96,97%	98,48%
Flautas	100,00%	100,00%	100,00%	100,00%
Palhetas	95,35%	93,02%	88,37%	97,67%
Cordas	100,00%	100,00%	98,10%	99,05%
	98,8%	98,4%	96,4%	98,8%
	98,5%	97,9%	95,9%	98,8%

Tabela B.3: Tabela contendo as taxas de acerto para o agrupamento FRBS.

Novamente podemos notar na Tabela B.3 que o classificador individual com a pior taxa de acerto obteve um desempenho superior à taxa de acerto relatada em [3] (95%). A diferença encontrada não é muito significativa. No entanto, todos os classificadores apresentados na Tabela B.3 tiveram desempenhos superior a 96%, enquanto que todas as soluções obtidas por Krishna e Sreenivas em [3] para esse agrupamento foram sempre inferiores a 95%.

O agrupamento FRBS a princípio teve um desempenho similar ao agrupamento MFPC, conforme pode ser visto na Tabela B.4. Sendo assim, a preferência de se usar nessa dissertação o agrupamento MFPC mostra-se adequada, já que esse agrupamento (MFPC) se encontra mais coerente com literatura sobre organologia.

Além da base de dados MIS, Krishna e Sreenivas usaram a base de dados RWC para entre outras coisas, classificar notas provenientes de 19 Instrumentos. Foi relatada para esse caso uma taxa de acerto global de 74%. Como o artigo não discrimina os instrumentos usados para a RWC, não foi possível fazer uma comparação direta entre os resultados obtidos por Krishna e Sreenivas e os resultados

Família	LSF-1	LSF-6	MFCC-5	Prop
Metais	96,67%	96,67%	100,00%	96,67%
Flautas	100,00%	100,00%	100,00%	100,00%
Palhetas	98,73%	98,73%	98,73%	100,00%
Cordas	100,00%	100,00%	98,10%	99,05%
	99,2%	99,2%	98,8%	99,2%
	98,9%	98,9%	99,2%	98,9%

Tabela B.4: Tabela contendo as taxas de acerto para o agrupamento MFPC.

obtidos nessa dissertação para essa mesma base de dados. No entanto, fica claro que o modelo apresentado por Krishna e Sreenivas teve uma variação de aproximadamente 16% para um acréscimo de 5 instrumentos (mantido o modelo obtido com a base de dados MIS). Já o trabalho apresentado nessa dissertação obteve para a base de dados RWC com 17 instrumentos (mantido o modelo obtido com a base de dados MIS) uma taxa de acerto de 97,1%, praticamente a mesma obtida para a base de dados MIS com 14 instrumentos. Um segundo resultado foi obtido para 20 instrumentos, com o mesmo modelo. Nesse caso as amostras das 3 bases de dados foram misturadas, e a taxa de acerto foi superior a 94%. Percebe-se que a solução proposta nessa dissertação apresenta maior adaptabilidade à mudança de base de dados do que a solução apresentada por Krishna e Sreenivas, inclusive sofrendo pouca variação com o acréscimo do número de classes.