UNIVERSIDADE FEDERAL DE SANTA CATARINA

PROGRAMA DE PÓS-GRADUAÇÃO
EM ENGENHARIA ELÉTRICA

CLASSIFICAÇÃO DE SINAIS DE ÁUDIO COM ÊNFASE NA SEGMENTAÇÃO DO CANTO DENTRO DE SINAIS DE MÚSICA BASEADA EM ANÁLISE HARMÔNICA

Dissertação submetida à
Universidade Federal de Santa Catarina
como parte dos requisitos para a
obtenção do grau de Mestre em Engenharia Elétrica.

PHABIO JUNCKES SETUBAL

Florianópolis, Julho de 2004.

CLASSIFICAÇÃO DE SINAIS DE ÁUDIO COM ÊNFASE NA SEGMENTAÇÃO DO CANTO DENTRO DE SINAIS DE MÚSICA BASEADA EM ANÁLISE HARMÔNICA

Phabio Junckes Setubal

'Esta Dissertação foi julgada adequada para obtenção do Título de Mestre em Engenharia Elétrica, Área de Concentração em *Comunicações e Processamento de Sinais*, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina.'

_____
Prof. Sidnei Noceti Filho, D.Sc.
Orientador

_____
Prof. Jefferson Luiz Brum Marques, Ph.D.
Coordenador do Programa de Pós-Graduação em Engenharia Elétrica

Banca Examinadora:

_____
Prof. Sidnei Noceti Filho, D.Sc.
Presidente

_____
Prof. Rui Seara, Dr.

_____
Eng. Eduardo Márcio de Oliveira Lopes, Ph.D.

_____
Prof. Bartolomeu Ferreira Uchôa Filho, Ph.D.

_____
Prof. Walter Pereira Carpes Jr., Dr.

AGRADECIMENTOS

À minha família, pelo apoio incondicional.

À Miriam, pelo companheirismo, incentivo e exemplo.

Ao Prof. Rui, por ter me incentivado a trilhar este caminho e pelas valiosas orientações.

Ao Prof. Sidnei, pela idéia do tema inicial e também pelas valiosas orientações.

Aos amigos André e Richard, pelas contribuições e pela amizade.

Aos membros da banca, pelas valiosas correções e sugestões.

Aos colegas do LINSE.

Ao CNPq, pelo apoio financeiro.

A todos que, de alguma forma, contribuíram para a realização deste trabalho.

Resumo da Dissertação apresentada à UFSC como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Elétrica.

CLASSIFICAÇÃO DE SINAIS DE ÁUDIO COM ÊNFASE NA SEGMENTAÇÃO DO CANTO DENTRO DE SINAIS DE MÚSICA BASEADA EM ANÁLISE HARMÔNICA

Phabio Junckes Setubal
Julho/2004

Orientador: Prof. Sidnei Noceti Filho, D.Sc.
Co-orientador: Prof. Rui Seara, Dr.
Área de Concentração: Comunicações e Processamento de Sinais.
Palavras-chave: classificação de sinais de áudio, segmentação do canto dentro de sinais de música, análise harmônica, vibrato, variação de *pitch*.
Número de Páginas: 77.

RESUMO: A área de pesquisa conhecida como classificação de sinais de áudio busca realizar a identificação automática das classes de áudio (fala, música, ruído, canto, dentre outras). Inicialmente, o objetivo deste trabalho é apresentar o estado-da-arte nessa área de pesquisa e discutir a sua estrutura padrão de diagrama em blocos. Atenção especial é dada à etapa de extração de parâmetros. Posteriormente, o objetivo do trabalho adquire caráter de inovação científica, concentrando-se no tema específico de segmentação do canto dentro de sinais de música. A abordagem proposta baseia-se na diferença entre o conteúdo harmônico dos sinais de canto e de instrumentos musicais, observadas através de análise visual do espectrograma. Os resultados obtidos são comparados com os de outra técnica proposta na literatura, usando o mesmo banco de dados. Mesmo considerando um método de medida de desempenho mais criterioso, a taxa de acerto obtida situa-se na mesma faixa da técnica usada como comparação, em torno de 80%. Como vantagem, a abordagem aqui proposta apresenta menor complexidade computacional. Adicionalmente, permite discriminar os diferentes tipos de erro envolvidos no processo de segmentação, sugerindo alternativas para reduzi-los, quando possível. Finalmente, a partir do algoritmo proposto, é realizado um primeiro experimento com o objetivo de separar os sinais de canto de instrumentos musicais dentro de um sinal de música. Os resultados subjetivos obtidos indicam que o processo de separação proposto opera satisfatoriamente.

Abstract of Dissertation presented to UFSC as a partial fulfillment of the requirements for the degree of Master in Electrical Engineering.

AUDIO SIGNAL CLASSIFICATION WITH EMPHASIS ON SEGMENTATION OF SINGING VOICE WITHIN MUSIC SIGNALS BASED ON HARMONIC ANALISYS

Phabio Junckes Setubal
July/2004

Advisor: Prof. Sidnei Noceti Filho, D.Sc.
Co-advisor: Prof. Rui Seara, Dr.
Area of Concentration: Communications and Signal Processing.
Keywords: audio signal classification, segmentation of singing voice within music signals, harmonic analysis, vibrato, *pitch* variation.
Number of Pages: 77.

ABSTRACT: The research area known as audio signal classification seeks to accomplish the automatic identification of the audio classes (speech, music, noise, song, etc.). Initially, the goal of this research work is to present the state-of-art in this research area, and to discuss its standard structure of block diagram. Special attention is given to the feature extraction stage. Next, the aim of the research follows a scientific innovation character, concentrating on a specific subject, which is the segmentation of singing voice within music signals. The proposed approach is based on the difference between the harmonic content of the signals of singing voice and musical instruments, observed through visual analysis in the corresponding spectrogram. The results obtained are compared with those of another technique proposed in the literature, using the same database. Even considering a more rigorous performance measure method, the obtained accuracy rate stays at the same level of the technique used as a comparison, around 80%. As an advantage, the proposed approach presents a lower computational complexity. In addition, it allows to discriminate the different error types involved in the segmentation process, suggesting alternatives to reduce them, when it is possible. Finally, from the proposed algorithm, a first experiment is accomplished, aiming to separate singing voice signals from those of musical instruments within music signals. The subjective results obtained indicate that the proposed separation process operates satisfactorily.

SUMÁRIO

Lista de Figuras ix

Lista de Tabelas xi

Introdução 1

1.1 Objetivos e Motivação do Trabalho 3

1.2 Aplicações da Segmentação do Canto Dentro de Sinais de Música 5

1.2.1 Organização em um Banco de Dados de Sinais de Música 6

1.2.2 Transcrição Automática da Letra de uma Música 6

1.2.3 Identificação do Cantor de uma Música 6

1.2.4 Separação entre Canto e Instrumentos Musicais 6

1.2.5 Classificação Completa de uma Peça Musical 7

1.3 Estrutura da Dissertação 7

Noções Básicas para a Classificação de Sinais de Áudio 9

2.1 Introdução 9

2.2 Produção da Fala 9

2.2.1 Aparelho Fonador 10

2.2.2 Unidades da Fala 12

2.3 Sistema Auditivo Humano 12

2.4 Espectrograma 14

2.5 Conclusões 16

Estado-da-Arte em Classificação de Sinais de Áudio 17

3.1 Introdução 17

3.2 Discriminação entre Sinais de Fala e Música 17

3.3 Outros Tipos de Classificação de Sinais de Áudio 19

3.4 Canto e Instrumentos Musicais Dentro de Sinais de Música 21

3.5 Aplicações da Classificação de Sinais de Áudio em Geral 22

3.5.1 Monitoração do Sinal em uma Estação de Rádio 22

3.5.2 Transcrição Automática da Fala 23

3.5.3 Organização de um Banco de Dados Multimídia 23

3.5.4 Segmentação de Sinais de Audiovisual 24

3.5.5 Compressão de Sinais de Áudio 24

3.5.6 Otimização do Desempenho de Aparelhos Auditivos 24

3.5.7 Monitoração em Sistemas de Segurança 24

3.5.8 Auxílio em Diagnósticos 25

3.5.9 Identificação do Falante 25

3.6 Conclusões 25

Processamento Automático de CSA: Abordagem Geral *26*

4.1 Introdução **26**

4.2 Extração de parâmetros **26**
4.2.1 Parâmetros no Domínio do Tempo 27
4.2.1.1 Energia de Curta Duração 27
4.2.1.2 Taxa de Cruzamentos por Zero (TCZ) 29
4.2.2 Parâmetros no Domínio da Freqüência 30
4.2.2.1 Centróide Espectral (CE) 30
4.2.2.2 Ponto de *Rolloff* Espectral (RE) 31
4.2.2.3 Fluxo Espectral 31
4.2.2.4 Freqüência Fundamental (f_0) 32
4.2.3 Parâmetros no Domínio Tempo-Freqüência 34
4.2.3.1 Energia de Modulação a 4 Hz (EM4) 34
4.2.3.2 Vibrato 34
4.2.4 Parâmetros Baseados em Probabilidade *a Posteriori* 35
4.2.4.1 Entropia 35
4.2.4.2 Dinamismo 36
4.2.5 Outros Parâmetros 36

4.3 Classificadores de Padrões **36**

4.4 Pós-processamento **37**

4.5 Conclusões **37**

Segmentação do Canto Dentro de Sinais de Música *38*

5.1 Introdução **38**

5.2 Modelo para a Segmentação do Canto Dentro de Sinais de Música **40**
5.2.1 Padrão para Identificação da Classe dos Instrumentos Musicais 40
5.2.2 Padrões Principais para Identificação da Classe de Canto 40
5.2.3 Modelo Simplificado para a Segmentação do Canto 41
5.2.4 Padrões Secundários Característicos do Canto 42
5.2.5 Parâmetro Secundário para a Segmentação do Canto 43
5.2.6 Modelo Completo para a Segmentação do Canto 46

5.3 Análise de Casos **47**
5.3.1 Caso 1: Canto caracterizado pelo vibrato 47
5.3.2 Caso 2: Canto caracterizado por pequena variação de *pitch* 48
5.3.3 Caso 3: Canto vozeado caracterizado por poucas trilhas 49
5.3.4 Caso 4: Canto puramente fricativo 50

5.4 Extração e seleção das trilhas de picos espectrais variáveis **51**
5.4.1 Representação do Sinal no Domínio Tempo-Freqüência 51
5.4.2 Análise da Amplitude e Largura dos Picos 53
5.4.3 Formação das Trilhas (Vetores Linha) 53
5.4.4 Seleção das Trilhas de Acordo com a Forma 56
5.4.5 Seleção das Trilhas Harmônicas 57

5.5 Conclusões **58**

Resultados Experimentais *60*

6.1 Introdução **60**

6.2 Comparação dos Métodos de Análise de Desempenho **61**

6.3 Comparação de Desempenho **62**

6.4 Classificador Automático & Classificador Manual **62**

6.5 Discriminação dos Tipos de Erro **63**

6.6 Análise das Causas dos Erros **64**

6.6.1 Seleção de Trilhas Variáveis de Forma Insatisfatória 64

6.6.2 Seleção de Trilhas Variáveis Produzidas por Instrumentos Musicais 65

6.6.3 Erros Provocados pelo Algoritmo 66

6.7 Filtragem das Trilhas de Picos Espectrais **67**

6.8 Conclusões **68**

Conclusões *69*

Referências Bibliográficas *72*

LISTA DE FIGURAS

Figura 1.1: Sinal de áudio (música) contendo segmentos de canto e puramente instrumentais. 2
Figura 1.2: Etapas de um processamento automático simulando as funções do sistema auditivo humano em uma situação real. 2
Figura 2.1: Modelo para a CSA realizada por seres humanos. 9
Figura 2.2: Aparelho fonador humano (adaptado de [12]). 10
Figura 2.3: Sistema Auditivo Periférico (adaptado de [16]). 13
Figura 2.4: Espectrograma de um sinal de áudio (música): (a) banda larga; (b) banda estreita. 15
Figura 3.1: Estrutura taxonômica do sinal de áudio, considerando as principais classes estudadas na literatura. 17
Figura 4.1: Diagrama de blocos do processamento automático de CSA. 26
Figura 4.2: (a) Sinal de áudio (fala); (b) parâmetro de energia de curta duração. 28
Figura 4.3: (a) Sinal de áudio (fala); (b) parâmetro de TCZ. 29
Figura 4.4: Comparação da TCZ de sinais de: (a) fala e (b) música. 30
Figura 4.5: Espectrograma: (a) sinal de fala; (b) sinal de música. 31
Figura 4.6: Comparação dos valores de CE: (a) sinal de fala; (b) sinal de música. 31
Figura 4.7: Comparação dos valores de FE: (a) sinal de fala; (b) sinal de música. 32
Figura 4.8: Extração do valor de f_0 a partir da autocorrelação temporal. 33
Figura 4.9: Extração do valor de f_0 no domínio da freqüência. 33
Figura 4.10: Comparação dos valores de EM4: (a) sinal de fala; (b) sinal de música. 34
Figura 4.11: Comparação da trilha de f_0 : (a) fala; (b) canto. 35
Figura 4.12: Definição das classes pelo limite de divisão determinado pelo classificador. 37
Figura 5.1: Diferenças entre classificação de sinais de áudio a partir de discriminação e de segmentação. 39
Figura 5.2: (a) Espectro da nota harmônica de um instrumento musical; (b) trilhas de picos espectrais constantes produzidas durante o intervalo de tempo em que a nota é produzida. 40
Figura 5.3: Trilhas de picos espectrais onduladas produzidas pelo vibrato. 41
Figura 5.4: Trilhas de picos espectrais caracterizadas por pequenas variações de pitch. 41
Figura 5.5: Modelo simplificado para a segmentação do canto dentro de sinais de música. 42
Figura 5.6: Espectrograma de um sinal de música destacando um segmento de canto fricativo, #1, e um segmento de canto vozeado que não produz trilhas variáveis, #2. 43
Figura 5.7: Modelo completo para a segmentação do canto em sinais de música. 46
Figura 5.8: (a) Espectrograma do sinal de canto (vibrato); (b) trilhas variáveis selecionadas pelo algoritmo. 47
Figura 5.9: (a) Espectrograma do sinal de música contendo canto (pequena variação de pitch); (b) trilhas variáveis selecionadas pelo algoritmo; (c) extensão dos limites do canto pela variação de MMB. 48
Figura 5.10: (a) Espectrograma do Sinal; (b) variação da MMB. 49
Figura 5.11: (a) Espectrograma de um segmento de canto fricativo; (b) MMA & MMB. 50

Figura 5.12: Diagrama em blocos da extração e seleção de trilhas de picos espectrais variáveis. 51
Figura 5.13: Comparação da magnitude do espectro obtido por: (a) FFT; (b) modelo AR. 52
Figura 5.14: Captura dos picos espectrais do sinal de música mostrado na Figura 5.9(a) (contendo canto caracterizado por uma pequena variação de pitch) obtidos com modelo AR de ordem: (a) 40; (b) 80. 52
Figura 5.15: Captura dos picos espectrais do sinal de música mostrado na Figura 5.8(a) (contendo canto caracterizado pelo vibrato) obtidos com modelo AR de ordem: (a) 40; (b) 80. 53
Figura 5.16: Valor limite acompanhando a trajetória da trilha. 54
Figura 5.17: Fluxograma para a formação das trilhas a partir de vetores linha. 55
Figura 5.18: Análise da forma das trilhas (esboço): (a) descarte completo; (b) seleção completa; (c) seleção parcial. 56
Figura 5.19: Análise da relação harmônica entre trilhas selecionadas pela forma variável. 57
Figura 5.20: Evolução da etapa de seleção de trilhas de picos espectrais variáveis: (a) captura de todos os picos; (b) seleção por amplitude e suavidade; (c) formação das trilhas a partir de vetores linha; (d) seleção das trilhas variáveis; (e) análise harmônica das trilhas variáveis. 58
Figura 6.1: Exemplo de erro não percebido pelo método de análise de desempenho usado em [4]. 61
Figura 6.2: Espectrograma de um sinal de música. 67
Figura 6.3: Filtragem das trilhas de picos espectrais por banco de filtros: (a) rejeita-faixa e (b) passa-faixa. 67

LISTA DE TABELAS

Tabela 6.1: Comparação do desempenho entre o algoritmo da referência [4] e o algoritmo proposto 62
Tabela 6.2: Análise de desempenho do algoritmo proposto aplicado ao banco de dados de treinamento e ao banco total 63
Tabela 6.3: Distribuição da taxa de erro conforme os seus tipos 63
Tabela 6.4: Desempenho do algoritmo aplicado aos sinais com nível de confiabilidade 1 (um) 65

Capítulo 1

Introdução

O som pode ser definido como um uma perturbação física que se propaga, de forma ondulatória, em um meio material elástico. Ele adquire grande importância quando é percebido pelo ser humano, permitindo a propagação da informação.

O sistema auditivo humano é treinado para extrair informações do som nas mais diversas, e muitas vezes adversas, condições. Por exemplo, duas pessoas conseguem conversar mesmo em um ambiente com alto nível de ruído de fundo. Apesar de o som que chega aos ouvidos ser formado pela soma de todos os sons produzidos no ambiente, cada uma das pessoas consegue isolar e perceber somente a fonte sonora desejada. Essa situação é conhecida como "*cocktail party effect*", e a área que estuda a separação de fontes sonoras é a análise da cena auditiva (*Auditory Scene Analysis* – ASA) [1].

Considerando a importância e o caráter subjetivo da percepção humana na audição, processar sinais de áudio (som audível) com o objetivo de simular funções do sistema auditivo não é tarefa fácil de se realizar automaticamente. Essa afirmação é válida mesmo para tarefas consideradas simples para o sistema auditivo humano, tal como a identificação das classes de sinais de áudio (fala, música, ruído, canto, instrumentos musicais, dentre outras). O sinal de áudio está disponível como a variação da onda sonora no tempo. Então, diretamente, muito pouca informação pode ser extraída de tal sinal a fim de realizar a identificação de sua classe. Como exemplo, a Figura 1.1 ilustra a representação de um sinal de música contendo um segmento da classe de canto e outros sem canto (classe de instrumentos musicais), de acordo com a marcação manual destacada. Observa-se que, a partir do sinal de áudio, não é possível identificar a localização das duas classes envolvidas. Portanto, é necessário extrair outros parâmetros do sinal. A área de pesquisa conhecida como classificação de sinais de áudio (CSA) estuda o desenvolvimento de parâmetros e estratégias de classificação, visando a identificação automática das classes de sinais de áudio.

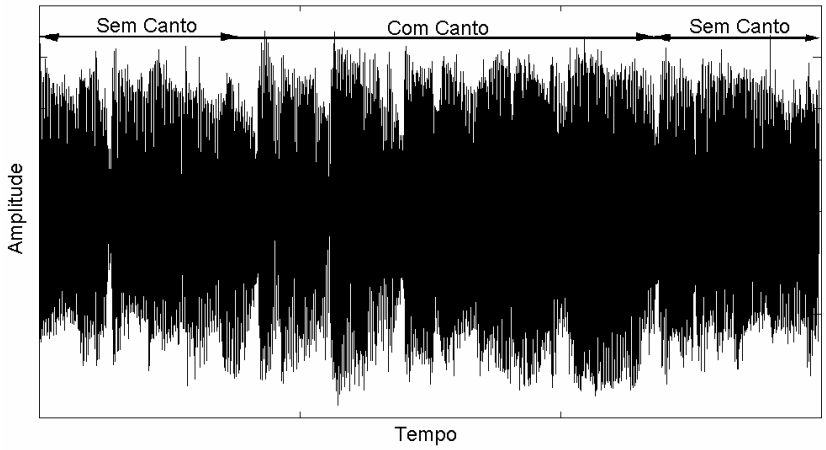


Figura 1.1: Sinal de áudio (música) contendo segmentos de canto e puramente instrumentais.

A CSA ocupa uma posição intermediária em um processamento automático que pretenda substituir todas as funções do sistema auditivo humano em uma situação real, como ilustrado na Figura 1.2. Por exemplo, na situação descrita anteriormente, supõe-se a substituição de uma das pessoas conversando por uma máquina capaz de ouvir. No extremo inicial do processamento automático, deve existir um sistema baseado em ASA, capaz de isolar a fonte desejada dos demais sons do ambiente. Na posição intermediária, deve haver um sistema que classifique o sinal como fala, para que possa ser encaminhado e tratado pelo reconhecimento automático de fala, que constitui a etapa final.

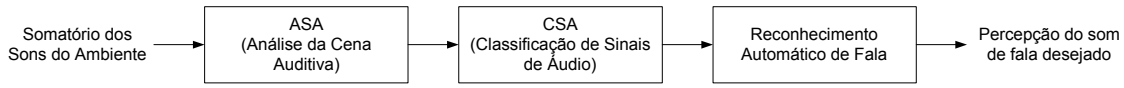


Figura 1.2: Etapas de um processamento automático simulando as funções do sistema auditivo humano em uma situação real.

O desenvolvimento de pesquisa na área de CSA está relacionado com a área de análise da cena auditiva, introduzida em 1990 por Bregman [1]. Relaciona-se, mais especificamente, com a análise da cena auditiva computacional (*Computational Auditory Scene Analysis* – CASA) [2], a qual pretende, através do estudo da percepção humana dos sinais de áudio, desenvolver algoritmos para realizar a segregação das fontes sonoras. Em CSA, pretende-se desenvolver algoritmos baseados na percepção humana, a fim de realizar a classificação.

1.1 Objetivos e Motivação do Trabalho

Este trabalho tem dois objetivos principais. Considerando a relevância da pesquisa na área de CSA e o seu reconhecimento relativamente menor em comparação com áreas diretamente relacionadas como, por exemplo, o reconhecimento automático de fala, um primeiro objetivo possui caráter informativo, e pretende apresentar e discutir aspectos relacionados à CSA em geral. Assim, é apresentado o estado-da-arte em CSA, através de uma revisão bibliográfica que contempla os principais trabalhos de pesquisa realizados na área. Grande parte desses trabalhos considera o problema clássico de discriminação entre sinais de fala e música. Outros trabalhos concentram-se em uma classificação mais geral, incluindo outras classes de sinais de áudio, tais como silêncio, som ambiente (ruído), canto, fala e música simultâneas, fala com som ambiente, música com som ambiente, dentre outras. Alguns trabalhos consideram uma classificação mais específica, tais como a identificação de sons característicos de algum evento (esportivo, indicativo de violência, dentre outros), a identificação dos gêneros de uma música e a identificação de instrumentos musicais em uma música. Finalmente, a revisão bibliográfica contempla os poucos trabalhos que se concentram no objetivo final a ser proposto.

Posteriormente, é apresentada e discutida a estrutura de diagrama em blocos para a solução de qualquer problema relacionado à CSA. O diagrama em blocos é composto pelas etapas de extração de parâmetros, classificador de padrões e pós-processamento. Destaque especial é dado à etapa de extração de parâmetros do sinal de áudio. Essa é a etapa mais importante do processamento e pretende traduzir a forma como o ser humano percebe o som. Parâmetros relevantes em CSA devem apresentar comportamento semelhante para sinais de mesma classe, e comportamento distinto para sinais de classes diferentes. Neste trabalho, são compilados os principais parâmetros encontrados na literatura. Em alguns casos, são realizadas simulações que comprovam o comportamento distinto dos parâmetros para sinais de classes diferentes.

Introduzidos os aspectos referentes à CSA em geral, o objetivo final deste trabalho adquire um caráter de inovação científica e considera um tema específico dentro da área de CSA. Assim, é proposto um algoritmo que realize a segmentação do canto dentro de sinais de música. Em relação a problemas clássicos em CSA, como a discriminação de sinais de fala e música, tal problema proposto como objetivo final é complexo, pois:

- Geralmente, em uma música, o canto está sobreposto aos instrumentos musicais. Portanto, ocupam o mesmo espaço no tempo e em freqüências. Entretanto, tais trechos devem ser identificados unicamente como canto;
- Há uma semelhança maior entre as classes de canto e instrumentos musicais do que entre as classes de fala e música. Assim, é mais difícil encontrar parâmetros do sinal que possam confiavelmente distinguí-las;
- Há uma dificuldade maior de realizar a classificação por segmentação, em relação à classificação por discriminação, como será visto no Capítulo 5;

A abordagem proposta neste trabalho para realizar a segmentação do canto baseia-se na diferença entre o conteúdo harmônico do canto e dos instrumentos musicais. Essa diferença, que pode ser observada por análise visual no correspondente espectrograma, é evidenciada pela ampliação dos padrões identificados por um parâmetro apresentado em [3], denominado trilha de pico espectral. Em [3], tal parâmetro é usado para extrair o vibrato, característico dos sinais de canto, com o objetivo principal de distinguí-los dos sinais de fala. No trabalho aqui desenvolvido, além do vibrato, propõe-se a extração de pequenas variações de *pitch,* que também caracterizam unicamente o canto quando comparados aos instrumentos musicais. A técnica discutida em [3] apresenta resultados baseados apenas em experimentos de discriminação. Assim, uma comparação direta de resultados entre a abordagem proposta e aquela apresentada em [3] fica prejudicada.

Em [4], é encontrado um dos poucos trabalhos na literatura que apresenta resultados na segmentação do canto dentro de sinais de música. Entretanto, a abordagem proposta apresenta um alto custo computacional. Necessita, inclusive, de um sistema de reconhecimento automático de fala.

Portanto, o presente trabalho propõe o desenvolvimento de um algoritmo baseado, sobretudo, na ampliação dos padrões identificados pelo parâmetro apresentado em [3], a fim de realizar o mesmo experimento de [4]. O desempenho do algoritmo proposto é avaliado e comparado usando o mesmo banco de dados de [4], gentilmente cedido pelos autores. Os resultados obtidos neste trabalho são semelhantes àqueles obtidos em [4],

fornecendo uma taxa de acerto de segmentação em torno de 80%. Entretanto, é possível destacar algumas diferenças e vantagens da solução proposta, tais como:

- Extensão dos testes sobre todo o banco de dados;
- Menor complexidade computacional;
- Adoção de um método de análise de desempenho mais rigoroso;
- Possibilidade de discriminar e analisar os tipos de erro envolvidos;
- Sugestão de alternativas para redução da taxa de erro como, por exemplo, a adição de um grau de confiabilidade ao algoritmo.

O objetivo final deste trabalho é propor um algoritmo que realize a segmentação do canto dentro de sinais de música. A escolha de tal proposta baseia-se em três motivos principais:

- O caráter desafiador do tema que, por sua complexidade, exige um estudo aprofundado que fornece uma base para o tratamento de problemas de CSA em geral;
- A existência de poucos trabalhos de pesquisa publicados com esse objetivo específico;
- A intenção inicial de propor como tema a separação entre canto e instrumentos musicais dentro de sinais de música. Como na literatura não se encontrou uma base satisfatória para tal pesquisa, adotou-se a segmentação do canto como uma possibilidade mais próxima, permitindo, inclusive, um primeiro experimento na tentativa da separação entre canto e instrumentos musicais.

1.2 Aplicações da Segmentação do Canto Dentro de Sinais de Música

A relevância da pesquisa na área de CSA é comprovada por suas muitas aplicações. Aplicações específicas da segmentação do canto dentro de sinais de música são discutidas a seguir. Mais adiante, será apresentada uma discussão sobre as aplicações de CSA em geral.

1.2.1 Organização em um Banco de Dados de Sinais de Música

Atualmente, com a facilidade de acesso a arquivos compactados (mp3 [5] e similares) via *internet*, tem-se a possibilidade de se dispor de extensas bibliotecas de áudio de música. Entretanto, uma quantidade limitada de informações pode ser extraída de tais arquivos como, por exemplo, o nome da música, do artista, título do álbum. Essas informações, conhecidas como etiquetas ID3 [6], são ainda inseridas de forma manual e baseadas em texto. Como aplicação direta, a segmentação de canto propõe a localização automática dos segmentos contendo canto em um sinal de música. Isso permite o acesso direto a esses segmentos e, portanto, facilita a organização em um banco de dados de música. O canto, normalmente, exerce uma função fundamental, ao carregar as informações da letra e da melodia de uma música.

1.2.2 Transcrição Automática da Letra de uma Música

A segmentação do canto dentro de sinais de música compõe uma etapa de pré-processamento, encaminhando somente os segmentos contendo canto para o tratamento posterior de uma etapa de transcrição automática da letra de uma música, como sugerido em [4].

1.2.3 Identificação do Cantor de uma Música

A partir da segmentação dos trechos contendo canto em um sinal de música, é selecionado um segmento de canto. Visando a identificação do cantor, tal segmento é comparado com um banco de dados contendo vozes de cantores, como proposto em [7].

1.2.4 Separação entre Canto e Instrumentos Musicais

A segmentação do canto dentro de sinais de música pode ser vista como uma etapa de pré-processamento e um ponto de partida para a separação entre canto e instrumentos musicais. Essa separação poderia ser utilizada, por exemplo, na recuperação de gravações históricas ou na implementação de um sistema de *karaokê* aperfeiçoado. Esse sistema permitiria ao usuário cantar com o acompanhamento da banda original, excluindo a voz do cantor na gravação. Por outro lado, também permitiria ao usuário executar a parte

instrumental da música com o acompanhamento da voz original do cantor, excluindo os instrumentos musicais.

Atualmente, existe uma função *karaokê* que permite atenuar a voz do cantor, utilizando a reversão de fase [8]. Nesse caso, para que se tenham resultados razoáveis, é necessário ter na entrada do sistema um sinal estéreo e cuja voz esteja gravada igualmente nos dois canais. Assim, aplicando-se a reversão de fase (180º) a um dos canais e somando-se os sinais em ambos, obtém-se na saída um sinal de áudio mono com a voz atenuada. Instrumentos musicais que, eventualmente, estejam gravados igualmente nos dois canais, sofrerão o mesmo processo de atenuação.

A separação entre canto e instrumentos musicais também seria necessária para obter bons resultados na transcrição automática da letra de uma música [4] e na identificação do cantor em uma música [7].

1.2.5 Classificação Completa de uma Peça Musical

A partir da segmentação de uma música em trechos com canto e sem canto, é possível imaginar a classificação completa deste sinal. Por exemplo, como sugerido em [9], os trechos identificados como canto podem ser classificados como canto solo ou em coro. Os trechos identificados como canto solo podem ser classificados como voz masculina ou feminina. Por outro lado, os trechos identificados como sem canto, ou seja, puramente instrumentais, podem ser classificados de acordo com a categoria de instrumentos (metais, cordas, dentre outros) ou instrumentos individuais.

Como visto, muitas vezes as aplicações em CSA constituem uma etapa de pré-processamento. Essa etapa é importante, pois, além de extrair somente a informação necessária à etapa seguinte do processo, reduzindo seu custo de processamento, também leva a uma redução na taxa de erro da próxima etapa, visto que são eliminadas informações desnecessárias.

1.3 Estrutura da Dissertação

No Capítulo 2, são apresentadas noções básicas para a CSA, incluindo aspectos relacionados à produção da fala, ao funcionamento do sistema auditivo humano, e à análise

no domínio tempo-freqüência, cuja representação gráfica através do espectrograma fornece informações importantes para o tratamento de problemas de CSA.

O Capítulo 3 descreve o estado-da-arte da CSA. Contempla uma revisão bibliográfica dos principais trabalhos na área, considerando diversas classes de sinais de áudio e seus tipos de classificação. São apresentados também alguns trabalhos existentes relacionados ao objetivo final do tema proposto, que é a segmentação do canto dentro de sinais de música. Algumas aplicações da CSA em geral também são discutidas.

O Capítulo 4 aborda o diagrama em blocos para o processamento automático de CSA em geral. Destaque especial é dado à extração de parâmetros do sinal de áudio. Uma coletânea dos principais parâmetros encontrados na literatura é apresentada.

A partir do Capítulo 5, o trabalho se concentra em seu objetivo final: uma abordagem para realizar a segmentação do canto dentro de sinais de música é proposta.

O Capítulo 6 descreve os resultados experimentais obtidos a partir da abordagem desenvolvida no capítulo anterior. Esses resultados são comparados com os resultados obtidos em [4], apresentando um experimento com objetivos idênticos. É também realizado um primeiro experimento buscando a separação entre canto e instrumentos musicais dentro de um sinal de música.

No Capítulo 7 são apresentadas as conclusões gerais e sugestões de trabalhos futuros.

Capítulo 2

Noções Básicas para a Classificação de Sinais de Áudio

2.1 Introdução

Como ilustrado na Figura 2.1, nos seres humanos, a CSA baseia-se no processamento de um sinal acústico realizado pelo sistema auditivo. O sinal acústico é, muitas vezes, produzido pelo aparelho fonador. Portanto, antes de estudar algoritmos que pretendam substituir o processamento humano, é interessante ter uma noção de como ocorre tal processamento. Assim, são discutidos neste capítulo, de forma sucinta, aspectos relacionados à produção da fala e ao funcionamento do sistema auditivo humano. Além disso, é apresentado o gráfico do espectrograma, que permite representar os sinais de áudio no domínio tempo-freqüência, e cuja análise visual torna possível vislumbrar a extração de parâmetros relevantes para o processamento automático de CSA. O objetivo deste capítulo é apresentar conceitos e características que serão importantes para o desenvolvimento deste trabalho.

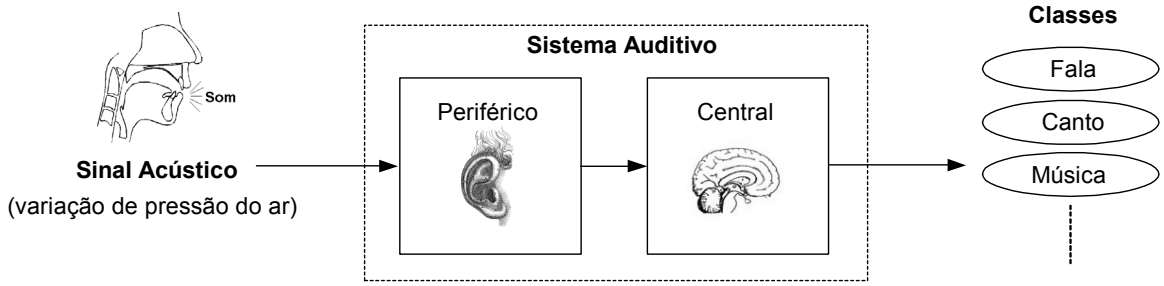


Figura 2.1: Modelo para a CSA realizada por seres humanos.

2.2 Produção da Fala

A fala compõe uma das classes dos sinais de áudio, consistindo na principal forma de comunicação entre os seres humanos.

O canto constitui outra importante classe de sinais de áudio, e é produzido pelo mesmo aparelho fonador que produz a fala. Considera-se o canto uma classe intermediária entre a fala e a música, possuindo características de ambas. As diferenças entre o canto sem o acompanhamento de instrumentos musicais e a fala, são muito sutis e subjetivas. Por analogia, em [10], essas diferenças são comparadas às diferenças que existem entre o ato de andar (falar) e dançar (cantar). Embora ambos possam realizar a mesma função de se movimentar (transmitir informação), há um certo "estilo" presente em um que o diferencia do outro. Foi mostrado em [10] que as próprias pessoas, algumas vezes, ao ouvir um dos sons, discordavam ao realizar a discriminação entre fala e canto.

2.2.1 Aparelho Fonador

O sinal acústico da fala é formado a partir do aparelho fonador, cujos principais componentes são: pulmões, traquéia, laringe, trato vocal (cavidade faringeal e cavidade oral) e trato nasal (cavidade nasal) [11]. Há ainda os componentes móveis (articuladores), cujas mudanças de posição permitem produzir diferentes tipos de sons. Os articuladores são compostos pelas cordas vocais, véu palatino, língua, dentes, lábios e mandíbula [11]. A Figura 2.2 ilustra os componentes do aparelho fonador. Os órgãos que compõem o aparelho fonador não são especializados e exclusivos para produzir a fala. Suas funções primárias incluem a respiração, alimentação e olfato [11].

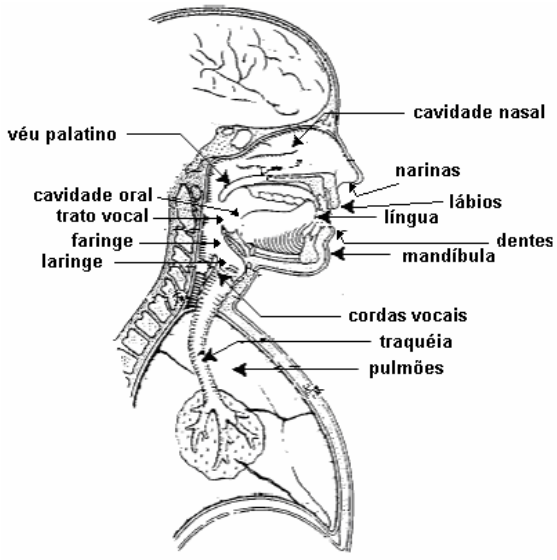


Figura 2.2: Aparelho fonador humano (adaptado de [12]).

A produção da fala ocorre a partir da excitação gerada pelo fluxo de ar emitido pelos pulmões. Esse fluxo atravessa a traquéia e alcança a laringe, onde se encontram as cordas vocais. A interação com a abertura das cordas vocais (glote) pode alterar a estrutura do fluxo de ar, definindo a forma da excitação [13]. Após essa interação, o fluxo de ar alcança o trato vocal, onde é processado. O trato vocal consiste em um tubo acústico de seção transversal não uniforme variante no tempo. As freqüências de ressonância desse tubo acústico são chamadas de formantes. No interior do trato vocal encontram-se os articuladores. A configuração da posição dos articuladores altera a forma da seção transversal do trato vocal, modificando os valores dos formantes, definindo o som específico a ser emitido pela cavidade oral e/ou nasal [13].

Os sons da fala podem ser classificados entre vozeados e não vozeados. Os sons vozeados são caracterizados pela forma invariavelmente periódica da excitação. Nesse caso, inicialmente, a glote encontra-se fechada, provocando um aumento da pressão do ar emitido pelos pulmões. Essa elevação de pressão provoca a passagem de ar através de uma vibração quase periódica da glote [11], gerando uma série de pulsos de ar (glotais) que caracterizam a forma periódica da excitação. O período do ciclo de abertura e fechamento da glote define o valor da freqüência fundamental (f_0). Os sons não vozeados, ao contrário dos vozeados, apresentam excitação aperiódica.

Os sons vozeados são de grande importância para o desenvolvimento do algoritmo proposto no Capítulo 5, que pretende realizar a segmentação do canto dentro de sinais de música, baseado na extração do parâmetro de trilhas de picos espectrais. Tal parâmetro é extraído nos segmentos vozeados do canto no domínio tempo-freqüência e é formado por valores de freqüências múltiplas da fundamental.

Outro tipo de classificação divide os sons da fala entre vogais e consoantes. Nas consoantes, a configuração do trato vocal forma constrições que obstruem a passagem do ar em um ou vários pontos [13]. Esses sons podem apresentar excitação periódica ou não. Por outro lado, as vogais são produzidas sem obstrução da passagem de ar [14] e apresentam excitação periódica.

Dentre os sons não vozeados, os mais interessantes para este trabalho são as consoantes fricativas, em razão dos padrões que formam no domínio tempo-freqüência, que também são explorados pelo algoritmo proposto no Capítulo 5. Nesses sons, o fluxo de

ar é forçado a passar por um canal estreito (formado pela constrição dos articuladores em algum ponto do trato vocal), criando uma turbulência e mudando a forma da excitação para aperiódica. Exemplos de consoantes fricativas incluem [s], [f], [ʃ].

2.2.2 Unidades da Fala

A menor unidade que compõe uma língua falada é o fonema. Os fonemas representam os menores segmentos que modificam o significado das palavras [13]. Por exemplo, nas palavras "tia", "mia" e "ria", os segmentos que distinguem os seus significados e que, portanto, constituem fonemas, são representados pelas letras "t", "m", e "r". Algumas letras podem representar mais de um fonema como, por exemplo, a letra "x" que, na língua portuguesa, representa o fonema [ʃ] em "xarope", [z] em "exato", [s] em "extensão" e os fonemas [ks] em "táxi". Cada uma das línguas faladas no mundo é composta pela combinação de 30 a 50 fonemas [13]. Na língua portuguesa, como citado em [14], são 33 ou 34 fonemas, e na língua inglesa são 42 [11].

Unidades da fala superiores ao fonema incluem [14]: difone, demissílaba, trifone, sílaba, palavra e frase.

2.3 Sistema Auditivo Humano

Como visto anteriormente, a produção da fala em seres humanos é efetuada por um conjunto de órgãos não exclusivos e não especializados na função. Situação oposta é observada em relação à audição. Nesse caso, há um sistema auditivo extremamente complexo e especializado.

Como visto na Figura 2.1, o sistema auditivo está dividido em duas partes:

- Sistema auditivo periférico. Responsável pela conversão do sinal acústico (ondas sonoras) em sinal neural (impulsos nervosos);
- Sistema auditivo central. Responsável pelo processamento dos impulsos nervosos recebidos da saída do sistema auditivo periférico. Localizado no cérebro.

Ainda não se tem completo conhecimento do funcionamento do sistema auditivo humano. Afirma-se em [15] que, geralmente, quanto mais distante da periferia, ou seja, mais próximo do sistema auditivo central, menos certeza se tem sobre sua função exata.

Em razão do maior conhecimento do sistema auditivo periférico e da possibilidade de descrever fisicamente os eventos que ali ocorrem, são discutidos a seguir aspectos relativos à sua composição e ao processo de conversão do sinal acústico em sinal neural. A Figura 2.3 ilustra os componentes do sistema auditivo periférico.

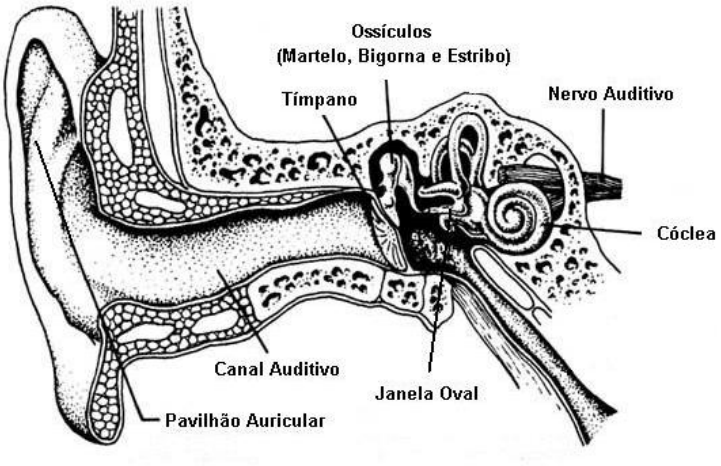


Figura 2.3: Sistema Auditivo Periférico (adaptado de [16]).

O sistema auditivo periférico é composto pelo ouvido externo, médio e interno:

- **Ouvido externo**

Sua função essencial é captar as ondas sonoras através do pavilhão auricular e encaminhá-las, pelo canal auditivo, até o ouvido médio. Além disso, o canal auditivo, por possuir a característica de um tubo ressonante, atua como um amplificador na faixa de freqüências entre 2 e 5,5 kHz. É obtida uma amplificação máxima (cerca de 11dB) em torno de 4 kHz [15]. O pavilhão auricular também exerce a função secundária de auxiliar o sentido de localização. Curiosamente, alguns morcegos possuem pavilhões auriculares altamente desenvolvidos que fornecem alta sensibilidade para essa função. Outros animais, tais como cachorros e gatos, podem mover seus pavilhões auriculares para localizar fontes sonoras [17].

- **Ouvido médio**

Sua função principal é realizar a conversão da energia acústica em energia mecânica, promovendo um acoplamento adequado entre o meio gasoso (ouvido externo) e o meio líquido (ouvido interno) [17]. É composto pelo tímpano, localizado na entrada do ouvido médio, por uma cadeia de ossículos, e pela janela oval, situada na saída do ouvido médio. O tímpano consiste em uma membrana que vibra ao ser excitada por ondas sonoras.

Essas vibrações são transmitidas pela cadeia de ossículos móveis interconectados (martelo, bigorna e estribo) até o ouvido interno, através da janela oval [16]. A cadeia de ossículos forma um sistema de alavanca que promove um aumento de força mecânica. Além disso, há uma diminuição da área de vibração no estribo (em contato com a janela oval) em relação ao tímpano. Como resultado, obtém-se um grande aumento de pressão, de cerca de 30 dB [18], que impede que grande parte da energia seja refletida ao passar de um meio gasoso (baixa impedância) para um meio líquido (alta impedância).

- **Ouvido interno**

Responsável pela transdução final da onda sonora em impulsos nervosos. O ouvido interno é formado pela cóclea [19] (acoplada à janela oval) que encontra-se preenchida pelo chamado fluído coclear. Possui a forma da concha de um caracol e é dividido longitudinalmente pela membrana basilar. De forma resumida, a energia mecânica transmitida ao ouvido interno (pelas vibrações dos ossículos) através da janela oval, causa uma flutuação no fluído coclear que, por sua vez, provoca uma deformação na membrana basilar [16]. O movimento da membrana basilar atinge as células ciliadas, cuja perturbação gera os impulsos nervosos que são transmitidos ao cérebro pelo nervo auditivo. O grau de perturbação da membrana basilar e, conseqüentemente, das células ciliadas, varia com a freqüência do som recebido. Por essa razão, afirma-se que a cóclea produz uma análise em freqüência do som [20].

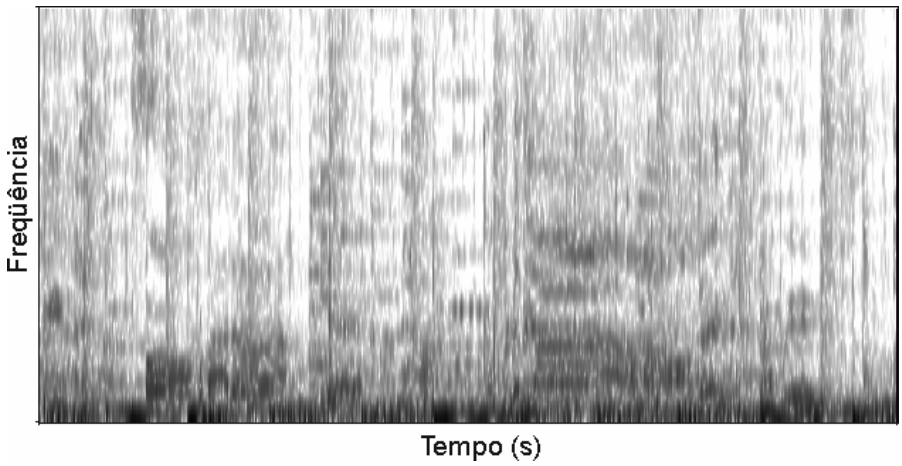
2.4 Espectrograma

Em CSA, a análise dos sinais de áudio no domínio da freqüência é extremamente relevante, haja vista que o próprio sistema auditivo humano a realiza [20]. Entretanto, considerando que os sinais de áudio apresentam propriedades variantes no tempo (são não-estacionários), a representação adequada desses sinais depende de uma análise no domínio tempo-freqüência. Assim, o sinal de áudio é segmentado no tempo em quadros nos quais ele seja considerado aproximadamente estacionário, e a representação em freqüência é obtida a cada quadro do sinal.

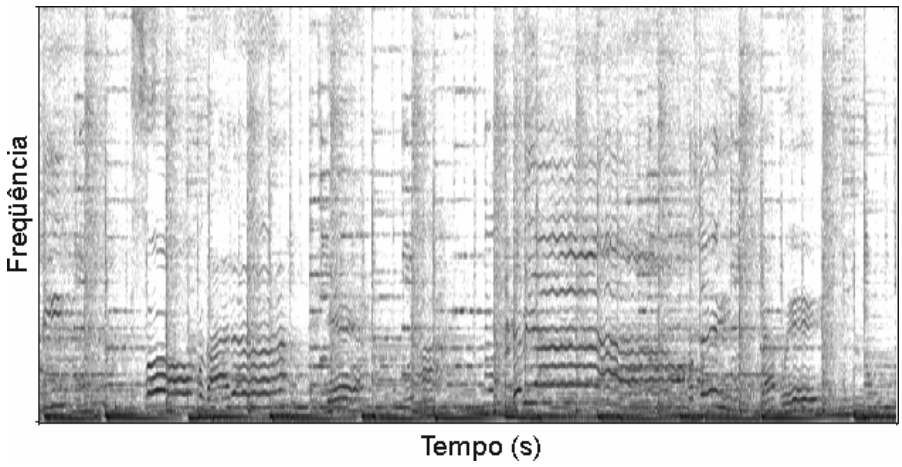
Uma forma usual de se obter a representação de um sinal de áudio no domínio tempo-freqüência é através da Transformada de Fourier de Curta Duração (*Short-Time Fourier Transform* – STFT) [21], que consiste na aplicação da Transformada de Fourier

Discreta (TFD) [21] a segmentos curtos do sinal no domínio do tempo, considerando o uso de uma dada janela de segmentação [21].

A representação gráfica de sinais no domínio tempo-freqüência é obtida através de um espectrograma [21], ilustrado na Figura 2.4. Os valores de magnitude do espectro são dados pelos diferentes tons de cinza. Quanto mais escuros, maior é a amplitude.



(a)



(b)

Figura 2.4: Espectrograma de um sinal de áudio (música): (a) banda larga; (b) banda estreita.

No espectrograma obtido através da STFT, há um compromisso entre a resolução tempo × freqüência. A Figura 2.4(a) mostra um espectrograma de banda larga [21], caracterizado pela aplicação de uma janela de curta duração (5 ms), promovendo uma melhor resolução temporal em detrimento da resolução em freqüência. Considerando o mesmo sinal, a Figura 2.4(b) mostra um espectrograma de banda estreita [21], caracterizado por uma janela de maior duração (30 ms), obtendo-se neste caso uma melhor resolução em freqüência em detrimento à resolução temporal.

No Capítulo 5 deste trabalho, pretende-se identificar o canto através do comportamento de trilhas formadas no domínio tempo-freqüência. Como observado na Figura 2.4 essas trilhas não são visíveis em espectrogramas de banda larga. Portanto, para tal aplicação, a segmentação do sinal no domínio do tempo deve ser obtida por janelas de maior duração, para que a resolução em freqüência permita a visualização e a seleção dos padrões que identificam o canto através das trilhas.

2.5 Conclusões

Este capítulo proporcionou uma base para o processamento automático de CSA, discutindo aspectos relativos ao modo como o ser humano realiza a CSA e à representação gráfica de sinais no domínio tempo-freqüência através de um espectrograma. A análise visual do espectrograma permite vislumbrar a extração de parâmetros relevantes para o processamento automático de CSA.

Em relação à produção da fala, é importante para o desenvolvimento do trabalho a definição da característica harmônica (periódica) dos sons vozeados e não-harmônica (não periódica) dos sons fricativos. Além disso, destaca-se a definição de formantes e fonenas, sobre os quais se baseia a solução apresentada em [4], cujo desempenho será comparado com a abordagem a ser proposta no Capítulo 5.

A descrição do sistema auditivo humano confirma a sua complexidade e a dificuldade em entender os eventos que ocorrem no processamento auditivo, sobretudo, aqueles que ocorrem no sistema auditivo central. Assim, compreende-se o quanto é complicado desenvolver um sistema automático que o substitua. Destaca-se a análise em freqüência promovida pelo sistema auditivo, a qual sugere a sua importância no estudo para viabilizar o processamento automático de CSA.

Capítulo 3

Estado-da-Arte em Classificação de Sinais de Áudio

3.1 Introdução

O problema clássico em CSA considera a discriminação das classes de sinais de fala e música (Figura 3.1). Entretanto, existem ainda outras classes e outros tipos de classificação estudados na literatura, como aqueles apresentados na estrutura taxonômica proposta na Figura 3.1. O objetivo deste capítulo é realizar uma revisão bibliográfica, apresentando trabalhos que contemplam essas diversas classes de sinais de áudio e tipos de classificação. Em primeiro lugar, são discutidos os trabalhos na área de discriminação de sinais de fala e música. Posteriormente, são abordados outros tipos de classificação presentes na Figura 3.1. Finalmente, são comentados os trabalhos relacionados ao objetivo final do tema proposto, ou seja, à segmentação do canto dentro de sinais de música.

Neste capítulo também são apresentadas e discutidas aplicações de CSA em geral.

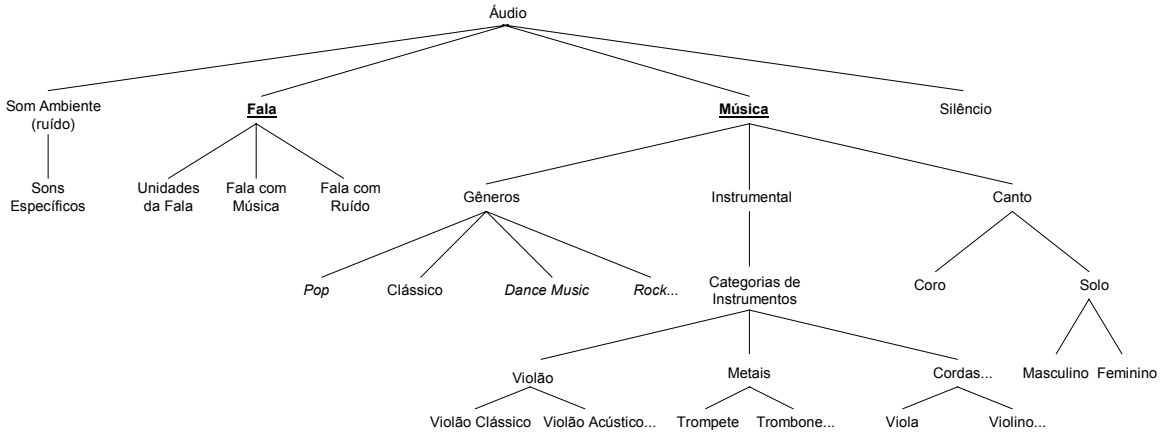


Figura 3.1: Estrutura taxonômica do sinal de áudio, considerando as principais classes estudadas na literatura.

3.2 Discriminação entre Sinais de Fala e Música

Um dos primeiros trabalhos na área de CSA, ou mais especificamente, na discriminação entre sinais de fala e música [22], extrai parâmetros do sinal no domínio do

tempo, como taxa de cruzamentos por zero (TCZ) e energia. Alcança 98% de taxa de acerto na discriminação, considerando janelas de 2,4 s.

Em outro trabalho pioneiro [23], são apresentados e comparados um total de 13 parâmetros, nos domínios do tempo, freqüência e cepstro [21]. No experimento realizado, os parâmetros que apresentaram os melhores resultados são: variância do fluxo espectral, energia de modulação a 4 Hz e pulso métrico [23]. São comparados também os desempenhos de quatro classificadores automáticos: *multidimensional Gaussian a posteriori* (MAP), Modelo de Mistura Gaussiana (GMM), *k-d trees* e *k-nearest-neighbor* (k-NN). Conclui-se em [23] que há muito pouca diferença nos resultados obtidos de cada um, ou mesmo entre diferentes configurações de um mesmo classificador. A taxa de acerto obtida é de 98,6%, para a discriminação de sinais com duração de 2,4 s.

Posteriormente, surgiram diversos trabalhos na área, utilizando diferentes técnicas e abordagens para a solução do problema de discriminação entre sinais de fala e música [24] – [29]. Em [30], é aplicado um sistema de reconhecimento automático de fala, o qual estima a probabilidade de que o sinal de entrada corresponda a um determinado fonema. Estatísticas dessa distribuição de probabilidades, tais como entropia e dinamismo, são extraídas para discriminar fala de não fala. O banco de dados utilizado é o mesmo de [23], e a taxa de acerto obtida é de 98,7% para a discriminação de trechos de 2,5 s.

Em [31], é utilizado um extenso banco de dados com arquivos de 10 s, somando aproximadamente 12 horas de fala e 8 horas de música, em diferentes línguas. São comparados os desempenhos individuais de quatro parâmetros e suas variações: coeficientes do cepstro e delta cepstro (1,2% de taxa de erro), amplitude e delta amplitude (1,7% de taxa de erro), *pitch* e delta *pitch* (4% de taxa de erro), e TCZ e delta TCZ (6% de taxa de erro).

Os métodos até aqui citados apresentam resultados satisfatórios para a discriminação entre sinais de fala e música pura. Entretanto, na presença de fala e música simultâneas, seus rendimentos são reduzidos. Por exemplo, em tal condição, a taxa de acerto da técnica desenvolvida em [23] diminui para apenas 65%. A fim de aperfeiçoar os resultados nessas condições de fala e música simultâneas, são propostas diferentes abordagens, que são discutidas a seguir.

A técnica desenvolvida em [32] utiliza dois classificadores GMM independentes: um classifica entre fala e não-fala, usando parâmetros do cepstro, e o outro classifica entre

música e não-música, utilizando parâmetros do espectro. Essa técnica apresenta uma taxa de acerto de 99,5% na detecção de fala e 93% na detecção de música. A mesma idéia de classificadores independentes é utilizada em [33].

O procedimento apresentado em [34] sugere dois estágios de classificação. O primeiro estágio classifica entre fala e não-fala. O segundo estágio classifica os sinais de fala entre fala pura e fala não-pura, e os sinais de não-fala entre som ambiente e música. São extraídos parâmetros como MFCC (*mel-frequency cepstral coefficients*) [35], TCZ, energia, fluxo espectral (FE), dentre outros. Obtêm-se resultados em torno de 96% de taxa de acerto, para trechos de 1 s.

Na técnica desenvolvida em [36], é sugerida a divisão da classe fala com música em duas: fala predominante e música predominante. Utiliza três parâmetros relativamente simples: variância da TCZ, energia e freqüência fundamental. Obtém 89% de taxa de acerto na classificação da mistura de fala e música.

A presença de canto em sinais de música é outro fator que contribui para a redução do desempenho da discriminação entre sinais de fala e música. Quando não identificado como uma classe específica, o canto deve ser classificado como música. Entretanto, como ele possui características semelhantes à fala, muitas vezes acaba sendo classificado, de forma errada, como tal. Para reduzir esses erros, em [37] é proposta a discriminação de fala e música em dois estágios. O primeiro estágio discrimina o sinal de entrada entre canto e não-canto. Para isso, emprega parâmetros com comportamento bastante distinto entre fala e canto como, por exemplo, a energia de modulação a 4 Hz. Excluídos os sinais de canto, no segundo estágio, são aplicadas técnicas tradicionais de discriminação de fala e música.

3.3 Outros Tipos de Classificação de Sinais de Áudio

De acordo com a Figura 3.1, a fala pode ser classificada em unidades de fala como, por exemplo, fonemas e palavras. A área de pesquisa de reconhecimento automático de fala trabalha com a classificação de unidades de fala.

O som ambiente, algumas vezes chamado de ruído, geralmente dá origem ao som de fundo de sinais de fala e, nesse caso, não forma uma classe em específico. Por exemplo, em [15] e [34], é considerada a classe de fala com ruído. Eventualmente, o som ambiente também pode formar uma única classe. Por exemplo, em [38], são consideradas um total

de 7 classes, dentre as quais ruído ambiente e fala com ruído. Em [3], a classe de som ambiente é reconhecida e é ainda subdividida em duas: som ambiente harmônico e som ambiente não-harmônico.

Alguns artigos exploram a classificação de sons específicos. Por exemplo, em [39] são classificados sons que caracterizam tiros, explosões e choro, com o objetivo de identificar cenas que indiquem a presença de violência. Em [40], há dois estágios de classificação. O primeiro estágio classifica o sinal de áudio em fala, música, som ambiente e silêncio. No segundo estágio, os sinais identificados como som ambiente são classificados nos seguintes sons específicos: aplausos, chuva, canto de pássaros, latido de cachorro, explosão, passos, risos, fluxo de um rio, trovão, vendaval. De um total de 50 sons específicos testados, 41 são identificados corretamente (taxa de acerto em torno de 80%).

Existem trabalhos na área de CSA que buscam a identificação do gênero de uma música. Em [35], são classificados três gêneros musicais: *heavy metal*, *dance music* e clássico. De um banco de testes composto por 189 músicas desses gêneros, a melhor taxa de acerto obtida é de cerca de 88%. Em [41], são classificados os gêneros *rock*, *pop*, *tecno* e clássico, com 86% de taxa de acerto. O banco de dados é composto por 360 trechos de 30 s de música dos quatro gêneros. Em [42], são classificados os gêneros *rock*, piano e *jazz*. São testados quatro arquivos de cada gênero. A taxa de acerto é de 91,67%.

Outros trabalhos exploram a classificação de categorias de instrumentos e de instrumentos individuais. Em [43], são classificados instrumentos específicos como, por exemplo, violão acústico, violão clássico, trompete, trombone, viola, violino. São classificadas também categorias de instrumentos como, por exemplo, violão (violão acústico, violão clássico), metais (trompete, trombone), cordas (viola, violino). A taxa de acerto obtida situa-se na faixa entre 75,73% e 79,73%, para classificação de instrumentos individuais, e entre 82,2% e 90,65%, para classificação de categorias de instrumentos.

A identificação da classe de silêncio pode ser interessante, por exemplo, para aplicação em monitoração de ambientes, como sugerido em [39].

A classificação da classe de canto, com ou sem acompanhamento musical, é relativamente pouco explorada na literatura. Alguns trabalhos nessa área visam a discriminação entre o canto e a fala [10], [37]. A partir da identificação da classe de canto, é possível imaginar a classificação entre canto solo (masculino ou feminino) ou em coro, como sugerido em [9].

3.4 Canto e Instrumentos Musicais Dentro de Sinais de Música

A classificação obtida a partir da segmentação das classes de canto e instrumentos musicais dentro de sinais de música compõe o objetivo final do presente trabalho, e será tratada nos Capítulos 5 e 6. Tais classes e tal tipo de classificação são relativamente pouco explorados na literatura. Como visto, grande parte dos trabalhos em CSA considera o problema de discriminação entre sinais de fala e música, em que os sinais de canto e de instrumentos musicais são tratados como pertencentes à classe de música. Em outras ocasiões, em que são considerados individualmente, a classificação é obtida a partir da discriminação e não da segmentação. Por exemplo, em [3], os autores realizam a classificação de sinais de áudio em diversas classes, dentre as quais música instrumental e canto (com ou sem acompanhamento musical). Os parâmetros extraídos do sinal são TCZ, energia de curta duração, freqüência fundamental e trilha de pico espectral. Para o classificador, é utilizado um procedimento baseado em regras (*rule based*), no qual o próprio programador, de acordo com a experiência e a observação do comportamento dos parâmetros, define os limiares a serem escolhidos para distinguir as classes. Elimina-se, portanto, a necessidade de classificadores automáticos (de maior complexidade computacional). Apesar do artigo sugerir a segmentação dos sinais de audiovisual, os resultados apresentados derivam de experimentos de discriminação. São testadas 200 amostras de músicas puras, com 189 acertos na classificação, ou seja, 94,5% de taxa de acerto, e 50 amostras de canto, com e sem acompanhamento musical, com 42 acertos, ou seja, 84% de taxa de acerto. Ressalta-se que a taxa de acerto de canto é a menor dentre todas as classes, comprovando a dificuldade de sua detecção.

Muito recentemente, ocorrendo inclusive em paralelo com o desenvolvimento deste trabalho, em [9], é sugerida uma alteração no parâmetro de trilha de pico espectral (proposto em [3]), visando detectar o canto vozeado através do vibrato e de pequenas variações de *pitch*. Além disso, sugere-se a extração do parâmetro de TCZ para detectar segmentos consonantais do canto. Entretanto, em [9], não há informações sobre a implementação da técnica. Coincidentemente, a abordagem a ser proposta no Capítulo 5 baseia-se em idéia similar à sugerida em [9], propondo alteração semelhante no parâmetro de trilha de pico espectral, além de propor a extração de outro parâmetro para auxiliar no processo de segmentação.

Em [7], é discutida uma outra técnica com o objetivo de detectar somente o início do canto em uma música, visando a identificação do cantor. Essa técnica é baseada na extração de quatro parâmetros: energia de curta duração, TCZ, coeficiente harmônico e FE.

Apesar de considerar as classes de canto e instrumentos musicais, os trabalhos anteriores não podem ser usados para comparação de desempenho com uma abordagem que pretende propor a segmentação do canto dentro de sinais de música. Isso porque ou não tratam o tipo de classificação a partir da segmentação, como [3] e [7], ou porque apenas sugerem uma técnica para tratá-lo, não discutindo sua implementação, como em [9]. Um dos únicos trabalhos que exploram tal problema de segmentação do canto é aquele publicado por Berenzweig e Ellis, em [4]. Eles propuseram a mesma aproximação utilizada em [30], a qual aplica um sistema de reconhecimento automático de fala (uma rede neural estimando as probabilidades *a posteriori*) para realizar a discriminação de sinais de fala e música. Em [4], a justificativa para o uso da mesma abordagem de [30] é a de que, embora os sinais de canto e fala sejam diferentes, ambos compartilham algumas características comuns, tais como a estrutura de formantes e a transição de fonemas. Portanto, conjectura-se que um modelo acústico treinado em fala poderia responder de forma similar a sinais de canto, e diferente a instrumentos musicais. Os resultados obtidos apontaram uma taxa de acerto de aproximadamente 80%, com o erro sendo calculado usando janelas com duração de 1,3 s, equivalente à soma de 81 quadros de 16 ms (tempo de um quadro do classificador automático utilizado). No Capítulo 6, esses resultados são comparados com os obtidos da abordagem a ser proposta no Capítulo 5 do presente trabalho.

3.5 Aplicações da Classificação de Sinais de Áudio em Geral

Além das aplicações apresentadas na introdução, que consideram o problema específico de segmentação do canto dentro de sinais de música, há diversas aplicações para a CSA em geral. Algumas delas são discutidas a seguir.

3.5.1 Monitoração do Sinal em uma Estação de Rádio

Um dos primeiros trabalhos em CSA [22], mais especificamente na discriminação entre sinais de fala e música, sugere a disponibilidade de uma função no aparelho de rádio que o mantenha apresentando música durante todo o tempo. Para isso, o sinal da estação

selecionada deve estar sendo monitorado constantemente. Assim, ao classificar o sinal como fala (de um intervalo comercial ou de notícias), automaticamente muda-se a estação, buscando outra que esteja apresentando música.

Se for considerada a classificação dos gêneros de uma música, é possível estender a função de disponibilizar música todo o tempo, incluindo ainda a escolha do seu gênero.

Outra aplicação da monitoração é a possibilidade de extrair, automaticamente, estatísticas sobre o tempo que a estação de rádio dedica a cada tipo de programação (música ou não) [22].

3.5.2 Transcrição Automática da Fala

A discriminação entre sinais de fala e música também tem o objetivo de permitir a utilização de sistemas de reconhecimento de fala em situações mais reais, em que o sinal de fala não é o único sinal presente. Assim, o objetivo da etapa de CSA é pré-processar o sinal, encaminhando somente trechos de fala a uma etapa seguinte de reconhecimento. Nesse sentido, uma aplicação pretendida é a transcrição automática da fala em telejornais [29], no qual trechos de vinhetas e inserções comerciais, classificados como música, devem ser desprezados para a transcrição.

3.5.3 Organização de um Banco de Dados Multimídia

Atualmente, com a facilidade da internet, tem-se acesso a um número muito grande de informações. Cada vez mais o valor de uma informação depende do grau de facilidade com que ela pode ser encontrada, recuperada, acessada, filtrada e gerenciada [44]. Nesse sentido, a CSA realiza a indexação das informações de acordo com as classes de sinais de áudio, facilitando a organização em um banco de dados multimídia.

Atualmente, o padrão MPEG-7 [44] está sendo desenvolvido com o objetivo de tornar a *internet* tão pesquisável em conteúdo multimídia como é pesquisável hoje em texto, através de sites de pesquisa como, por exemplo, o *Google*[1]. Em uma situação de pesquisa multimídia, supõe-se que alguém deseje encontrar uma determinada música em um banco de dados, mas não lembre seu título. Nesse caso, a pessoa poderia localizá-la cantando um trecho da música em um microfone.

[1] www.google.com.br

3.5.4 Segmentação de Sinais de Audiovisual

Normalmente, prefere-se realizar a segmentação de sinais de audiovisual através de características visuais, ou seja, analisando apenas o sinal de vídeo. Entretanto, muitas vezes, o áudio representa melhor uma cena do que o próprio vídeo. Por exemplo, cenas de tiro, choro, risos, ou uma seqüência com uma música de fundo, podem apresentar conteúdos de imagens muito distintos, mas são caracterizados pelos mesmos sinais de áudio. Assim, é sugerida a análise dos sinais de áudio e vídeo em conjunto [3].

3.5.5 Compressão de Sinais de Áudio

Dependendo da classe do sinal de áudio, pode existir um algoritmo específico que apresente melhores resultados para sua compressão [28]. Portanto, a CSA consiste em uma etapa de pré-processamento para a aplicação do melhor algoritmo de compressão a cada uma das classes identificadas. Como exemplo da importância da compressão de sinais de áudio, pode-se citar os arquivos mp3. Explorando características redundantes do sinal de áudio, levando em consideração o modo como o ser humano percebe o som, é possível alcançar uma alta taxa de redução no tamanho dos arquivos, com uma perda aceitável de qualidade.

3.5.6 Otimização do Desempenho de Aparelhos Auditivos

Aparelhos auditivos modernos utilizam diversos programas para diferentes situações, permitindo, por exemplo, mudar a resposta em freqüência, parâmetros de compressão, ativar um microfone direcional ou um redutor de ruído. Cada configuração visa otimizar a resposta do aparelho auditivo de acordo com o ambiente sonoro em que o usuário se encontra. Normalmente, é o próprio usuário quem deve reconhecer o ambiente e chavear o aparelho para a melhor situação. A classificação automática de sinais de áudio pode reconhecer o ambiente e configurá-lo automaticamente [15].

3.5.7 Monitoração em Sistemas de Segurança

Em algumas situações, pode ser mais interessante monitorar um ambiente através de microfones do que com câmeras. A captação do sinal por microfones não está restrita

pela linha de visão de uma câmera. Esse sistema pode trabalhar lado a lado com sistemas de vídeo, quando a captação de um som (quebra do silêncio) posiciona a câmera [39].

3.5.8 Auxílio em Diagnósticos

Como sugerido em [10], a classificação de sons específicos pode auxiliar em diagnósticos humanos baseado em sons como, por exemplo, o diagnóstico de alguma enfermidade realizado por um médico em um exame clínico, ou o diagnóstico de um defeito no carro realizado por um mecânico em uma inspeção do veículo.

3.5.9 Identificação do Falante

Os sinais classificados como fala podem ser segmentados de acordo com o falante, aplicando técnicas de identificação, como proposto em [45].

3.6 Conclusões

Neste capítulo foi apresentada uma revisão bibliográfica da área de CSA em geral. Foram discutidas sucintamente técnicas que tratam o problema clássico em CSA: discriminação entre sinais de fala e música. Posteriormente, foram discutidos problemas de classificações mais gerais, considerando diversas classes de áudio, e mais específicos, como identificação de sons característicos de algum evento, identificação de gêneros de uma música e reconhecimento automático de instrumentos musicais. Finalmente, foram abordados temas relacionados ao objetivo final deste trabalho, a ser apresentado nos Capítulos 5 e 6, e que trata da segmentação do canto dentro de sinais de música.

Também foram discutidas aplicações da CSA em geral, além daquelas apresentadas na introdução deste trabalho.

Capítulo 4

Processamento Automático de CSA: Abordagem Geral

4.1 Introdução

Há duas formas de se tratar o problema de CSA automática [39]. A primeira aproximação consiste na simulação do modelo do sistema auditivo humano. A segunda aproximação considera que, a partir do conhecimento do sinal de entrada e das classes de áudio que se desejam obter na saída, é possível investigar padrões que caracterizem e distingam as classes, baseados no mecanismo de percepção humana do som. Lembrando o fato de que não se tem o conhecimento completo do funcionamento do sistema auditivo humano, a segunda aproximação é geralmente a mais usada. Isso transforma a CSA automática em um problema de reconhecimento e classificação de padrões.

Este capítulo tem por objetivo descrever as principais etapas do diagrama de blocos para o processamento automático de CSA, ilustradas na Figura 4.1. É dispensada uma atenção especial à etapa principal, de extração de parâmetros. São reunidos os principais parâmetros encontrados na literatura. Em alguns casos, são realizadas simulações que comprovam a relevância dos parâmetros na distinção de classes.

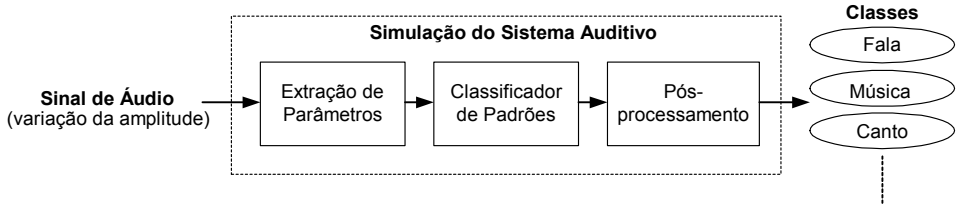


Figura 4.1: Diagrama de blocos do processamento automático de CSA.

4.2 Extração de parâmetros

A extração de parâmetros compreende a primeira e mais importante etapa no processamento automático de CSA. Consiste em extrair informações do sinal de entrada

para formar padrões que caracterizem cada uma das classes. Portanto, devem apresentar comportamento similar para sinais de mesma classe, e distinto para sinais de classes diferentes.

Os parâmetros dos sinais de áudio podem ser divididos entre físicos e perceptuais [35]. Parâmetros físicos são aqueles extraídos diretamente do sinal, por análise matemática ou estatística. Parâmetros perceptuais dependem do modo como o ser humano percebe um determinado som. Todo parâmetro perceptual está sempre associado, de alguma forma, a algum parâmetro físico do sinal. Por exemplo, a freqüência fundamental e a intensidade dos sons (parâmetros físicos) estão relacionadas, respectivamente, ao *pitch* e ao *loudness* (parâmetros perceptuais). Neste trabalho, considera-se indiferente usar o termo físico ou o termo perceptual associado.

Um exemplo típico para o entendimento da relação entre parâmetros físicos e perceptuais é o fenômeno da "falta da freqüência fundamental" [35]. Geralmente, o valor da freqüência fundamental está situado na faixa entre 100 e 150 Hz para os homens e entre 150 e 250 Hz para as mulheres [3]. Entretanto, geralmente, os sistemas telefônicos transmitem sinais com freqüências entre 300 e 3400 Hz. Apesar da qualidade relativamente baixa do sinal transmitido, não se percebe a falta da freqüência fundamental. A partir das freqüências presentes, o sistema auditivo preenche a freqüência fundamental perdida [35].

São apresentados a seguir alguns parâmetros importantes em CSA, divididos de acordo com o domínio do qual são extraídos. Muitas vezes, as estatísticas determinadas a partir de um parâmetro extraído do sinal, tais como média, variância e desvio padrão, são tão ou mais importantes do que o seu próprio valor em si.

4.2.1 Parâmetros no Domínio do Tempo

Os parâmetros extraídos no domínio do tempo são determinados diretamente do sinal, não necessitando de nenhum tipo de transformação. Portanto, apresentam a vantagem de maior velocidade de processamento.

4.2.1.1 Energia de Curta Duração

A medida de energia de curta duração consiste no parâmetro mais próximo do sinal de áudio em sua forma original, disponível como a variação da sua amplitude sobre o

tempo. Tal parâmetro tem a função de representar essa variação. É calculado a cada quadro do sinal segmentado, dado por [3]:

$$E_n = \frac{1}{N}\sum_m \left[x(m)w(n-m)\right]^2 \tag{4.1}$$

onde $x(m)$ é o sinal de áudio discreto, n é o índice de tempo da energia de curta duração e $w(m)$ é a janela retangular de largura N.

Uma das aplicações da energia de curta duração na CSA consiste em auxiliar na discriminação entre segmentos de fala vozeados e não-vozeados. Como ilustrado na Figura 4.2, os segmentos não-vozeados, geralmente, apresentam valores de energia de curta duração menores do que segmentos vozeados.

O parâmetro de energia de curta duração também pode ser usado para detectar silêncio. Além disso, comenta-se em [3] que esse parâmetro pode revelar propriedades do som como ritmo e periodicidade. Finalmente, tal parâmetro pode ser usado na discriminação entre sinais de fala e música. A distribuição de energia de curta duração em sinais de fala é formada por períodos de valores altos (segmentos vozeados) e baixos (segmentos não-vozeados e intervalos entre palavras). Enquanto isso, a distribuição de energia em sinais de música tende a ser mais uniforme [10].

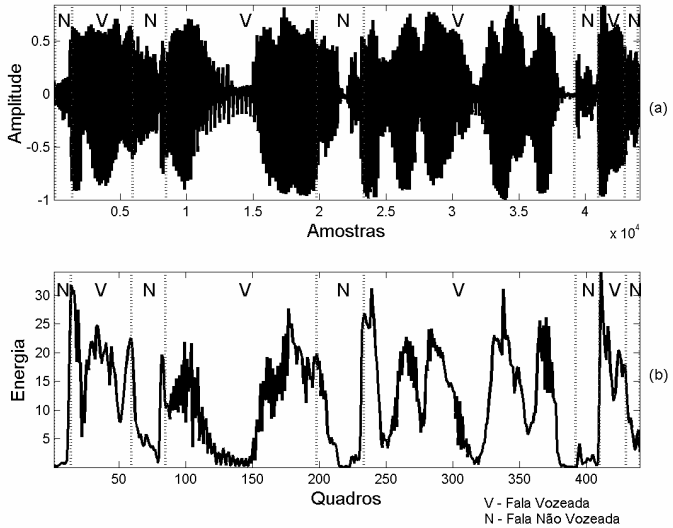


Figura 4.2: (a) Sinal de áudio (fala); (b) parâmetro de energia de curta duração.

4.2.1.2 Taxa de Cruzamentos por Zero (TCZ)

A TCZ determina o número de vezes que um sinal inverte sua polaridade por unidade de tempo. Seu valor é calculado a cada quadro do sinal segmentado, de acordo com a seguinte expressão [3]:

$$Z_n = 0.5 \cdot \sum_m \left| \text{sgn}\left[x(m)\right] - \text{sgn}\left[x(m-1)\right] \right| w(n-m) \tag{4.2}$$

onde

$$\text{sgn}\left[x(n)\right] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases} \tag{4.3}$$

O parâmetro de TCZ pode ser usado para discriminar segmentos de fala vozeado e não-vozeado. Geralmente, segmentos não-vozeados da fala possuem componentes de freqüências mais altas e, portanto, apresentam valores maiores de TCZ, como ilustrado na Figura 4.3.

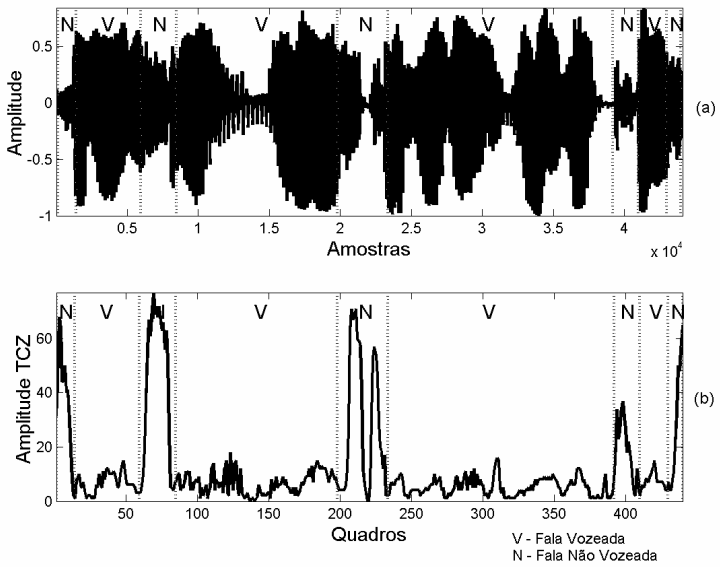


Figura 4.3: (a) Sinal de áudio (fala); (b) parâmetro de TCZ.

O sinal de música, por outro lado, apresenta uma distribuição de freqüências mais uniforme por mais tempo. Portanto, sua curva de TCZ possui uma maior região de estabilidade, apresentando uma menor variância e amplitude média, Assim, a TCZ pode ser usada também para discriminar sinais de fala e música, como ilustrado na Figura 4.4.

Finalmente, em [35], a TCZ é usada na classificação de gêneros musicais.

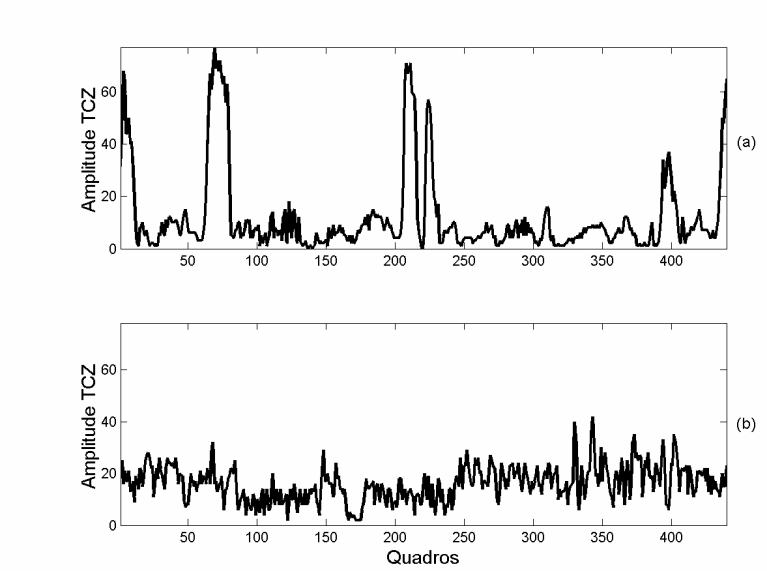


Figura 4.4: Comparação da TCZ de sinais de: (a) fala e (b) música.

4.2.2 Parâmetros no Domínio da Freqüência

Ao contrário dos parâmetros extraídos no domínio do tempo, os parâmetros no domínio da freqüência necessitam de uma etapa de transformação em freqüência. Geralmente, essa transformação é obtida através da STFT.

4.2.2.1 Centróide Espectral (CE)

O CE é definido como o centro de gravidade da magnitude do espectro da STFT [46]. Consiste em uma medida da forma do espectro e tem seu valor dado por

$$CE_q = \frac{\sum_{k=1}^{K} \left| X_q(k) \right| \cdot k}{\sum_{k=1}^{K} \left| X_q(k) \right|} \tag{4.4}$$

onde $X_q(k)$ é o valor da TFD para a freqüência k no quadro q.

O CE apresenta resultados distintos para segmentos de fala vozeado e não-vozeado. Como os segmentos não-vozeados apresentam componentes de freqüências mais altas, seu valor de CE é maior quando comparado aos segmentos vozeados. Portanto, o CE também pode ser usado na discriminação de sinais de fala e música. A distribuição mais uniforme de freqüências no espectro dos sinais de música (ver Figura 4.5) conduz, em média, a valores de CE mais altos para tais sinais, em comparação a sinais de fala (ver Figura 4.6).

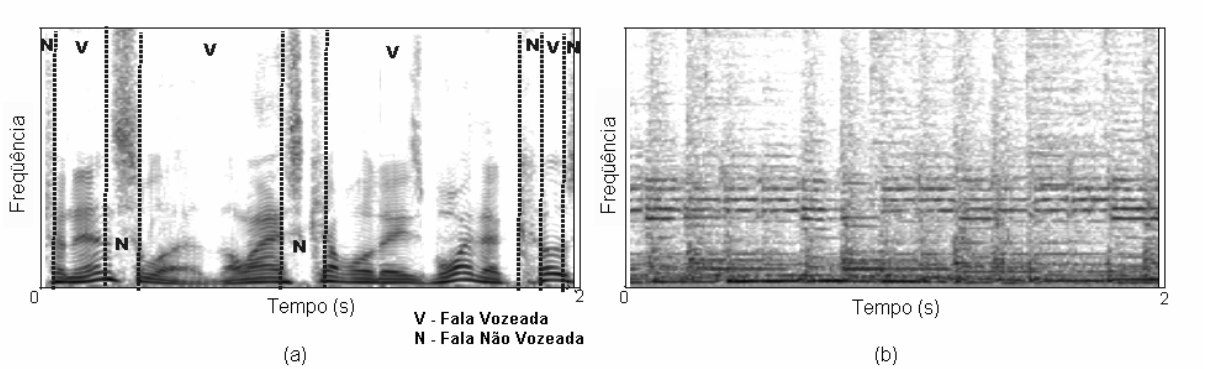


Figura 4.5: Espectrograma: (a) sinal de fala; (b) sinal de música.

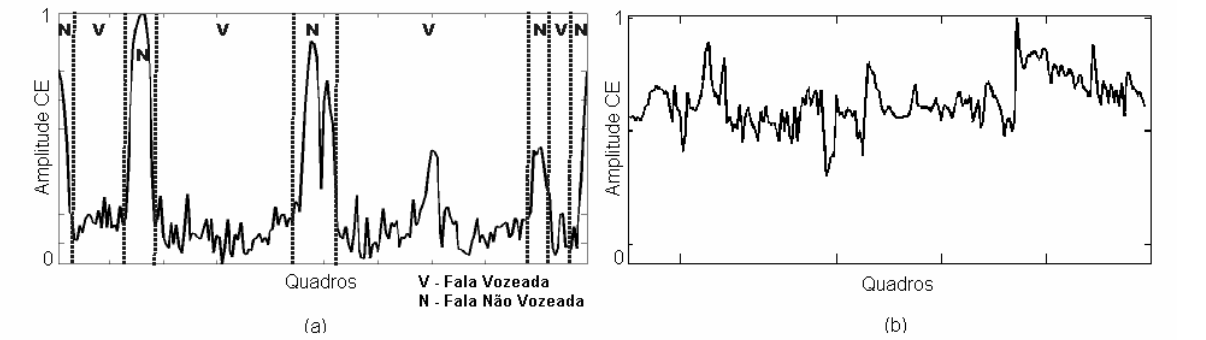


Figura 4.6: Comparação dos valores de CE: (a) sinal de fala; (b) sinal de música.

Em [35], o parâmetro de CE é usado na classificação de gêneros musicais.

4.2.2.2 Ponto de *Rolloff* Espectral (RE)

Assim como o CE, o *rolloff* espectral é uma medida da forma do espectro [46]. É definido como a freqüência F que corresponde a $x\%$ da distribuição de magnitude:

$$\sum_{k=1}^{F}|X(k)| = \frac{x}{100}\cdot\sum_{k=1}^{K}|X(k)| \tag{4.5}$$

Diversos valores são adotados para x como, por exemplo, 80% em [35], 85% em [46] e 95% em [23].

Assim como o CE e a TCZ, o parâmetro de RE é usado para distinguir segmentos vozeados e não-vozeados em sinais de fala, bem como para discriminar sinais de fala e música. Em [35], também é usado na classificação de gêneros musicais.

4.2.2.3 Fluxo Espectral

É uma medida da diferença espectral quadro a quadro. Caracteriza a mudança na forma do espectro. Seu valor é determinado por [46]:

$$FE_q = \sum_{k=1}^{K} \left[N_q(k) - N_{q-1}(k) \right]^2 \tag{4.6}$$

onde $N_q(k)$ e $N_{q-1}(k)$ denotam a magnitude normalizada da TFD no quadro q e $q-1$, respectivamente.

Os sinais de fala alternam períodos de grande variação no fluxo espectral, na transição entre segmentos vozeados e não-vozeados, com períodos aproximadamente constantes (segmentos vozeados). Por outro lado, a música apresenta uma taxa de mudança mais constante [23]. Portanto, em média, o valor de FE é maior para sinais de música, como ilustrado na Figura 4.7.

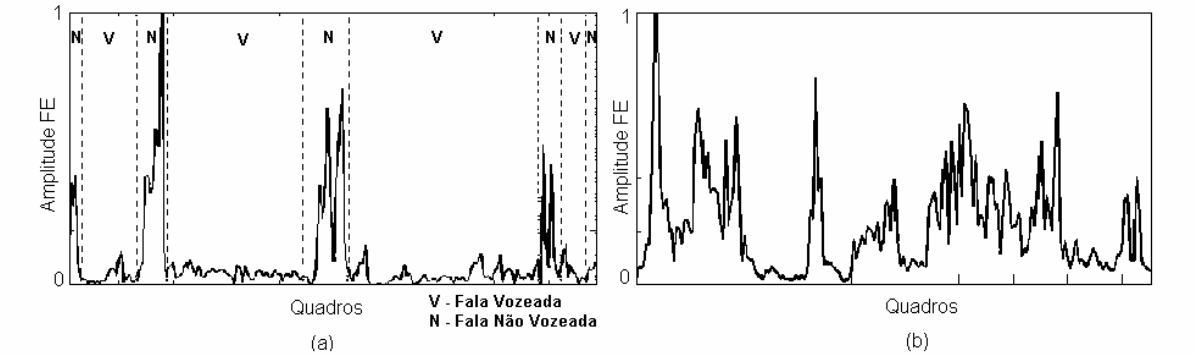


Figura 4.7: Comparação dos valores de FE: (a) sinal de fala; (b) sinal de música.

4.2.2.4 Freqüência Fundamental (f_0)

Algumas vezes, na literatura, a extração da f_0 (parâmetro físico) é também chamada de extração do *pitch* (parâmetro perceptual). Como são parâmetros associados, as duas formas são consideradas corretas para aplicações em CSA.

Um sinal é dito harmônico quando seu espectro é formado, principalmente, por uma série de picos na magnitude do espectro, compostos pela f_0 e múltiplos inteiros de seu valor. A extração da f_0 pode ser usada para medir a característica harmônica de um sinal [3]. Assim, quando é possível extrair um valor de f_0, o sinal é dito harmônico. Quando isso não é possível, o valor de f_0 é definido como zero e o sinal é dito não harmônico.

O som da maioria dos instrumentos musicais é harmônico. Por outro lado, o sinal de fala possui característica mista, alternando períodos harmônicos, em segmentos

vozeados, e não-harmônicos, em segmentos não-vozeados. Em sinais de canto, tende-se a prolongar a duração dos segmentos vozeados, quando comparados aos sinais de fala.

A extração da f_0 em sinais de áudio pode ser obtida por diversos métodos em diferentes domínios. Não existe método que produza resultados excelentes para todos os tipos de sinais [3]. Portanto, o melhor método depende da aplicação a qual ele se destina.

Uma aproximação para extração do valor de f_0 é baseada na autocorrelação no domínio do tempo (ver Figura 4.8). A autocorrelação de um sinal periódico é também periódica, com o mesmo período do sinal. Portanto, o valor de f_0 é determinado como o inverso do período obtido na função autocorrelação. No caso da Figura 4.8, $f_0 \approx 588Hz$.

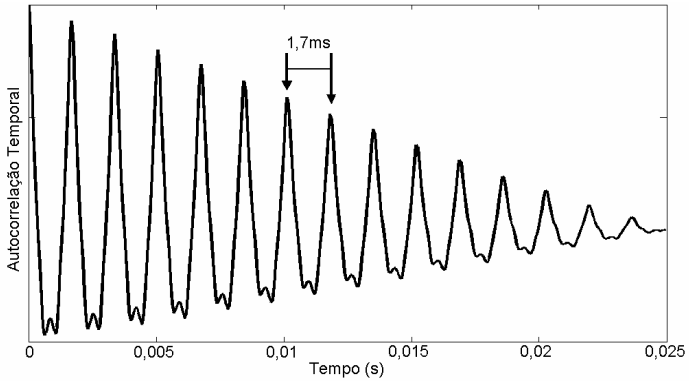


Figura 4.8: Extração do valor de f_0 a partir da autocorrelação temporal.

Uma outra aproximação para extração do valor de f_0 baseia-se na magnitude no domínio da freqüência. A Figura 4.9 apresenta a representação da magnitude no domínio da freqüência do mesmo sinal usado na Figura 4.8. O valor da f_0 é determinado pela diferença entre duas freqüências harmônicas consecutivas.

Outros métodos para extração da f_0 podem ser encontrados em [10].

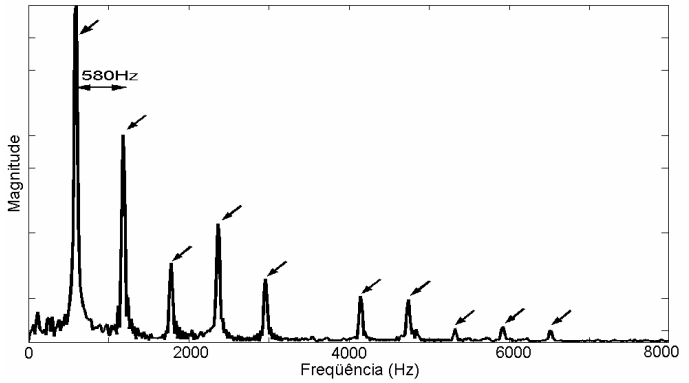


Figura 4.9: Extração do valor de f_0 no domínio da freqüência.

4.2.3 Parâmetros no Domínio Tempo-Freqüência

4.2.3.1 Energia de Modulação a 4 Hz (EM4)

O sinal de fala apresenta um pico na energia de modulação em torno da taxa silábica de 4 Hz [23]. Esse comportamento, geralmente, não ocorre para outros sinais, como música e canto. Portanto, o valor deste parâmetro tende a ser significativamente maior para os sinais de fala, como ilustrado na Figura 4.10.

O procedimento clássico para a extração da EM4 [26] consiste em filtrar cada quadro do sinal segmentado por um banco de 40 filtros na escala *mel*[2]. A energia é extraída em cada canal e filtrada por um filtro passa-faixa, com freqüência central de 4 Hz. A energia filtrada é somada para todos os canais e normalizada pela energia média no quadro. A modulação é obtida computando a variância da energia filtrada em 1 s de sinal.

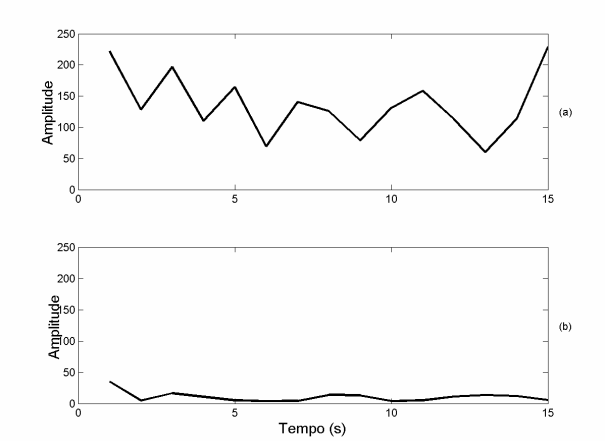


Figura 4.10: Comparação dos valores de EM4: (a) sinal de fala; (b) sinal de música.

4.2.3.2 Vibrato

O vibrato é um parâmetro característico de sinais de canto. Consiste em um *pitch* (f_0) estacionário, modulado na faixa de freqüências entre 4 e 8 Hz [10]. O vibrato não está presente em todos os segmentos em sinais de canto, mas está presente em poucos trechos de fala [47]. Assim, tal parâmetro é importante na discriminação de sinais de canto e fala.

A identificação do vibrato depende do comportamento da f_0 ao longo do tempo, ou

[2] A escala *mel* segue um modelo de percepção auditiva humana. As freqüências centrais dos 13 primeiros filtros na escala *mel* são espaçadas linearmente (distantes de 133,33Hz). As freqüências centrais dos demais 27 filtros são espaçadas logaritmamente (fator multiplicativo de 1,071)

seja, da forma da trilha da f_0. Trilhas da f_0 caracterizadas pelo vibrato possuem uma forma ondulada. A Figura 4.11 apresenta o exemplo de trilhas da f_0 para a vogal "I" falada e cantada por uma voz feminina em inglês. Em [48], são apresentados diversos métodos para extração das trilhas da f_0.

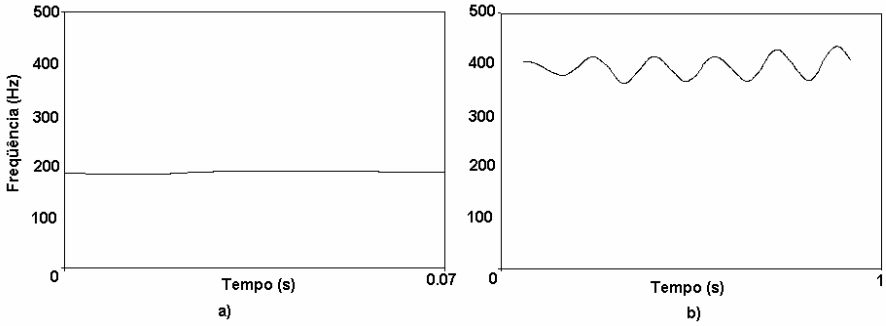


Figura 4.11: Comparação da trilha de f_0 : (a) fala; (b) canto.

O parâmetro do vibrato também é importante para distinguir o canto dos instrumentos musicais. Portanto, tal parâmetro será explorado no Capítulo 5.

4.2.4 Parâmetros Baseados em Probabilidade *a Posteriori*

Introduzidos em [30], estes parâmetros são calculados na saída de um modelo acústico treinado em fala e são normalmente utilizados em sistemas de reconhecimento automático de fala. Dado o segmento de um sinal de áudio na entrada do modelo acústico, em sua saída é obtida a probabilidade *a posteriori* de que aquele sinal pertença a cada uma das classes de saída. Geralmente, cada classe de saída representa um fonema.

4.2.4.1 Entropia

Entropia é uma medida da incerteza ou do grau de desordem em uma dada distribuição [29]. Ela é definida como:

$$H_q = -\sum_{k=1}^{K} P(q_k | x_q) \cdot \log_2 P(q_k | x_q) \tag{4.7}$$

onde x_q representa o segmento do sinal de áudio para o quadro q, q_k é a k-ésima classe (fonema) de saída do modelo acústico, e $P(q_k | x_q)$ é a probabilidade *a posteriori* da classe q_k, dado o sinal de entrada x_q.

Geralmente, no caso de um sinal de fala, o valor da probabilidade *a posteriori* de um fonema em particular (o fonema reconhecido) é muito maior do que os outros. Portanto, o valor da entropia nesse instante, dado pelo somatório das probabilidades, é baixo. Por outro lado, em um sinal de música, por não haver nenhum fonema que possa ser reconhecido, os valores das probabilidades serão distribuídos de forma mais uniforme, resultando em um valor de entropia maior.

4.2.4.2 Dinamismo

Dinamismo é uma medida da taxa de mudança de uma quantidade [29]. Ele mede o comportamento dinâmico dos valores de probabilidades, definido como:

$$D_q = \sum_{k=1}^{K} \left[P(q_k | x_n) - P(q_k | x_{n+1}) \right]^2 \tag{4.8}$$

Normalmente, os sinais de fala apresentam transições regulares nos valores de probabilidades, correspondendo à transição de fonemas durante a fala. Em contrapartida, as probabilidades de sinais de música variam menos e apresentam transições menos regulares, uma vez que, em nenhum momento, o modelo consegue decidir qual o fonema presente. Portanto, esse parâmetro apresenta valores maiores para sinais de fala.

4.2.5 Outros Parâmetros

Existem ainda outras formas de se obter parâmetros em CSA. Por exemplo, aqueles baseados em MFCC, calculados no domínio do cepstro. Outros parâmetros são extraídos do sinal de áudio em sua forma compactada, de acordo com padrões de compressão MPEG [46]. Em [49], são comparados parâmetros baseados em MFCC e MPEG, para a classificação de sinais característicos de eventos esportivos. É possível também extrair parâmetros no domínio da freqüência usando outras transformações, além da tradicional TFD. Por exemplo, a partir da transformada wavelet [50], como em [46] e [42].

4.3 Classificadores de Padrões

Um determinado parâmetro extraído de uma série de sinais define o espaço de parâmetros. O classificador de padrões tem a função de realizar o mapeamento do espaço de parâmetros para o espaço de decisão [15]. Para isso, deve encontrar limites que dividam

o espaço de parâmetros em regiões que correspondam a classes individuais. A Figura 4.12 ilustra a função de um classificador.

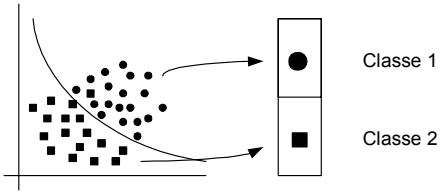


Figura 4.12: Definição das classes pelo limite de divisão determinado pelo classificador.

Alguns dos principais classificadores utilizados em CSA incluem: distância mínima [15], k-*nearest-neighbour* (k-NN) [51], modelo de mistura Gaussiana (GMM) [46], rede neural [15], modelo oculto de Markov (*Hidden Markov Model* – HMM) [52] e baseado em regras [3].

As vantagens e desvantagens de cada tipo de classificador em aplicações baseadas em CSA ainda não são óbvias [15]. Entretanto, na literatura, concorda-se que o uso de classificadores diferentes não altera significativamente os resultados. A extração dos parâmetros mais relevantes consiste na etapa decisiva para se obter um melhor desempenho em CSA [15], [23].

4.4 Pós-processamento

Esta etapa final é opcional e pretende corrigir pequenos erros de classificação, além de controlar o comportamento dinâmico do algoritmo [15], através da aplicação de uma suavização dos resultados. O algoritmo apresentado no Capítulo 5 possui uma etapa de pós-processamento, considerada em mais detalhes na seção 5.2.5.

4.5 Conclusões

Neste capítulo foi discutida a estrutura padrão para o processamento automático de CSA em geral, composta pelas etapas de extração de parâmetros, classificador de padrões e pós-processamento (opcional). Foi apresentada uma coletânea dos principais parâmetros considerados na literatura. A seleção dos parâmetros mais relevantes, ou seja, aqueles que apresentam comportamento mais semelhante para sinais de mesma classe, e mais distinto para sinais de classes diferentes, consiste na etapa mais importante do processamento automático de CSA.

Capítulo 5

Segmentação do Canto Dentro de Sinais de Música

5.1 Introdução

Após tratar o problema de CSA em geral, este capítulo está direcionado ao estudo de um algoritmo que realize a segmentação do canto dentro de sinais de música.

Comparado a problemas clássicos em CSA, como a discriminação entre sinais de fala e música, o problema em questão é um pouco mais complexo, pelos seguintes motivos:

- **Canto Sobreposto aos Instrumentos Musicais**

Geralmente, em uma música, os segmentos contendo canto estão sobrepostos aos instrumentos musicais, ocupando o mesmo espaço no tempo e no espectro de freqüências. Entretanto, tais segmentos devem ser identificados unicamente como canto.

- **Distinção das Classes**

As classes de canto e instrumentos musicais não dispõem da mesma facilidade de padrões tão distintos que existem entre as classes de fala e música, vistos nos diversos parâmetros apresentados no Capítulo 4. É mais difícil encontrar padrões para distinguir confiavelmente canto e instrumentos musicais dentro de sinais de música.

- **Discriminação & Segmentação**

Na classificação a partir da discriminação, os sinais de entrada já se encontram pré-segmentados. Assim, a solução do problema contempla apenas a identificação correta das classes. Na condição de classificação a partir da segmentação, o sinal de entrada pode pertencer a uma ou mais classes. Nesse caso, a solução do problema é mais complexa, contemplando também a localização das possíveis transições de classes. A Figura 5.1 ilustra a diferença entre ambas as condições de classificação, para um mesmo sinal de música com duração de 10 s. Na condição de discriminação, o sinal de entrada já se encontra pré-segmentado em arquivos de 2 s de duração, os quais contêm apenas uma das classes a serem identificadas. Portanto, para se obter sucesso na classificação, é necessário

obter, em cada classe de saída, um número igual à quantidade de arquivos de entrada referentes àquela classe. Por outro lado, na condição de segmentação, é analisado o sinal de entrada em sua forma completa, contendo os 10 s de música. Aqui, para se obter o mesmo sucesso na classificação, devem-se obter na saída os intervalos exatos de ocorrência de cada uma das classes.

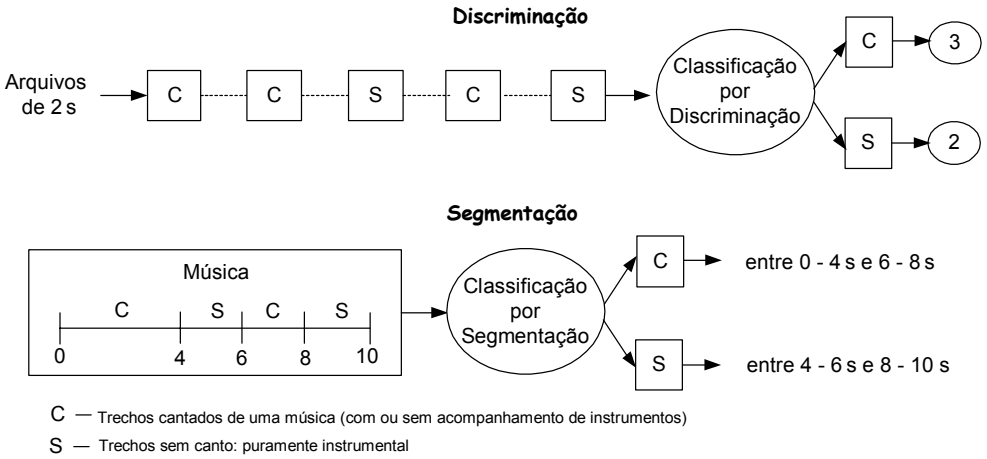


Figura 5.1: Diferenças entre classificação de sinais de áudio a partir de discriminação e de segmentação.

A abordagem proposta explora ao máximo a diferença no conteúdo harmônico do canto e dos instrumentos musicais, através da ampliação dos padrões identificados pelo parâmetro de trilhas de picos espectrais. Esse parâmetro é proposto em [3], em que o objetivo principal é o de discriminar os sinais de fala e canto. Então, o canto é identificado pelo vibrato, caracterizado por grandes variações das trilhas. Aqui, pretende-se identificar o canto não somente pelo vibrato, como em [3], mas também pelas pequenas variações das trilhas, correspondendo a pequenas variações de *pitch* que ocorrem na transição de fonemas. Além disso, é proposto um parâmetro secundário para auxiliar no processo de identificação do canto. A análise visual do espectrograma é de grande importância para a observação dos padrões extraídos a partir dos dois parâmetros propostos.

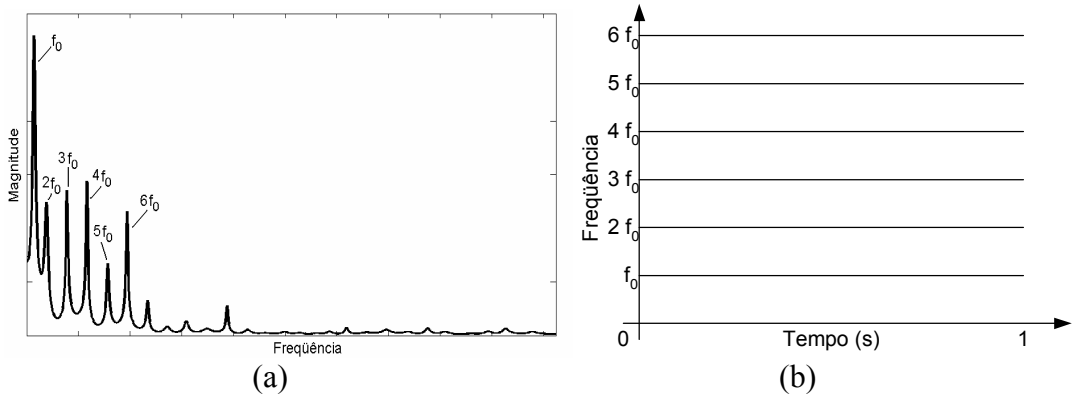
No Capítulo 6, o desempenho do algoritmo proposto é comparado com a técnica proposta em [4]. Esta última apresenta resultados de um experimento com objetivos idênticos de segmentação do canto dentro de sinais de música, utilizando uma abordagem diferente, discutida no Capítulo 3.

5.2 Modelo para a Segmentação do Canto Dentro de Sinais de Música

Como em qualquer problema de CSA, grande parte da solução depende da extração de parâmetros relevantes para a distinção das classes de áudio requeridas. No problema aqui proposto, as classes em questão são o canto e os instrumentos musicais.

5.2.1 Padrão para Identificação da Classe dos Instrumentos Musicais

Geralmente, um instrumento musical emite sons harmônicos. A magnitude da TFD do som harmônico é caracterizada por picos em valores múltiplos da freqüência fundamental, como ilustrado na Figura 5.2(a). Como, na maior parte dos casos, a nota de um instrumento é fixa, a magnitude da TFD se repete durante o trecho em que ela é produzida, dando origem a trilhas de picos espectrais constantes [3], como mostrado na Figura 5.2(b).



(a) (b)

Figura 5.2: (a) Espectro da nota harmônica de um instrumento musical; (b) trilhas de picos espectrais constantes produzidas durante o intervalo de tempo em que a nota é produzida.

5.2.2 Padrões Principais para Identificação da Classe de Canto

Os segmentos vozeados do canto também são harmônicos. Assim, da mesma forma, são formadas trilhas de picos espectrais. Entretanto, na maior parte dos casos, elas não são constantes, apresentando variações em razão do vibrato ou de pequenas variações de *pitch*.

- **Vibrato**

Como discutido no capítulo anterior, o vibrato produz uma trilha ondulada na freqüência f_0 e, por conseqüência, seus harmônicos como ilustrado na Figura 5.3.

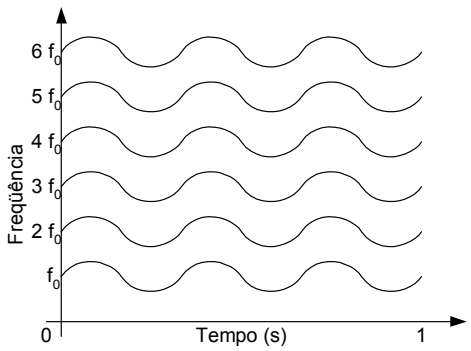


Figura 5.3: Trilhas de picos espectrais onduladas produzidas pelo vibrato.

- **Pequenas Variações de *Pitch***

Apesar de importante na identificação do canto, o vibrato não está presente em todas as músicas. Ainda que presente em uma música, com certeza não se encontra em todos os trechos cantados. Entretanto, mesmo na ausência do vibrato, a forma das trilhas em trechos com canto tende a ser variável. A transição dos fonemas durante o canto de uma palavra provoca pequenas variações no valor de *pitch*, como ilustrado na Figura 5.4.

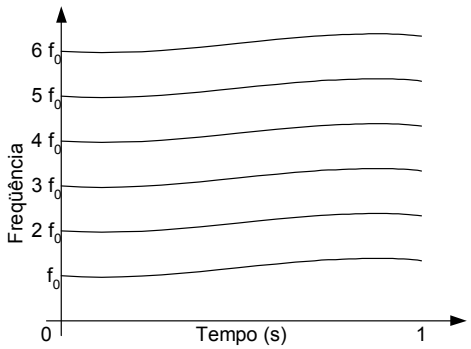


Figura 5.4: Trilhas de picos espectrais caracterizadas por pequenas variações de *pitch*.

Em [3], o canto é identificado apenas pelas trilhas variáveis em forma de ondas (vibrato). Entretanto, pequenas variações nas trilhas também caracterizam o canto e não identificam os instrumentos musicais, caracterizados por trilhas constantes. Assim, propõe-se ampliação dos padrões identificados pelo parâmetro de trilhas de picos espectrais apresentado em [3], de modo a identificar o canto não somente pelas grandes variações características do vibrato, mas também pelas pequenas variações.

5.2.3 Modelo Simplificado para a Segmentação do Canto

O modelo simplificado para a segmentação do canto dentro de sinais de música (ver Figura 5.5) baseia-se na extração do parâmetro de trilhas de picos espectrais e na seleção,

por um classificador, das trilhas variáveis que caracterizam o canto, em contraste às trilhas constantes que caracterizam os instrumentos musicais. Todo trecho não classificado como canto é considerado puramente instrumental.

Assim como em [3], o classificador utilizado neste trabalho é do tipo baseado em regras. Ao contrário dos classificadores automáticos, cuja fase de treinamento define automaticamente os limites que distinguem as classes, no classificador baseado em regras é o próprio programador quem os define, baseado em experiência adquirida na observação do comportamento dos parâmetros para diferentes sinais. Esse tipo de classificador pode apresentar menor complexidade computacional [15].

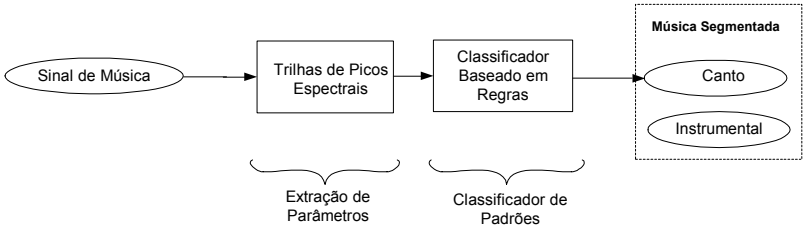


Figura 5.5: Modelo simplificado para a segmentação do canto dentro de sinais de música.

Um exemplo da aplicação direta do modelo simplificado do algoritmo pode ser verificado no Caso 1 da Seção 5.3: Análise de Casos.

A extração do parâmetro de trilhas de picos espectrais e a seleção das trilhas variáveis compreendem grande parte do esforço no desenvolvimento do algoritmo proposto neste trabalho. Portanto, sua descrição será tratada em mais detalhes na Seção 5.4.

5.2.4 Padrões Secundários Característicos do Canto

O modelo simplificado é capaz de identificar os segmentos de canto vozeado (harmônico) que formam trilhas variáveis no domínio tempo-freqüência. Grande parte dos segmentos contendo canto é identificada por tal padrão predominante. Por exemplo, na Figura 5.6 é apresentado o espectrograma do sinal de música mostrado na Figura 1.1, contendo canto e instrumentos musicais segmentados de acordo com a marcação manual destacada. De forma geral, é possível observar as trilhas constantes características dos instrumentos musicais e as trilhas variáveis características do canto. Entretanto, no início do trecho com canto, nas regiões marcadas com os números 1 e 2, são observados outros padrões que podem ser produzidos pelo canto. Na região #1, não se observa a presença de

trilhas variáveis no espectrograma. Contudo, é possível notar um padrão onde há uma maior magnitude nas altas freqüências, identificado pelos tons de cinza mais escuros. Esse é o padrão característico de segmentos fricativos do canto. Na região #2, composta por um segmento de canto vozeado, também não são formadas trilhas variáveis. Assim, propõe-se a extração de um segundo parâmetro que pretende detectar esses padrões secundários do canto, auxiliando no seu processo de segmentação.

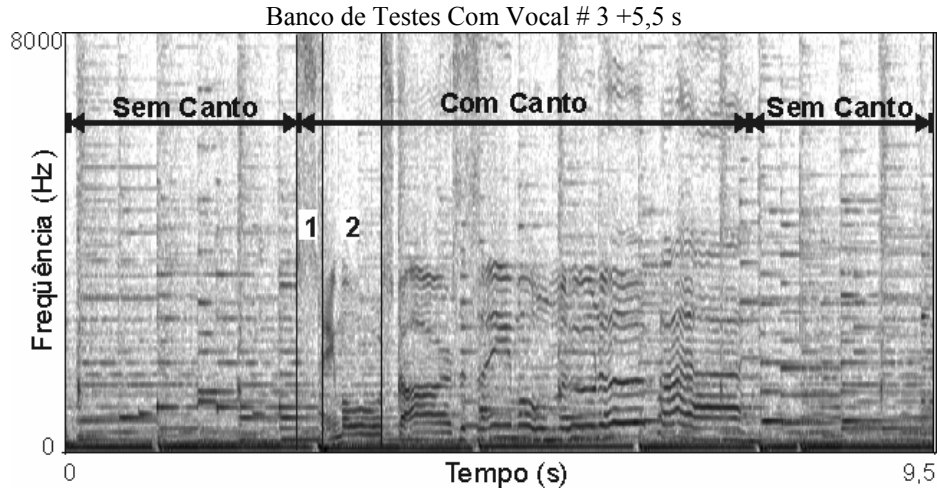


Figura 5.6: Espectrograma de um sinal de música destacando um segmento de canto fricativo, #1, e um segmento de canto vozeado que não produz trilhas variáveis, #2.

5.2.5 Parâmetro Secundário para a Segmentação do Canto

O parâmetro secundário proposto é chamado média da magnitude (MM). Consiste no cálculo do valor médio da magnitude da TFD, a cada quadro do sinal de entrada. Seu valor é calculado para duas bandas de freqüências. Assim,

$$MMB_q = \frac{\sum_{n=k_1}^{k_2} \left| X(k) \right|}{k_2 - k_1} \tag{5.1}$$

onde MMB_q é o valor da MM em baixas freqüências para o quadro q, calculada para a banda de freqüência entre k_1 e k_2, com $k_2 > k_1$.

O limite inferior k_1 = 500 Hz evita a influência de ruídos de baixa freqüência. Como discutido em [3], esse limite também é considerado na extração das trilhas de picos espectrais. O limite superior k_2 = 4 kHz considera que a maior parte da energia do canto vozeado está localizada até esse limite de freqüência.

Para altas freqüências, o valor da MM é calculado por:

$$MMA_q = \frac{\sum_{n=k_3}^{K} |X(k)|}{K - k_3} \tag{5.2}$$

onde MMA_q é o valor da MM em altas freqüências para o quadro *q*, calculada para a banda de freqüência entre k_3 e K, com $K > k_3$.

O limite superior K = 8 kHz equivale ao limite teórico da largura de banda para a taxa de amostragem adotada, de 16 kHz. O limite inferior k_3 = 5 kHz é definido como o início da banda de altas freqüências.

Os valores da MMA e MMB são normalizados pelos seus respectivos valores máximos considerando-se todos os quadros do sinal analisado.

O parâmetro secundário MM auxilia no processo de segmentação do canto em sinais de música, detectando segmentos de canto fricativo e atuando como uma etapa de pós-processamento:

- **Detecção do Canto Fricativo**

Trechos com canto produzido por segmentos fricativos não produzem trilhas no espectrograma. Entretanto, esses segmentos são caracterizados por sua maior magnitude em altas freqüências. Esse padrão é detectado pelo maior valor da MMA em relação à MMB, respeitando limites mínimos dessa relação e tempos mínimos de duração, de acordo com regras do tipo: $t_{\min} > 75$ ms e $MMA_q > 1,65 \cdot MMB_q$ e $MMA_q \geq 0,55$, onde $t_{\min}$ é o tempo mínimo de duração do segmento de canto fricativo.

Exemplos de aplicação da detecção de canto fricativo podem ser verificados nos Casos 3 e 4 da Seção 5.3: Análise de Casos.

- **Pós-Processamento**
  - **Ajuste de Limites do Canto Vozeado**

O parâmetro principal, de trilhas de picos espectrais variáveis, identifica o canto somente pela seleção das trilhas variáveis. Algumas vezes, essa seleção pode não identificar corretamente os limites de início e final dos segmentos contendo canto. Isso porque um segmento cantado pode produzir um trecho de *pitch* constante, imediatamente seguido por um trecho onde há variação das trilhas. Além disso, pode ocorrer também que

um mesmo segmento de canto simplesmente não produza trilhas durante toda a sua duração. Finalmente, pelo elevado número de restrições que a seleção das trilhas variáveis deve considerar (ver Seção 5.4), há chances de que a sua seleção não identifique corretamente os limites de início e fim dos trechos cantados. Assim, é proposto que a variação de valor da MMB seja usada para estender tais limites. Isso é justificado pelo fato de que, geralmente, o início e o final de um segmento contendo canto vozeado tendem a produzir uma variação maior do valor da MMB. Essa característica é explorada segundo as regras apresentadas a seguir:

i) Analisando toda a extensão do sinal, são escolhidos quadros candidatos a iniciar um segmento contendo canto de acordo com variações positivas da MMB;

ii) A partir do quadro onde se inicia uma trilha variável, é realizada uma varredura em direção ao início do sinal, buscando encontrar o primeiro quadro escolhido por (i). Esse valor é considerado como o novo quadro de início do segmento contendo canto;

iii) O mesmo procedimento é adotado para identificar o final de um trecho com canto. Nesse caso, a varredura segue em direção ao final do sinal, buscando encontrar o primeiro candidato a finalizar o segmento contendo canto, baseado em variações negativas da MMB.

Exemplos de aplicação do ajuste de limites de canto vozeado podem ser verificados nos Casos 2 e 3 da Seção 5.3: Análise de Casos.

- **Suavização dos Resultados**

A suavização dos resultados tem a função de alterar o estado de um trecho intermediário, baseado na predominância de um outro estado em trechos vizinhos. Tal procedimento é aplicado à localização de segmentos intermediários de canto não identificados por trilhas variáveis ou por canto fricativo. Assim, respeitando-se certos limites, tais como tempo de duração do segmento e valores médios da MMA ou MMB, esse trecho intermediário tem seu estado alterado, dando origem a um único segmento de canto. As principais regras de suavização são apresentadas a seguir:

i) $t_t < 1{,}9\,\text{s}$ e $\left(\overline{\text{MMB}_t} > 0{,}39 \text{ ou } \overline{\text{MMB}_t} > 0{,}88 \cdot \overline{\text{MMB}} \text{ ou } \overline{\text{MMB}_t} < \overline{\text{MMA}_t}\right)$, onde o índice t corresponde ao trecho intermediário, e $\overline{\text{MMB}_t}$ consiste no valor médio da MMB no trecho intermediário;

ii) ou $t_t < 1{,}25\,\text{s}$ e $\overline{\text{MMB}_t} > 0{,}741 \cdot \overline{\text{MMB}}$;

iii) ou $t_t < 0{,}7$.

A suavização dos resultados adquire maior importância em condições mais adversas para o tratamento do algoritmo, quando não é possível selecionar trilhas variáveis de forma satisfatória, como nos Casos 3 e 4 da Seção 5.3: Análise de Casos.

No banco de dados de música usado, composto por arquivos com 15 s de duração, existem sinais de música puramente instrumental. Portanto, são também aplicadas regras de suavização específicas para esses casos. Assim, se durante toda a extensão do sinal, o número e a duração dos trechos identificados como canto forem menores do que os limites predefinidos, conclui-se que não há canto nesta extensão do sinal. Os principais limites considerados para classificar um sinal de música como puramente instrumental são apresentados a seguir:

i) Se $\overline{t_{\text{trilhas}}} < 115\,\text{ms}$ e $n_{\text{t\_trilhas}} \leq 6$, onde $\overline{t_{\text{trilhas}}}$ é o tempo médio de duração das trilhas de picos espectrais variáveis selecionadas, e $n_{\text{t\_trilhas}}$ é o número de trechos de canto obtido pela seleção de trilhas;

ii) ou se $\overline{t_C} \leq 215\,\text{ms}$ e $n_{\text{t\_canto}} \leq 2$, onde $\overline{t_C}$ é o tempo médio de duração do canto detectado após o ajuste de limites do canto vozeado e detecção de segmentos fricativos, $n_{\text{t\_canto}}$ é o número de trechos de canto detectados.

5.2.6 Modelo Completo para a Segmentação do Canto

Considerando-se os padrões detectados pelo parâmetro secundário, o modelo completo proposto para a segmentação do canto dentro de sinais de música é o ilustrado na Figura 5.7. Esse modelo contempla as etapas de extração de parâmetros, classificador de padrões e pós-processamento, comuns a todos os problemas relativos à CSA, como visto no capítulo anterior.

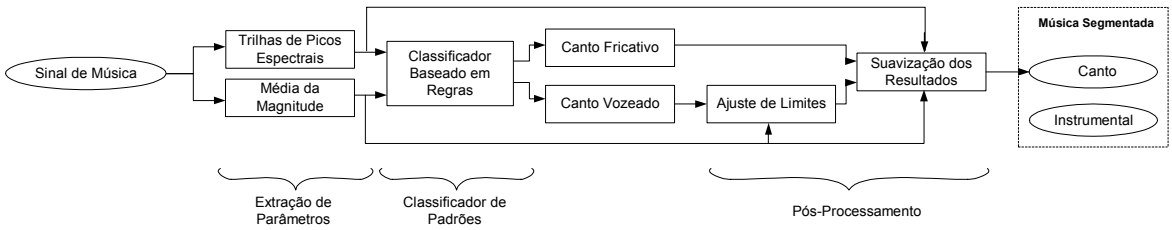


Figura 5.7: Modelo completo para a segmentação do canto em sinais de música.

5.3 Análise de Casos

Os sinais de música são compostos pela combinação de diferentes estilos de canto e dos mais diversos instrumentos musicais. O algoritmo proposto baseado no modelo completo propõe uma solução robusta para a segmentação do canto dentro de sinais de música, considerando várias situações possíveis. O desempenho do algoritmo é avaliado para algumas condições apresentadas nos casos discutidos a seguir. Os Casos 3 e 4 compreendem as situações mais adversas para o tratamento do algoritmo.

5.3.1 Caso 1: Canto caracterizado pelo vibrato

O canto caracterizado pelo vibrato representa a situação ideal para o tratamento do algoritmo, uma vez que as ondulações nas trilhas de picos espectrais são facilmente identificadas. Normalmente, casos desse tipo podem ser tratados pelo modelo simplificado.

A Figura 5.8(a) apresenta o espectrograma de um sinal de música que contém canto (caracterizado pelo vibrato) em toda a sua extensão. A Figura 5.8 (b) apresenta o resultado obtido após a seleção das trilhas de picos espectrais variáveis. Foram selecionadas trilhas durante todo o trecho, que confirmam a presença do canto em toda a sua extensão.

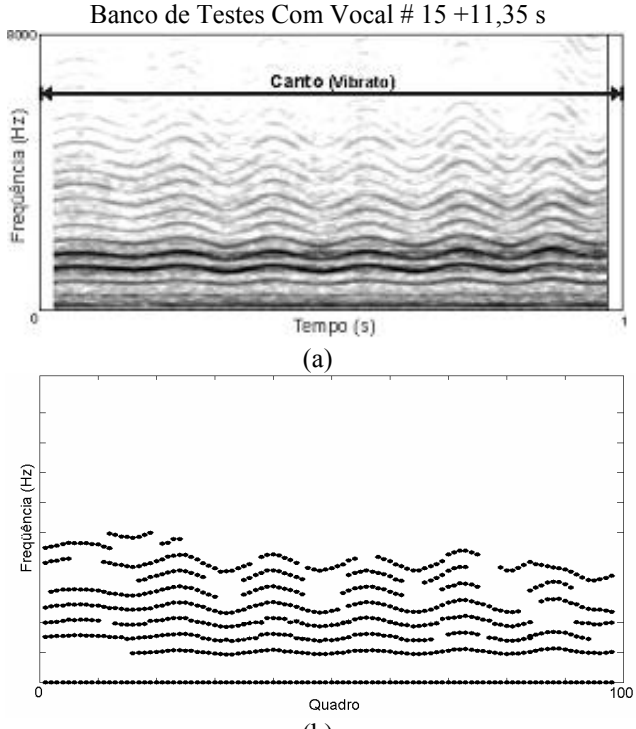


Figura 5.8: (a) Espectrograma do sinal de canto (vibrato); (b) trilhas variáveis selecionadas pelo algoritmo.

5.3.2 Caso 2: Canto caracterizado por pequena variação de *pitch*

Uma pequena variação de *pitch* não pode ser considerada uma situação ideal para o tratamento do algoritmo proposto, pois não é tão facilmente identificável como o vibrato. Apesar disso, a existência de um certo número de trilhas de picos espectrais, que apresentem um certo grau de variação de *pitch*, torna viável a extração e seleção de trilhas no trecho. Nesse caso, a melhor precisão no desempenho do algoritmo pode ser garantida pelas funções desempenhadas pelo parâmetro secundário.

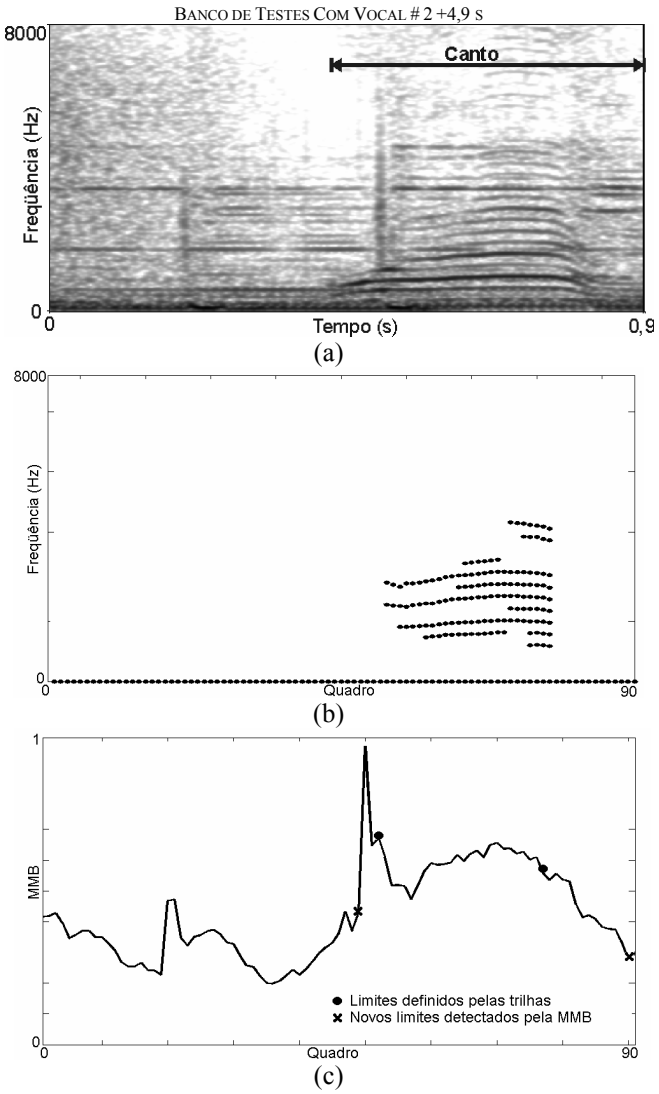


Figura 5.9: (a) Espectrograma do sinal de música contendo canto (pequena variação de *pitch*); (b) trilhas variáveis selecionadas pelo algoritmo; (c) extensão dos limites do canto pela variação de MMB.

A Figura 5.9(a) apresenta o espectrograma de um sinal de música, contendo o segmento de canto destacado, caracterizado por uma pequena variação de *pitch*. A Figura 5.9(b) mostra o resultado final da seleção das trilhas de picos espectrais variáveis. É

possível verificar que há uma divergência na localização dos limites de início e final do canto, comparado com aqueles apresentados na Figura 5.9(a). Contudo, como mostrado na Figura 5.9(c), tais limites são estendidos pelo ajuste realizado pelo parâmetro secundário, aproximando o resultado do algoritmo aos tempos obtidos usando marcação manual.

5.3.3 Caso 3: Canto vozeado caracterizado por poucas trilhas

A Figura 5.10(a) apresenta o espectrograma de um sinal de música contendo o segmento de canto destacado. Pela análise visual do espectrograma, percebe-se que as condições disponíveis para a análise do algoritmo são bastante adversas, uma vez que há poucas trilhas variáveis. Assim, a seleção das trilhas não pode fornecer muita informação sobre a presença do canto. O êxito na segmentação é garantido pelo parâmetro secundário.

Na Figura 5.10(b), os limites dos segmentos contendo canto, selecionados pelas trilhas, são estendidos pela análise da variação da MMB. Além disso, a relação entre MMA e MMB determina o segmento indicado de canto fricativo. Resta, portanto, um trecho intermediário onde existe canto, mas não se conseguiu selecionar nenhuma trilha. Entretanto, pelas regras de suavização, emendam-se os dois trechos vizinhos, identificando o segmento de canto apresentado, bastante próximo do obtido usando marcação manual.

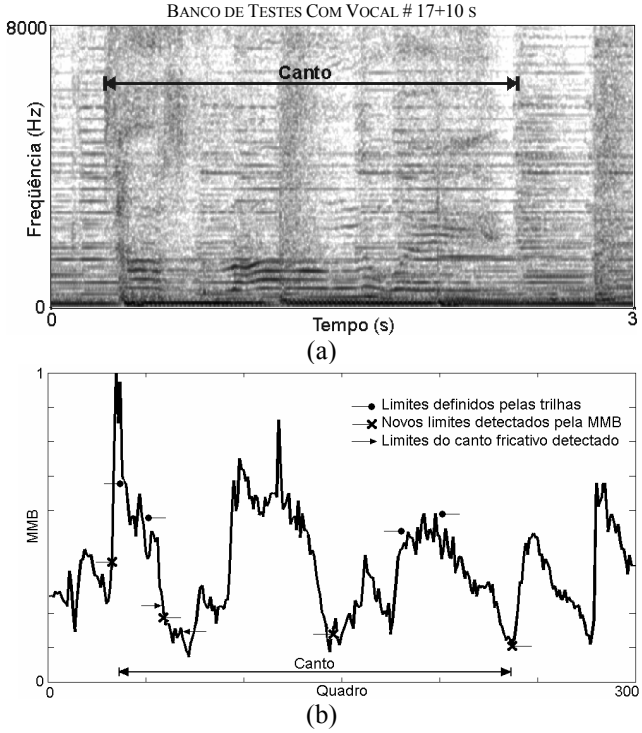


Figura 5.10: (a) Espectrograma do Sinal; (b) variação da MMB.

5.3.4 Caso 4: Canto puramente fricativo

No caso anterior, mesmo dependendo fortemente do parâmetro secundário, houve a necessidade de se localizar corretamente algumas trilhas variáveis, uma vez que o ajuste dos limites e a suavização dos resultados dependem da informação inicial de um trecho detectado a partir da seleção de trilhas. Há ainda outra situação em que, mesmo não sendo selecionada nenhuma trilha variável, ainda se obtém sucesso na localização do canto, como ilustrado na Figura 5.11. Nesse exemplo, o sinal de música é composto por um segmento de 1 s de canto puramente fricativo. São localizados dois trechos a partir da relação entre a MMA e MMB, como mostrado na Figura 5.11(b). Os dois trechos são emendados pelas regras de suavização aproximando o resultado final àquele obtido da marcação manual. Nesse caso, o sucesso na segmentação depende unicamente da localização de segmentos de canto fricativo e da suavização dos resultados, ou seja, exclusivamente do parâmetro secundário. Tal situação não é comum, pois, geralmente, não se encontram longos segmentos de canto puramente fricativo. Comumente, segmentos desse tipo caracterizam apenas alguns fonemas que compõem uma palavra e, portanto, auxiliam na localização dos limites do canto de uma palavra ou de algum trecho intermediário.

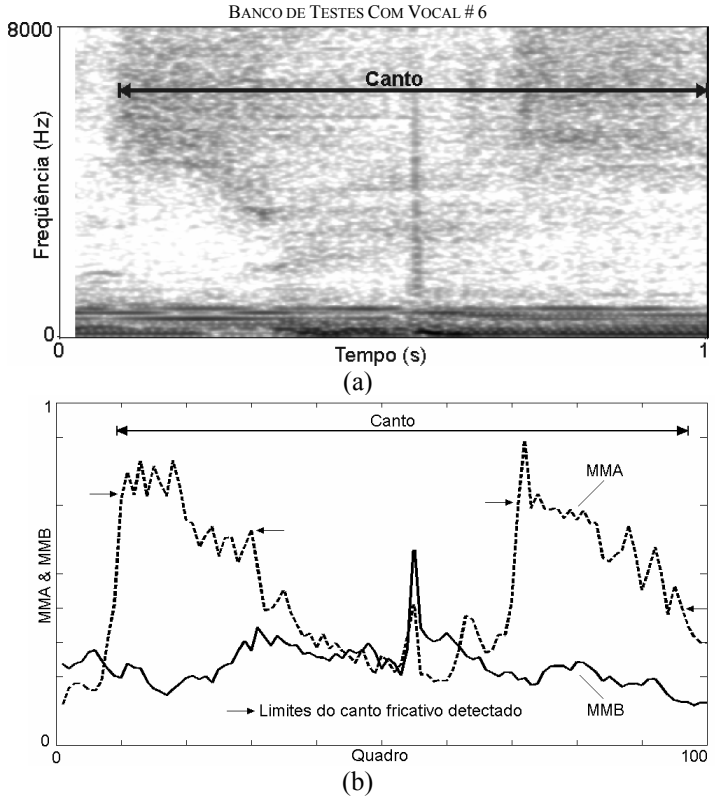


Figura 5.11: (a) Espectrograma de um segmento de canto fricativo; (b) MMA & MMB.

5.4 Extração e seleção das trilhas de picos espectrais variáveis

A abordagem proposta neste trabalho baseia-se na diferença do conteúdo harmônico do canto e dos instrumentos musicais. Tal diferença é evidenciada pela extração do parâmetro de trilhas de picos espectrais e seleção das trilhas variáveis, as quais identificam o canto pelo vibrato ou por pequenas variações de *pitch*. Grande parte do esforço no desenvolvimento do algoritmo concentra-se nessa etapa, cujo diagrama em blocos é mostrado na Figura 5.12 e discutido a seguir. Ao final desta Seção, na Figura 5.20, é apresentada a evolução de cada etapa do processamento para o sinal de entrada considerado no Caso 2 da Seção 5.3: Análise de Casos. O resultado final da seleção das trilhas variáveis para esse sinal já havia sido ilustrado na Figura 5.9(b) e é reapresentado na Figura 5.20(e).

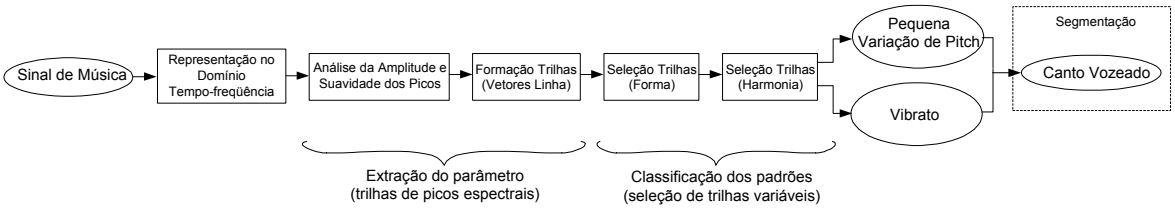


Figura 5.12: Diagrama em blocos da extração e seleção de trilhas de picos espectrais variáveis.

5.4.1 Representação do Sinal no Domínio Tempo-Freqüência

Uma vez que o parâmetro de trilha de pico espectral é formado no domínio tempo-freqüência, inicialmente, o sinal é segmentado no domínio do tempo, em quadros de 25 ms, com 15 ms de recobrimento (tais tempos também são considerados no cálculo do parâmetro secundário), usando uma janela de Hamming [21]. Em seguida, é necessário definir o tipo de transformação em freqüência a ser aplicado a cada quadro. Nesse caso, é considerado o fato de que as trilhas de picos espectrais buscam identificar as freqüências harmônicas de um sinal, caracterizadas por valores de pico na magnitude do espectro. Portanto, assim como em [3], a sua representação é obtida através de um modelo autoregressivo (AR), constituído somente por pólos, privilegia a localização dos picos no espectro [3]. O espectro gerado pelo modelo AR consiste em uma versão suavizada daquele gerado pela Transformada de Fourier, como ilustrado na Figura 5.13.

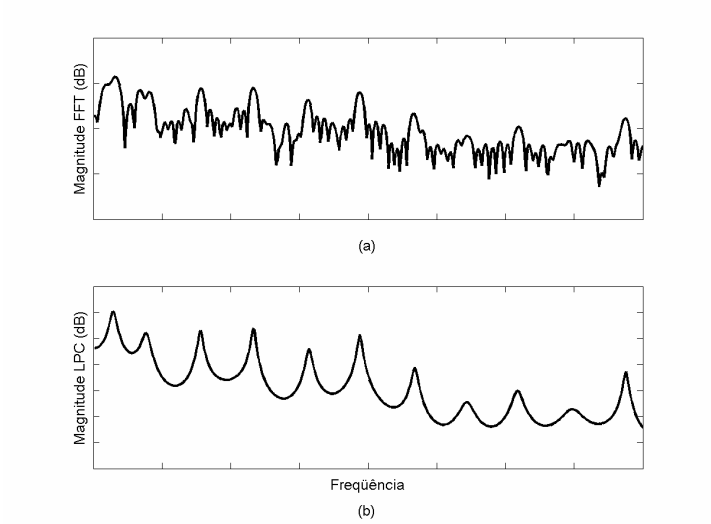


Figura 5.13: Comparação da magnitude do espectro obtido por: (a) FFT; (b) modelo AR.

Em [3] as trilhas de picos espectrais têm o objetivo principal de distinguir a fala do canto. Assim, são adotados modelos AR de três ordens diferentes: 40, 80 e 100. Comenta-se em [3] que a ordem 40 é suficiente para identificar a maioria dos segmentos contendo canto; ordem 80 é mais confiável para identificar segmentos com *pitch* entre 150–250 Hz, característico da fala feminina; e ordem 100 para identificar trechos com *pitch* entre 100–150 Hz, característico da fala masculina. Apesar deste capítulo considerar somente os sinais de canto, é adotado um modelo AR de ordem 80. Justifica-se essa opção pela proposta de identificar o canto não somente pelo vibrato, mas também por pequenas variações de *pitch*. A Figura 5.14(a) mostra que o modelo de ordem 40 não é suficiente para capturar as trilhas do canto caracterizado por pequenas variações de *pitch*. Em contraste, o modelo de ordem 80 [ver Figura 5.14(b)] apresenta resultados satisfatórios. Para segmentos de canto caracterizado pelo vibrato, [ver Figura 5.15(a)], é possível notar que o modelo de ordem 40 é suficiente, confirmando o observado em [3].

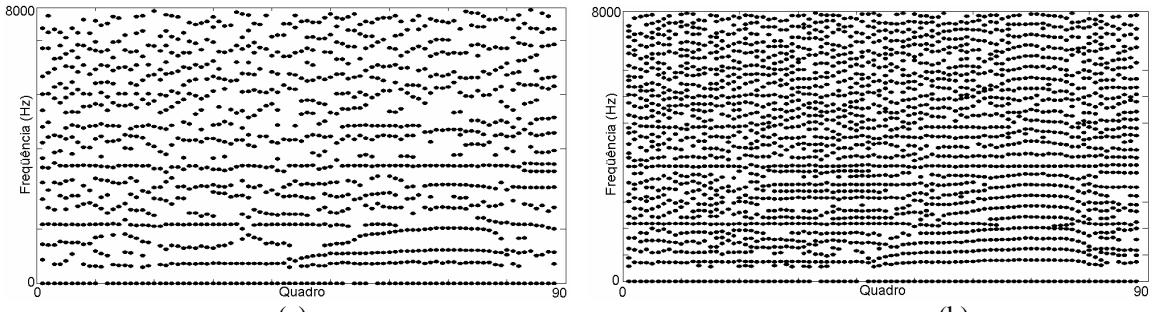
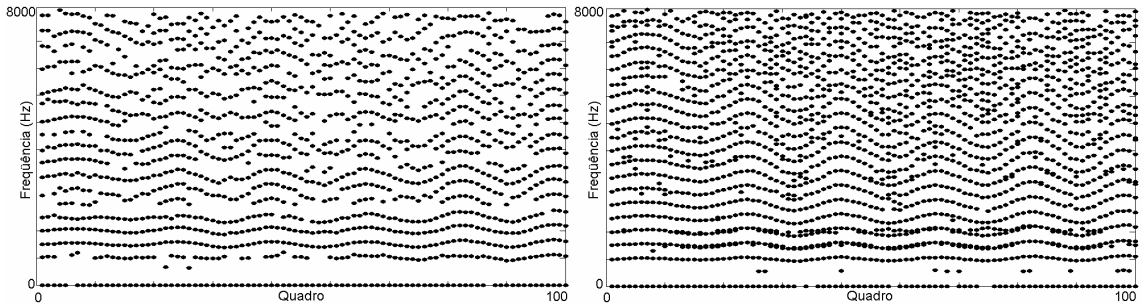


Figura 5.14: Captura dos picos espectrais do sinal de música mostrado na Figura 5.9(a) (contendo canto caracterizado por uma pequena variação de *pitch*) obtidos com modelo AR de ordem: (a) 40; (b) 80.



(a) (b)

Figura 5.15: Captura dos picos espectrais do sinal de música mostrado na Figura 5.8(a) (contendo canto caracterizado pelo vibrato) obtidos com modelo AR de ordem: (a) 40; (b) 80.

5.4.2 Análise da Amplitude e Largura dos Picos

Obtida a representação do sinal no domínio tempo-freqüência, uma primeira etapa para a extração das trilhas de picos espectrais, como em [3], consiste em excluir alguns picos detectados. Esses picos são descartados por apresentarem um reduzido valor de amplitude e serem muitos suaves. Em [3], não são apresentados os valores limites desta etapa. Neste trabalho, foram adotados os seguintes valores para descarte:

- Amplitude: $Ap_k \leq \overline{A}$, onde Ap_k é a amplitude de pico na freqüência k, e $\overline{A}$ é o valor médio da amplitude do espectro para o quadro atualmente analisado;
- Largura: $A_{k-4} \geq 0.81 \cdot Ap_k$ e $A_{k+4} \geq 0.81 \cdot Ap_k$.

Ao contrário do realizado em [3], esta etapa não considera a relação harmônica dos picos. No algoritmo proposto neste trabalho, essa análise é realizada na última etapa. Justifica-se essa opção pela dificuldade de obter, algumas vezes, um valor predominante de freqüência fundamental analisando apenas um quadro. Portanto, prefere-se extrair o máximo de picos nessa etapa, baseado somente na análise da amplitude e suavidade. O resultado ao final desta primeira etapa é mostrado na Figura 5.20(b).

5.4.3 Formação das Trilhas (Vetores Linha)

As trilhas visíveis na Figura 5.20(b) estão alinhadas na forma de vetores coluna, onde cada coluna é formada a cada quadro do sinal pelos valores das freqüências em que ocorrem picos espectrais selecionados por análise de amplitude e suavidade. Tais vetores fornecem uma informação estática no tempo, equivalente a um quadro do sinal. Entretanto,

a forma das trilhas é determinada pela variação dos valores ao longo do tempo. Portanto, é necessário alinhar essas trilhas na forma de vetores linha.

O espectrograma de um sinal de música apresenta um cenário complexo, sendo composto pela combinação da voz cantada (em diferentes estilos) e dos mais diversos tipos de instrumentos musicais. Assim, a transformação dos vetores coluna em vetores linha, através da concatenação correta entre os correspondentes valores de freqüências que formam as trilhas, consiste em uma etapa delicada. Em [3], essa etapa é apenas citada, não informando o método usado na sua implementação. Neste trabalho, optou-se pela formação das trilhas a partir de um critério de diferença e de duração mínima da trilha. O fluxograma da Figura 5.17 descreve o método desenvolvido.

A matriz $A_{m\times n}$ é formada por *m* linhas, onde *m* equivale ao quadro em que foi selecionado o maior número de freqüências, e *n* colunas, onde *n* equivale ao horizonte da análise (1 s), que inicia pela primeira linha. A validade do critério da diferença e, portanto, a validação de um elemento da matriz como pertencente a uma trilha, ocorre quando $\left(\left|a_q - a_{q\pm1}\right| \text{ ou } \left|a_q - a_{q\pm2}\right| \text{ ou } \left|a_q - a_{q\pm3}\right|\right) \le L$, onde a_q é o elemento analisado ($a_q = a_{\min}$ no início da análise), e $L = 78\,\text{Hz}$ para $f \le 2{,}35\,\text{kHz}$ e $L = 118\,\text{Hz}$ para $f > 2{,}35\,\text{kHz}$. Como ilustrado na Figura 5.16, o valor limite *L* acompanha a trajetória da trilha, sendo aplicado sempre em relação ao elemento a_q atualmente analisado.

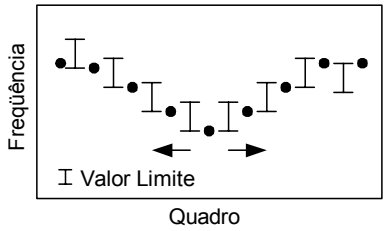


Figura 5.16: Valor limite acompanhando a trajetória da trilha.

É permitida uma falha de até dois quadros na formação das trilhas. Nesses casos, os valores são preenchidos com o valor médio dos quadros vizinhos. Finalizada a formação das trilhas da matriz $A_{m\times n}$, é reiniciado o processamento para o próximo intervalo de 1 s.

O resultado ao final desta etapa é ilustrado na Figura 5.20(c). Comparando-se tal resultado com aquele obtido da etapa anterior [ver Figura 5.20(b)], observa-se que alguns elementos isolados (que não formavam trilhas) presentes na Figura 5.20(b) são descartados, por não respeitarem o critério de duração mínima da trilha. Entretanto, o

resultado principal obtido nessa etapa é o de preservar, de uma forma geral, as trilhas visíveis na Figura 5.20(b), alinhadas em vetores coluna. Como, ao final da etapa atual, as trilhas estão alinhadas em vetores linha, é possível processá-las visando a seleção das trilhas variáveis que identificam o canto.

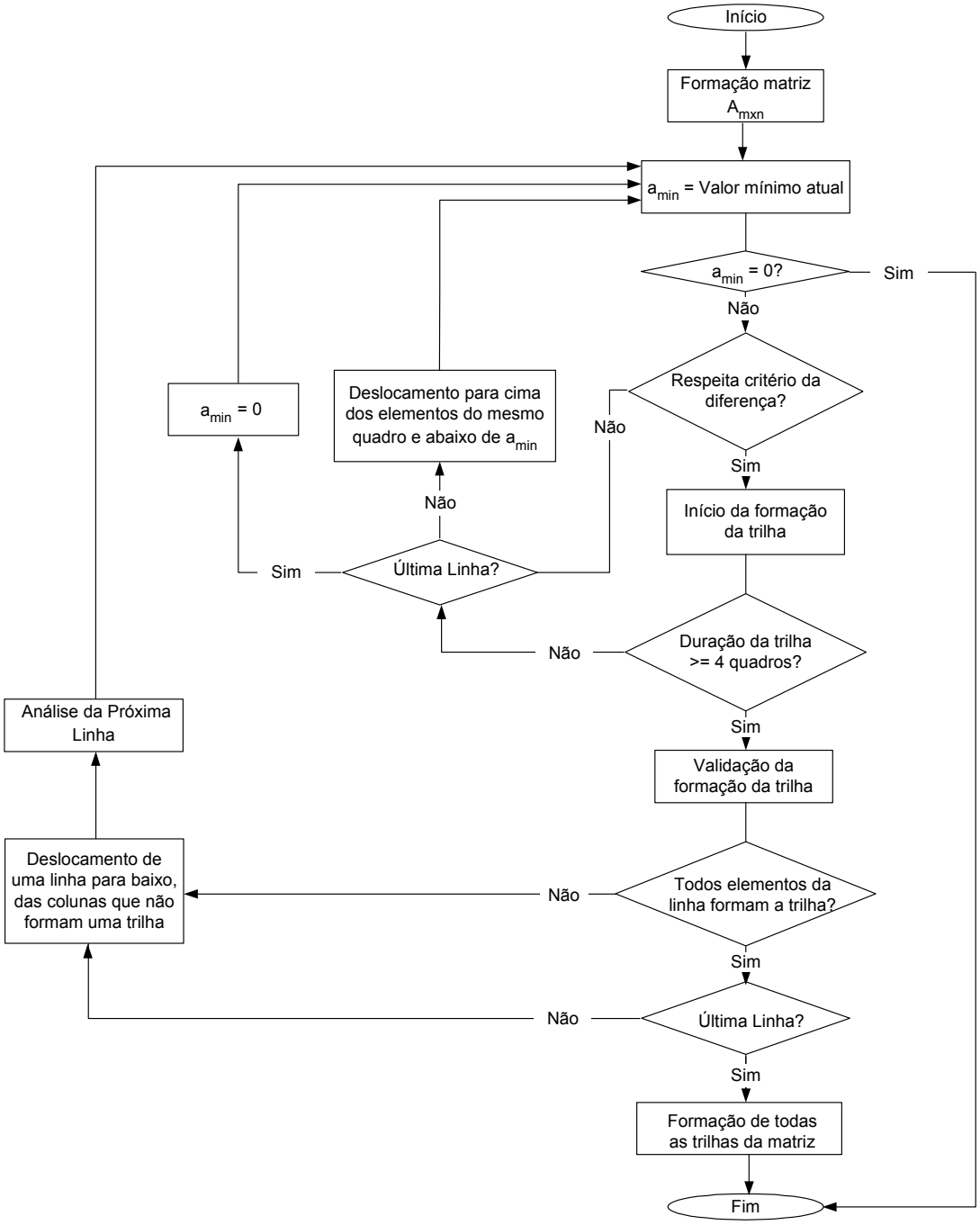


Figura 5.17: Fluxograma para a formação das trilhas a partir de vetores linha.

5.4.4 Seleção das Trilhas de Acordo com a Forma

A análise da forma das trilhas visa selecionar as trilhas variáveis que identificam o canto, seja pelo vibrato ou por pequenas variações de *pitch*. Nesta etapa se encontra a principal diferença em relação à seleção de padrões a partir das trilhas de picos espectrais proposta em [3], cuja análise da forma seleciona apenas as grandes variações de *pitch*, características da forma ondulada do vibrato.

Em [10], é observado que a taxa de variação das trilhas resultante do vibrato situa-se, geralmente, na faixa entre 4 e 8 Hz. É verificado também que uma taxa de variação entre 1 e 3 Hz é mais provável de ocorrer na transição entre fonemas, caracterizando uma pequena variação de *pitch*. Finalmente, é ainda observado que taxas de variação maiores do que um limite superior de 15 Hz estão além do limiar audível e são muito difíceis de serem produzidas pelo trato vocal humano [10]. Portanto, aqui também foram consideradas tais restrições. A análise da forma das trilhas é baseada em um critério de diferenças. A medida da taxa de variação depende da mudança de sinal no valor da diferença entre quadros consecutivos. É considerada tanto a forma completa de uma trilha quanto segmentos parciais. Assim, uma trilha pode ser selecionada ou excluída completa ou parcialmente, como mostrado pelo exemplo da Figura 5.18.

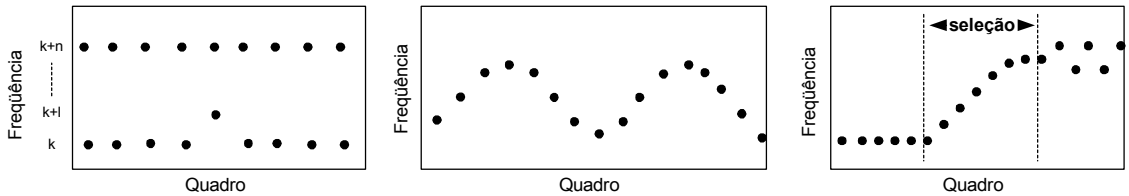


Figura 5.18: Análise da forma das trilhas (esboço): (a) descarte completo; (b) seleção completa; (c) seleção parcial.

Pequenas variações de amplitude de trilhas [ver Figura 5.18(a), $l \leq 8\,\mathrm{Hz}$], ainda são consideradas trilhas constantes e, portanto, também são descartadas. Trilhas caracterizadas pelo vibrato são, geralmente, selecionadas em sua forma completa, como ilustrado na Figura 5.18(b). A Figura 5.18 (c) apresenta o exemplo de uma seleção parcial, caracterizando uma pequena variação de *pitch*. Nesse caso, o primeiro trecho é descartado por ser constante, e o segundo, por apresentar uma taxa de variação maior do que 15 Hz.

Comparando a Figura 5.20(c) e Figura 5.20(d), é possível perceber que após a seleção de acordo com a forma, muitas trilhas consideradas constantes foram totalmente ou parcialmente descartadas.

5.4.5 Seleção das Trilhas Harmônicas

A seleção de uma simples trilha pela sua forma variável não garante a classificação do correspondente segmento como canto. Devido à complexidade dos espectrogramas de sinais de música, é possível se obter, artificialmente no processo de análise, trilhas variáveis para segmentos onde não existe informação de canto. Entretanto, como o canto possui característica harmônica, a forma variável de uma trilha deve se repetir para múltiplos inteiros da freqüência fundamental. Assim, para validar um segmento de canto, é necessário obter um mínimo de 3 trilhas harmonicamente relacionadas durante um tempo de duração mínima equivalente a 4 quadros, como ilustrado na Figura 5.19. Obviamente, são permitidas pequenas variações nas relações, da ordem de 30–40 Hz.

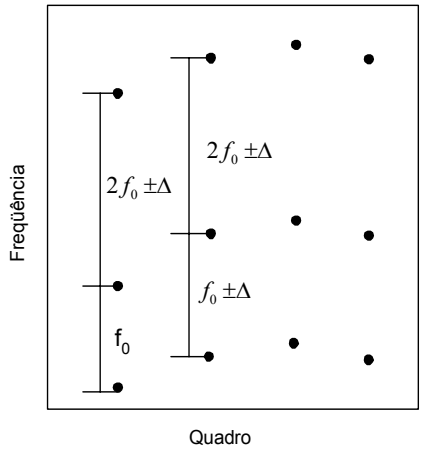


Figura 5.19: Análise da relação harmônica entre trilhas selecionadas pela forma variável.

O resultado da seleção harmônica, que consiste na última etapa da extração e seleção das trilhas de picos espectrais variáveis, é mostrado na Figura 5.20(e). Observa-se que foram selecionadas trilhas variáveis somente no segmento do sinal contendo canto.

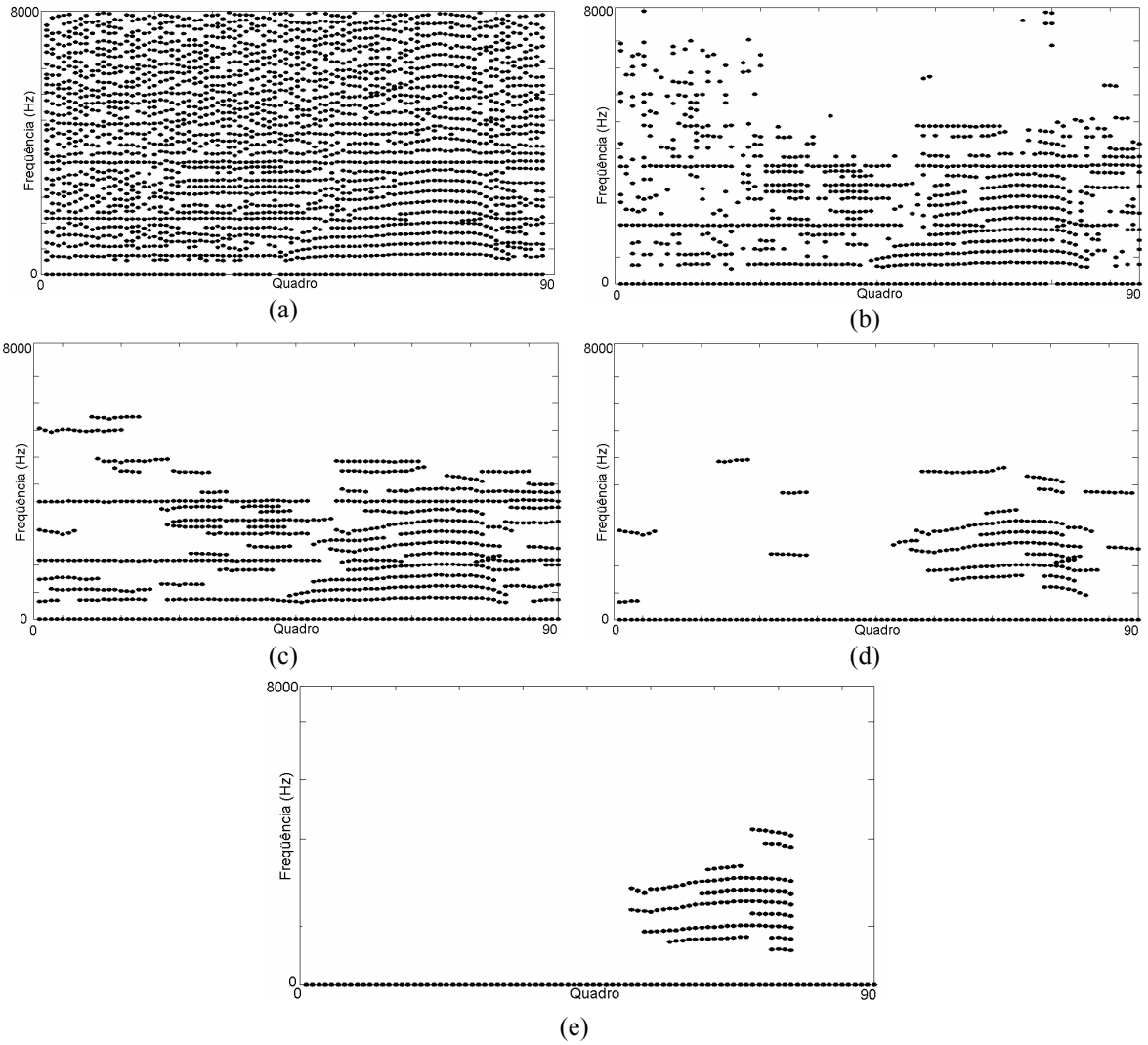


Figura 5.20: Evolução da etapa de seleção de trilhas de picos espectrais variáveis: (a) captura de todos os picos; (b) seleção por amplitude e suavidade; (c) formação das trilhas a partir de vetores linha; (d) seleção das trilhas variáveis; (e) análise harmônica das trilhas variáveis.

5.5 Conclusões

De acordo com o objetivo final do trabalho, este capítulo propôs um algoritmo para realizar a segmentação do canto dentro de sinais de música. A abordagem apresentada baseia-se na diferença do conteúdo harmônico do canto e dos instrumentos musicais. Essa diferença é evidenciada a partir dos padrões obtidos do parâmetro de trilhas de picos espectrais, proposto em [3]. O método desenvolvido neste capítulo sugeriu uma ampliação dos padrões obtidos de tal parâmetro, a fim de caracterizar melhor o canto, identificando-o não somente pelo vibrato (grandes variações das trilhas), como realizado em [3], mas

também por pequenas variações de *pitch*, ou seja, pequenas variações nas trilhas. A partir dessa modificação, foi proposto um modelo simplificado para segmentação do canto. Tal modelo aplica-se melhor para segmentos de canto vozeado caracterizados por trilhas de picos espectrais que apresentem um bom grau de variação, como no Caso 1 da Seção 5.3: Análise de Casos, cujo segmento de canto é caracterizado pelo vibrato.

Adicionalmente, foi proposto um parâmetro secundário para auxiliar no processo de segmentação do canto. A partir desse parâmetro é realizada a detecção de segmentos do canto fricativo e, como etapas de pós-processamento, são efetuados o ajuste dos limites do canto vozeado e a suavização dos resultados. Somando-se esses padrões secundários ao modelo simplificado, foi proposto um modelo completo para segmentação do canto dentro de sinais de música. O desempenho do modelo completo foi analisado nos Casos 2, 3 e 4 da Seção 5.3: Análise de Casos.

Finalmente este capítulo descreveu em detalhes as etapas necessárias para a extração do parâmetro de trilhas de picos espectrais e para a seleção das trilhas variáveis. Essa seção do processamento concentra grande parte do esforço para o desenvolvimento do algoritmo proposto, além de ser o maior responsável pelo seu bom ou mau desempenho, o qual será avaliado no próximo capítulo

Capítulo 6

Resultados Experimentais

6.1 Introdução

Este capítulo pretende avaliar e comparar o desempenho do algoritmo desenvolvido com aquele proposto em [4], que apresenta resultados em um experimento idêntico de segmentação do canto dentro de sinais de música. Para tal, é usado o mesmo banco de dados adotado em [4]. Esse banco foi utilizado inicialmente em [23] e posteriormente em [30], em experimentos de discriminação entre sinais de fala e música. Obviamente, para a avaliação proposta no presente trabalho, bem como em [4], foram considerados somente os sinais de música que compõem o banco, totalizando 101 arquivos com duração de 15 s cada um, amostrados originalmente a uma taxa de 22,05 kHz. Esses sinais foram gravados aleatoriamente de uma rádio FM no ano de 1996, contendo diversos gêneros musicais, com predominância da língua inglesa e incluindo alguns arquivos com sinais de música puramente instrumental.

A comparação dos resultados obtidos com aqueles apresentados em [4] indicam que em ambos a taxa de acerto situa-se na mesma faixa, em torno de 80%, mesmo considerando um método de medida de desempenho mais rigoroso na avaliação do desempenho do algoritmo proposto. Como vantagem, a abordagem atual apresenta menor complexidade computacional. Adicionalmente, ela permite discutir novos resultados, tais como a extensão dos testes para todo o banco de dados e a discriminação dos tipos de erro envolvidos no processo de segmentação, sugerindo alternativas para reduzi-los, quando possível.

Finalmente, a partir do algoritmo proposto, é apresentado o resultado de um primeiro experimento na tentativa de separar o canto de instrumentos musicais dentro de sinais de música.

6.2 Comparação dos Métodos de Análise de Desempenho

Neste trabalho, o método utilizado para analisar o desempenho do algoritmo é mais rigoroso do que o proposto em [4]. Em [4], para alcançar a melhor taxa de acerto, o erro é calculado usando janelas de 1,3 s do sinal analisado. Assim, dentro desse intervalo, não importam os tempos exatos da localização dos trechos com canto, e sim o tempo total de canto detectado. Portanto, não são percebidos os erros cometidos dentro do intervalo de 1,3 s, como ilustrado no exemplo da Figura 6.1.

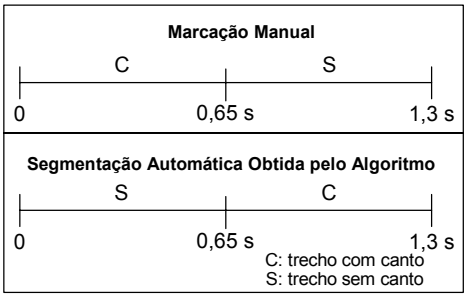


Figura 6.1: Exemplo de erro não percebido pelo método de análise de desempenho usado em [4].

Neste exemplo, que constitui uma situação limite, para não cometer erro o método de análise de desempenho proposto em [4] deve considerar que existe canto em 0,65 s do trecho. Como se percebe pelo resultado da segmentação automática, realmente obtém-se um trecho de canto com duração de 0,65 s e, portanto, a taxa de erro do método proposto em [4] é 0%. Entretanto, como é possível notar, a localização do trecho está totalmente incorreta. Outra desvantagem desse método é que ele não permite saber qual dos três tipos de erro está sendo cometido:

- Erro na detecção dos limites inferior ou superior do canto $\left(E_{\mathrm{L}} = E_{\mathrm{LI}} + E_{\mathrm{LS}}\right)$;
- Erro ao não detectar um segmento completo de canto, chamado de falso-negativo $\left(E_{\mathrm{FN}}\right)$;
- Erro ao detectar um segmento completo de canto que não existe na marcação manual, chamado de falso-positivo $\left(E_{\mathrm{FP}}\right)$;

O método para análise de desempenho proposto neste trabalho não apresenta nenhuma das desvantagens citadas. Na saída do algoritmo, obtêm-se os tempos limites dos segmentos contendo canto, com precisão da ordem de centésimos de segundo. Esses tempos são diretamente comparados com os tempos obtidos usando marcação manual.

A expressão que define a taxa de erro para um determinado sinal de entrada com duração de 15 s é dada pelo somatório dos tipos de erro. Assim:

$$E(\%) = \frac{|E_{\mathrm{LI}}| + |E_{\mathrm{LS}}| + E_{\mathrm{FN}} + E_{\mathrm{FP}}}{15} \bullet 100 \tag{6.1}$$

onde E_{LI} é a diferença entre o limite inferior de um trecho de canto detectado pelo algoritmo e o limite inferior da marcação manual; E_{LS} é a diferença entre o limite superior de um trecho de canto detectado pelo algoritmo e o limite superior da marcação manual.

Portanto, a expressão que avalia o desempenho do algoritmo (taxa de acerto) é:

$$\eta(\%) = 100 - E \tag{6.2}$$

6.3 Comparação de Desempenho

Aplicando o algoritmo proposto aos 40 sinais do banco de dados de teste, a taxa de acerto [obtida por (6.2)] é comparada com a melhor taxa de acerto atingida em [4] (calculada usando janelas de 1,3 s), como apresentado na Tabela 6.1.

Tabela 6.1: Comparação do desempenho entre o algoritmo da referência [4] e o algoritmo proposto

Algoritmo	*Taxa de Acerto* (%)
Referência [4]	**81,2**
Proposto	**81,7**

Mesmo usando um método de medida de desempenho mais rigoroso na avaliação de desempenho do algoritmo proposto, observa-se que os resultados obtidos situam-se na mesma faixa de taxa de acerto.

6.4 Classificador Automático & Classificador Manual

Na abordagem realizada em [4], dos 101 arquivos do banco de dados, 61 são usados como banco de treinamento e 40 como banco de testes. Em [4], os limites que distinguem as classes de canto e instrumentos musicais são obtidos através de um classificador automático HMM considerando o banco de treinamento. No trabalho aqui proposto, não há necessidade de uma etapa de treinamento, uma vez que o classificador usado é do tipo baseado em regras. Nesse caso, alguns dos arquivos do banco de dados são selecionados

para definir, de forma manual, todos os limites a serem considerados. Assim, é possível estender a realização dos testes para os 61 arquivos do banco de treinamento e, por conseqüência, para todos os 101 sinais do banco de dados. Como mostrado na Tabela 6.2, a mesma taxa de acerto é (aproximadamente) mantida.

Tabela 6.2: Análise de desempenho do algoritmo proposto aplicado ao banco de dados de treinamento e ao banco total

Banco de Dados	*Taxa de Acerto* (%)
Treinamento (61 sinais)	**82,6**
Total (101 sinais)	**82,2**

Além de permitir a extensão dos testes para todo o banco de dados, a solução a partir do classificador baseado em regras é de menor complexidade computacional quando comparada à da proposta em [4], que necessita, inclusive, de um sistema de reconhecimento automático de fala.

6.5 Discriminação dos Tipos de Erro

Diferentemente do método de análise de desempenho descrito em [4], o método aqui proposto permite discriminar os tipos de erro envolvidos. A Tabela 6.3 apresenta a distribuição da taxa de erro de acordo com o seu tipo.

Tabela 6.3: Distribuição da taxa de erro conforme os seus tipos

Tipo de Erro	*Taxa de Erro* (%)
E_{L}	**7,7**
E_{FN}	**4,4**
E_{FP}	**5,7**
Total	**17,8**

Os tipos de erro mais grosseiros são os E_{FN} e E_{FP}. Nesses casos, obtém-se ou despreza-se um segmento completo de canto que está em oposição à situação real. Entretanto, analisando a distribuição da taxa de erro conforme os seus tipos (ver Tabela 6.3), verifica-se que o erro de maior ocorrência é do tipo E_{L}. Realmente, a detecção automática dos limites do canto consiste no ponto mais crítico, pelas seguintes razões:

- Maior grau de subjetividade nas suas definições;
- Não apresenta, algumas vezes, trilhas de picos espectrais variáveis;
- Não é possível utilizar a suavização, realizada em trechos intermediários;

Apesar de maiores dificuldades em sua detecção, do ponto de vista da compreensão da informação segmentada, os resultados não são tão críticos. Lembre-se que o valor de E_{L} ainda deve ser distribuído entre o erro na detecção dos limites superior e inferior. Assim, geralmente, na audição do trecho segmentado, o erro na detecção dos limites não compromete a compreensão de todo o seu conteúdo.

6.6 Análise das Causas dos Erros

Basicamente, duas causas principais contribuem para a ocorrência de erros na segmentação de canto a partir da abordagem proposta: a seleção de trilhas variáveis de forma insatisfatória (sinais com baixo conteúdo harmônico), e a seleção de trilhas variáveis produzidas por instrumentos musicais. Há ainda uma terceira causa, referente aos erros provocados por análise incorreta do algoritmo. A seguir, essas causas são discutidas e, adicionalmente, são propostas alternativas para reduzir seus efeitos, quando possível.

6.6.1 Seleção de Trilhas Variáveis de Forma Insatisfatória

Apesar de representar um padrão relevante na distinção da classe de canto, as trilhas de picos espectrais variáveis não são produzidas em todos os sinais dessa classe. Na ausência de tal padrão, não há como melhorar o desempenho do algoritmo, pois, na verdade, ele não está cometendo erros ao não selecionar trilhas variáveis que não existem. Nesse caso, o erro obtido na análise de desempenho, do tipo E_{FN}, deve-se à aplicação a qual o algoritmo se destina, que é a segmentação do canto dentro de sinais de música. A única situação em que a total ausência de trilhas variáveis em um segmento de canto não implicará em erros do tipo E_{FN}, é aquela apresentada no Caso 4 da Seção 5.3: Análise de Casos, em que um raro segmento completo de canto puramente fricativo pode ser detectado apenas pelo parâmetro secundário.

No amplo universo dos sinais presentes no banco de dados, há poucos exemplos de segmentos contendo canto que não possuam, em algum momento, o padrão de trilhas

variáveis. Mais comumente são encontradas situações onde há pouca variação das trilhas e o conteúdo harmônico, ou seja, o número de trilhas harmonicamente selecionadas, é muito pequeno. Apesar de terem sido obtidos resultados corretos em algumas situações desse tipo, como nos Casos 3 e 4 da Seção 5.3: Análise de Casos do Capítulo 5, com certeza, o nível de acertos para uma série de sinais desse tipo será menor. Portanto, visando excluir da análise de desempenho os sinais que apresentem um baixo conteúdo harmônico, é proposta a adoção de um nível de confiabilidade binária aos sinais de entrada, definido como uma medida da certeza de que o resultado produzido pelo algoritmo esteja correto, baseado em atender um conteúdo harmônico mínimo. Assim, para os sinais que não alcancem, em algum trecho nos 15 s de duração, um valor mínimo de 7 trilhas variáveis harmonicamente relacionadas, é atribuído um nível de confiabilidade igual a zero. Por exemplo, no segmento de canto identificado na Figura 5.20(e), obtém-se um trecho com até 9 trilhas harmonicamente relacionadas, garantindo um nível de confiabilidade 1 (um). Nessa análise não são considerados os sinais que tenham sido classificados como puramente instrumentais.

Excluindo os sinais do banco de dados definidos com nível de confiabilidade zero, comprova-se que o desempenho do algoritmo tende a aumentar, conforme mostrado na Tabela 6.4. Em relação ao desempenho obtido na Tabela 6.1 e Tabela 6.2, verifica-se um ganho de quase 5% na taxa de acerto quando o algoritmo é aplicado somente aos sinais com confiabilidade um. Grande parte desse ganho deve-se à eliminação do erro tipo E_{FN}. Aos demais sinais (confiabilidade zero) é sugerida a aplicação de outras técnicas associadas à abordagem proposta.

Tabela 6.4: Desempenho do algoritmo aplicado aos sinais com nível de confiabilidade 1 (um)

Banco de Dados	*Taxa de Acerto* (%)
Testes (32 sinais)	**86,6**
Total (81 sinais)	**86,9**

6.6.2 Seleção de Trilhas Variáveis Produzidas por Instrumentos Musicais

Geralmente, os erros tipo E_{FP} estão associados a instrumentos musicais que produzem trilhas variáveis em certos momentos. Particularmente, no banco de dados analisado, alguns tipos de instrumentos (com predominância do saxofone) contribuem com

a maior parte dos 5,7% da taxa de erro de E_{FP} (ver Tabela 6.3). Nesse caso, sugere-se a adição de técnicas de reconhecimento de instrumentos, como a proposta em [43].

6.6.3 Erros Provocados pelo Algoritmo

O algoritmo proposto neste trabalho visa extrair automaticamente a informação da presença de canto em um sinal de música, baseado principalmente na diferença entre o conteúdo harmônico do canto e dos instrumentos musicais, observada através da análise visual do espectrograma. Portanto, pode-se afirmar que o algoritmo computacional pretende automatizar a análise visual do espectrograma de uma música. Obviamente, erros são inseridos ao longo de todo o processo, desde a análise visual até a finalização da automatização da análise. A complexidade e a grande variabilidade do espectrograma dos sinais de música impõem um elevado número de condições e etapas a serem analisadas para a realização do algoritmo, tais como:

- Escolha da ordem do modelo AR para o cálculo da magnitude do espectro;
- Definição dos limites de amplitude e largura dos picos a serem selecionados na magnitude do espectro;
- Processo de formação das trilhas a partir de vetores linha, que deve traduzir o mais próximo possível a imagem vista em um espectrograma;
- Processo de seleção das trilhas variáveis;
- Análise da relação harmônica entre as trilhas variáveis selecionadas;
- Detecção de padrões secundários e proposta de um parâmetro para extraí-los;
- Adaptação do algoritmo para tratar sinais de música com duração de 15 s.

Embora tenham sido exaustivamente investigadas diversas variações de muitos dos pontos destacados, pode-se tentar ainda aperfeiçoá-los, tanto individualmente como as suas combinações, visando o objetivo final de melhorar o desempenho do algoritmo. Tal aperfeiçoamento é considerado relevante, uma vez que os resultados obtidos são bastante animadores.

6.7 Filtragem das Trilhas de Picos Espectrais

Além de localizar os trechos com canto, é possível imaginar esta outra aplicação para a seção do algoritmo que extrai e seleciona as trilhas de picos espectrais. A partir da localização de um trecho com canto e da determinação do valor de *pitch*, é sugerida a filtragem das trilhas de picos espectrais, através de um banco de filtros seletivos, com o objetivo de realizar um primeiro passo na busca pela separação do canto e instrumentos musicais dentro de uma música. Assim, é proposto um primeiro experimento com esse objetivo, a partir do sinal cujo espectrograma é apresentado na Figura 6.2. Por questões de simplicidade, tal experimento considera apenas o trecho de canto destacado na Figura 6.2, com *pitch* constante. O valor do *pitch* no trecho, determinado pela distância entre as trilhas, é de aproximadamente 550 Hz. Esse experimento baseia-se no fato de que em um sinal harmônico, como o canto vozeado, grande parte da energia está concentrada em múltiplos da freqüência fundamental.

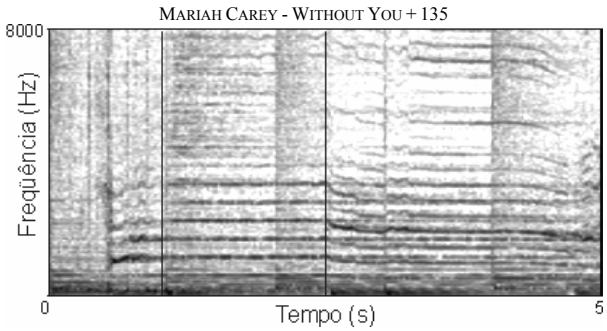


Figura 6.2: Espectrograma de um sinal de música.

Aplicando um banco de filtros rejeita-faixa com freqüência de corte de múltiplos de 550 Hz, as trilhas de picos espectrais do canto são excluídas do espectrograma, como mostrado na Figura 6.3(a). Aplicando um banco de filtros passa-faixa com os mesmos valores de freqüência de corte, as trilhas são isoladas, como ilustrado na Figura 6.3(b).

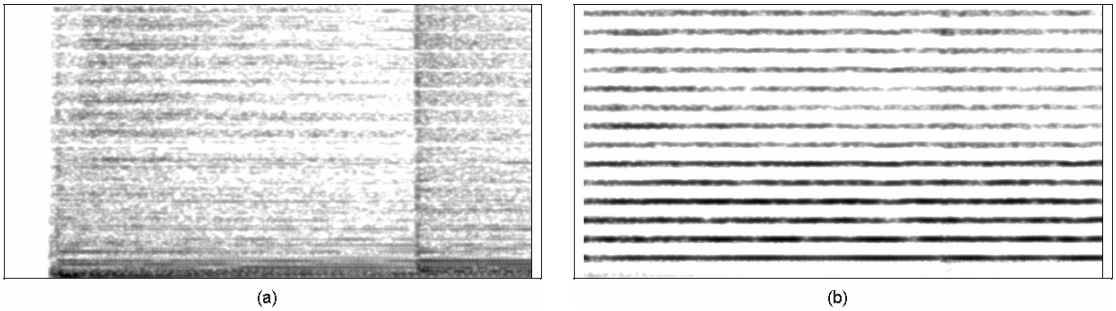


Figura 6.3: Filtragem das trilhas de picos espectrais por banco de filtros: (a) rejeita-faixa e (b) passa-faixa.

A avaliação subjetiva do sinal obtido na Figura 6.3(a), que representa o segmento com instrumentos musicais isolados, e do sinal obtido na Figura 6.3(b), que representa o segmento com canto isolado, indica um resultado satisfatório, aproximando-se da separação entre o canto e os instrumentos musicais do trecho de música destacado na Figura 6.2. Entretanto, este experimento consiste apenas em um primeiro passo. A validação dessa técnica dependeria de um estudo extenso sobre as condições em que poderia ser aplicada, dos artefatos produzidos na filtragem e do projeto dos filtros.

6.8 Conclusões

Este capítulo apresentou a análise de desempenho do algoritmo proposto e a comparação dos resultados obtidos com a abordagem apresentada em [4], considerando um experimento com objetivos idênticos. Para maior validade da comparação, foi usado o mesmo banco de dados adotado em [4]. O desempenho obtido em ambos os casos é similar (em torno de 80%), mesmo considerando um método de medida de desempenho mais rigoroso na avaliação do algoritmo aqui proposto. Como vantagem, a solução proposta (baseada em regras) apresenta menor complexidade computacional do que aquela descrita em [4] (necessita de um sistema de reconhecimento automático de fala). Adicionalmente, ela permite discutir novos resultados, tais como a extensão dos testes para todo o banco de dados, mantendo aproximadamente a mesma taxa de acerto, e a discriminação dos tipos de erro envolvidos no processo de segmentação, sugerindo alternativas para reduzi-los, quando possível. Uma das alternativas propostas para a redução dos erros é a adoção de um nível de confiabilidade binária aos sinais de entrada. Assim, a taxa de acerto obtida somente sobre os sinais com nível de confiabilidade 1 (um) aumenta em quase 5%.

Finalmente, o algoritmo proposto permitiu realizar um experimento de filtragem das trilhas de picos espectrais. Tal experimento foi realizado com o objetivo de separar o canto dos instrumentos dentro de sinais de música, considerando que as trilhas de picos espectrais concentram a maior parte da energia do canto vozeado. Após a audição dos sinais separados, verifica-se que o resultado obtido é satisfatório.

Capítulo 7

Conclusões

A área de pesquisa de CSA busca identificar automaticamente as classes de sinais de áudio. Para isso, são aplicadas técnicas de processamento digital de sinais, a fim de simular o funcionamento do sistema auditivo humano, de caráter biológico, através do estudo da percepção humana do som, de caráter psicológico. Portanto, a solução de problemas referentes à CSA possui caráter multidisciplinar.

A CSA situa-se em uma posição intermediária em um sistema automático que pretenda substituir completamente a função do sistema auditivo humano em situações normalmente vividas no mundo real. Nos extremos inicial e final de tal sistema deve existir, respectivamente, uma etapa que isole a fonte sonora que se deseja "ouvir" [baseado em técnicas de análise da cena auditiva (ASA)], e uma etapa de tratamento da classe de áudio identificada pela CSA, como por exemplo, reconhecimento automático de fala. Na verdade, todas essas áreas encontram-se relacionadas. Por exemplo, a ASA estuda a percepção humana do som para realizar a segregação de fontes sonoras. A CSA utiliza-se desses princípios para realizar a identificação das classes. Por outro lado, as técnicas de reconhecimento automático de fala tratam da identificação de unidades da fala (tais como, fonemas e palavras), que podem ser consideradas como classes de áudio.

É possível imaginar muitas aplicações para a CSA. Em muitas delas, a classificação de sinais de áudio constitui uma etapa de pré-processamento. Essa etapa é importante, pois, além de extrair somente a informação necessária à etapa seguinte do processo, reduzindo seu custo de processamento, também leva a uma redução na taxa de erro da próxima etapa, visto que são eliminadas informações desnecessárias.

Com um objetivo inicial de caráter informativo, foi apresentado o problema de CSA geral. Assim, foi descrito o estado-da-arte em CSA através de uma revisão bibliográfica que contempla os principais trabalhos de pesquisa realizados na área. Foi também discutida a estrutura padrão de diagrama em blocos para o processamento automático de CSA geral. Tal estrutura é comum a problemas de reconhecimento e classificação de padrões. A extração de parâmetros relevantes do sinal consiste na etapa

principal do processamento. Portanto, foi concedida uma atenção especial para a descrição e análise dos principais parâmetros encontradas na literatura.

Posteriormente, este trabalho buscou um caráter de inovação científica, focando no desenvolvimento de um algoritmo para realizar a segmentação do canto dentro de sinais de música. A abordagem proposta é baseada na diferença entre o conteúdo harmônico da voz cantada e dos instrumentos musicais. Os padrões que evidenciam essa diferença são visíveis na análise do correspondente espectrograma, com o canto sendo caracterizado pela formação de trilhas variáveis, e os instrumentos musicais pela formação de trilhas constantes. Esses padrões foram extraídos a partir do parâmetro principal de trilha de pico espectral discutido em [3]. A abordagem desenvolvida neste trabalho propôs uma ampliação dos padrões obtidos desse parâmetro, a fim de caracterizar melhor o sinal de canto, identificando-o canto não somente pelo vibrato (grandes variações das trilhas em forma de ondas), como em [3], mas também por pequenas variações de *pitch*, ou seja, pequenas variações nas trilhas.

Adicionalmente, foi proposto um parâmetro secundário (baseado no cálculo do valor médio da magnitude da TFD) para auxiliar no processo de segmentação do canto. A partir desse parâmetro foi realizada a detecção de segmentos de canto fricativo e, como etapas de pós-processamento, foram efetuados o ajuste dos limites do canto vozeado e a suavização dos resultados.

Como a técnica desenvolvida em [3] não apresenta resultados em experimento de segmentação do canto dentro de sinais de música, o desempenho do algoritmo proposto foi comparado (usando o mesmo banco de dados) com o descrito em [4], que possui os mesmos objetivos de segmentação do trabalho aqui proposto. Os resultados obtidos em ambas as técnicas são bastante próximos, na faixa de 80%. Entretanto, algumas diferenças e vantagens da solução proposta neste trabalho são destacadas e discutidas. Inicialmente, o método de análise de desempenho proposto é mais rigoroso do que aquele adotado em [4]. Como vantagem, a solução proposta apresenta menor complexidade computacional por adotar um classificador do tipo baseado em regras, enquanto a abordagem apresentada em [4], que usa um classificador automático HMM, necessita, inclusive, de um sistema de reconhecimento automático de fala. Além dessa vantagem, o uso de um classificador baseado em regras permitiu a extensão dos testes para todo o banco de dados, uma vez que não se faz necessária uma etapa de treinamento. Os resultados obtidos sobre todo o banco

de dados mantiveram a taxa de acerto em torno de 80%. Adicionalmente, a abordagem proposta permitiu a discriminação da taxa de erro de acordo com o tipo, possibilitando a sua análise e a sugestão de alternativas para reduzi-los, quando possível. Uma das alternativas considerou a adoção de um nível de confiabilidade aos sinais de entrada, baseado em atender um conteúdo harmônico mínimo. Assim, aplicando o algoritmo somente aos sinais do banco de dados com nível de confiabilidade um, foi obtida uma diminuição de quase 5% na taxa de erro total.

A pesquisa em CSA é relativamente recente e ainda há muitas questões em aberto. Ainda não houve a consagração de técnicas definitivas. Constantemente, estão sendo lançadas novas publicações na área. Em relação ao problema de segmentação do canto dentro de sinais de música, a pesquisa é mais recente e há ainda menos publicações. Portanto, há uma grande possibilidade de trabalhos futuros, tanto no desenvolvimento e aperfeiçoamento de técnicas de segmentação, quanto na sua integração com outras aplicações citadas neste trabalho. Especificamente, sobre o algoritmo proposto, sugere-se a investigação dos pontos destacados na análise de erros do Capítulo 6, buscando o seu possível aperfeiçoamento. Outras sugestões contemplam a busca de novos parâmetros a serem extraídos do sinal, utilização de classificadores automáticos, e associação da abordagem proposta com outras existentes como, por exemplo, a utilização de técnicas de reconhecimento automático de instrumentos musicais para reduzir o erro na seleção de trilhas variáveis produzidas por instrumentos.

Finalmente, é sugerida a investigação do problema de separação entre canto e instrumentos dentro de sinais de música. Inicialmente, pretendia-se propor esse tema para a dissertação em questão. Entretanto, investigando a literatura pertinente, verificou-se que tal problema encontra-se totalmente em aberto. Assim, a segmentação do canto dentro de sinais de música foi adotada como a possibilidade mais próxima da intenção inicial. Além disso, a abordagem proposta para a segmentação permitiu realizar um primeiro experimento na tentativa de separação dos sinais. Sugere-se o estudo da área de pesquisa de análise da cena auditiva [1], [2], para investigar formas de se obter a segregação do canto e instrumentos musicais dentro de sinais de música.

Referências Bibliográficas

[1] BREGMAN, A. S. **Auditory Scene Analysis**. Cambridge, MA: MIT Press, 1990.

[2] BROWN, G. J., COOKE, M. Computational Auditory Scene Analysis. **Computer Speech and Language**, v. 8, n. 2, p. 297-336, 1994.

[3] ZHANG, T.; KUO, C.-C. J. Audio Content Analysis for Online Audiovisual Data Segmentation and Classification. **IEEE Transactions on Speech and Audio Processing**, v. 9, n. 4, p. 441-457, May 2001.

[4] BERENZWEIG, A. L.; ELLIS, D. P. Locating Singing Voice Segments Within Music Signals. In: IEEE WORKSHOP ON APPLICATIONS OF SIGNAL PROCESSING TO AUDIO AND ACOUSTICS. Out. 2001, Nova York, **Proceedings**… p. 119-122.

[5] HACKER, Scot. **MP3: The Definitive Guide**. O'Reilly & Associates, 2000.

[6] M. Nilsson. **ID3v2**. Disponível em <http://www.id3.org/>. Acesso em 25 fev. 2004.

[7] ZHANG, T. Automatic Singer Identification. In: IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO, Jul. 2003, Baltimore. **Proceedings**… v. 1, p. 33–36.

[8] BERENZWEIG, A. L.; PANOMRUTTANARUG, B. **Voice Removal from Music**.

[9] ZHANG, T. Semi-Automatic Approach for Music Classification. In: SPIE CONFERENCE ON INTERNET MULTIMEDIA MANAGEMENT SYSTEMS IV, Sept. 2004, Orlando. **Proceedings**... v. 5242, p. 81-91.

[10] GERHARD, D. B. **Computationally Measurable Temporal Differences Between Speech and Song**. Burnaby, BC, Canadá, 2003. 207 f. Tese (Doutorado em Ciência da Computação) – Escola de Ciência da Computação, Universidade Simon Fraser.

[11] DELLER JR., J. R.; HANSEN, J. H. L.; PROAKIS, J. G. **Discrete-Time Processing of Speech Signals**. New York: Macmillan, 1993 (IEEE Press Classic Reissue).

[12] ZUE, V. **Acoustic Properties of Speech Sounds**. Disponível em: <http://www.clsp.jhu.edu/ws2000/presentations/preliminary/victor_zue/Zue-lecture2.pdf>. Acesso em 18 fev. 2004.

[13] BRAID, Antonio César Morant. **Fonética Forense**. Porto Alegre: Editora Sagra Luzzato, 1999.

[14] PACHECO, F. S. **Técnicas de Processamento de Sinais para Alteração de Parâmetros Prosódicos Aplicadas a um Sistema de Conversão Texto-Fala para a Língua Portuguesa Falada no Brasil**. Florianópolis, 2001. 100 f. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica, Universidade Federal de Santa Catarina.

[15] BÜCHLER, M. C. **Algorithms for Sound Classification in Hearing Instruments**. Zurique, Suíça, 2002. 136 f. Tese (Doutorado) – Instituto Federal Suíço de Tecnologia.

[16] MAGGIOLO, D. **Sistema Auditivo Periférico**. Disponível em <http://www.eumus.edu.uy/docentes/maggiolo/acuapu/sap.html>. Acesso em 10 fev. 2004.

[17] LAMANCUSA, J. S. **Fundamentals of Hearing**. Disponível em <http://www.me.psu.edu/lamancusa/me458/2_hearing.pdf>. Acesso em 10 fev. 2004.

[18] BOER, E. Auditory Physics. Physical Principles in Hearing Theory. 1. **Physics Reports**, v. 62, n. 2, p. 87-174, Jun. 1980.

[19] PERDIGÃO, F. M. S. **Modelo do Sistema Auditivo Periférico no Reconhecimento Automático de Fala**. Coimbra, Portugal, 1997. 258 f. Tese (Doutorado em Engenharia Eletrotécnica) – Faculdade de Ciências e Tecnologia, Universidade de Coimbra.

[20] GERHARD, D. B. **Audio Signal Classification: History and Current Techniques**. Disponível em <http://www2.cs.uregina.ca/~gerhard/publications/TRdbg-Audio.pdf >. Acesso em 10 fev. 2004.

[21] OPPENHEIM, A.V.; SCHAFER, R.W. **Discrete-Time Signal Processing**. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[22] SAUNDERS, J. Real-Time Discrimination of Broadcast Speech/Music. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, May 1996, Atlanta. **Proceedings**… v. 2, p. 993-996.

[23] SCHEIRER, E.; SLANEY, M. Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, Apr. 1997, Munique. **Proceedings**… v. 2, p. 1331-1334.

[24] SRINIVASAN, S.; PETKOVIC, D.; PONCELEON, D. Towards Robust Features for Classifying Audio in the CueVideo System. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA (PARTE 1), 1999, Orlando. **Proceedings**... p. 393-400.

[25] ROSSIGNOL, S. et al. Feature Extraction and Temporal Segmentation of Acoustic Signals. In: INTERNATIONAL COMPUTER MUSIC CONFERENCE, Oct. 1998, Ann Arbor. **Proceedings**...

[26] PINQUIER, J; ROUAS, J.-L.; ANDRÉ-OBRECHT, R. Robust Speech/Music Classification In Audio Documents. In: 7th INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (ICSLP), Sept. 2002, Denver. **Proceedings**... p. 2005-2008.

[27] LU, G.; HANKINSON, T. An Investigation of Automatic Audio Classification and Segmentation. In: 5th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING WCCC-ICSP, Aug. 2000, Beijin. **Proceedings**... v. 2, p. 776-781.

[28] SAAD, E. M. Et al. A Multifeature Speech/Music Discrimination System. In: IEEE CANADIAN CONFERENCE ON ELETRICAL & COMPUTER ENGINEERING, May 2002, Winnipeg. **Proceedings**… v. 2, p. 1055-1058.

[29] AJMERA, J.; MCCOWAN, I.; BOURLARD, H. Speech/Music Segmentation using Entropy and Dynamism Features in a HMM Classification Framework. **Speech Communication**, v. 40, n. 3, p. 351-363, May 2003.

[30] WILLIAMS, G. ELLIS, D. Speech/music discrimination based on posterior probability features. In: EUROSPEECH, Set 1999, Budapeste, Hungria. **Proceedings**… p. 687-690.

[31] CAREY, M. J.; PARRIS, E.; LLOYD-THOMAS, H. A Comparison of Features for Speech, Music Discrimination. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTIC, SPEECH, SIGNAL PROCESSING, Mar. 1999, Phoenix. **Proceedings**... v. 1, p. 149-152.

[32] PINQUIER, J; SÉNAC, C.; ANDRÉ-OBRECHT, R. Speech and Music Classification In Audio Documents. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTIC, SPEECH, SIGNAL PROCESSING, May. 2002, Orlando. **Proceedings**... v. 4, p. 41-44.

[33] PINQUIER, J; ROUAS, J.-L.; ANDRÉ-OBRECHT, R. A Fusion Study in Speech/Music Classification. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTIC, SPEECH, SIGNAL PROCESSING, Apr. 2003, Hong Kong. **Proceedings**... v. 2, p. 17-20.

[34] LU, L.; LI, S. Z.; ZHANG, H-J. Content-Based Audio Segmentation Using Support Vector Machines. In: IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO, Aug. 2001, Tokyo. **Proceedings**... p. 956-959.

[35] KOSINA, K. **Music Genre Recognition**. Hagenberg, Austria, 2002. 84 f. Dissertação (Mestrado) Fachhochschule Hagenberg for Media Technology and Design.

[36] RIZVI, S.; CHEN, L.; ÖZSU, M. T. **MADClassifier: Content-Based Continuous Classification of Mixed Audio Data**. Technical Report, Oct. 2002. Disponível em

<http://darwell.uwaterloo.ca/~ddbms/publications/multimedia/cs2002-34.pdf> Acesso em 22 maio 2003.

[37] CHOU, W.; GU, L. Robust Singing Detection in Speech/Music Discriminator Design. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTIC, SPEECH, SIGNAL PROCESSING, May. 2001, Salt Lake City. **Proceedings**... v. 2, p. 865-868.

[38] LI, D. et al. Classification of General Audio Data for Content-Based Retrieval. **Pattern Recognition Letters**, v. 22, n. 5, p. 533-544, Apr. 2001.

[39] PFEIFFER, S.; FISCHER, S.; EFFELSBERG, W. Automatic Audio Content Analysis. In: ACM Multimedia 96, Nov. 1996, Boston. **Proceedings**…. p. 21-30.

[40] ZHANG, T.; K. C.-C. J. Hierarchical Classification of Audio Data for Archiving and Retrieving. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTIC, SPEECH, SIGNAL PROCESSING, Mar. 1999, Phoenix. **Proceedings**... v. 6, p. 3001-3004.

[41] SOLTAU, H. et al. Recognition of Music Types. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTIC, SPEECH, SIGNAL PROCESSING, May. 1998, Seattle. **Proceedings**... v. 2, p. 1137-1140.

[42] LAMBROU, T. et al. Classification of Audio Signals Using Statistical Features on Time and Wavelet Transform Domains. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTIC, SPEECH, SIGNAL PROCESSING, May. 1998, Seattle. **Proceedings**... v. 6, p. 3621-3624.

[43] T. KITAHARA, M. GOTO E H. OKUNO, "Musical Instrument Identification Based On F0-Dependent Multivariate Normal Distribution," In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTIC, SPEECH, SIGNAL PROCESSING, Apr. 2003, Hong Kong. **Proceedings**... v. 5, p. 421-424.

[44] **MPEG-7 Overview (version 8)**. Disponível em <http://vilab.hit.edu.cn/~bozhang/MPEG/mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm>. Acesso em 29 maio 2003.

[45] LU, L.; ZHANG, H.-J. Content Analysis for Audio Classification and Segmentation. **IEEE Transactions on Speech and Audio Processing**, v. 10, n. 7, p. 504-516, Oct. 2002.

[46] TZANETAKIS, G. **Manipulation, Analysis and Retrieval Systems for Audio Signals**. Princeton, NJ, USA, 2002. 184 f. Tese (Doutorado em Ciência da Computação) – Universidade de Princeton.

[47] GERHARD, D. Pitch-Based Acoustic Feature Analysis for the Discrimination of Speech and Monophonic Singing. **Journal of the Canadian Acoustical Association**, v. 30, n. 3, p. 152-153, Sept. 2002.

[48] ROSSIGNOL, S.; DESAIN, P.; HONING, H. State-of-the-art in fundamental frequency tracking. In: WORKSHOP ON CURRENT RESEARCH DIRECTIONS IN COMPUTER MUSIC, 2001, Barcelona. **Proceedings**… v. …, p. 244-254.

[49] XIONG, Z. et al. Comparing MFCC and MPEG-7 Audio Features for Feature Extraction, Maximum Likelihood and Entropic Prior HMM for Sports Audio Classification. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTIC, SPEECH, SIGNAL PROCESSING, Apr. 2003, Hong Kong. **Proceedings**... v. 5, p. 628-631.

[50] SARKAR, T. K. et al. A Tutorial on Wavelets from a Eletrical Engineering Perspective, Part 1: Discrete Wavelets Techniques. **IEEE Antennas and Propagation Magazine**, v. 40, n. 5, p. 49-70, Out. 1998.

[51] DUDA, R. O.; HART, P. E. **Pattern Classification and Scene Analysis**. California: Wiley-Interscience, 1973.

[52] RABINER, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. **Proceedings of IEEE**, v. 77, n. 2, p. 257-285, Feb. 1989.