

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO
DEPARTAMENTO DE TELECOMUNICAÇÕES

**Transcrição Automática do Baixo em Músicas
Populares com Processamento de Sinais
Baseado em Predição Linear**

Dissertação de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica. Área de concentração: Telecomunicações.

Autor: Tiago Fernandes Tavares
Orientador: Amauri Lopes
Co-orientador: Jayme Garcia Arnal Barbedo

Campinas, SP
Janeiro, 2010

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

T197tt	<p>Tavares, Tiago Fernandes</p> <p>Transcrição automática do baixo em músicas populares com processamento de sinais baseado em predição linear / Tiago Fernandes Tavares. –Campinas, SP: [s.n.], 2010.</p> <p>Orientadores: Amauri Lopes; Jayme Garcia Arnal Barbedo.</p> <p>Dissertação de Mestrado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.</p> <p>1. Processamento de Sinais - Técnicas digitais. 2. Musica por computador. 3. Aprendizado do computador. I. Lopes, Amauri. II. Barbedo, Jayme Garcia Arnal. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.</p>
--------	---

Título em Inglês:	Automatic transcription of the bass in popular music with signal processing based on linear prediction
Palavras-chave em Inglês:	Signal processing - Digital techniques, Computer music, Machine learning
Área de concentração:	Telecomunicações e Telemática
Titulação:	Mestre em Engenharia Elétrica
Banca Examinadora:	Jônatas Manzolli e Romis Ribeiro de Faissol Attux.
Data da defesa:	15/03/2010
Programa de Pós Graduação:	Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidato: Tiago Fernandes Tavares

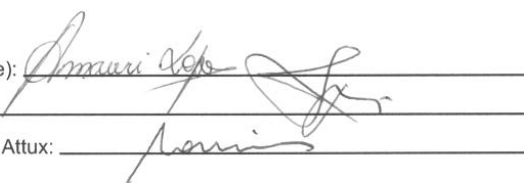
Data da Defesa: 15 de março de 2010

Título da Tese: "Transcrição Automática do Baixo em Músicas Populares com Processamento de Sinais Baseado em Predição Linear"

Prof. Dr. Amauri Lopes (Presidente):

Prof. Dr. Jônatas Manzolli:

Prof. Dr. Romis Ribeiro de Faissol Attux:



Resumo

Este trabalho aborda o problema da transcrição automática de música aplicado à transcrição do baixo em músicas populares. Conceitos teóricos básicos em música e acústica são apresentados. Um método de transcrição existente na literatura científica é descrito e implementado. Na tentativa de melhorar a resolução das análises realizadas no domínio da frequência, são utilizadas técnicas de predição linear. Verifica-se que o uso de tais técnicas traz ganhos consideráveis ao desempenho do transcritor automático implementado.

Palavras-chave: Processamento digital de sinais, Cognição computacional, Computação musical.

Abstract

In this work, the problem of automatic transcription of the bass in popular music is studied. Basic theoretical concepts are presented. An automatic transcription method, obtained in scientific literature, is described and implemented. In order to improve the resolution of necessary frequency domain analysis, linear prediction techniques are used. It is observed that the use of such techniques brings sensible improvements to the accuracy of the implemented transcriptor.

Keywords: Digital signal processing, computer cognition, music computing.

**ESTE PROJETO FOI FINANCIADO PELA FAPESP
PROCESSO NÚMERO 2008/52797-1**

Agradecimentos

Ao meu orientador e meu co-orientador, agradeço pela orientação e pela inspiração, não só em relação ao trabalho científico mas também à vida, que foram determinantes no meu desejo em continuar a atuar no campo da pesquisa;

À minha família, pelo amor incondicional e especialmente pela disposição em compartilhar do crescimento que veio da execução deste trabalho;

Aos meus amigos, cada um por um motivo especial, simplesmente por serem pessoas especiais, me entenderem também como uma pessoa especial e isso me fazer feliz;

À minha namorada, por me mostrar o tempo todo que há maneiras simples e nem por isso menos profundas de se lidar com a vida;

Aos colegas da graduação, por mostrarem que o mundo está cheio de pessoas brilhantes, de forma que, mesmo que eu não consiga resolver todos os problemas do mundo, sei que os outros problemas estarão em boas mãos;

À FEEC e à UNICAMP, pelo apoio institucional e pela oportunidade de participar e atuar num ambiente democrático;

À FAPESP, pelo apoio financeiro;

A todas as pessoas do mundo que, direta ou indiretamente, farão uso dos conhecimentos compartilhados nesta dissertação, por darem sentido à realização deste trabalho.

Dedico este trabalho a toda a humanidade.

Sumário

1	Introdução	1
1.1	Aplicações	1
1.2	Revisão histórica	3
1.2.1	Áudio monofônico	3
1.2.2	Piano solo	4
1.2.3	Acordes	5
1.2.4	Melodia	5
1.2.5	Baixo	6
1.3	Objetivos	7
1.4	Estrutura da dissertação	8
1.5	Publicações	9
2	Conceitos gerais	11
2.1	Física de tons musicais	11
2.2	Regiões de uma nota musical	12
2.3	Intervalos	13
2.4	Nomenclatura de notas musicais	14
2.5	Notação	16
2.6	Composição musical	17
2.7	Discretização e processamento digital de áudio	18
2.8	Análise em quadros de tempo curto e a Transformada Discreta de Fourier	19
2.9	Fisiologia do ouvido e bandas críticas	19
3	Cognição e inferência computacional	21
3.1	Misturas gaussianas	21
3.1.1	Definição	21
3.1.2	Estimativa	23
3.1.3	Heurísticas usadas junto ao método EM	24
3.1.4	Inicialização	24
3.2	Modelos ocultos de Markov	25
3.2.1	Modelos esquerda-para-a-direita	26
3.2.2	Unindo HMMs	27
3.2.3	Distribuição e dimensionalidade de emissões	29
3.2.4	Seqüência de estados mais provável	30

3.2.5	Estimativa de parâmetros	32
4	Transcrição automática de música	37
4.1	Eventos musicais e cadeias de eventos discretos	37
4.2	Algoritmos de processamento digital de sinais	41
4.2.1	Detecção de múltiplas frequências fundamentais	42
4.2.2	O sinal de acento	50
4.3	Modelos de inferência	53
4.3.1	O modelo acústico para notas	54
4.3.2	O modelo acústico para silêncio	55
4.3.3	Treinamento de modelos acústicos	55
4.3.4	O modelo musical	56
4.3.5	O modelo completo	56
4.3.6	Limite superior de F0	56
4.4	Treinamento	57
4.4.1	Estratégia de treino	58
4.5	Resultados	59
5	Melhoria da resolução em frequência usando predição linear	63
5.1	Preditores	63
5.2	Predição linear no domínio do tempo	64
5.2.1	Síntese aditiva e preditores lineares	65
5.2.2	Predição para expansão de quadros de áudio	66
5.2.3	Estimativa de modelos AR	67
5.2.4	A ordem do preditor	68
5.3	Detecção de múltiplas frequências fundamentais usando predição linear	70
5.3.1	Modificações propostas ao método original	71
5.3.2	Base de dados e testes realizados	72
5.3.3	Resultados e discussão	73
6	Aplicação de predição linear na transcrição automática	75
6.1	Estimação espectral por modelo AR	77
6.2	Modelo AR com pré-filtro	78
6.3	Pré-filtro aplicado ao sinal inteiro	79
6.4	Melhorias no modelo	80
7	Discussões	83
7.1	Categorias de transcritores de música	83
7.1.1	Conversão para partitura	83
7.1.2	Obtenção de curva melódica	84
7.1.3	Conversão wave-MIDI	85
7.2	O início de uma nota	86
7.3	O fim de uma nota	87
7.4	Avaliação de transcritores	88

8 Conclusão	91
--------------------	-----------

Referências bibliográficas	93
-----------------------------------	-----------

Capítulo 1

Introdução

Transcrever um sinal de áudio significa encontrar qual é a sequência de ações que foi executada para gerar esse mesmo sinal. No contexto de música, essa tarefa consiste em descrever os movimentos executados pelos músicos na forma de símbolos, que devem conter toda a informação julgada relevante acerca da música. Transcrição automática é o nome que se dá à realização da transcrição por um dispositivo, sem que haja interferência humana.

A transcrição musical pode ser entendida como um processo de extração de informações aplicada sobre um conjunto de dados que é o sinal correspondente à música. Desse conjunto de dados, é possível enumerar uma série de possíveis informações que podem ser extraídas. De uma simples execução de uma nota musical, um músico pode inferir diversas informações, como de qual nota se trata, a que instante ela foi tocada, a que instante ela parou de soar, qual foi o instrumento e a técnica utilizada para tocar, o quão forte a nota foi tocada, e assim por diante. Ao se lidar com uma música inteira, novas informações podem ser encontradas, como ritmo e tempo.

O conjunto de informações que devem ser obtidas pelo sistema de transcrição automática depende da aplicação específica a que o transcritor se destina.

1.1 Aplicações

Transcritores automáticos podem ser utilizados para diversos fins. A transcrição, ao converter uma música para um formato simbólico, já é uma ferramenta de grande valia para músicos, permitindo a conversão de sinais musicais para o formato MIDI [1, 2] ou a obtenção automática de partituras [3]. A conversão para um formato simbólico permite a um músico realizar análises e transformações diretamente no conteúdo da

música. Além disso, formatos simbólicos podem ser utilizados para registro em papel de peças, para fins documentais.

Sobre um transcritor automático, é possível construir ferramentas para aprendizado musical auxiliado por computador [4, 5]. Nesse tipo de aplicação, um programa de computador entende o quão boa foi uma determinada performance realizada pelo estudante a partir da análise de determinadas características do sinal sonoro e realimenta o usuário com informações sobre como melhorar sua técnica. A transcrição automática de música, nesse caso, tem a função de identificar o que foi tocado pelo músico, de forma que sua execução possa ser comparada com um gabarito.

Também, transcritores automáticos são utilizados como componente essencial em sistemas de busca musical orientada a conteúdo. Tais sistemas, chamados de Query-by-Humming [6, 7], recebem como entrada uma música cantada por um usuário e buscam por essa música no banco de dados. Esse tipo de buscador é especialmente importante em grandes bancos de dados de mídia, já que é mais comum que o usuário se lembre do conteúdo do arquivo que dos seus identificadores formais (nome, artista, etc.) [8]. O processo de transcrição automática tem a função de identificar qual foi a melodia cantada pelo usuário, numa notação que permita sua comparação com outras melodias.

Determinados sistemas inteligentes de processamento musical baseiam-se em transcritores automáticos. A tecnologia Direct Note Access [9], que permite a edição de arquivos de áudio gravados através da alteração das notas musicais que o compõem, tem como componente essencial um transcritor que detecta as notas musicais lá presentes. As informações sobre a existência de notas musicais são passadas a um sistema de filtragem que tem por objetivo separar as notas musicais em diferentes arquivos.

Através de processamento inteligente, também é possível desenhar uma aplicação que elimina os vocais de uma faixa de áudio, transformando-a num acompanhamento para karaokê [10]. Também nesse caso, os resultados de um processo de transcrição são utilizados como fonte primária para um sistema de filtragem que busca eliminar as frequências correspondentes à voz.

Outra aplicação interessante é um sistema para alinhamento temporal automático de uma música com sua respectiva letra [11]. Nessa aplicação, uma mescla de transcrição automática de música com reconhecimento de fala, o sistema mostra ao usuário o trecho da letra da música que está sendo cantada em um determinado momento.

Por fim, transcritores automáticos de música podem ser utilizados como parte de outros sistemas cognitivos aplicados à música, como classificadores, já que determinadas formas de organização de figuras musicais podem caracterizar determinados estilos, ou

sistemas de separação de fontes, que podem se beneficiar das informações obtidas pelo transcritor.

1.2 Revisão histórica

A dificuldade para a transcrição automática de música é bastante variável, não só pelo grande número de perguntas que podem ser respondidas ou relevadas pelo sistema de transcrição, mas também pelo grande número de comportamentos diferentes que podem ser esperados de um sinal musical. Um solo de flauta sem acompanhamento, por exemplo, é uma música cujo sinal acústico é mais simples que o de uma música executada por uma orquestra. Da mesma forma, é mais simples encontrar somente as notas executadas por um determinado instrumento que encontrar a técnica específica utilizada para sua execução.

Assim, cada transcritor é desenvolvido para operar sobre um conjunto específico de músicas, provendo um conjunto específico de informações consideradas mais relevantes a cada aplicação.

1.2.1 Áudio monofônico

Uma música monofônica é aquela na qual apenas uma nota soa de cada vez, sem acompanhamento. Podemos encontrar, nessa categoria, vozes desacompanhadas, instrumentos de sopro sem acompanhamento e algumas peças específicas para instrumento solo, como algumas peças para violoncelo e violino, em que não há sobreposição de notas.

Trata-se do caso mais simples tratado, e geralmente é utilizado como ponto de partida para pesquisas subseqüentes [12, 13]. Os resultados já atingidos, com índices de acerto na faixa dos 90%, fizeram com que as atenções das pesquisas se voltassem, nos últimos anos, para a transcrição de outros tipos de áudio. Apesar disso, ainda há algumas iniciativas no sentido de experimentar novas formas de análise de sinal nesse tipo de transcritor [14, 15].

Devido à sua simplicidade, há um certo número de técnicas que podem ser utilizadas com sucesso nesse caso. Em [12], [16] e [15], a transcrição é realizada através da aplicação de análises baseadas em conjuntos de regras pré-determinadas. Em [13], um modelo estatístico é empregado, de forma a calcular a seqüência de notas que mais provavelmente gerou o sinal analisado. Por fim, em [14] a transcrição é realizada através

da simples aplicação de um banco de filtros cujas saídas são analisadas considerando modelos físicos de produção de notas musicais.

1.2.2 Piano solo

O piano solo é, também, um caso especial de transcrição. Como não há contato direto do músico com as cordas vibrantes do instrumento, pode-se considerar que as notas musicais emitidas têm características acústicas mais previsíveis que, por exemplo, um violino, que está sujeito a diversas variações no uso do arco. Ainda assim, a possibilidade de serem encontradas diversas notas soando ao mesmo tempo faz com que a transcrição do piano solo seja uma tarefa mais difícil que a transcrição de áudio monofônico.

Um dos primeiros transcritores para piano solo baseou-se num sistema do tipo *black-board* (quadro-negro), no qual diversos agentes se realimentam com informações sobre as possibilidades de transcrição da música [17, 18]. Tratou-se de um sistema limitado a transcrever peças para coral de Bach executadas no piano, uma categoria de música com características estruturais bem definidas.

Outra tentativa de transcrição baseou-se no uso de redes neurais [19]. O processo proposto inicia com a aplicação, no sinal musical, de um banco de filtros que buscam simular o ouvido humano. A saída desse banco de filtros alimenta duas redes neurais artificiais: uma para reconhecimento do início de notas e outra para discriminação de notas. Um algoritmo especializado combina as informações fornecidas pelas redes neurais e termina o processo de transcrição.

A estabilidade espectral do piano solo permite a comparação de duas peças musicais através da subtração espectral, ou seja, através da subtração de uma da outra no domínio do espectro. Assim, quanto mais próximo de zero for esse resultado, mais próximas são as peças. Esse processo de comparação foi utilizado no desenvolvimento de um algoritmo evolutivo [20]. O algoritmo funciona sintetizando um certo conjunto de peças musicais, chamadas candidatas, que são comparadas com a peça a ser transcrita. As peças candidatas passam, então, por uma série de processos de mutação, aproximando-se gradativamente da peça desejada. Esse algoritmo foi ampliado posteriormente [21], adicionando-se a capacidade de evolução também do timbre sintetizado.

Modelos estatísticos foram, também, utilizados na solução desse tipo de transcrição [22]. O modelo é encarregado de detectar a sequência de acordes mais provavelmente tocada, dado um conjunto de características numéricas, chamadas *features*, calculadas ao longo do tempo.

Por fim, os espectros das notas do piano podem ser considerados bases de um espaço vetorial [23]. Nesse processo, uma determinada matriz é fatorada como o produto de duas outras matrizes, sendo uma delas uma matriz de bases e a outra uma matriz de combinação [24]. Como ambas as matrizes são não-negativas, e se a matriz de bases corresponder às notas do piano, é possível saber qual nota está soando a cada momento através da análise da matriz de combinações.

1.2.3 Acordes

Para determinadas aplicações, não é necessário saber exatamente quais notas estão soando a cada momento. Para tocar o violão popular, por exemplo, basta saber qual acorde está soando. A transcrição de acordes é utilizada em aplicações comerciais como o iChords [25], embora os sistemas de transcrição de acordes ainda estejam sujeitos a diversas limitações. A principal delas diz respeito a quão genérico o sistema deverá ser: uma vez que um acorde é uma combinação de notas, é preciso delimitar quais acordes serão passíveis de transcrição pelo sistema.

Esse problema pode ser resolvido pela aplicação de modelos estatísticos [26, 27]. Em ambos os casos, *features* capazes de discriminar acordes são calculadas. Um algoritmo de programação dinâmica se encarrega, então, de determinar a sequência mais provável de acordes correspondente à música.

Embora poderosos, os modelos estatísticos não levam em consideração as estruturas e repetições da música popular. Ignoram, assim, que os dois refrões de uma música geralmente terão a mesma sequência de acordes. Essa característica foi utilizada em [28], através da aplicação de um algoritmo especializado.

1.2.4 Melodia

Na música popular, melodia é uma definição que diz respeito à voz principal de uma peça. Embora geralmente corresponda à voz do cantor, também pode se referir ao solo de guitarra ou a outro instrumento que faça esse mesmo papel. Transcrever a melodia, na prática, significa transcrever um instrumento monofônico em um contexto polifônico, ou seja, embora várias notas estejam soando ao mesmo tempo, só uma delas é desejada.

Uma das soluções propostas para transcrição da melodia [29] fez uso do fato de que, em geral, a melodia é um som que predomina sobre os demais sons da música. O sistema construído baseia-se na simples detecção do som predominante na faixa de

freqüência entre 250 Hz e 10 kHz. Um trabalho semelhante [30] utiliza algoritmos mais eficazes para realizar as mesmas tarefas.

Em determinadas aplicações, deseja-se apenas encontrar na música um determinado trecho de melodia, para fins de busca por conteúdo [8]. Nesse caso, um modelo estatístico correspondente à melodia a que se busca é montado, e o sistema calcula a probabilidade de o modelo fazer parte da música analisada.

É interessante perceber que a tarefa de transcrição pode ser descrita como uma forma de classificação em períodos curtos. Nessa abordagem [31], o sinal de áudio é dividido em quadros de duração curta - por volta de 100ms - sendo que em cada um deles se calcula um certo conjunto de *features* que permite, posteriormente, sua classificação em função da nota que está soando.

Ainda, um trabalho que combina processos heurísticos com a idéia de classificação foi apresentado em [32]. Nesse trabalho, a etapa de classificação é sucedida pela aplicação de algumas regras que determinam explicitamente o funcionamento da maior parte das melodias da música contemporânea.

Por fim, também na tentativa de realizar a transcrição da melodia de músicas polifônicas, experimentou-se a aplicação de modelos estatísticos [33]. Trata-se, essencialmente, de um trabalho semelhante a [31] e [32], diferenciando-se dos anteriores por alguns detalhes na etapa de classificação.

1.2.5 Baixo

Baixo é uma palavra que, no contexto da transcrição automática, refere-se ao contrabaixo, contrabaixo elétrico ou qualquer instrumento semelhante. O baixo, na música popular contemporânea, tem função tanto harmônica como rítmica. Também, tal como no caso da melodia, trata-se da transcrição de um instrumento monofônico num contexto polifônico.

Junto à melodia, o baixo compõe parte bastante importante da música popular contemporânea [29]. Quando consideram-se apenas as faixas de freqüências mais graves - até por volta de 200 Hz - o baixo é, em geral, o som predominante, da mesma forma que a melodia o é se for considerado todo o espectro da música. Assim, o baixo pode ser transcrito da mesma forma que a melodia [29].

Num trabalho desenvolvido por Ryyanen, a abordagem de classificação estatística é retomada [34]. Nesse trabalho, a forma de classificação já utilizada em [33] é adaptada para o contexto do baixo.

A transcrição do baixo não foi tão extensivamente estudada quanto a transcrição

da melodia por tratar-se de um problema com aplicações de mais restritas. O público em geral tende a se lembrar mais das melodias das músicas que das linhas de baixo, de forma que aplicações de busca tendem a buscar mais a transcrição da melodia que do baixo. Transcritores de baixo têm aplicação principalmente em programas de computador especializados para músicos [35].

1.3 Objetivos

Neste trabalho, foram aplicadas modificações a um transcritor automático de música já existente na tentativa de obter melhorias de desempenho.

Como ponto de partida, foi utilizado o transcritor de baixo proposto em [34]. Esse transcritor foi utilizado por ter melhor desempenho em comparação a outros métodos analisados [35]. A escolha do baixo veio pelo fato de tratar-se de um instrumento pouco explorado na transcrição automática, para o qual há, ainda, bastante espaço para a melhoria de resultados.

O método de transcrição automática proposto pode ser dividido em três partes, como mostra a Figura 1.1. Inicialmente, o arquivo de áudio passa por uma série de algoritmos de processamento digital de sinais, que têm por função calcular, ao longo do tempo, um conjunto de *features* capazes de discriminar notas musicais. Um modelo estatístico toma essas *features* e infere a sequência de notas que mais provavelmente as gerou.

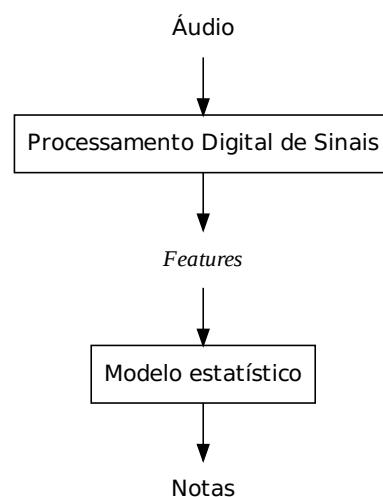


Figura 1.1: Visão geral do sistema de transcrição implementado.

A etapa de processamento digital de sinais, tal como proposto em [34], baseia-se no cálculo da Transformada Discreta de Fourier (TDF) de quadros de curta duração da peça de áudio analisada. Na escolha da duração do quadro analisado, tem-se um compromisso entre a resolução obtida no domínio da frequência, que aumenta com o aumento da duração do quadro analisado, e do tempo, que aumenta com a diminuição dessa mesma duração [36].

Como adição ao método implementado, foram aplicadas técnicas de predição que buscam expandir artificialmente os quadros analisados, aumentando a resolução no domínio da frequência sem perda da resolução temporal.

Os objetivos deste trabalho podem ser, portanto, discriminados como a seguir:

1. Estudo dos conceitos essenciais de física, música e percepção utilizados no método de transcrição implementado;
2. Estudo e implementação do modelo estatístico para cognição;
3. Estudo, implementação e validação dos algoritmos de processamento digital de sinais necessários;
4. Implementação e validação do transcritor de áudio proposto em [34];
5. Estudo e implementação de algoritmos de predição para aumento da resolução no domínio da frequência;
6. Incorporação de algoritmos de predição no transcritor de áudio implementado;
7. Comparação dos resultados obtidos com os originais.

1.4 Estrutura da dissertação

Essa dissertação se organiza como se segue. Inicialmente, no Capítulo 2 são abordados os conceitos de música, física e psico-acústica que serão necessários para a compreensão dos capítulos seguintes. Após, no Capítulo 3, os algoritmos de cognição computacional utilizados são detalhados de forma correspondente à implementação realizada. A seguir, no Capítulo 4, é descrito o transcritor automático utilizado como base para a realização deste trabalho. No Capítulo 5, são discutidos conceitos ligados à predição linear, técnica incorporada no transcritor automático implementado anteriormente conforme descrito no Capítulo 6. O Capítulo 7 traz discussões e, por fim, o Capítulo 8 traz a conclusão do trabalho.

1.5 Publicações

- TAVARES, T. F. ; BARBEDO, J. G. A. ; LOPES, A., *Transcrição Automática de Sinais de Áudio Monofônico Baseada em Quadros de Tamanho Variável*. In: V Congresso de Engenharia de Audio, 2007, São Paulo. Anais do V Congresso de Engenharia de Áudio, 2007. p. 47-50.
- TAVARES, T. F. ; BARBEDO, J. G. A. ; LOPES, A. *Towards the Evaluation of Automatic Transcription of Music*. In: VI Congresso de Engenharia de Àudio, 2008. Anais do VI Congresso de Engenharia de Áudio, 2008. p. 96-99.
- TAVARES, T. F. ; BARBEDO, J. G. A. ; LOPES, A. *Performance Evaluation of Fundamental Frequency Estimation Algorithm*. In: International Workshop on Telecommunications, 2009, São Paulo. Proceedings of the International Workshop on Telecommunications, 2009. v. 1. p. 94-97.
- TAVARES, T. F. ; BARBEDO, J. G. A. ; LOPES, A. *Research on Automatic Transcription of Music*. Submetido ao VIII Congresso de Engenharia de Áudio, 2010.

Capítulo 2

Conceitos gerais

A relação entre o fenômeno físico da vibração do ar e as sensações auditivas é estudada por um campo da ciência chamado psicoacústica. A psicoacústica, historicamente, trouxe uma série de conceitos ligados à física e à percepção humana que posteriormente inspiraram algoritmos relacionados à cognição computacional na área de áudio. Tais algoritmos beneficiam-se, também, de conhecimentos sobre teoria musical e sobre a fisiologia do ouvido humano. Este capítulo apresenta os conceitos desses três campos que serão essenciais para a compreensão dos processos computacionais que serão discutidos posteriormente.

2.1 Física de tons musicais

A acústica clássica mostra que corpos tendem a vibrar em um padrão harmônico, ou seja, composto pela somatória de N padrões senoidais múltiplos de uma frequência fundamental (F_0) f_0 . Cada um desses padrões, chamados de parciais, apresenta amplitude g_m e fase ϕ_m próprias, de forma que a variação da pressão do ar correspondente a uma onda acústica proveniente de uma única fonte pode ser descrita pelo seguinte modelo harmônico:

$$\frac{dp_h(t)}{dt} = \sum_{m=1}^M g_m \cos(2\pi m f_0 t + \phi_m). \quad (2.1)$$

Uma das obras mais antigas sobre a relação entre a física e a audição foi publicada no século XIX por Helmholtz [37]. Nessa obra, Helmholtz diferencia três sensações ligadas à audição de tons musicais: a qualidade, ou timbre, que é a peculiaridade que permite a diferenciação entre sons de diferentes instrumentos; a intensidade, que se relaciona ao volume do tom; e a altura, que é a posição do tom numa escala que

vai do grave ao agudo. Nos modelos estudados por Helmholtz, a qualidade do som é dependente das relações entre as amplitudes g_m das parciais harmônicas, a intensidade é dependente da potência total relacionada ao sinal sonoro e a altura é dependente da F0. A interdependência entre a altura e a frequência fundamental é tal que Helmholtz utiliza a nomenclatura “número de altura” (*pitch number*) como um sinônimo da F0 de um tom. Sons agudos têm F0 elevada, enquanto sons graves têm F0 menor.

O auxílio de aparelhagem eletrônica e o desenvolvimento de metodologias para avaliação subjetiva de sensações ligadas a tons musicais permitiu, a partir da década de 1960, que avanços no sentido de incrementar os modelos propostos por Helmholtz fossem realizados [38]. Mostrou-se, por exemplo, que a intensidade do som pode influenciar a sensação de qualidade e altura, principalmente quando em tons muito potentes. Apesar disso, a F0 continua sendo utilizada como única referência para medida da altura de um tom em situações comuns como a afinação de instrumentos.

2.2 Regiões de uma nota musical

Embora a Expressão 2.1 seja válida para ondas acústicas estacionárias, é importante salientar que, como a maior parte dos sistemas dinâmicos, instrumentos musicais assumem estados transitórios quando acionados e quando desligados. O comportamento de fontes sonoras durante os estados transitórios varia de acordo com o tipo de instrumento musical utilizado e com a técnica de execução da nota.

Para fins de análise, uma nota musical pode ser dividida, no domínio do tempo, em três momentos ou regiões, como descrito na Figura 2.1, que mostra esses três momentos em uma nota musical isolada. À região transitória que inicia a nota, dá-se o nome de ataque (em inglês, *attack*). Essa região corresponde ao início da nota, em que a onda sonora está se formando no instrumento. A região que vem a seguir é chamada de sustentação (em inglês, *sustain*), e corresponde ao período em que a nota soa livremente, mantendo-se a fricção do arco de um violino, a vibração da corda de um violão ou o sopro em uma flauta. Em comparação com as demais regiões, trata-se daquela na qual a onda sonora se comporta de maneira mais próxima de um regime estacionário. Por fim, a região transitória que finaliza a nota é chamada de soltura (em inglês, *release*), e corresponde aos gestos que removem a nota de um instrumento, como abafar a corda ou cessar o sopro.

As iniciais dos termos em inglês dão o nome ASR (*attack - sustain - release*) ao modelo que descreve essas três fases.

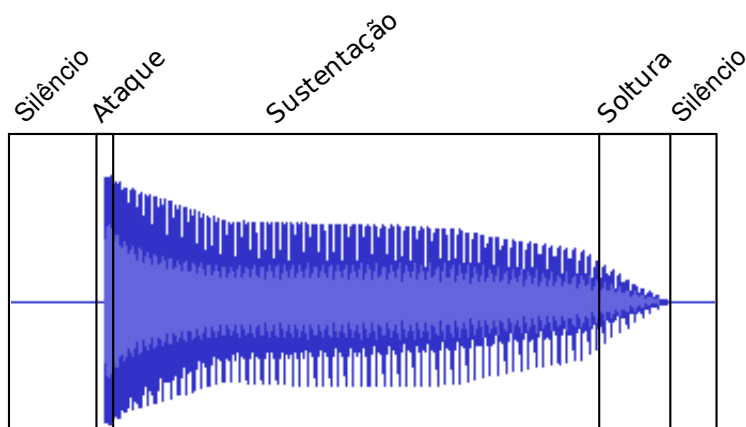


Figura 2.1: Momentos de uma nota musical: ataque, sustentação e soltura, numa nota que começa e termina com silêncio.

2.3 Intervalos

Quando dois tons soam simultaneamente, diz-se que há entre eles um intervalo, relacionado à razão entre as alturas de ambos [37]. É possível observar que, assim como mostram experimentos com a percepção subjetiva, algumas relações entre alturas geram sons agradáveis, chamados consonantes, enquanto que outros geram sons desagradáveis, chamados dissonantes. A medida do quão desagradável é um intervalo é chamada nível de dissonância. Experimentos perceptivos mostram que níveis especialmente baixos de dissonância são encontrados em intervalos de 2:1, 3:2 e 5:4, conhecidos respectivamente como oitava perfeita, quinta perfeita e terça maior [39].

A sensação de baixa dissonância permitiu a Pitágoras, na Grécia antiga, o desenvolvimento de um método para afinação de instrumentos [40]. A afinação pitagórica pressupõe uma escala em que há doze níveis de altura discretos, chamados notas, a cada intervalo de oitava perfeita. Partindo de uma primeira nota inicial, afina-se a oitava nota da escala de tal forma que se forme um intervalo de quinta perfeita. Após, a décima quinta nota é obtida utilizando-se um intervalo de quinta perfeita em relação à oitava. Obtém-se, então, a terceira nota da escala através do intervalo de oitavas. Repetindo-se essa seqüência, é possível afinar todo um instrumento para um certo tom de referência. A Tabela 2.1 evidencia esse procedimento, numerando em ordem crescente de altura as notas envolvidas, de forma a relacionar cada nota à sua ordinalidade no processo de afinação pitagórica. Assim, é possível verificar, por exemplo, que a nona nota da escala será a décima terceira a ser afinada.

É importante observar que, ao fim do processo de afinação pitagórico, o intervalo

Tabela 2.1: Processo de afinação de instrumento pela escala pitagórica.

Nota	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Ordem	1	12	4	15	7	18	10	2	13	5	16	8	19	11	3	14	6	17	9

entre a primeira nota da escala e a décima terceira não é de uma oitava perfeita, mas sim de $1062882:524288^1$, o que representa um desvio de aproximadamente 1,3% em relação ao valor esperado.

Um desvio da mesma natureza é encontrado quando a afinação é realizada levando em conta outras combinações de intervalos. J. S. Bach, porém, consagrou um sistema de afinação no qual o desvio inerente à afinação seria igualmente distribuído por todos os tons da escala. A chamada “escala de igual temperamento” é obtida a partir de três pressupostos. O primeiro deles é que o intervalo de uma oitava deve sempre corresponder a uma razão de alturas de exatamente 2:1. O segundo é que devem existir doze notas para cada oitava, sendo que o intervalo entre duas notas consecutivas é constante para toda a escala. Assim, o intervalo entre duas notas consecutivas - o intervalo de um semitom - é igual a $\sqrt[12]{2}$. O terceiro pressuposto é a existência de uma altura de referência para pelo menos uma das notas da escala, de forma que todas as outras alturas são automaticamente definidas. Essa altura é definida por simples convenção.

2.4 Nomenclatura de notas musicais

Historicamente, as doze notas que compõem uma oitava foram batizadas como dó, dó sustenido (ou ré bemol), ré, ré sustenido (ou mi bemol), mi, fá, fá sustenido (ou sol bemol), sol, sol sustenido (ou lá bemol), lá, lá sustenido (ou si bemol) e si. Uma nomenclatura comumente utilizada é o padrão ABC, que refere-se às notas utilizando letras. Nessa notação, utiliza-se a letra A para a nota lá, B para si, C para dó e assim por diante. O símbolo “#” discrimina o sustenido e o símbolo “b” discrimina o bemol. Assim, nessa notação Eb significa “mi bemol”, F# significa “fá sustenido” e assim por diante. O padrão ABC inclui ainda um número que designa a oitava à qual pertence uma nota, sendo que oitavas superiores contêm notas mais agudas. Dessa forma, C#4 significa “dó sustenido na quarta oitava”. Além disso, é comum o uso da altura de 440 Hz para o lá central do piano (A4) como referência para a construção da escala.

A chegada dos dispositivos eletrônicos compatíveis com o padrão MIDI [41] trouxe

¹Essa diferença é chamada de *coma pitagórica*.

Tabela 2.2: Intervalos comuns na escala MIDI.

Diferença na escala MIDI	Intervalo
0	Uníssono
1	Semi-tom, ou segunda menor
2	Tom, ou segunda maior
3	Terça menor
4	Terça maior
5	Quarta perfeita
6	Quinta diminuta
7	Quinta perfeita
8	Sexta menor
9	Sexta maior
10	Sétima menor
11	Sétima maior
12	Oitava perfeita

uma nova forma de nomenclatura, desta vez voltada à simplificação, em termos computacionais, de mensagens referentes a estruturas musicais. No padrão MIDI, cada nota musical recebe um número próprio, chamado número MIDI, de tal forma que o número 69 corresponde ao A4 e um incremento unitário corresponde a um intervalo de um semi-tom. Dessa forma, se a F0 de um tom é igual a f_0 , o número MIDI da nota correspondente é dado por:

$$p = 69 + 12 \log_2 \frac{f_0}{440}. \quad (2.2)$$

O número MIDI, também chamado de frequência na escala MIDI, é especialmente interessante para aplicações computacionais porque descreve alturas de notas musicais em uma escala na qual intervalos podem ser medidos simplesmente pela subtração de dois p relacionados a duas notas, uma vez que uma diferença unitária nessa escala corresponde a um intervalo de um semi-tom. A Tabela 2.2 mostra a nomenclatura comumente dada a alguns intervalos e sua correspondência com a diferença de frequência na escala MIDI.

A notação MIDI torna fácil a aplicação da operação conhecida como transposição. Nessa operação, todas as notas de uma música devem ser deslocadas de um mesmo intervalo. Numa transposição de um semi-tom, por exemplo, todo C4 deverá se tornar um C#4, todo B3 se tornará em C4 e assim por diante. A notação MIDI permite que a transposição seja realizada simplesmente somando ao p de todas as notas o número de semi-tons correspondente ao intervalo de transposição desejado.

Embora, a rigor, a escala MIDI seja discreta, é importante perceber que a Expressão 2.2 modela um p contínuo. Desta forma, afinações imperfeitas também são projetadas na escala logarítmica.

2.5 Notação

Uma música pode ser descrita como uma seqüência finita de movimentos que devem ser realizados durante a execução. A descrição desses movimentos é feita através de notações que representam uma seqüência de instruções de forma conveniente. Usualmente, a música ocidental é representada através de partituras, como mostrado na Figura 2.2. Na partitura, estão representadas, sequencialmente, as notas (com suas respectivas durações) e pausas que compõem uma determinada peça.



Figura 2.2: Partitura de uma música.

A Figura 2.3 mostra a representação por tablatura da mesma música representada na Figura 2.2. A representação por tablatura mostra, diretamente, a posição dos dedos no braço de um instrumento, de forma que sua leitura é mais simples quando o instrumento tocado possui um braço com trastes, como é o caso do violão, da guitarra e do contrabaixo.

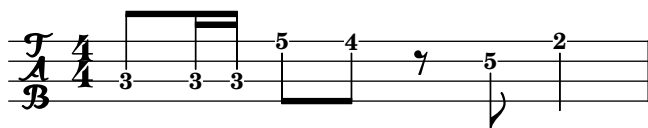


Figura 2.3: Tablatura de uma música.

É importante perceber que a representação por qualquer notação leva a alguma perda de informações. Através da partitura da Figura 2.2 é possível saber quais são as notas tocadas na peça, embora tenha sido desconsiderado o posicionamento dos dedos no instrumento. Na tablatura da Figura 2.3, é necessário conhecer de antemão a afinação de cada corda do instrumento para que a peça possa ser executada corretamente. Os detalhes específicos de cada execução da peça - um eventual deslize do músico, por exemplo - não são registrados na notação.

Assim, a forma da notação utilizada depende apenas da conveniência e da relevância dos dados para execução. O formato MIDI prevê um tipo de notação no qual notas são descritas por seu tempo de início e final (em segundos) e por sua altura. Essa notação pode ser representada graficamente em uma forma conhecida como *piano-roll*, que se assemelha a uma representação em computador digital dos rolos de papel que alimentam as antigas pianolas. Nessa notação, retângulos denotam as instruções para tocar notas: a altura (vertical) em que o retângulo é posicionado corresponde à altura da nota, e as posições do início e do fim do retângulo na horizontal denotam os tempos de início e fim de uma nota. É possível verificar uma equivalência imediata entre um *piano-roll* e o sinal musical correspondentes a uma mesma música. A Figura 2.4 mostra a essa equivalência sobrepondo a notação *piano-roll* e o sinal musical sintetizado correspondentes ao trecho musical exibido nas Figuras 2.2 e 2.3.

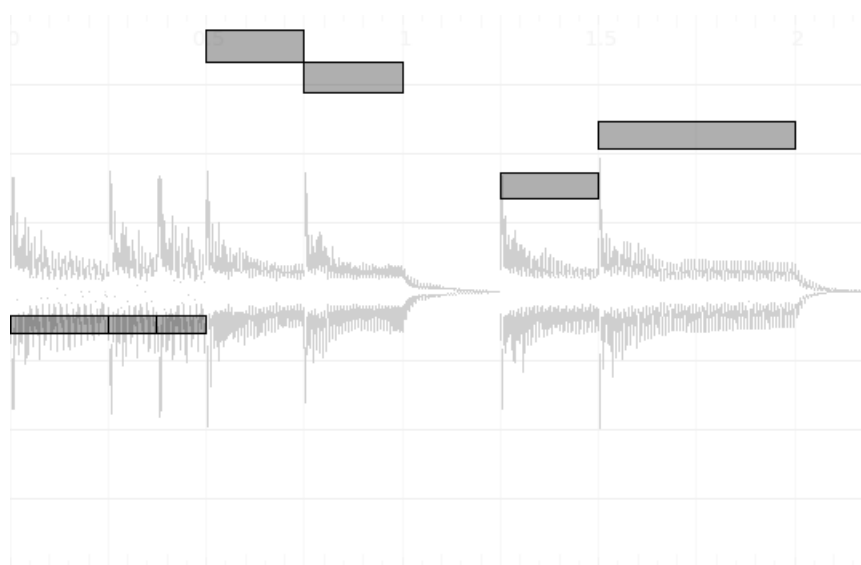


Figura 2.4: Representação *piano-roll* de uma música sobreposta ao sinal musical correspondente.

2.6 Composição musical

Essencialmente, peças musicais são conjuntos de estruturas que descrevem notas incluindo sua altura, seu tempo de início (quando ela deve ser tocada) e sua duração (por quanto tempo ela deve ser tocada). É evidente que, embora seja possível um número praticamente infinito de combinações de durações, alturas e tempos de início, apenas um subconjunto destas combinações constitui-se de peças agradáveis ao ouvido

humano.

Historicamente, foram consolidados conjuntos de regras para composição que fornecem referências sobre como criar estruturas musicais agradáveis ao ouvinte. Um conjunto de regras muito utilizado na música ocidental refere-se à criação do contraponto [42]. Esse conjunto de regras, bastante estudado no período renascentista e aplicado extensivamente na música barroca do século XVII e XVIII, permite inferir possíveis prosseguimentos a uma composição tomando por base as estruturas musicais já escritas.

Uma importante definição que surge no contexto da composição musical é o tom de uma música. O tom de uma música é uma nota utilizada como referência para sua composição. Uma vez que é definido, determina que algumas transições e relações entre notas são mais adequadas que outras no contexto da peça.

2.7 Discretização e processamento digital de áudio

O processo de discretização de áudio consiste da conversão de um sinal acústico para uma seqüência de números compatíveis com processadores. Nesse processo, o sinal acústico é transformado em uma tensão elétrica analógica através de microfones, e um dispositivo amostrador toma amostras dessa tensão igualmente espaçadas no tempo. Assim, um sinal amostrado com frequência de amostragem f_s é definido como uma seqüência $x[n]$ tal que:

$$x[n] = x(n/f_s). \quad (2.3)$$

Em particular, um sinal descrito pela Expressão 2.1, ao ser discretizado, assume a forma:

$$x_h[n] = \sum_{m=1}^M g_m \cos(2\pi m f_0 \frac{n}{f_s} + \phi_m). \quad (2.4)$$

O Teorema de Nyquist mostra que, se a frequência de amostragem f_s for superior ao dobro da frequência mais aguda presente no sinal contínuo, é possível recuperá-lo completamente a partir do sinal amostrado [36]. Em outras palavras, nessas condições não existe perda de informação ao se aplicar o processo de discretização dado pela Expressão 2.3, de forma que os sinais descritos pelas Expressões 2.1 e 2.4 são equivalentes.

Devido a essa forte equivalência, é comum a utilização de variáveis de tempo contínuo para referir-se a um sinal discreto. Embora seja matematicamente incorreta, essa substituição simplifica a descrição de eventos sonoros discretizados. Por exemplo, é mais simples dizer que uma nota tem duração de 1 segundo que dizer que sua duração

é de 48000 amostras.

Sinais acústicos, de maneira geral, ocupam uma faixa de banda muito larga. Apesar disso, experimentos psico-acústicos mostram que a audição humana é limitada em banda para frequências aproximadamente entre 20Hz e 20kHz, de forma que convencionou-se, em processos de discretização para música, o uso de um filtro passa-baixas que elimina do sinal analógico a informação correspondente a sons ultrassônicos. Assim, sem perda de informação audível, o sinal passa a ser considerado limitado em faixa e, portanto, passível de discretização conforme a Equação 2.3.

A discretização de áudio permitiu o desenvolvimento de um campo da ciência chamado de recuperação de informações da música (RIM). Estudos relacionados à RIM buscam obter, a partir de algoritmos de processamento digital de sinais, um conjunto de fatos que descrevem arquivos de áudio em termos ligados a aspectos culturais humanos. Entre as aplicações encontradas nesse campo, encontramos dispositivos automáticos que buscam a classificação de músicas por gênero, a identificação de instrumentos musicais presentes ou a transcrição, a qual consiste da obtenção de notações que descrevem as estruturas musicais que compõem o arquivo analisado.

2.8 Análise em quadros de tempo curto e a Transformada Discreta de Fourier

Neste trabalho, assim como na maior parte dos outros trabalhos ligados a áudio digital, será utilizada em grande escala a análise de sinais em quadros de tempo curto. Esse tipo de análise se baseia no fato de que determinadas características de sinais acústicos usualmente se mantêm constantes em intervalos curtos de tempo [16, 12].

Ao se dividir um sinal $x[n]$ em quadros de tempo curto, é preciso definir o tempo (ou número de amostras) de cada quadro e o tempo (ou número de amostras) entre o início de cada quadro. A partir dessa definição, denota-se $x_q[n]$ o sinal contido no q -ésimo quadro.

Nesse contexto, pode-se calcular a Transformada Discreta de Fourier (TDF) [36] de cada quadro, denotada $X_q[k]$, um recurso que será bastante utilizado posteriormente.

2.9 Fisiologia do ouvido e bandas críticas

Uma vez que a RIM busca simular percepções inerentemente humanas em computadores digitais, é natural que muitas vezes algoritmos sejam inspirados na fisiologia do

processo auditivo. Em especial, o fenômeno das bandas críticas recebe grande atenção.

Tal fenômeno tem sua origem numa estrutura do ouvido interno chamada membrana basilar, que é responsável pela transformação de vibrações acústicas do ambiente em estímulos internos. Estes estímulos darão origem, posteriormente, a impulsos nervosos relacionados à audição [43]. Cada região da membrana basilar é mais sensível a uma determinada frequência.

Em torno da frequência central relacionada a um ponto da membrana basilar, existe uma faixa de frequências conhecida como banda crítica [44]. Quando dois tons se situam em uma mesma banda crítica, é possível que apenas o tom mais intenso seja ouvido, resultando num fenômeno chamado de mascaramento. Dessa forma, a resolução em frequência da audição humana é diretamente relacionada às bandas críticas.

Embora seja possível tabelar com precisão os limites das bandas críticas em torno de cada frequência central [45], é comum, no contexto de RIM, o uso de bancos de filtros para simulação desse fenômeno de forma aproximada [46, 47].

Tais bancos de filtros podem ser definidos como conjuntos de filtros FIR passa-banda [46], cada um relacionado a uma banda crítica $b \in \mathbb{N}$ e referido como $H_b[k]$, sendo que sua frequência central (em Hz) é dada por $c_b = 229(10^{(b+1)/21.4} - 1)$ e sua resposta em frequência é triangular, se estendendo de c_{b-1} a c_{b+1} , com máximo valor igual à unidade em c_b . A Figura 2.5 mostra a resposta em frequência das primeiras 20 bandas de $H_b[k]$.

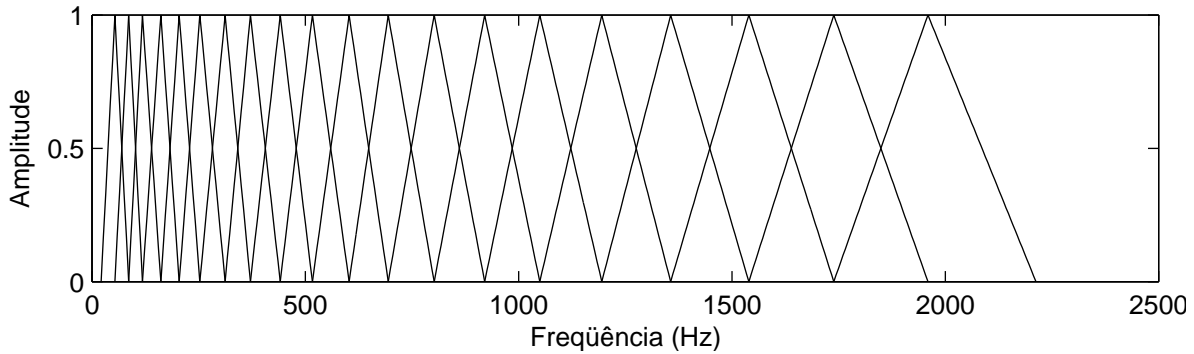


Figura 2.5: Banco de filtros FIR no domínio da frequência para 20 bandas críticas.

Capítulo 3

Cognição e inferência computacional

Uma música pode ser considerada como uma seqüência de eventos discretos $\mathbf{e} = [e_1, e_2, \dots, e_Q]$ - a execução de notas musicais - que gera um sinal acústico contínuo $x(t)$. Quando esse sinal é amostrado com frequência f_s , obtém-se um sinal discreto $x[n]$ que se relaciona a $x(t)$ conforme a Expressão 2.3. A transcrição automática de música, nesse contexto, representa a análise do sinal discreto $x[n]$ - uma seqüência de números - na tentativa de inferir a seqüência de eventos \mathbf{e} que o gerou.

Neste capítulo, serão explicados os fundamentos do modelo computacional utilizado para a transcrição de áudio. Inicialmente, serão mostrados alguns conceitos básicos relacionados a distribuições e misturas gaussianas. Posteriormente, os conceitos ligados aos modelos ocultos de Markov serão abordados. Por fim, serão discutidas particularidades do sistema implementado.

3.1 Misturas gaussianas

3.1.1 Definição

Uma curva gaussiana de média μ e variância σ^2 pode ser definida, para qualquer ponto x , como:

$$\psi(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{|x - \mu|^2}{2\sigma^2}\right). \quad (3.1)$$

Curvas gaussianas frequentemente são as mais adequadas para descrever conjuntos de dados que se organizam em torno de um certo valor, como será discutido a seguir. Para esses casos, $\psi(x, \mu, \sigma)$ pode ser considerada a densidade de probabilidade que determina a distribuição da variável aleatória x [48]. Dessa forma, a probabilidade do

resultado de um experimento descrito por uma gaussiana se situar entre os pontos x_a e x_b é dada por:

$$P(x_a \leq x \leq x_b) = \int_{x_a}^{x_b} \psi(x, \mu, \sigma) dx. \quad (3.2)$$

Há determinadas distribuições de dados que podem ser descritas razoavelmente bem por gaussianas simples. Um exemplo é a distância alcançada por um projétil disparado por um canhão. Embora o canhão mire sempre num ponto à mesma distância, o projétil é aleatoriamente desviado devido a condições ambientais (por exemplo, o atrito com o ar) de forma que a distância que atinge passa a ser uma variável aleatória com média e variância mensuráveis.

Existem casos, porém, em que os dados disponíveis se organizam em torno de um certo número de valores. Tomando por base o exemplo acima, se houverem dois canhões diferentes, é natural que os projéteis disparados por cada um deles sejam variáveis aleatórias com média e variância diferentes. Ao se analisar as distâncias referentes a um número grande de disparos, sem se considerar o canhão que os realizou, pode-se verificar que os dados se organizam em torno de dois valores, cada um deles relacionado a um canhão. Essa distribuição é modelada, matematicamente, como uma mistura de gaussianas.

Misturas gaussianas são somas finitas de M curvas gaussianas, sendo cada curva descrita por média μ_m e variância σ_m^2 . No somatório, as curvas gaussianas são ponderadas por um fator g_m , cujo significado será explicado adiante. Por simplicidade, o conjunto dos parâmetros μ , σ e g relativos a uma certa mistura serão referidos por θ . Assim, uma curva caracterizada por uma mistura de gaussianas é calculada por:

$$\rho(x, \theta) = \sum_{m=1}^M g_m \psi(x, \mu_m, \sigma_m). \quad (3.3)$$

É importante perceber que, para manter a área de $\rho(x, \theta)$ igual a 1, a soma de todos os ganhos $g(m)$ deve ser igual a 1.

Dada uma mistura gaussiana θ , é possível gerar um conjunto de dados que obedeça essa distribuição. A cada dado gerado, a m -ésima componente da mistura é selecionada com probabilidade g_m . Uma vez realizada essa escolha, μ_m e σ_m são utilizadas para gerar um ponto com distribuição gaussiana através de um algoritmo pseudo-aleatório existente. O processo é repetido por quantas iterações se deseje.

3.1.2 Estimativa

Embora seja relativamente simples gerar pontos aleatórios que obedeçam aos parâmetros θ de uma mistura gaussiana, não existe processo matematicamente exato que forneça os parâmetros θ a partir de um conjunto de pontos observados $\mathbf{o} = [o_1, o_2, \dots, o_Q]$. Para realizar essa estimativa, são utilizados métodos iterativos. Neste trabalho, o método de maximização de expectativa (*expectation maximization* - EM) foi escolhido por ser amplamente documentado e utilizado em trabalhos semelhantes. O método busca maximizar um parâmetro Λ correspondente à somatória dos valores da mistura gaussiana para cada ponto o_q , ou seja:

$$\Lambda = \sum_{q=1}^Q \rho(o_q, \theta). \quad (3.4)$$

Uma dedução completa das equações relativas ao método EM pode ser encontrada em [49]. Algumas adaptações foram realizadas neste trabalho visando manter o paralelismo de notação e a clareza de raciocínio.

O primeiro passo do método EM é o cálculo da expectativa - chamado de passo E. Nesse passo, calcula-se $\gamma_{q,m}$, que representa a probabilidade de o ponto observado o_q ter sido gerado pela m -ésima componente da mistura. O cálculo de $\gamma_{q,m}$ será discutido posteriormente.

O segundo passo do método EM é a re-estimação dos parâmetros - chamado de passo M. Nesse ponto, um novo modelo $\bar{\theta}$ é estimado de forma que Λ é aumentado. As equações de re-estimativa de modelo são as seguintes:

$$\bar{\mu}_m = \frac{\sum_{q=1}^Q \gamma_{q,m} o_q}{\sum_{q=1}^Q \gamma_{q,m}}. \quad (3.5)$$

$$\bar{\sigma}_m = \sqrt{\frac{\sum_{q=1}^Q \gamma_{q,m} (\bar{\mu}_m - o_q)^2}{\sum_{q=1}^Q \gamma_{q,m}}}. \quad (3.6)$$

$$\bar{g}_m = \frac{\sum_{q=1}^Q \gamma_{q,m}}{\sum_{q=1}^Q \sum_{m=1}^M \gamma_{q,m}}. \quad (3.7)$$

Os dois passos são repetidos até que algum critério de convergência seja atingido.

3.1.3 Heurísticas usadas junto ao método EM

Um aspecto que não pode ser ignorado durante a implementação do método EM é que, com o aumento de $|\mu_m - x|$, o valor de $\psi(x, \mu, \sigma)$ dado pela Expressão 3.1 decai exponencialmente para zero. Embora em domínios de precisão infinita isso não seja realmente relevante, em computadores digitais a precisão limitada da máquina pode levar o valor de uma gaussiana a ser arredondado para zero. Na prática, isso faz com que algumas amostras do conjunto \mathbf{o} sejam desconsideradas durante a estimativa de $\bar{\theta}$.

Ao desconsiderar alguns elementos da mistura, uma gaussiana tenderá a diminuir sua variância. Em uma situação extrema, um dos elementos da mistura gaussiana pode chegar a ter variância quase nula e situar-se sobre um único ponto do conjunto \mathbf{o} . Assim, embora as equações de re-estimativa, de fato, maximizem o parâmetro Λ a cada iteração, a mistura obtida não é adequada para uso posterior. Para resolver esse problema, uma série de medidas pode ser adotada.

Rabiner [50] propõe que sejam adicionadas restrições para as variâncias das gaussianas da mistura. Assim, o valor de uma gaussiana nunca ultrapassa um determinado piso pré-determinado. Na implementação HMM Toolbox for MatLab [51], esse problema é contornado somando um valor fixo a todas as variâncias da mistura a cada iteração.

Na implementação Hidden Markov Model Toolbox [52], uma solução diferente é adotada. Como a redução exagerada da variância de uma gaussiana pode significar que não há dados suficientes para estimar uma mistura de gaussianas com tantos elementos quantos se deseje, quando uma gaussiana assume variância ou fator de ponderação menores que dois pisos pré-determinados ela é simplesmente removida da mistura. Essa remoção, realizada atribuindo-se fator de ponderação zero para o elemento em questão, não pode ser realizada caso só haja um elemento na mistura e, além disso, deve ser observado que a soma dos fatores de ponderação da mistura deve se manter sempre igual a 1. Assim, para cada elemento m restante na mistura, deve-se atribuir $g_m = g_m / \sum_{m|g_m \neq 0} g_m$. Embora a remoção de uma gaussiana da mistura leve à diminuição momentânea do fator Λ , ela permite a estimativa de uma mistura estatisticamente mais válida para os dados do conjunto \mathbf{o} .

3.1.4 Inicialização

O método EM, junto às heurísticas citadas, é capaz de realizar a maximização local de uma dada mistura gaussiana. É um fato conhecido, porém, que muitos problemas reais são multimodais, ou seja, possuem diversos máximos locais [50]. Por isso, é preciso

inicializar os parâmetros da mistura gaussiana de forma a buscar idealmente um máximo global. Neste trabalho, as misturas gaussianas foram inicializadas de forma semelhante à inicialização de redes neurais do tipo RBF descrita por Haykin [53]: as médias das gaussianas iniciais são determinadas pelo algoritmo K-Means, conforme descrito no Algoritmo 1, a variância inicial de cada gaussiana é igual à menor distância entre sua média e as médias das outras gaussianas, conforme a expressão 3.8 e o ganho é igual ao inverso do número de gaussianas da mistura, conforme a expressão 3.9.

```

// Inicialização
para cada gaussiana m faça
|  $\mu_m \leftarrow o_m$ 
fim
para cada elemento a ser classificado j faça
|  $\bar{G}_j \leftarrow \arg \min_m |o_j - \mu_m|$ 
fim
// Passo
enquanto houver mudança em algum  $\mu_m$  em relação à iteração anterior faça
| para cada gaussiana m faça
| |  $\mu_m \leftarrow$  média de  $o_j$  para todo  $j$  tal que  $\bar{G}_j = m$ 
| fim
| para cada elemento a ser classificado j faça
| |  $\bar{G}_j \leftarrow \arg \min_m |o_j - \mu_m|$ 
| fim
fim

```

Algoritmo 1: Algoritmo K-Means.

$$\sigma_m = \min_j |\mu_m - \mu_j|, j \neq m, m = [1, 2, \dots, M]. \quad (3.8)$$

$$g_m = \frac{1}{M}, m = [1, 2, \dots, M]. \quad (3.9)$$

3.2 Modelos ocultos de Markov

Na seção anterior, foi discutido um caso hipotético em que dois canhões diferentes disparam projéteis. Supõe-se que cada um dos disparos seja independente do anterior, de forma que é possível saber imediatamente qual canhão mais provavelmente disparou o último projétil. É possível, porém, alterar as hipóteses de forma que a escolha do canhão que irá disparar o próximo projétil leve em conta o canhão que disparou o

último. Assim, pode-se dizer que o sistema gerador de disparos possui dois estados: s_1 , para o caso do disparo do primeiro canhão, e s_2 , para o caso do disparo do segundo. Ainda, no modelo, a probabilidade do canhão j ser escolhido para o próximo disparo dado que o canhão escolhido anteriormente foi o canhão i é igual a $a_{i,j}$. Além disso, ao assumir um estado s_1 ou s_2 o sistema passa a gerar disparos que dependem dos parâmetros de cada um dos canhões. A Figura 3.1 mostra uma representação gráfica do sistema gerador de tiros.

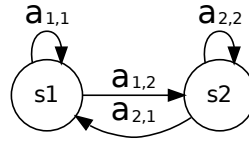


Figura 3.1: Representação do sistema gerador de tiros de canhão tal que a escolha de um canhão para um tiro depende do canhão que disparou o último tiro.

Esse tipo de modelo estatístico se chama modelo oculto de Markov (*hidden Markov model* - HMM). Um HMM é completamente descrito por uma matriz de probabilidades de transições entre estados $\mathbf{A} \in \mathbb{R}^{J \times J}$, onde J é o número de estados do sistema, um conjunto de funções probabilísticas b_j tal que $b_j(o_q)$ é a probabilidade de ter sido observada a saída o_q dado que o estado do sistema é s_j , e um vetor de inicialização \mathbf{I} , onde I_j é a probabilidade de o sistema se iniciar no estado s_j na primeira observação discreta. O conjunto de todos esses parâmetros é denotado de forma sintética por λ . A saída do sistema em um dado instante q é uma variável aleatória chamada emissão (no exemplo acima, a emissão do sistema é a distância percorrida pelo projétil disparado).

Através de HMMs cujos parâmetros foram corretamente ajustados, é possível calcular a sequência de estados que mais provavelmente gerou uma sequência de observações \mathbf{o} . Se os estados do HMM forem construídos de forma a modelar eventos, então é possível saber qual é a sequência de eventos \mathbf{e} que mais provavelmente gerou a sequência de observações \mathbf{o} .

Embora seja possível descrever HMMs de forma generalizada [50], para este trabalho foram realizadas algumas especializações que serão descritas a seguir.

3.2.1 Modelos esquerda-para-a-direita

Neste trabalho, serão utilizados essencialmente modelos HMM do tipo esquerda-para-a-direita (*left-to-right*, LR). Nesse tipo de modelo, $I_1 = 1$ e $I_j = 0$ para todo

$j \neq 1$. Além disso, $a_{i,j}$ só é diferente de zero se $i = j$ ou se $i = j - 1$. Para uma análise consistente dos HMMs do tipo LR, são adicionados a ele dois estados que não produzem emissão (pseudo-estados). O primeiro deles é um estado inicial, que tem probabilidade de transição 1 para o estado s_1 . O segundo é o estado final, que representa o fim das emissões decorrentes do modelo. Esses estados não têm significado físico, existindo apenas como apoio lógico e matemático ao sistema, como será visto a seguir. A Figura 3.2 mostra um HMM do tipo LR com três estados.

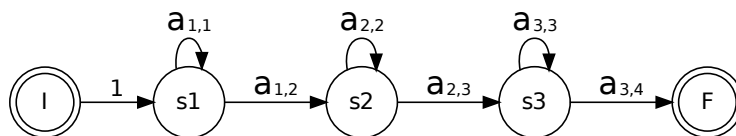


Figura 3.2: HMM do tipo esquerda-para-a-direita com três estados, incluindo os pseudo-estados inicial (I) e final (F). A probabilidade de transição entre o terceiro estado e o estado final é denotada $a_{3,4}$.

3.2.2 Unindo HMMs

Uma característica interessante de HMMs do tipo LR é que estes podem ser acoplados facilmente como blocos de um HMM mais complexo. Quando K HMMs do tipo LR são unidos, é necessária uma matriz de transições entre os modelos $\bar{\mathbf{A}} \in \mathbb{R}^{K \times K}$ que indica a probabilidade de transição do último estado do modelo i para o primeiro estado do modelo j .

Nesse processo, é interessante denotar as probabilidades de transição de cada um dos modelos k como $a_{k,i,j}$, representando a probabilidade de transição do estado s_i para o estado s_j no contexto do modelo k .

Embora a implementação da união de HMMs seja razoavelmente complexa, o conceito de união de HMMs permite a sintetização dos significados dos estados, bem como a redução da complexidade da representação dos HMMs gerados. A Figura 3.3 mostra a união de dois HMMs do tipo LR de três estados. Pode-se verificar que a complexidade do sistema aumenta e sua interpretação torna-se menos intuitiva.

Na Figura 3.4, o mesmo HMM é mostrado utilizando o conceito de união de modelos. Não é importante, nesse contexto, mostrar o comportamento do sistema no interior de cada modelo, mas somente as relações entre os modelos do tipo LR que geraram o modelo final.

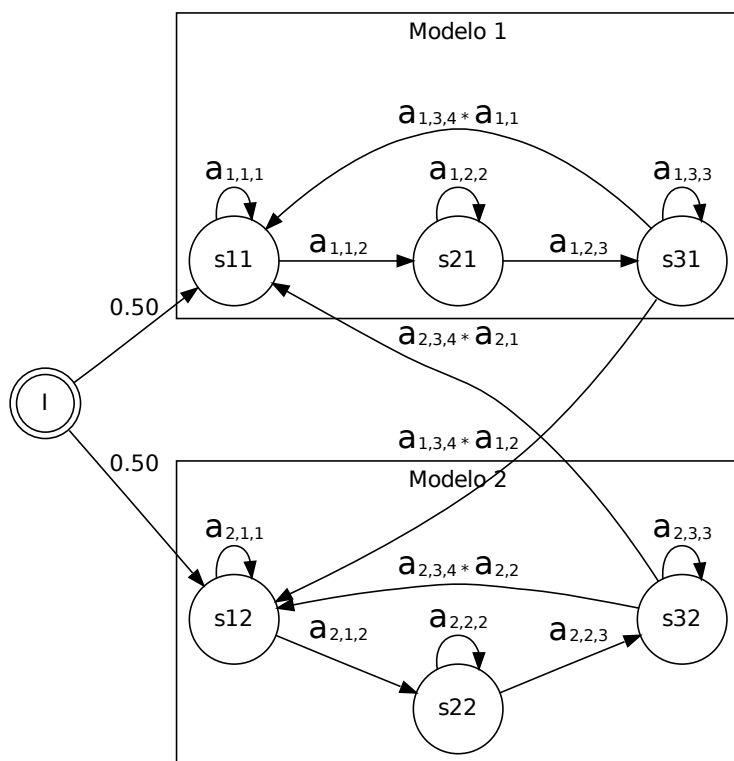


Figura 3.3: HMM resultante da união de dois HMMs do tipo LR com três estados cada um.

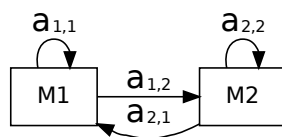


Figura 3.4: Representação sintética do HMM resultante da união de dois HMMs do tipo LR com três estados cada um. A caixa M1 representa o primeiro HMM e a caixa M2 representa o segundo HMM.

O conceito de união de HMMs, embora não faça diferença em termos computacionais, simplifica a modelagem da lógica de sistemas mais complexos. Esse conceito será utilizado posteriormente.

3.2.3 Distribuição e dimensionalidade de emissões

Como já foi discutido, a emissão de um HMM é uma variável aleatória. Sendo assim, tem uma certa densidade de probabilidade. Neste trabalho, a densidade de probabilidade das emissões relacionadas a cada um dos estados será modelada como uma mistura de gaussianas.

Deve-se considerar, também, que é possível que um HMM produza emissões em diversas dimensões. No caso dos canhões, é possível que, por exemplo, além da distância percorrida pelo projétil, seja medida a intensidade do ruído gerado pelo disparo, de forma que a emissão passa a ser uma variável aleatória com duas dimensões independentes.

Adota-se uma mistura de gaussianas distinta para cada dimensão d em cada estado j . Assim, se há D dimensões independentes, para cada um dos estados haverá D misturas independentes, cada uma com um certo número de gaussianas. Cada uma dessas misturas é denotada $\theta_{j,d}$ e é composta dos elementos $[g_{j,d,m}, \mu_{j,d,m}, \sigma_{j,d,m}]$, com d variando entre 1 e D e m variando entre 1 e o número $M_{j,d}$ de componentes na mistura.

Para representar uma seqüência de Q observações de dimensão D , utiliza-se uma matriz $\mathbf{O} \in \mathbb{R}^{Q \times D}$, de forma que $o_{q,d}$ representa a observação no tempo discreto q relacionada à dimensão d . Para se referir ao conjunto de observações $o_{q,d}$ com d variando de 1 a D , ou seja, ao vetor $[o_{q,1}, o_{q,2}, \dots, o_{q,D}]$, utiliza-se a notação \mathbf{o}_q . Para calcular a probabilidade $b_j(o_q)$, é preciso recorrer à Expressão 3.3, na forma:

$$b_j(o_q) = \prod_{d=1}^D \rho(o_{q,d}, \theta_{j,d}) = \prod_{d=1}^D \sum_{m=1}^{M_{j,d}} g_{j,d,m} \psi(o_{q,d}, \mu_{j,d,m}, \sigma_{j,d,m}). \quad (3.10)$$

Os valores $b_j(o_q)$ constituem uma matriz $\mathbf{B} \in \mathbb{R}^{Q \times J}$ de probabilidades *a priori* relacionadas ao modelo e às observações. A expressão *a priori* denota o fato de que essas são as probabilidades de cada um dos estados ter gerado cada uma das observações sem levar em consideração outras características do modelo.

É interessante perceber que, embora $\sum_{j=1}^J b_{q,j} \neq 1$, o que é matematicamente impossível, uma vez que as medidas observadas com certeza foram geradas pelo sistema, todas as operações que seguem lidam com razões entre probabilidades e nunca com

valores absolutos, de forma a corrigir automaticamente esse erro.

3.2.4 Seqüência de estados mais provável

Uma vez que sejam conhecidos os parâmetros λ de um HMM e uma seqüência de observações \mathbf{o} , é possível calcular uma seqüência de estados que melhor explica (no sentido de maximizar a probabilidade) a ocorrência de \mathbf{o} . Para isso, é utilizado o algoritmo de Viterbi [50], que, conforme descrito no Algoritmo 2, se baseia na construção de duas matrizes: $D_{q,j}$, que contém a probabilidade de se estar no estado j durante a observação q , e $G_{q,j}$, que contém o número do estado da observação $q - 1$ que gerou a melhor probabilidade de se estar no estado j durante a iteração q .

```

// Inicialização
para cada estado  $i$  faça
|    $d_{1,i} \leftarrow I_i b_{1,i}$ 
|    $g_{1,i} \leftarrow 0$ 
fim
// Recursão
para cada Tempo discreto  $q$  entre 2 e  $Q$  faça
|   para cada estado  $j$  faça
|   |    $d_{q,j} \leftarrow \max_i d_{q-1,i} \times a_{i,j} \times b_{q,j}$ 
|   |    $g_{q,j} \leftarrow \arg \max_i d_{q-1,i} \times a_{i,j}$ 
|   fim
|   para cada estado  $j$  faça
|   |   // Normalização
|   |    $d_{q,j} \leftarrow \frac{d_{q,j}}{\max_i d_{q,i}}$ 
|   fim
fim
// Término
 $e_Q \leftarrow \arg \max_i d_{Q,i}$ 
// Cálculo do caminho ótimo
para cada Tempo discreto  $q$ , em ordem decrescente a partir de  $Q - 1$  faça
|    $e_q \leftarrow g_{q+1, g_{q+1}}$ 
fim

```

Algoritmo 2: Algoritmo de Viterbi.

Atenção especial deve ser dada à etapa de normalização. É interessante perceber que, rapidamente, $d_{q,j}$ tende a valores muito pequenos - tão pequenos que deixam de ser representáveis pela maior parte dos computadores. A etapa de normalização busca manter os valores de $d_{q,j}$ próximos à unidade. Embora $d_{q,j}$ deixe de conter, de fato,

a probabilidade de se estar no estado j durante a iteração q , o que é relevante para o algoritmo é a proporção entre os valores de cada linha de D , e não seu valor absoluto. Assim, o algoritmo continua correto.

Em determinadas situações, pode ser interessante implementar esse algoritmo usando a escala logarítmica. Algumas sequências de observações podem conter pontos situados bastante distantes da média dos elementos de uma ou mais misturas, levando algum elemento $b_{q,j}$ a zero. Esse arredondamento pode levar o algoritmo a apresentar comportamentos inadequados. Para evitar erros dessa natureza, assume-se que a matriz D será calculada em escala logarítmica, de forma que o algoritmo de Viterbi passa a ser formulado como descrito no Algoritmo 3.

```

// Inicialização
para cada estado  $i$  faça
|    $d_{1,i} \leftarrow \log I_i + \log b_{1,i}$ 
|    $g_{1,i} \leftarrow 0$ 
fim
// Recursão
para cada Tempo discreto  $q$  entre 2 e  $Q$  faça
|   para cada estado  $j$  faça
|   |    $d_{q,j} \leftarrow \max_i d_{q-1,i} + \log a_{i,j} + \log b_{q,j}$ 
|   |    $g_{q,j} \leftarrow \arg \max_i d_{q-1,i} + \log a_{i,j}$ 
|   fim
|   fim
// Término
 $e_Q \leftarrow \arg \max_i d_{Q,i}$ 
// Cálculo do caminho ótimo
para cada Tempo discreto  $q$ , em ordem decrescente a partir de  $Q - 1$  faça
|    $e_q \leftarrow g_{q+1, g_{q+1}}$ 
fim

```

Algoritmo 3: Formulação logarítmica do algoritmo de Viterbi.

É interessante perceber que, na formulação logarítmica, não é necessária a realização da etapa de normalização, uma vez que o decaimento dos valores $d_{q,j}$ é, agora, linear em relação a q e não mais exponencial, de forma que é necessário um número muito grande de iterações até que haja arredondamento de $d_{q,j}$ para zero¹. Apesar disso, se

¹Num exemplo simples, se as probabilidades *a priori* de cada observação forem da ordem de 10^{-1} , são necessárias por volta de 300 iterações para que os valores $d_{q,j}$ sejam arredondados para zero caso seja aplicado o algoritmo de Viterbi conforme descrito no Algoritmo 2, ao passo que, na versão logarítmica descrita no Algoritmo 3 seriam necessárias por volta de 10^{300} iterações, considerando números de precisão dupla das máquinas de 32 bits da atualidade.

necessária, a normalização pode ser realizada somando-se um valor comum a todo $d_{q,j}$ a cada iteração q .

3.2.5 Estimativa de parâmetros

Para que o algoritmo de Viterbi funcione adequadamente, é preciso que os parâmetros do modelo λ sejam devidamente ajustados de forma que o modelo se aproxime do fenômeno analisado. Matematicamente, esse problema pode ser enunciado como maximizar a probabilidade relacionada a uma dada matriz \mathbf{O} de observações, ou seja:

$$\max_{\lambda} P(\mathbf{O}|\lambda). \quad (3.11)$$

Embora não exista uma maneira analítica de maximizar λ na Expressão 3.11, é possível, de maneira iterativa, encontrar um máximo local de λ [50]. Através da inicialização correta de λ , é possível que esse máximo local seja suficientemente próximo ao máximo global, de forma a garantir a funcionalidade do modelo para fins práticos. O algoritmo de maximização local, conhecido como algoritmo de Baum-Welch [50], inicia-se com o cálculo da matriz de probabilidades *a priori* \mathbf{B} conforme a Expressão 3.10.

Após, deve ser calculada a matriz de probabilidades adiante (*forward*) $\alpha \in \mathbb{R}^{Q \times J}$, que é definida matematicamente como:

$$\alpha_{q,j} = P(\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \dots, \mathbf{o}_q, e_q = s_j | \lambda). \quad (3.12)$$

Assim como o Algoritmo de Viterbi, o cálculo da matriz α também demanda a normalização progressiva dos coeficientes encontrados, de forma a evitar que sejam atingidos valores não representáveis pelos formatos de ponto flutuante dos computadores digitais. O processo completo para o cálculo da matriz α está descrito no Algoritmo 4.

Também deve ser calculada a matriz de probabilidade reversa (*backwards*) $\beta \in \mathbb{R}^{Q \times J}$, definida matematicamente como:

$$\beta_{q,j} = P(\mathbf{o}_{q+1}, \mathbf{o}_{q+2}, \mathbf{o}_{q+3}, \dots, \mathbf{o}_Q, e_q = s_j | \lambda). \quad (3.13)$$

O cálculo de $\beta_{q,j}$, bastante semelhante ao cálculo de $\alpha_{q,j}$, também observa a questão da normalização, conforme descrito no Algoritmo 5.

Uma vez obtidos $\alpha_{q,j}$ e $\beta_{q,j}$, passa-se ao cálculo de $\xi(q, i, j)$, que é a probabilidade de se estar no estado i durante a observação q e no estado j durante a observação $q + 1$,

```

// Inicialização
para cada estado  $j$  de 1 a  $J$  faça
|    $r_\alpha(1) \leftarrow \frac{1}{\sum_{i=1}^J I_i b_{1,i}}$ 
|    $\alpha_{1,j} \leftarrow 1$ 
fim
// Indução
para cada Tempo discreto  $q$  faça
|   para cada estado  $j$  faça
|   |    $\alpha_{q+1,j} \leftarrow \sum_{i=1}^J \alpha_{q,i} a_{i,j} b_{q+1,j}$ 
|   fim
|   // Normalização
|    $r_\alpha(q+1) \leftarrow \frac{1}{\sum_{j=1}^J \alpha_{q+1,j}}$ 
|   para cada estado  $j$  faça
|   |    $\alpha_{q+1,j} \leftarrow \alpha_{q+1,j} r_\alpha(q+1)$ 
|   fim
fim

```

Algoritmo 4: Cálculo da matriz de probabilidades adiante.

ou, matematicamente:

$$\xi(q, i, j) = P(e_q = s_i, e_{q+1} = s_j | \mathbf{O}, \lambda). \quad (3.14)$$

A partir dos parâmetros do modelo e das matrizes $\alpha_{q,j}$ e $\beta_{q,j}$, $\xi(q, i, j)$ é obtido pela expressão:

$$\xi(q, i, j) = \frac{\alpha_{q,j} a_{i,j} b_{q+1,j} \beta_{q+1,j}}{\sum_{i=1}^J \sum_{j=1}^J \alpha_{q,i} a_{i,j} b_{q+1,j} \beta_{q+1,j}}. \quad (3.15)$$

É interessante perceber que $\xi(q, i, j)$ só pode ser calculado, pela Expressão 3.15, para valores de q variando de 1 a $Q - 1$. Essa característica será abordada mais adiante.

O próximo passo no algoritmo é o cálculo de $\gamma_{q,j}$, que representa a probabilidade *a posteriori*, ou seja, levando em conta as transições entre estados do modelo, de se estar no estado j durante a observação q . Esse cálculo se dá pela expressão:

$$\gamma_{q,i} = P(e_q = s_i) = \sum_{j=1}^J \xi(q, i, j). \quad (3.16)$$

Em HMMs nos quais a emissão de cada estado é modelada por uma mistura de gaussianas, é possível calcular $\gamma_{q,j,m,d}$, que representa a probabilidade de a m -ésima gaussiana da mistura do estado j ter gerado a observação q na dimensão d . Esse

```

// Inicialização
para cada estado  $j$  de 1 a  $J$  faça
|    $r_\beta(Q) \leftarrow \frac{1}{J}$ 
|    $\beta_{Q,j} \leftarrow 1$ 
fim
// Indução
para cada Tempo discreto  $q$  de  $Q - 1$  a 1 faça
|   para cada estado  $i$  faça
|   |    $\beta_{q,i} \leftarrow \sum_{j=1}^J \beta_{q+1,j} a_{i,j} b_{q+1,j}$ 
|   fim
|   // Normalização
|    $r_\beta(q) \leftarrow \frac{1}{\sum_{j=1}^J \beta_{q,j}}$ 
|   para cada estado  $j$  faça
|   |    $\beta_{q,j} \leftarrow \beta_{q,j} r_\beta(q)$ 
|   fim
fim

```

Algoritmo 5: Cálculo da matriz de probabilidades reversa.

parâmetro é calculado por:

$$\gamma_{q,j,d,m} = \gamma_{q,j} \frac{g_{j,d,m} \psi(o_{q,d}, \mu_{j,d,m}, \sigma_{j,d,m})}{\sum_{m'=1}^M g_{j,d,m'} \psi(o_{q,d}, \mu_{j,d,m'}, \sigma_{j,d,m'})}. \quad (3.17)$$

No atual ponto da execução do algoritmo, já é possível re-estimar os parâmetros do modelo de forma que o modelo re-estimado $\bar{\lambda}$ seja melhor que o modelo atual λ no sentido de maximizar a probabilidade de que a seqüência de observações utilizada como entrada tenha sido gerada pelo modelo.

Inicialmente, é possível re-estimar a matriz de transições \mathbf{A} conforme proposto por Rabiner [50] na forma:

$$\bar{a}_{i,j} = \frac{\sum_{q=1}^{Q-1} \xi(q, i, j)}{\sum_{q=1}^{Q-1} \gamma_{q,i}}. \quad (3.18)$$

A re-estimativa dos parâmetros relacionados às probabilidades de observações baseia-se nas Expressões 3.5, 3.6 e 3.7. Tais expressões devem ser calculadas para cada uma das misturas gaussianas existentes no modelo, o que significa calculá-las para cada estado j e para cada dimensão d , obtendo:

$$\bar{\mu}_{j,d,m} = \frac{\sum_{q=1}^Q \gamma_{q,j,d,m} o_{q,d}}{\sum_{q=1}^Q \gamma_{q,j,d,m}}. \quad (3.19)$$

$$\bar{\sigma}_{j,d,m} = \sqrt{\frac{\sum_{q=1}^Q \gamma_{q,j,d,m} (\bar{\mu}_{j,d,m} - o_{q,d})^2}{\sum_{q=1}^Q \gamma_{q,j,d,m}}}. \quad (3.20)$$

$$\bar{g}_{j,d,m} = \frac{\sum_{q=1}^Q \gamma_{q,j,d,m}}{\sum_{q=1}^Q \sum_{m=1}^M \gamma_{q,j,d,m}}. \quad (3.21)$$

Múltiplas seqüências de observação

Assim como em outros casos de algoritmos de aprendizado supervisionado, é interessante que diversas seqüências de observação sejam utilizadas para o treino de um modelo. Nesse tipo de caso, o procedimento é bastante simples. Basta calcular, para cada uma das K seqüências de observações, os parâmetros $\xi^{(k)}(q, i, j)$, $\gamma_{q,j}^{(k)}$ e $\gamma_{q,j,d,m}^{(k)}$. Ao aplicar as equações de re-estimativa, utiliza-se o conjunto total dos parâmetros calculados e observações, tomando-se o cuidado de manter as seqüências disjuntas ao desconsiderar as transições entre o último estado relacionado à seqüência k e o primeiro estado relacionado à seqüência $k + 1$.

Re-estimativa em modelos esquerda-para-a-direita

No caso da re-estimativa de parâmetros em modelos HMM tipo LR, é preciso levar em consideração o evento de saída do modelo, que supostamente ocorre ao fim de cada seqüência de observações. Para isso, utiliza-se o pseudo-estado final, como mostrado na Figura 3.2, ao qual se atribui o seguinte comportamento especial: a probabilidade *a priori* de se estar no estado final é 1 para $q = Q + 1$ (ou seja, após o fim das observações) e 0 para qualquer outro q . Isso significa adicionar um estado e uma observação à matriz \mathbf{B} , de tal forma que $b_{Q+1,J+1} = 1$, e, portanto, uma linha e uma coluna são adicionados também às matrizes α e β .

Assim, essa operação permite o cálculo de $\xi(q, i, j)$ para todo q entre 1 e Q , e incorpora ao cálculo da matriz de transições a probabilidade de saída do modelo (representado por $a_{3,4}$ no exemplo da Figura 3.2).

Capítulo 4

Transcrição automática de música

Neste capítulo, o algoritmo de transcrição utilizado como base neste trabalho será descrito em detalhes, tal qual foi implementado. O capítulo se iniciará com a explicação dos conceitos elementares utilizados para transcrição de áudio. Após, os algoritmos de processamento digital de sinais serão abordados, seguidos do detalhamento das particularidades do modelo cognitivo utilizado. Depois disso, as estratégias e parâmetros utilizados para treino serão mostrados e, por fim, os resultados do processo de avaliação serão discutidos.

4.1 Eventos musicais e cadeias de eventos discretos

Como discutido na Seção 2.5, qualquer forma de notação musical pode ser considerada como uma sequência de instruções que devem ser cumpridas para a execução correta de uma música. A obtenção de uma forma de notação correspondente a um sinal acústico pode ser vista como um processo de parametrização desse mesmo sinal, no qual a sequência de eventos que o gerou deve ser descoberta.

É possível, de maneira geral, modelar um instrumento musical como um sistema que assume estados discretos através do tempo, que variam de acordo com a nota musical que soa. Dentro do contexto desse modelo, são excluídos casos em que o instrumento emite mais de uma nota ao mesmo tempo - o instrumento modelado é, portanto, monofônico. Assim, ou o instrumento assume um estado de silêncio ou assume um entre os estados $s_1, s_2 \dots s_p$ correspondentes às P notas da tessitura do instrumento. A Figura 4.1 mostra um modelo de instrumento desse tipo capaz de emitir uma entre duas notas ou permanecer em silêncio.

Ainda, é possível assumir que, em intervalos de tempo suficientemente curtos, o

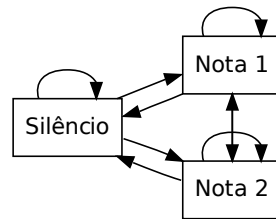


Figura 4.1: Modelo máquina de estados de um instrumento capaz de emitir duas notas ou silêncio.

sistema permanece em apenas um estado, uma vez que notas musicais usualmente têm duração da ordem de alguns décimos de segundo. A duração desse intervalo de tempo é, em princípio, arbitrária, e dependente apenas da precisão temporal requerida do instrumento, embora existam razões físicas (apresentadas posteriormente) pelas quais esse intervalo não pode ser infinitamente curto.

Dado o problema e adotado esse conceito de instrumento, é possível modelar um transcritor da seguinte forma: um sinal $x[n]$ é inicialmente dividido em Q quadros $x_q[n]$ de forma conveniente. Para cada quadro q , um certo conjunto de *features* \mathbf{o}_q , presumidamente capaz de caracterizar cada um dos estados de um instrumento, é calculado. O problema da transcrição automática pode ser, assim, interpretado como o problema de encontrar o caminho entre os $P + 1$ estados do sistema que mais provavelmente gerou a seqüência calculada de vetores \mathbf{o}_q . Se a relação entre \mathbf{o}_q e cada um dos estados s_p for uma função densidade de probabilidade, então a máquina de estados correspondente ao instrumento pode ser considerada um Modelo Oculto de Markov e, portanto, o cálculo desta seqüência mais provável de estados pode ser realizado através do algoritmo de Viterbi, conforme visto no Capítulo 3.

O cálculo de \mathbf{o}_q será abordado posteriormente. Por ora, é necessário rever a definição da máquina de estados correspondente ao instrumento musical. Com o modelo de instrumento discutido até agora (uma máquina de estados com um estado para cada nota), é possível que uma execução do algoritmo de Viterbi retorne uma seqüência em que um estado apareça repetidas vezes, como mostrado na Figura 4.2, gerando uma situação de interpretação dúbia na qual é impossível saber se a nota correspondente ao estado s_1 foi tocada quatro vezes, se foi tocada uma única vez com duração quatro vezes maior, se foi tocada duas vezes com duração duas vezes maior ou se alguma combinação desses quatro casos ocorreu. Isso poderia ocorrer durante a análise das três primeiras notas de uma execução da partitura da Figura 2.2, por exemplo.

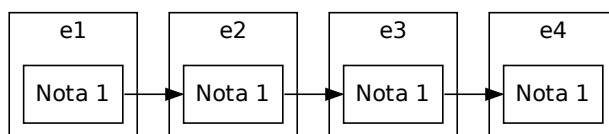


Figura 4.2: Sequência de estados cuja interpretação é duvidosa.

Numa abordagem um pouco mais complexa, cada uma das notas de um instrumento musical poderia ser modelada como uma sequência de dois estados: liga e desliga. Denota-se “modelo acústico” de uma determinada nota o conjunto de estados correspondentes a essa mesma nota do instrumento. Um modelo acústico, se analisado isoladamente, é um HMM do tipo esquerda-para-a-direita. Considerando um modelo acústico de dois estados, o modelo para o mesmo instrumento musical da Figura 4.1 passa a ser como na Figura 4.3.

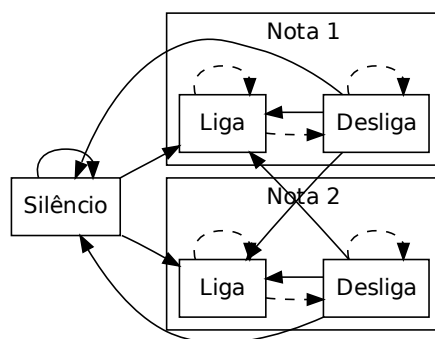


Figura 4.3: Modelo máquina de estados de um instrumento com dois estados por nota musical. As transições de estado que não implicam na finalização de uma nota estão destacadas como setas tracejadas.

Nessa nova configuração, a ambiguidade entre os possíveis inícios e finais de notas musicais deixa de existir, pois os eventos de início e fim de nota estão muito bem definidos.

A física dos instrumentos musicais mostra, porém, que notas podem ser divididas em três momentos essenciais, conforme discutido na Seção 2.2. O primeiro deles é um momento transitório chamado de *ataque*, no qual a nota é iniciada. Após, há um momento estacionário, chamado de *sustentação*, no qual a nota soa livremente (a tecla do piano permanece pressionada, o arco continua friccionando o violino ou a corda de

um violão vibra livremente). Por fim, há um momento transitório final, chamado de *soltura*, correspondente aos gestos que finalizam a execução da nota.

Ao se considerar essas características físicas, obtém-se um modelo mais fielmente relacionado à realidade, de forma que a construção do vetor de *features* observáveis \mathbf{o}_q pode se basear com mais segurança em conceitos já bastante estudados na psico-acústica. A Figura 4.4 evidencia que o modelo com três estados por nota é análogo a um modelo com dois estados por nota com a adição de um estado intermediário no modelo acústico. É interessante perceber que os estados *ataque* e *soltura* são análogos, respectivamente, aos estados *liga* e *desliga* do modelo na Figura 4.3. O estado *sustentação* poderia, na linguagem de instruções adotada na Figura 4.3, ser chamado de *mantém*.

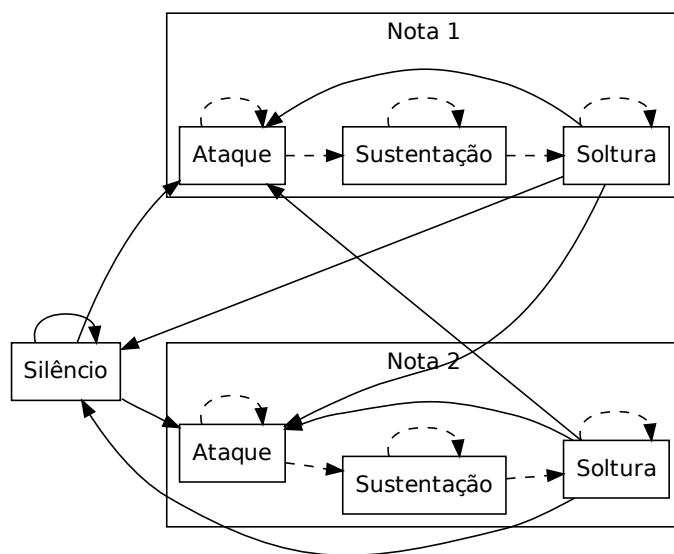


Figura 4.4: Modelo máquina de estados de um instrumento com três estados por nota musical. As transições de estado que não implicam na finalização de uma nota estão destacadas como setas tracejadas.

Uma vez que a função de cada estado do modelo de transcrição foi definida, resta determinar quais são as características psico-acústicas que devem inspirar as *features* que irão compor o vetor de observações \mathbf{o}_q .

4.2 Algoritmos de processamento digital de sinais

Considerando o modelo da Figura 4.4, é necessária a definição de um vetor de *features* \mathbf{o}_q capaz de distinguir entre os estados. Essa definição, tal qual realizada por Ryymanen [34], inspirou-se em algoritmos de processamento de sinais já existentes.

Inicialmente, é preciso discernir entre dois modelos acústicos. Cada um dos modelos se relaciona a uma nota musical, e, portanto, é razoável supor que a frequência fundamental emitida pelo instrumento faça parte do vetor de observações. O sinal analisado neste trabalho, porém, é um sinal de difícil análise, resultante da soma de vários sinais simples, como será visto na Seção 4.2.1. Assim, faz-se necessário o uso de um algoritmo capaz de detectar múltiplas frequências fundamentais presentes em um sinal de áudio. O algoritmo escolhido foi aquele proposto por Klapuri [46].

Esse algoritmo opera calculando uma função chamada *saliência* para cada frequência fundamental possivelmente presente no sinal. A *saliência*, descrita em detalhes na Seção 4.2.1, pode ser vista como a soma ponderada das amplitudes das harmônicas de uma certa frequência, e fornece uma medida do quão forte uma certa F0 está em um certo sinal. Assim, quando o algoritmo é executado, fornece um certo número de F0s e suas respectivas *saliências*.

Ryymanen utiliza, em seu trabalho, modelos acústicos idênticos para cada nota musical [34]. Isso significa que o mesmo conjunto de parâmetros é utilizado para cálculos de probabilidades referentes a todos os modelos acústicos. Como será visto adiante, essa igualdade permite maior simplicidade na análise de dados e a obtenção de maior número de pontos na execução de algoritmos para otimização dos parâmetros do HMM construído. Como os modelos acústicos são idênticos, é preciso diferenciá-los no cálculo do vetor de *features* \mathbf{o}_q . Essa diferenciação ocorre na utilização de um parâmetro chamado Δf . Esse parâmetro, explicado em detalhes na Seção 4.3, corresponde à imprecisão da estimativa de frequências fundamentais em relação à nota mais próxima. Espera-se que, se alguma das tentativas de detecção de frequências fundamentais for bastante próxima à frequência fundamental nominal da nota analisada, o valor de Δf seja próximo a zero, ao passo que se nenhuma das estimativas de frequência fundamental apresentar essa proximidade, o valor de Δf terá módulo maior.

Embora o valor de Δf seja razoavelmente eficiente para discernir entre modelos acústicos, a distinção de estados no interior de modelos acústicos também é importante. Para isso, utilizam-se dois outros parâmetros.

Primeiro, utiliza-se a *saliência* da frequência fundamental tomada como base para o cálculo de Δf , conforme descrito na Seção 4.3. Espera-se que a *saliência* tenha um

valor pequeno nas regiões transitórias - ataque e soltura - de uma nota musical, pois a vibração harmônica está, supostamente, se formando ou se apagando do corpo do instrumento.

Também, é calculado o valor do *acento* do sinal acústico. O acento é um valor que dá uma idéia do quão forte é a sensação de batida (no sentido musical) em um determinado quadro, e foi inicialmente utilizado em [47] com o objetivo de detectar a métrica de sinais acústicos. Explicado em detalhes na Seção 4.2.2, é utilizado por indicar regiões que potencialmente contêm ataques em um determinado sinal.

Os algoritmos até agora citados serão descritos em detalhes a seguir.

4.2.1 Detecção de múltiplas frequências fundamentais

Um sinal periódico com frequência fundamental (F0) f_0 é aquele que se repete a cada período de $1/f_0$ segundos. Matematicamente, essa relação pode ser formulada como:

$$x(t) = x(t - k \frac{1}{f_0}), k \in \mathbb{Z}. \quad (4.1)$$

Como já discutido na Seção 2.1, a F0 de um sinal está intimamente relacionada com a sensação psico-acústica de altura, sendo que sinais com F0 elevada são percebidos como agudos e sinais com F0 baixa são percebidos como graves. A frequência fundamental é, assim, a propriedade física que permite a distinção entre os sinais derivados de duas notas musicais diferentes.

A detecção da frequência fundamental de um sinal é um problema já amplamente estudado [54, 55]. No contexto de áudio, é comum assumir que dentro de um curto intervalo de tempo - da ordem de alguns centésimos de segundo - um sinal acústico contendo uma nota musical isolada - chamado, portanto, de monofônico - pode ser considerado periódico.

O problema se torna mais difícil quando o sinal analisado não é simplesmente um sinal periódico, mas sim a soma de um certo número de sinais periódicos, que em suas formas discretas são descritos pela Expressão 2.4, dando origem a um sinal polifônico. A sensação psico-acústica relacionada a esse tipo de sinal é a audição de um certo número de notas musicais. Detectar múltiplas frequências fundamentais, assim, é o processo de estimar quais são as frequências fundamentais f_j dos J sinais harmônicos que, somados,

resultam no sinal $x_p[n]$:

$$x_p[n] = \sum_{j=1}^J \sum_{m=1}^{M_j} g_{m,j} \cos(2\pi m f_j \frac{n}{f_s} + \phi_{m,j}). \quad (4.2)$$

Embora a Expressão 4.1 não seja válida para sinais polifônicos de maneira geral, a hipótese de que as características do sinal analisado não se modificam ao longo de um curto intervalo de tempo é mantida. Assim, sinais polifônicos são quebrados, no domínio do tempo, em quadros de duração conhecida. Cada quadro, tal qual no caso monofônico, é analisado individualmente.

A Transformada Discreta de Fourier (TDF) torna-se, nesse caso, uma importante ferramenta, pois o sinal analisado passa a ser descrito no domínio do tempo como uma somatória de sinais senoidais, cada um com módulo e fase próprias, uma forma mais próxima do modelo da Expressão 4.2.

Como os seres humanos são, em geral, insensíveis à fase de um sinal no processo de identificação de tons [38], todo o processamento no domínio da frequência é realizado sobre o módulo da TDF do sinal. Ainda, como sinais acústicos são sinais reais, é possível descartar a segunda metade dos elementos dessa TDF devido à sua simetria com a primeira metade. Assim, sem perda de generalidade é possível analisar a TDF do sinal descrito pela Expressão 4.2 na forma:

$$X_p[k] = \sum_{j=1}^J \sum_{m=1}^M G_{m,j} W[k - m f_j K / f_s], \quad (4.3)$$

onde K é o comprimento da TDF, $W[k]$ é a TDF da janela utilizada e $0 \leq k < (K/2)$.

É importante observar que $G_{m,j} \neq g_{m,j}$, uma vez que a soma dos espectros de trechos de senóides com frequências muito próximas pode resultar em interferências tanto construtivas quanto destrutivas.

O processo de detecção de múltiplas frequências fundamentais proposto por Klapuri [46] e utilizado aqui baseia-se no cálculo de uma função - a saliência - que indica o quão plausível é a hipótese de uma determinada frequência ser uma frequência fundamental do sinal.

Uma vez conhecida uma das frequências fundamentais que compõem o sinal analisado, o espectro correspondente é calculado, assumindo-se um modelo espectral semelhante ao da Expressão 4.3. Esse espectro calculado é subtraído do espectro do sinal analisado, e então nova estimativa é realizada. Esse processo, explicado em detalhes

nas seções seguintes, é repetido por tantas vezes quantas se deseje, embora saiba-se que a probabilidade de acerto do sistema caia a cada nova iteração [46].

Branqueamento espectral

Um problema que deve ser enfrentado na estimativa de uma frequência fundamental através da função de saliência é que as amplitudes $G_{m,j}$ dos picos podem variar muito para uma mesma mistura, dependendo das interferências entre as senóides que as compõem. Essas diferenças tendem a tornar mais difícil o cálculo da saliência através de uma única função, pois esta teria que se adaptar aos muitos possíveis fatores de ponderação $G_{m,j}$ através de uma operação matematicamente mais complexa. Para contornar essa dificuldade, antes do cálculo da saliência é realizado um processo chamado branqueamento espectral.

O branqueamento espectral é um processo que busca homogeneizar a energia entre as bandas críticas, de forma que relevância semelhante seja dada a cada uma delas. Como a maior parte da contribuição energética de cada banda é dada pelos picos presentes, o branqueamento espectral termina por diminuir as diferenças entre as amplitudes $G_{m,j}$.

Nesse processo, os coeficientes no domínio da frequência de um filtro branqueador $H_w[k]$ são calculados de forma que sua saída $Y_q[k] = X_q[k]H_w[k]$ apresente conteúdo espectral distribuído de forma aproximadamente uniforme por todo o espectro.

Inicialmente, deve-se calcular a energia do sinal em cada banda crítica:

$$\sigma_b = \sqrt{\frac{1}{K} \sum_{k=1}^{K/2} H_b[k] |X_q[k]|^2}, \quad (4.4)$$

onde K é o comprimento da TDF e $H_b[k]$ foi definido na Seção 2.9.

A seguir, o índice de compressão de banda é calculado como $\gamma_b = \sigma_b^{u-1}$, onde $u = 0,33$ [46] é um parâmetro que determina o quanto a energia em cada uma das bandas críticas será homogeneizada. Trata-se de um índice obtido heurísticamente, sem motivações fisiológicas. Cada parâmetro γ_b refere-se à frequência central de uma banda crítica dentre as analisadas.

Cada um dos coeficientes de $H_w[k]$ é obtido através da interpolação linear dos parâmetros γ_b adjacentes, levando-se em conta que γ_b refere-se à frequência central da banda crítica b e $H_w[k]$ refere-se à frequência $k f_s / K$. O espectro branqueado, $Y_q[k] = X_q[k]H_w[k]$, é passado adiante para as etapas seguintes de processamento.

A título de exemplo, foram sintetizadas 4096 amostras de um sinal gerado pelo mo-

delo $x(t) = 10 + 2 \sin(2\pi t 100) + 5 \sin(2\pi t 400) + 10 \sin(2\pi t 1000) + 20 \sin(2\pi t 2000)$, considerando frequência de amostragem de 44100 Hz, a mesma utilizada em CDs comerciais. O sinal foi ampliado com 4096 zeros complementares após as amostras já existentes, de forma a aumentar o número de pontos disponíveis no domínio da frequência, e o módulo de sua TDF, $|X[k]|$, foi calculado, obtendo-se o resultado mostrado na Figura 4.5(a). Após, o filtro branqueador $H_w[k]$ foi calculado, obtendo-se a resposta em frequência com módulo mostrado na Figura 4.5(b). Pode-se perceber que $H_w[k]$ busca atenuar as bandas críticas que contêm mais energia, ao passo que amplifica as bandas com menos energia. O filtro foi então aplicado a $|X[k]|$, obtendo-se o resultado mostrado na Figura 4.5(c).

É importante realizar algumas análises complementares na Figura 4.5. Inicialmente, é possível verificar que a aplicação do filtro branqueador realizou, de fato, seu trabalho de aproximar as amplitudes das parciais presentes no sinal analisado. Ainda, como $H_w[0] = 0$, a componente DC do sinal foi eliminada.

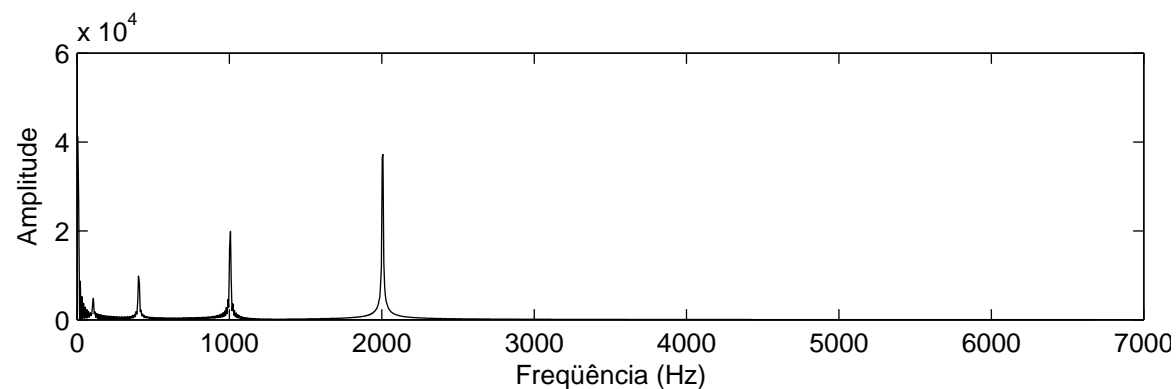
Apesar disso, a Figura 4.5(c) apresenta determinadas distorções que podem vir a prejudicar a análise de sinais mais complexos. Percebe-se, por exemplo, que houve ampliação significativa do sinais de frequência inferior a 400 Hz. Nessa região, supostamente só deveriam haver elementos espúrios decorrentes do processo de janelamento no domínio do tempo. Também, é possível verificar que o formato do pico relacionado à senoide de 2 kHz foi alterado em sua base.

Assim, verifica-se que o branqueamento espectral tende a homogeneizar as amplitudes das parciais que o compõem, enquanto que, ao mesmo tempo, gera distorções no espectro. Embora o branqueamento simplifique a análise no domínio da frequência, seu uso indiscriminado pode distorcer excessivamente o sinal, destruindo as propriedades que se deseja analisar.

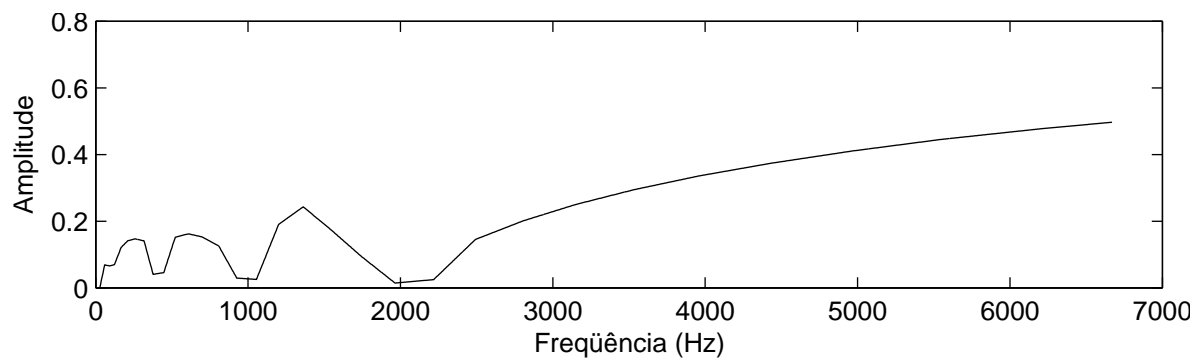
A função saliência

Um sistema capaz de estimar, a partir de $|X_p[k]|$, as J frequências fundamentais f_j presentes é um sistema tomador de decisões. Como tal, aplica alguma função sobre $|X_p[k]|$ que assuma valores altos para frequências fundamentais presentes e valores baixos para frequências fundamentais não presentes. Essa característica é conseguida, no sistema implementado, através do cálculo de uma função chamada saliência.

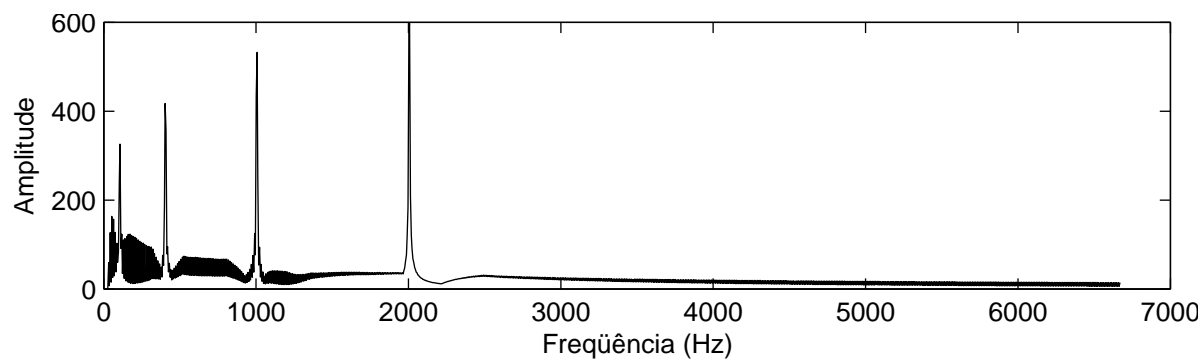
Saliência é uma função calculada a partir de uma frequência f_c candidata a ser uma das frequências fundamentais presentes no quadro analisado. O valor da saliência é uma medida do quão forte é a hipótese de que f_c seja uma das f_j no referido quadro, e deve



(a) Espectro gerado por modelo de sinal com grande variação de energia entre bandas críticas



(b) Filtro branqueador calculado



(c) Espectro branqueado

Figura 4.5: Demonstração de um processo de branqueamento espectral.

ser calculado levando em conta todas as frequências candidatas a serem fundamentais de uma série harmônica. Esse conjunto de frequências candidatas é, em princípio, constituído de todas as frequências fundamentais que podem ser emitidas por um baixo, levando em conta uma resolução arbitrária que pode ser tão fina quanto a diferença entre as frequências relacionadas a dois coeficientes vizinhos da TDF que originou a análise.

A Expressão 4.5 a seguir define a saliência de f_c como a soma das amplitudes das harmônicas de f_c ponderadas por uma função peso dependente tanto da frequência f_c quanto do número m da harmônica em questão:

$$\hat{s}(f_c) = \sum_{m=1}^M h(f_c, m) \max_{k \in k_{f_c, m}} |Y_q[k]|, \quad (4.5)$$

onde o conjunto $k_{f_c, m}$ envolve todos os valores de k para os quais a diferença entre a frequência relacionada a $Y_q[k]$, dada por $k f_s / K$, e a frequência da m -ésima parcial de f_c , dada por $m f_c$, é inferior a meio semi-tom.

A função peso $h(f_c, m)$, obtida em [46], busca maximizar o número de acertos do algoritmo de detecção de múltiplas F0s sobre uma grande base de dados. No processo de obtenção, alguns pontos de referência foram espalhados pelo domínio da função e os pontos restantes foram obtidos por interpolação. Ao fim de um processo de busca exhaustiva, Klapuri obteve um resultado descrito pela expressão:

$$h(f_c, m) = \frac{f_c + \alpha}{m f_c + \beta}, \quad (4.6)$$

onde α e β são parâmetros livres, fixados posteriormente, que controlam o comportamento da função peso no domínio de interesse.

Para valores de α e β fixos, $h(f_c, m)$ cai com o aumento de m , o que significa que o aumento do índice m de uma parcial na série harmônica implica em redução de seu peso para fins de cálculo de saliência. Também se observa aumento da função peso com o aumento de f_c , que significa que frequências fundamentais mais baixas são naturalmente atenuadas em relação ao cálculo da saliência. Na Figura 4.6, o comportamento da função peso para diferentes valores de frequência fundamental candidata e número de harmônica avaliada pode ser visualizado. Assintoticamente, $h(f_c, m)$ se aproxima de $\frac{\alpha}{\beta}$ para valores muito baixos de f_c e de $\frac{1}{m}$ para valores altos de f_c .

Para fins de cálculo, é interessante perceber que a natureza discreta dos cálculos que levam à construção de $\hat{s}(f_c)$ permite que a saliência seja calculada como um sinal discreto $\hat{s}[k]$, onde $k = f_c K / f_s$.

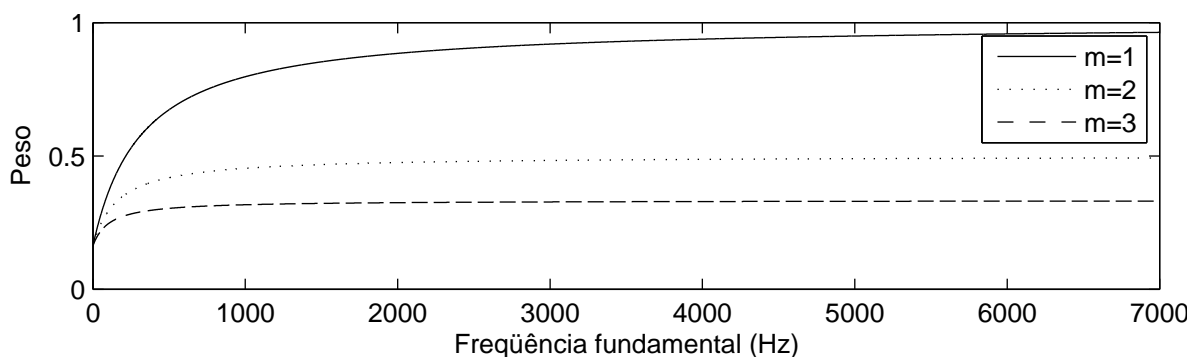


Figura 4.6: Função de ponderação $h(f_c, m)$ considerando m fixado em três diferentes valores e f_c como variável livre.

O caso mais simples de aplicação é o cálculo da função saliência relacionada a um sinal senoidal puro. A Figura 4.7 é uma demonstração dessa aplicação. Nela, pode-se verificar o cálculo da saliência tomando por base a TDF de uma senóide de 300 Hz e amplitude unitária amostrada a 44,1 kHz. Essa TDF foi calculada considerando um quadro de 4096 amostras multiplicado no tempo por uma janela de Hanning e posteriormente expandido no tempo até o dobro de seu tamanho. Como esperado, o módulo da TDF, como mostrado na Figura 4.7, possui de um pico de energia em 300 Hz. A função saliência foi, então, calculada considerando dois diferentes comprimentos de série harmônica. Inicialmente, o cálculo foi realizado considerando uma série com três harmônicas. Pode-se verificar claramente que a função saliência possui picos relacionados às sub-harmônicas de 300 Hz, tendo estes, porém, amplitude menor que o pico relacionado a 300 Hz. Após, o cálculo da saliência foi efetuado novamente, levando em consideração uma série com 20 harmônicas. Nesse caso, o comportamento encontrado para 3 harmônicas é simplesmente expandido para sub-harmônicas ainda mais graves, e, como pode-se verificar, o comportamento desse cálculo é idêntico ao anterior nas frequências superiores a 100 Hz. É importante verificar que, embora o pico da função saliência não esteja tão bem definido quando o pico do módulo da TDF do sinal, ele se encontra em algum ponto dentro do erro de meio semi-tom esperado para a estimativa. Além disso, é possível verificar que a função saliência não possui picos relacionados às harmônicas da senóide analisada.

Cancelamento

Como será visto adiante, embora o pico de maior valor da função saliência corresponda a uma das frequências fundamentais presentes na onda sonora, o pico com se-

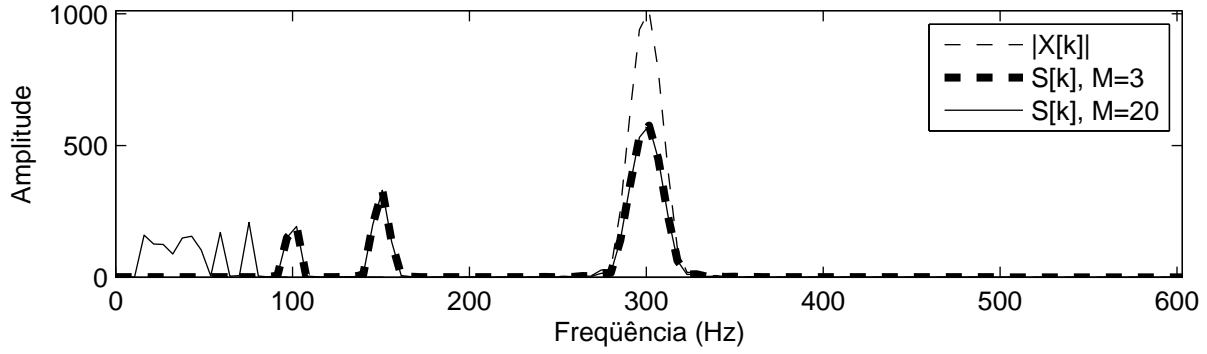


Figura 4.7: Demonstração do cálculo da saliência um sinal senoidal com frequência de 300 Hz amostrado a 44,1 kHz com parâmetros α e β iguais a 52 Hz e 320 Hz, respectivamente.

gundo maior valor usualmente corresponde a uma sub-harmônica dessa mesma frequência, e não a uma outra frequência fundamental presente no quadro.

Por esse motivo, após a estimativa de uma frequência fundamental f_j , é preciso retirá-la do quadro analisado, permitindo novo cálculo da saliência. Isso é realizado através de um processo chamado *cancelamento*. Esse processo, mostrado no Algoritmo 6, se inicia com a estimativa de uma frequência fundamental f_j . O espectro correspondente à série harmônica é calculado posicionando picos com o formato do módulo da TDF da janela (hanning) nas posições correspondentes ao máximo valor de $|X[k]|$ dentro de um alcance de meio semi-tom em relação às frequências $2f_j$, $3f_j$ e assim por diante. A amplitude de cada pico é igual à amplitude da harmônica em questão multiplicada pela função peso dada pela Expressão 4.6. O espectro estimado é, então, subtraído de $X[k]$, e nova estimativa de frequência fundamental é realizada.

Nesse algoritmo, é interessante verificar que $Y_D[k]$ não é inicializado a cada iteração. Assim, cada um dos espectros já cancelados anteriormente é novamente cancelado em cada iteração, até que sejam completamente removidos da mistura. Também, é importante perceber que a ponderação de cada uma das harmônicas adicionadas a $Y_D[k]$ pela função peso da Expressão 4.6 evita que sons harmonicamente relacionados à frequência fundamental estimada, possivelmente provenientes de outras fontes, sejam completamente cancelados, permitindo que sejam detectados posteriormente.

Como exemplo numérico, gravações de um violoncelo tocando a nota D#2 (77Hz) e um trombone tocando a nota G2 (100Hz) foram misturados. Um quadro aleatório da mistura foi fornecido como entrada ao algoritmo de detecção de frequências fundamentais. Inicialmente, o quadro foi multiplicado no domínio do tempo por uma janela de

```

// Inicialização
 $Y_R[k] \leftarrow |X[k]|$  (espectro residual)
 $Y_D[k] \leftarrow 0$  (espectro detectado)
// Passo
para cada frequência fundamental  $j$  a ser estimada faça
     $s[k]$ , a função saliência, é calculada tomando  $Y_R[k]$  por base
     $f_j$ , a  $j$ -ésima frequência fundamental, é estimada
    O espectro correspondente a  $f_j$  é adicionado a  $Y_D[k]$ 
     $Y_R[k] \leftarrow \max(Y_R[k] - dY_D[k], 0)$ 
fim
// Fim
Algoritmo retorna  $j$  frequências fundamentais e  $j$  saliências correspondentes

```

Algoritmo 6: Algoritmo de cancelamento iterativo.

Hanning. Após, foi executado o processo de branqueamento espectral conforme descrito na Seção 4.2.1. O algoritmo de detecção e cancelamento foi, então, executado 4 vezes. Os resultados de cada uma das iterações são mostrados na Figura 4.8.

Na Figura 4.8(a), pode-se verificar que a frequência fundamental de 77Hz foi corretamente detectada como um pico na saliência. Após o cancelamento, porém, a mesma frequência fundamental foi encontrada, como pode-se verificar na Figura 4.8(b). Isso ocorreu porque entre as notas D#2 e G2 há uma relação harmônica, o que significa que algumas parciais coincidem, e, portanto, podem ser canceladas erroneamente. Na Figura 4.8(c), pode-se verificar que, após novo processo de cancelamento, a frequência fundamental de 100Hz foi encontrada corretamente. É interessante verificar como os valores do espectro detectado $Y_D[k]$ sempre aumentam, ao passo que os valores do espectro originalmente analisado e da saliência sempre diminuem. Na Figura 4.8(d), uma frequência fundamental de 199Hz foi encontrada. Nessa frequência, espera-se encontrar a segunda harmônica de G2. Assim, verifica-se que, mesmo que todas as frequências fundamentais já tenham sido canceladas da mistura, o algoritmo encontra, utilizando o conteúdo espectral existente, algum valor de frequência de maior saliência.

4.2.2 O sinal de acento

O acento a_q é um valor associado a um quadro (formando, portanto, um sinal ao longo de todo o arquivo de áudio) que busca determinar numericamente o quanto este mudou, em termos de percepção acústica, em relação ao anterior. Espera-se que o acento de um quadro com conteúdo sonoro percussivo ou transitório (como quadros contendo sinais de tambores ou de troca de notas musicais) seja alto, ao passo que,

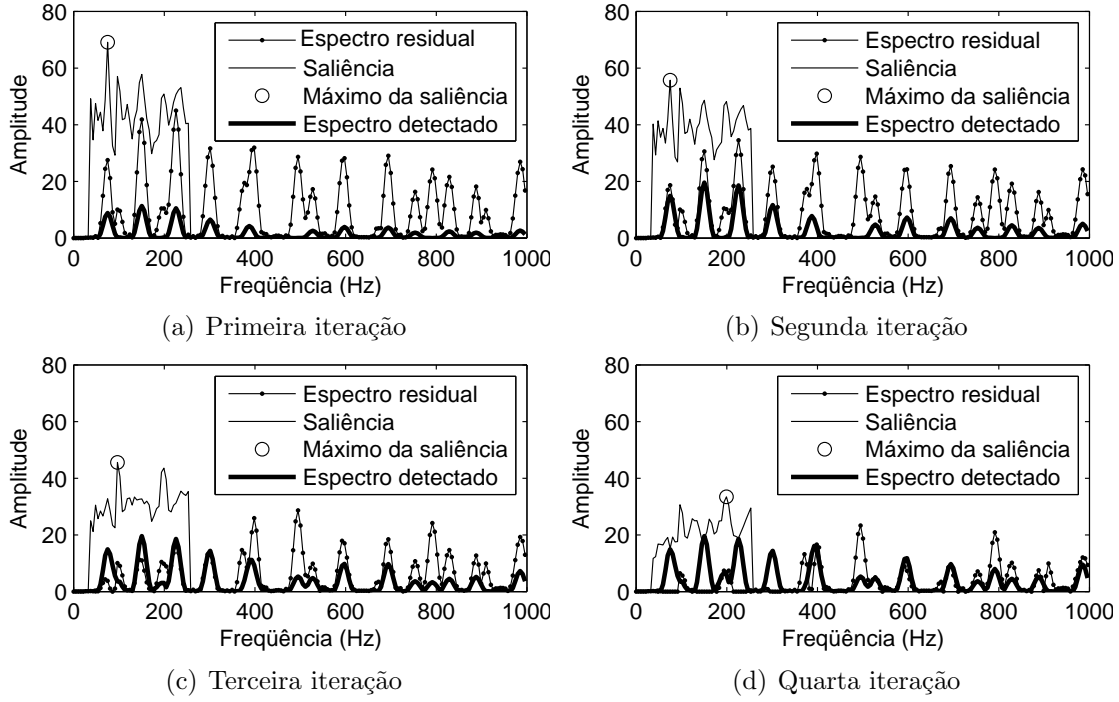


Figura 4.8: Demonstração de um processo de cancelamento iterativo para detecção de múltiplas frequências fundamentais.

em um quadro com conteúdo harmônico estacionário (como a região de sustentação de notas musicais), espera-se obter valores de acento mais baixos.

O processo de cálculo do sinal de acento foi proposto inicialmente em [47] visando a obtenção de uma característica variante no tempo que permita a detecção de ritmos em sinais acústicos discretizados. No contexto da transcrição automática de áudio, tal característica é utilizada em [33] e [34] como referência para a detecção de regiões transitórias - início e final - de notas musicais.

O cálculo do acento é realizado dividindo-se o sinal de entrada $x[n]$ - amostrado a 44,1kHz, já que o método é utilizado para gravações de CDs comerciais - em quadros de 1024 amostras, com superposição de 512 amostras. Em cada quadro $x_q[n]$, realiza-se o cálculo da energia presente em cada banda crítica b através da soma do resultado da aplicação de um filtro FIR no domínio da frequência:

$$X_{q,b} = \sum_{k=0}^K H_b[k] |X_q[k]|^2. \quad (4.7)$$

Neste trabalho, a sugestão em [47] para uso do total de $B = 36$ sub-bandas foi seguido.

Uma vez que a acuidade perceptiva humana para detecção da variação da intensidade sonora é proporcional ao inverso da intensidade sonora em si, $X_{q,b}$ é, então, processado por uma equação de compressão que é logarítmica para valores altos, e gradualmente se aproxima de um comportamento linear para valores baixos, resultando em:

$$Y_{q,b} = \frac{\ln(1 + \mu X_{q,b})}{\ln(1 + \mu)}, \quad (4.8)$$

onde μ é o coeficiente de compressão, estabelecido em 100 como sugerido em [47].

Buscando aumentar a resolução temporal de $Y_{q,b}$, o sinal é sobreamostrado e interpolado por um fator de dois. A sobreamostragem é realizada gerando-se um sinal $Z_{q,b}$ igual a $Y_{q/2,b}$ para valores pares de q e 0 para valores ímpares de q . $Z_{q,b}$ é, então, filtrado por um filtro Butterworth passa-baixas de sexta ordem com frequência de corte de 10 Hz¹. É importante perceber que a aplicação desse filtro leva a um atraso de por volta de 12 amostras, o qual deve ser compensado posteriormente.

Após, o sinal diferencial $Z'_{q,b}$ é calculado através de:

$$Z'_{q,b} = \max(0, (Z_{q,b} - Z_{q-1,b})). \quad (4.9)$$

O vetor $Z'_{q,b}$ contém a informação sobre o quanto $Z_{q,b}$ variou positivamente em relação ao quadro analisado anteriormente. Por fim, a soma ponderada de $Z_{q,b}$ e $Z'_{q,b}$ é calculada:

$$\alpha_q = \sum_{b=0}^{B-1} (1 - \lambda) Z_{q,b} + \lambda (\Delta t) Z'_{q,b}, \quad (4.10)$$

onde $\lambda = 0,8$ é o fator de ponderação obtido heurísticamente [47] e Δt é a diferença, em segundos, entre o início do quadro atual (q) e do anterior ($q - 1$).

Espera-se que o vetor α_q apresente picos em regiões transitórias do sinal de áudio, que, no contexto da transcrição automática de áudio, refletem regiões nas quais notas se iniciam.

Pós-processamento do sinal de acento

No contexto da transcrição automática, o logaritmo do sinal α_q é calculado, gerando os valores de acento a_q que serão utilizados posteriormente [34].

A Figura 4.9 demonstra o sinal de acento frente à onda sonora gerada pela sín-

¹Considerando a frequência de amostragem de $Z_{q,b}$, o filtro tem função de transferência $H(z) = 10^{-5} \frac{0.0414z^{-1} + 0.2486z^{-2} + 0.6214z^{-3} + 0.8286z^{-4} + 0.6214z^{-5} + 0.2486z^{-6} + 0.0414z^{-7}}{1.0000 + 4.5901z^{-1} + 8.9103z^{-2} + 9.3419z^{-3} + 5.5700z^{-4} + 1.7885z^{-5} + 0.2414z^{-6}}$.

tese da melodia da partitura mostrada na Figura 2.2. Pode-se perceber que picos do acento, em geral, coincidem com o início de notas, embora possam existir casos em que essa coincidência não ocorre, como próximo a 0,125 s, em que observa-se um aumento localizado no acento embora nenhuma nota esteja iniciando.

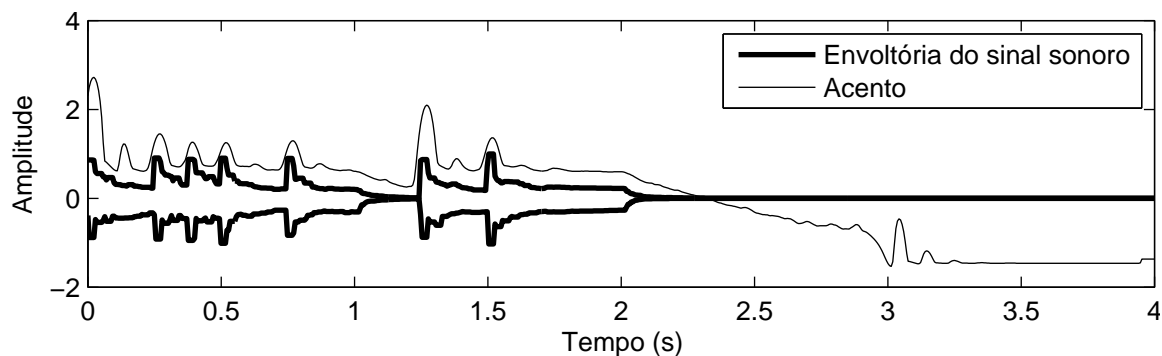


Figura 4.9: Demonstração do sinal de acento

É importante perceber, adicionalmente, que no intervalo de tempo em que é realizada uma estimativa de múltiplas frequências fundamentais são obtidos quatro valores de acento. Isso ocorre porque o intervalo de tempo entre os quadros utilizados para cálculo de múltiplas frequências fundamentais é de 1024 amostras, enquanto que, no cálculo do acento, esse intervalo é de 256 amostras [34, 47]. Por esse motivo, o sinal de acento é sub-amostrado através da seleção do máximo valor a cada quatro amostras.

4.3 Modelos de inferência

O modelo de inferência utilizado neste trabalho segue um conceito que vem sendo aprimorado desde 2004 por Ryyanen [56, 57, 33, 34], o qual consiste na obtenção de um HMM no qual determinados conjuntos de estados são, sabidamente, relacionados a determinadas notas musicais. Esse modelo é construído de forma que a seqüência de estados mais provável, dado um conjunto de observações, seja correspondente à seqüência de notas que gerou essas mesmas observações. Dessa forma, o processo de transcrição automática passa a depender, essencialmente, do algoritmo de Viterbi, já abordado no Capítulo 2.

Nesta seção, o modelo de inferência descrito de forma genérica na Seção 4.1 será explicado em maiores detalhes.

4.3.1 O modelo acústico para notas

Conforme descrito na Seção 4.1, o HMM utilizado para transcrição de áudio corresponde, de maneira aproximada, ao modelo da Figura 4.4. Excluindo-se o estado correspondente ao silêncio, restam, no modelo, seções referentes a cada uma das notas musicais participantes do processo de inferência. Cada uma dessas seções é chamada de *modelo acústico da nota p* , onde p é o número MIDI da nota em questão. O modelo acústico, de forma genérica, é um HMM do tipo esquerda-para-a-direita com três estados, conforme mostra a Figura 4.10.

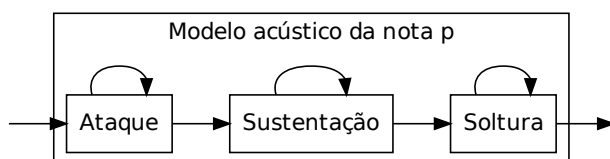


Figura 4.10: Modelo acústico para uma nota genérica com número MIDI p .

A informação da nota p a que cada um dos modelos acústicos se refere é importante, conforme será visto a seguir.

Vetor de observações para cada nota

O vetor de observações, como discutido na Seção 4.1, é composto de três dimensões.

A primeira delas é a diferença de frequência fundamental, Δf , que representa o quão boa foi a estimativa realizada pelo algoritmo de extração de frequências fundamentais abordado na Seção 4.2.1 em relação à nota p . Se, dentre as F0s estimadas, houver alguma f_j tal que $|f_j - p| < 1$ (assumindo a escala MIDI), então Δf será calculado como a diferença entre p e a F0 estimada mais próxima a p . Caso contrário, Δf será calculado como a diferença entre p e a F0 estimada de maior saliência² [34].

É importante perceber que o cálculo de Δf deve ser realizado de forma independente para cada um dos modelos acústicos, uma vez que para cada um deles o valor de p é diferente. Apesar disso, a existência de Δf permite que HMMs idênticos sejam utilizados no modelo acústico de diversas notas, apenas sendo necessária a variação de p para diferenciá-las.

²Por exemplo, se o algoritmo de detecção de múltiplas F0s retorna as F0s (na escala MIDI): 45.1, 50, 60 e 70, com saliências iguais a 50, 40, 30 e 70, respectivamente, Δf para a nota 45 será calculado em relação a 45.1, sendo, portanto, igual a 0.1, mas o mesmo parâmetro em relação à nota 55 será calculado em relação a 70, pelo critério da maior saliência, sendo, portanto, igual a 15.

A segunda dimensão do vetor de observações é o valor da saliência da F0 utilizada para o cálculo de Δf . A terceira dimensão é o valor do acento no quadro analisado. Assim, cada um dos modelos acústicos presentes no HMM tem seu próprio vetor de observações, dado por:

$$\mathbf{o}_{q,p} = [\Delta f_{q,p}, s_{q,p}, a_q]. \quad (4.11)$$

Assume-se que as emissões dos modelos acústicos para notas são independentes para cada dimensão, sendo cada dimensão modelada por uma mistura de quatro gaussianas. Esse número é, porém, ajustado pelo algoritmo de treino que elimina as gaussianas desnecessárias da mistura.

4.3.2 O modelo acústico para silêncio

O silêncio, também importante no contexto da transcrição automática, se relaciona a quadros em que o baixo não está soando. É modelado como um estado simples cujo vetor de observações é composto simplesmente do máximo valor de saliência no quadro. A emissão relacionada ao silêncio é modelada, também, como uma mistura de quatro gaussianas.

4.3.3 Treinamento de modelos acústicos

O processo de treinamento para modelos acústicos é bastante importante. Para realizar o treinamento, é preciso dispor de uma base de dados contendo arquivos de músicas digitalizadas e anotações correspondentes indicando o tempo de início, tempo de final e a altura p da nota tocada. Para cada uma das notas marcadas na anotação, é calculado o vetor de observações considerando o modelo acústico referente à altura anotada manualmente. Assim, espera-se obter um certo número de seqüências de observações. Devido ao uso do parâmetro Δf_q , todas as seqüências de observações são utilizadas em conjunto para treinar os parâmetros do modelo acústico. Ryyanen [34] propõe, ainda, que as seqüências cuja mediana do módulo de Δf for superior a 1 sejam descartadas. Dessa forma, evita-se que seqüências especialmente difíceis sejam usadas no treinamento. É importante evitá-las porque a existência de dados incorretos pode prejudicar o processo de otimização dos parâmetros.

Para cada seqüência de observações, só é possível dizer com certeza que a primeira observação corresponde a um estado de ataque e a última observação corresponde a um estado de soltura. Essa informação alimentará o algoritmo de Baum-Welch, conforme descrito na Seção 3.2.5.

Informações sobre a aplicação do algoritmo no trabalho de Ryyanen não estão disponíveis, de forma que foram necessários alguns testes empíricos na tentativa de obter boas configurações. Esses testes serão abordados posteriormente.

4.3.4 O modelo musical

Cada nota musical que o HMM é capaz de transcrever deve ter seu próprio modelo acústico. Assim, modelos acústicos para cada nota musical a ser transcrita, mais o silêncio, devem ser unidos seguindo o procedimento descrito na Seção 3.2.2. Para tal, é necessário definir uma matriz de transições entre modelos chamada de modelo musical. O modelo musical busca modelar o fato de que, em músicas reais, algumas seqüências de notas são mais prováveis que outras [42].

Ryyanen [34] utiliza três versões de modelos musicais. O primeiro deles é o modelo uniforme, ou inativo, que garante probabilidades de transição uniforme entre as notas. O segundo é um modelo apresentado previamente em [58], no qual probabilidades de transição são calculadas considerando diversas peças musicais clássicas. O terceiro é um modelo obtido por Ryyanen [34] através da análise de um vasto conjunto de arquivos MIDI. Neste trabalho, foram utilizados modelos musicais artificiais que atribuem probabilidade alta de transição para intervalos de oitava, quarta ou quinta, e probabilidade baixa de transição para outros intervalos.

Após a construção do modelo, torna-se evidente a semelhança entre o HMM para transcrição musical aqui implementado e os modelos HMM utilizados para transcrição de fala, uma vez que os modelos acústicos correspondem a modelos fonéticos e o modelo musical corresponde a um modelo linguístico [56].

4.3.5 O modelo completo

O HMM utilizado neste trabalho é constituído de modelos acústicos para as notas com número MIDI de 26 a 59, correspondentes à tessitura do baixo. Esse alcance implica no uso de 34 modelos acústicos para notas (cada um deles com três estados), mais um para o silêncio, num total de 103 estados ($34 \times 3 + 1$) que se organizam como uma expansão do modelo mostrado na Figura 4.4.

4.3.6 Limite superior de F0

Através de um cálculo heurístico, Ryyanen [34] propõe que, para cada quadro, seja inferido um limite superior de frequência fundamental permitida. Esse limite é calcu-

lado, inicialmente, pela média das quatro frequências fundamentais calculadas ponderada por suas respectivas saliências:

$$x_q = \frac{\sum_{i=1}^4 f_{q,i} s_{q,i}}{\sum_{i=1}^4 s_{q,i}}. \quad (4.12)$$

Após, x_q é filtrado por um processo de ataque-soltura, obtendo y_q através das Expressões 4.13 a 4.15:

$$\tau = \begin{cases} 0, 14 & \text{se } x_q > y_{q-1} \\ 2, 3 & \text{caso contrário} \end{cases} \quad (4.13)$$

$$g = \exp(-1/\tau f_r), \quad (4.14)$$

onde f_r é a frequência de quadros, igual a $\frac{1024}{44100}$, e

$$\begin{aligned} y_1 &= 59 \\ y_q &= (1 - g)x_q + gy_{q-1} \end{aligned} \quad (4.15)$$

Por fim, o limite superior é normalizado como $\lceil y_q - c \rceil$, onde c foi empiricamente definido como 4. É importante ressaltar que os valores de saliência são utilizados anteriormente a qualquer normalização.

Determinar o limite superior de F0 em cada quadro significa que a probabilidade de uma nota com F0 superior a esse limite ocorrer naquele quadro é nula. Dessa forma, busca-se não só reduzir o espaço de busca de notas, mas também melhorar o desempenho do transcritor, como será visto a seguir.

4.4 Treinamento

Após exaustivos testes com diferentes configurações, foram determinados os parâmetros de treinamento que levaram aos melhores resultados. Tais testes se fizeram necessários porque os parâmetros utilizados originalmente não foram disponibilizados. Os parâmetros mostraram ser de grande importância para o funcionamento correto do transcritor, uma vez que pequenas variações podem levar a grande diferença de desempenho no processo transcritivo.

No treinamento de gaussianas, foi determinado que gaussianas cuja variância ou fator de ponderação se tornem inferiores a limites pré-estabelecidos deveriam ser eliminadas da mistura, desde que essa operação não deixe a mistura vazia, seguindo a heurística utilizada em [52]. Também, foi necessário fixar o número de iterações do treino,

definindo uma condição de parada. Por fim, o estágio de inicialização foi definido de forma que um certo comportamento do modelo acústico fosse suposto, a princípio, de forma que o algoritmo *k-means* pudesse ser utilizado para inicializar gaussianas, conforme visto na Seção 3.1.4. Nesse estágio, um certo número de observações iniciais (um parâmetro chamado de tamanho inicial) é atribuído forçadamente ao estado de ataque, um certo número de observações finais é atribuído ao estado de soltura e o restante é atribuído ao estado de condução. Os valores para esses parâmetros são mostrados na Tabela 4.1.

Tabela 4.1: Melhores parâmetros para treinamento

Parâmetro	Valor
Menor variância permitida	10^{-2}
Menor fator de ponderação permitido	10^{-1}
Número de iterações de treino	30
Tamanho inicial do ataque	1
Tamanho inicial da soltura	1

4.4.1 Estratégia de treino

O treinamento do HMM relacionado à transcrição seguiu a estratégia explicada por Ryyanen em conversas por e-mail. Inicialmente, conforme descrito em [34], a base de dados RWC-POP MUSIC DATABASE [59] foi adquirida. Essa base de dados é composta de arquivos de áudio acústico e de arquivos MIDI correspondentes. Como visto na Seção 2.5, a notação MIDI tem uma correspondência imediata com a localização de notas musicais no domínio do tempo. O programa Sonic Visualizer [60] foi utilizado para sincronizar os arquivos MIDI com as gravações acústicas presentes na base de dados, de forma a obter um conjunto de transcrições de referência nos quais é possível dizer, em princípio, em que instante de tempo cada nota começa e termina. Em alguns casos, porém, foi impossível obter sincronia satisfatória entre o arquivo MIDI e a gravação acústica. Trechos de 68 peças foram utilizadas, totalizando 45 minutos de música. Verificou-se, durante as etapas de testes, que para cada peça deveria ser selecionado um trecho musical com comportamento mais uniforme - em especial, excluindo-se o início, o final e eventuais interlúdios - e, claro, que contenha notas do baixo.

O treinamento seguiu um esquema de validação cruzada, como descrito a seguir. A base de dados foi dividida em dois conjuntos de igual tamanho, um para treino e um

para validação. As notas musicais do conjunto de treino cujo módulo da mediana do parâmetro Δf calculado em relação à frequência nominal (aquela apontada pelo arquivo MIDI) não excede 1 foram utilizadas para treinamento do modelo acústico de notas, conforme descrito na Seção 3.2.5, utilizando os parâmetros da Tabela 4.1. O modelo obtido é, então, executado sobre o conjunto de validação. Após, os conjuntos de teste e validação são invertidos e todo o processo é realizado novamente. Os resultados das duas execuções são unidos em um só conjunto para posterior análise.

4.5 Resultados

Os resultados de um transcritor automático dependem da comparação entre a transcrição automática, inferida pelo sistema de transcrição, e uma transcrição de referência, a qual se assume como verdadeira. Nesse trabalho, a transcrição de referência é o arquivo MIDI existente na base de dados [59] e a transcrição automática é a saída do transcritor automático. Para cada nota na transcrição de referência, um algoritmo [61] encontra a nota da transcrição automática que melhor a descreve, num pareamento um-a-um. Conforme sugerido em [34], uma nota é considerada transcrita corretamente se, além de sua altura (na escala MIDI) ter sido estimada corretamente, a diferença entre os instantes de início da nota presentes na transcrição automática e na referência não exceder 150 ms (a escolha desse limite será explicada mais adiante).

Para cada música, três índices são calculados. Inicialmente, calcula-se a revocação (*recall*). Conforme a Expressão 4.16, trata-se da fração entre o número de notas corretamente transcritas e o número de notas da transcrição de referência. Um índice alto de revocação significa que o transcritor encontrou muitas das notas que estão presentes no gabarito.

$$\text{revocação} = \frac{\text{Notas corretamente transcritas}}{\text{Total de notas da referência}}. \quad (4.16)$$

Após, calcula-se a precisão (*precision*). Como mostra a Expressão 4.17, trata-se da fração de notas que foram estimadas corretamente pelo transcritor. Um alto índice de precisão significa que muitas das notas estimadas pelo transcritor estão corretas.

$$\text{precisão} = \frac{\text{Notas corretamente transcritas}}{\text{Total de notas inferidas pelo transcritor}} \quad (4.17)$$

Por fim, calcula-se o fator F pela Expressão 4.18. Trata-se de uma ponderação entre revocação e precisão, utilizada como medida de qualidade para sistemas de transcrição

automática.

$$F = \frac{2 \times \text{Revocação} \times \text{Precisão}}{\text{Revocação} + \text{Precisão}} \quad (4.18)$$

A média desses índices sobre todas as músicas é calculada a cada experimento, conforme o processo realizado em [34], e mostrada na saída. Assim, se um determinado experimento teve um fator F igual a 50%, este diz respeito à média dos fatores F s calculados para cada uma das músicas nesse experimento.

Durante o processo de validação, foram considerados três experimentos diferentes. No primeiro deles, foram utilizados simplesmente os modelos acústicos. No segundo, a heurística para determinação de limite superior de F0 foi adicionada ao sistema. Por fim, um modelo musical foi ativado. Os resultados obtidos estão mostrados na Tabela 4.2.

Tabela 4.2: Resultados com implementação própria.

Experimento	Revocação	Precisão	F
Modelo acústico	49,40%	50,56%	48,90%
Lim. sup. F0	51,23%	55,33%	52,09%
Modelo musical	51,79%	54,96%	52,29%

É importante, nesse momento, comparar os resultados obtidos através da implementação própria com os publicados originalmente

Para comparar os resultados, foi preciso calcular, também, o desvio padrão dos resultados mostrados na Tabela 4.2. Os desvios padrões estão mostrados na Tabela 4.3.

Tabela 4.3: Desvio padrão dos resultados obtidos por implementação própria.

Experimento	Revocação	Precisão	F
Modelo acústico	19,93%	19,77%	18,80%
Lim. sup. F0	20,37%	21,57%	20,04%
Modelo musical	19,83%	21,70%	19,87%

Como se pode observar, o valor do desvio padrão é significativo. Isso ocorre porque a transcrição funciona melhor em alguns trechos musicais que em outros, especialmente aqueles com comportamento mais uniforme.

Comparando os resultados obtidos para o modelo acústico, verifica-se uma diferença de desempenho de apenas 2% quando se considera o fator F . Trata-se de uma diferença

aceitável, dada a variação da base de dados. Em ambos os casos, obteve-se ganho considerável ao incorporar-se o limite superior de F0. O ganho, porém, foi pequeno em ambos os casos quando se considera a incorporação do modelo musical.

Uma vez que nem a descrição do HMM obtido originalmente por Ryyanen, nem os parâmetros utilizados para seu treinamento, foram disponibilizados, torna-se difícil a reprodução exata dos resultados divulgados em seu trabalho [34]. Dessa forma, a implementação realizada foi considerada validada. Os resultados obtidos nas modificações realizadas a seguir serão comparados com os da Tabela 4.2, verificando-se os ganhos e perdas relativos de cada uma das propostas.

Capítulo 5

Melhoria da resolução em frequência usando predição linear

Como foi visto nos capítulos anteriores, a detecção de múltiplas frequências fundamentais é limitada em resolução, no domínio da frequência, pela resolução oferecida pela Transformada Discreta de Fourier. O aumento da resolução da análise no domínio da frequência implica em maior tempo de observação, e, portanto, na diminuição da resolução dessa mesma análise no domínio do tempo. Neste capítulo, serão abordadas técnicas de predição que permitem expandir a resolução da análise no domínio da frequência considerando tempo de observação constante.

Inicialmente, uma visão geral do conceito de predição será abordada. Após, será discutida a predição linear no domínio do tempo. Por fim, será demonstrado como as técnicas de predição podem melhorar o desempenho do algoritmo de estimação de múltiplas frequências fundamentais abordado na Seção 4.2.1.

5.1 Preditores

Uma série temporal é uma seqüência de números. Uma série pode ser usada para representar diversos acontecimentos do mundo real, como a altura de uma criança a cada dia, a intensidade do vento em uma determinada região ou o número de pessoas em uma fila de banco.

Estudos sobre séries temporais são especialmente importantes se for levado em consideração o teorema de Nyquist [36]. Conforme visto na Seção 2.7, para um sinal analógico limitado em banda $x(t)$ existe uma dada frequência de amostragem f_s para o qual o processo de discretização $x[n] = x(n/f_s)$ ocorre sem distorção. Isso significa que toda

a informação contida no sinal analógico $x(t)$ está, também, contida no sinal discreto $x[n]$, de forma que a predição de novas amostras do sinal discreto $x[n]$ é equivalente à predição do comportamento do sinal contínuo $x(t)$.

A habilidade de prever valores de $x[n]$ está, portanto, diretamente ligada à habilidade de prever acontecimentos reais. Assim, desenvolvem-se teorias tanto para prever a altura de crianças quanto a intensidade do vento em uma determinada região e o número de pessoas em filas de banco, de forma que as pessoas envolvidas com tais questões possam tomar, antecipadamente, as providências cabíveis em cada caso.

Quando um preditor é aplicado sobre uma série temporal, assume que $x[n]$ obedece a algum tipo de modelo de comportamento e, a partir das N amostras disponíveis ($x[0] \dots x[N-1]$) retorna um valor estimado $\hat{x}[N]$ coerente com esse modelo. Matematicamente, esse processo pode ser enunciado como:

$$\hat{x}[N] = f(x[0], x[1], \dots, x[N-1]). \quad (5.1)$$

A função de predição $f(\cdot)$ é definida caso a caso, levando em consideração as características específicas da série temporal $x[n]$ e outros conhecimentos especialistas que possam refinar ainda mais o modelo.

5.2 Predição linear no domínio do tempo

Um preditor linear é, por definição, um sistema que estima valores futuros de uma série temporal a partir de uma combinação linear dos últimos $M + 1$ valores passados dessa mesma série e das últimas L estimativas realizadas [62]. Assim, a saída $\hat{x}[n]$ de um preditor linear num dado instante de tempo é dada por:

$$\hat{x}[n] = G \sum_{m=0}^M v_m x[n-m] - \sum_{l=1}^L u_l \hat{x}[n-l], \quad (5.2)$$

onde os parâmetros v_m e u_l são os coeficientes de ponderação e G é o ganho do sistema.

Define-se o erro de predição da amostra n como a diferença entre o valor verdadeiro da série temporal e o valor estimado pelo preditor:

$$\epsilon_n = x[n] - \hat{x}[n]. \quad (5.3)$$

Ao se tomar a Transformada Z da Expressão 5.2, o sistema passa a ser escrito como:

$$\hat{X}(z) = G \frac{\sum_{m=0}^M v_m z^{-m}}{1 + \sum_{l=1}^L u_l z^{-l}}. \quad (5.4)$$

A análise de $\hat{X}(z)$ será de grande utilidade posteriormente, por permitir análise de resposta em frequência e estabilidade do preditor a partir dos coeficientes v_m e u_l [36].

5.2.1 Síntese aditiva e preditores lineares

Sinais acústicos, em especial os emitidos por instrumentos musicais, podem ser analisados como a saída de um sistema físico (o corpo do instrumento) frente a uma excitação recebida como entrada (a atuação do músico). Um dos possíveis modelos matemáticos para um sinal acústico proveniente de um instrumento musical é a soma ponderada de senóides harmonicamente relacionadas, que pode ser calculado no domínio discreto, amostra a amostra, tal qual descrito na Expressão 2.4, assumindo-se valores pré-determinados para a frequência fundamental da série harmônica e para a amplitude e a fase de cada uma das componentes senoidais. Esse modelo de síntese é chamado de *síntese aditiva* [38].

É uma propriedade bem conhecida que um sistema que sintetiza uma senóide de frequência arbitrária f/f_s , onde f é a frequência da senóide e f_s é sua frequência de amostragem, possui, no domínio da Transformada Z, dois pólos complexo-conjugados correspondentes com módulo unitário e argumento $\pm 2\pi f/f_s$ [36]. Assim, um sistema que sintetiza duas senóides tem quatro pólos e assim por diante. Portanto, é possível modelar um sistema de síntese de sons razoavelmente complexos através da alocação adequada de pólos para a resposta ao impulso desse mesmo sistema, de forma que um sistema de síntese aditiva pode ser modelado, no domínio da Transformada Z, por:

$$X_h(z) = \frac{1}{\prod_{l=1}^L g_l(z - e^{j\pi l f_0/f_s})(z - e^{-j\pi l f_0/f_s})}, \quad (5.5)$$

onde f_0 é a frequência fundamental (em Hz) da série harmônica e os coeficientes g_l são as amplitudes das senóides presentes na mistura.

O modelo da Expressão 5.5 corresponde a um sistema sem zeros fora da origem do plano Z e com pólos distribuídos por todo o plano Z. Por esse motivo, é chamado de modelo inteiramente de pólos (*all-pole*). Quando a multiplicação do denominador é efetuada, obtém-se um polinômio em z^{-1} de forma que a expressão pode ser re-escrita

como:

$$X_h(z) = g \frac{z^{-L}}{1 + \sum_{l=1}^L u_l z^{-l}}, \quad (5.6)$$

onde g é um fator de ganho.

Essa nova forma imediatamente remete a um caso especial da Expressão 5.4 no qual encontramos o ganho $G = 1$ e todos os coeficientes v_m nulos, exceto por $v_0 = 1$. Sem perda de generalidade, os zeros de $X_h(z)$ podem ser ignorados, de forma que a expressão no domínio do tempo para a obtenção de um novo valor do sinal sintetizado é:

$$x_h[n] = - \sum_{l=1}^L u_l x_h[n-l]. \quad (5.7)$$

Assim, o valor de um sinal harmônico discretizado passa a ser escrito como uma combinação linear de valores passados. Nessa forma, torna-se evidente a motivação que levou à nomenclatura *Modelo Auto-regressivo* - ou simplesmente Modelo AR - para esse tipo de modelo¹.

Esse resultado pode ser expandido para um sinal correspondente à soma de sinais harmônicos, conforme a Expressão 4.2. Embora seja possível percorrer todo o raciocínio descrito considerando uma soma arbitrária de harmônicas, é interessante perceber que, se um sinal harmônico pode ser obtido a partir de um sistema linear, então a soma de sinais harmônicos pode, também, ser obtida por sistemas lineares convenientemente acoplados - que darão origem, assim, a um sistema linear único capaz de sintetizar um sinal polifônico.

5.2.2 Predição para expansão de quadros de áudio

O objetivo da aplicação de predição linear, neste trabalho, é obter um modelo AR para o sinal contido em um quadro de forma que, ao se analisar a resposta ao impulso desse modelo AR, uma estimativa espectral de maior resolução possa ser obtida, como discutido em [62].

Dessa forma, resta apenas determinar os coeficientes u_l que definirão o modelo AR de predição. É interessante ressaltar que, quando leva-se em conta que os coeficientes u_l são obtidos como função do sinal de entrada $x[n]$, então pode-se considerar que o sinal estimado $\hat{x}[n]$ é uma função desse mesmo sinal, assumindo assim a forma de preditor

¹“Modelos *all-pole*” e “modelo AR” são nomenclaturas que se referem ao mesmo tipo de sistema, embora geralmente sejam usadas em contextos diferentes, dando mais ênfase ao comportamento do modelo no domínio da Transformada Z ou no domínio do tempo, respectivamente

proposta na Expressão 5.1.

5.2.3 Estimativa de modelos AR

A estimativa de modelos AR a partir de N amostras de uma série temporal $x[n]$ é um processo bem conhecido [63], que parte da definição do Erro Quadrático Total:

$$\varepsilon = \sum_n \epsilon_n^2 = \sum_n (x[n] + \sum_{l=1}^L u_l x[n-l])^2, \quad (5.8)$$

onde os limites da somatória serão discutidos posteriormente.

A partir dessa definição, são definidos os coeficientes u_l que minimizam ε . Para isso, a Expressão 5.8 é derivada em relação a u_l e então igualada a zero:

$$\frac{\partial \varepsilon}{\partial u_l} = 0. \quad (5.9)$$

Após alguns passos algébricos, encontra-se a equação normal:

$$\sum_{l=1}^L u_l \sum_n x[n-l]x[n-i] = - \sum_n x[n]x[n-i], \quad (5.10)$$

que deve ser calculada para todo $i \in [1, L]$, levando-se em conta que $x[n] = 0$ para $n < 0$ e $n > N - 1$.

Obter uma solução direta para essa equação através da fatoração de Gauss-Jordan é um processo computacionalmente custoso. Isso motivou o desenvolvimento da Recursão de Levinson-Durbin [63], mostrada no Algoritmo 7. O algoritmo recebe como entrada a autocorrelação do sinal $x[n]$, o número de amostras disponíveis N e a ordem do preditor a ser estimado L . É importante lembrar que os valores de autocorrelação $r_{xx}[k]$ utilizados no algoritmo são estimados pela expressão:

$$r_{xx}[k] = \sum_{n=0}^{N-k-1} x[n]x[n+k], \quad (5.11)$$

levando-se em conta, assim, a limitação de existência do sinal $x[n]$ para $0 \leq n \leq N - 1$.

A Recursão de Levinson-Durbin, na tentativa de estimar o melhor preditor de ordem L para as entradas recebidas, estima também todos os estimadores de ordem inferior a L . A cada passo, a ordem do estimador é aumentada e os novos coeficientes são calculados baseando-se nos cálculos anteriores.

```

// Inicialização
 $\varepsilon^{(0)} \leftarrow r_{xx}[0]$ 
// Passo
para cada  $i$  de 1 a  $L$  faça
     $k_i \leftarrow -r_{xx}[i] + \sum_{j=1}^{i-1} u_l^{(i-1)} r_{xx}[i-j] / \varepsilon^{(i-1)}$ 
     $u_i^{(i)} \leftarrow k_i$ 
    para cada  $j$  de 1 a  $i-1$  faça
         $u_j^{(i)} \leftarrow u_j^{(i-1)} + k_i u_{i-j}^{(i-1)}$ 
    fim
     $\varepsilon^{(i)} \leftarrow (1 - k_i^2) \varepsilon^{(i-1)}$ 
fim
// Fim
Solução: coeficientes  $u_l^{(L)}$ 

```

Algoritmo 7: Recursão de Levinson-Durbin.

É um resultado conhecido que o limite superior para o erro quadrático total de um preditor linear de ordem L é o erro quadrático total do preditor de ordem $L-1$, considerando o mesmo sinal de entrada [63]:

$$\varepsilon^{(L)} \leq \varepsilon^{(L-1)}, \forall L < N. \quad (5.12)$$

Como pode ser verificado, a ordem máxima de preditor que pode ser calculada solucionando-se as equações normais dadas pela Expressão 5.10 é igual ao número de amostras disponíveis menos um. O uso da Recursão de Levinson-Durbin permite calcular um preditor dessa ordem de forma eficiente.

5.2.4 A ordem do preditor

O preditor linear obtido pela solução da Expressão 5.10 é um modelo AR que busca ter uma resposta em frequência tão próxima quanto possível da Transformada de Fourier do sinal $x[n]$ utilizado como base para sua construção [62], de forma que espera-se que os pólos do modelo sejam posicionados em frequências onde haja grande concentração de energia. Isso significa que as tarefas de síntese e de análise espectral se confundem, já que o modelo AR que melhor sintetiza um determinado sinal no domínio do tempo é também o modelo AR que melhor descreve esse mesmo sinal no domínio da frequência.

O número de pólos do preditor está, portanto, ligado ao número de senóides necessário para descrever $x[n]$. Cada par de pólos complexo-conjugados é capaz de modelar

corretamente uma senóide - assim, um sinal harmônico composto de 20 senóides pode ser descrito por um modelo AR de 40 pólos.

Quando se considera a presença de ruído, a necessidade por pólos aumenta significativamente. A solução da Expressão 5.10, em sinais ruidosos, leva à obtenção de um modelo AR no qual alguns pólos relacionam-se não às senóides que compõem o sinal, mas às senóides que melhor modelam o ruído. Dessa forma, o aumento da ordem do modelo AR não necessariamente implica na melhoria do modelo de sinal, embora, de fato, o erro quadrático total diminua conforme a Expressão 5.12.

Além disso, pode-se verificar que o aumento da ordem do modelo AR implica na utilização de mais amostras da autocorrelação. Analisando-se a expressão 5.11, pode-se verificar que $r_{xx}[k]$, com o aumento de k , é calculado levando-se em consideração menos amostras reais de $x[n]$. Dessa forma, a estimativa de $r_{xx}[k]$ degrada com o aumento de k , e, portanto, a qualidade dos coeficientes obtidos decai.

Isso significa que o aumento da ordem do modelo AR permite a obtenção de sinais mais complexos (em relação ao número de senóides que o compõe), embora isso implique na degradação da qualidade da estimativa. Encontramos, assim, um compromisso de acordo com o qual um número de pólos muito pequeno não é capaz, matematicamente, de modelar o sinal, ao passo que um número excessivo de pólos modela o sinal de forma inadequada.

Na maior parte das aplicações, a melhor ordem possível do preditor linear é um parâmetro desconhecido [62], e, portanto, deve ser determinado de forma empírica. Em trabalhos anteriores, já foram utilizados preditores de ordem 12 para predição de voz [64], ordem 18 para síntese de violoncelo, violino e viola [65], ordem entre 20 e 50 para síntese de instrumentos de forma geral [66]. Em todos esses trabalhos, foi realizada uma investigação extensiva sobre o melhor número de pólos para obter resultados específicos.

É necessário perceber que o modelo AR obtido pela aplicação da recursão de Levinson-Durbin sobre um pequeno quadro de áudio busca reproduzir não uma simples soma de senóides, mas uma soma de senóides multiplicadas por uma janela no domínio do tempo. Dessa forma, pode ser necessário o uso de um número bastante elevado de pólos, capaz de modelar um sinal que é nulo para todas as amostras tais que $n < 0$ ou $n > N - 1$ - o que envolve não só pólos que modelam o comportamento harmônico do sinal, mas também pólos que modelam a multiplicação por uma janela de duração finita. Neste trabalho, simulações mostraram ser necessário um número bastante elevado de pólos para obter resultados satisfatórios na estimação de frequências.

Como demonstração, um intervalo de 93 ms de um sinal composto da soma de

sinusóides de amplitude $\frac{1}{2}$ e frequências iguais a 50, 95, 100 e 105 Hz foi sintetizado e amostrado a 44100 Hz. Três formas de análise foram aplicadas a esse sinal, como pode ser visto na Figura 5.1. A primeira forma foi a TDF expandida com zeros, na qual o sinal foi multiplicado por uma janela de Hanning e expandido com zeros até dez vezes o seu comprimento original. A seguir, foram obtidos dois modelos AR para esse mesmo sinal, de ordem 2000 e 4000, gerando duas respostas em frequência que também foram utilizadas para analisar o sinal.

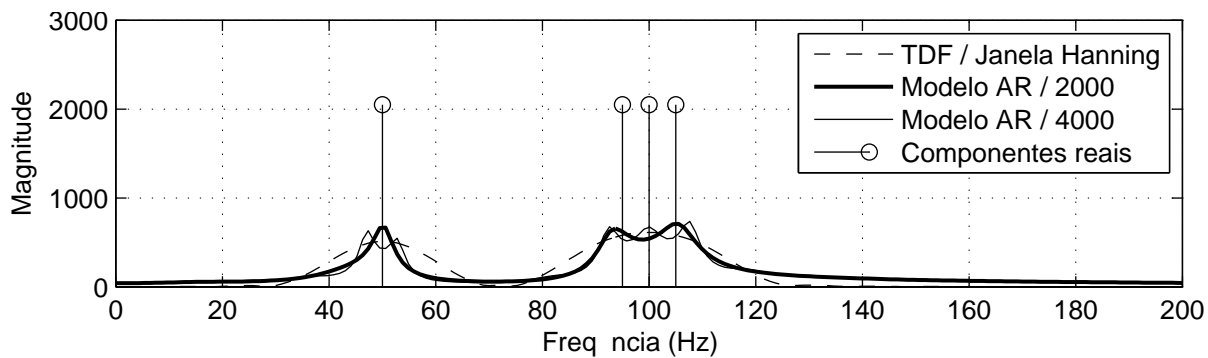


Figura 5.1: Comparação de diferentes formas de análise de um mesmo sinal: TDF do sinal multiplicado por janela Hanning e resposta em frequência de modelos AR de ordem 2000 e 4000. Impulsos correspondentes às componentes reais do sinal são mostradas.

Inicialmente, é possível perceber que os dois preditores levam à geração de picos mais agudos que a TDF correspondentes às componentes senoidais desejadas. Também, na região próxima a 50 Hz, verifica-se que o preditor de ordem 2000 gerou uma resposta em frequência com um único pico sobre a frequência esperada para a senoide, enquanto o preditor de ordem 4000 gerou uma resposta em frequência com dois picos. Na região próxima a 100 Hz, é importante perceber que, embora o modelo AR de ordem 4000 tenha sido capaz de modelar corretamente os três picos esperados, os dois picos apresentados pelo modelo de ordem 2000 são mais próximos dos valores reais.

Assim, a ordem de modelo escolhida para aplicação num caso real deverá ser por volta de 2000, embora esse valor dependa de testes empíricos posteriores.

5.3 Detecção de múltiplas frequências fundamentais usando predição linear

Como já mostrado ao longo deste capítulo, técnicas de predição linear podem ser utilizadas para o aumento da resolução espectral para análise de um determinado si-

nal. Nesta seção, serão mostrados experimentos que demonstram as vantagens desse aumento de resolução para a tarefa de extração de múltiplas frequências fundamentais em sinais musicais.

5.3.1 Modificações propostas ao método original

Tomando por base o método proposto por Klapuri [46], tal qual descrito na Seção 4.2.1, diferentes versões do estimador de múltiplas frequências fundamentais foram testadas. O método original consiste, basicamente, em tomar o módulo da TDF de um determinado sinal de N amostras previamente multiplicado por uma janela tipo Hanning e expandido com zeros até duas vezes seu tamanho original, realizar um processo de branqueamento que busca equalizar a energia ao longo de todo o espectro e realizar uma soma ponderada de harmônicas para cada frequência fundamental candidata, a qual indica um fator chamado saliência. A F0 de maior saliência é considerada como parte do sinal, tem seu espectro estimado através da soma ponderada da transformada da janela utilizada para o cálculo da TDF deslocada de forma a coincidir com as harmônicas estimadas. Esse espectro estimado é removido da mistura, dando início a uma nova iteração de cálculo de saliência.

As modificações propostas ao método giram em torno da substituição da estimativa espectral através da TDF pela estimativa derivada da resposta ao impulso de um modelo AR gerado a partir do sinal analisado.

Substituir a TDF por algum outro método de análise de alta resolução implica em obter um sinal, no domínio da frequência, com mais amostras, o que significa maior custo computacional nos processos de branqueamento e cancelamento inerentes ao método. Testes preliminares mostraram que uma representação espectral $X[k]$ com cinco vezes mais amostras que o sinal $x[n]$ original produz uma melhora significativa nos resultados, mantendo um tempo computacional aceitável para a aplicação.

Além das diferentes ordens de modelos AR, também foi testada a estimativa espectral baseada na TDF do sinal, expandido com zeros até cinco vezes o seu comprimento original. Esse teste foi realizado para verificar qual foi o ganho de desempenho trazido pelo uso do modelo AR e qual foi o ganho trazido pelo simples aumento do número de pontos disponíveis para operações matemáticas.

É importante perceber que o processo de cancelamento iterativo relacionado à detecção de múltiplas frequências fundamentais depende da estimativa do formato de um pico espectral, como já discutido na Seção 4.2.1. O formato de um pico espectral relacionado a uma parcial num modelo AR depende da posição dos pólos desse mesmo

modelo, mas, para modelos AR de ordem elevada, o alto custo computacional torna impraticável o cálculo da localização dos pólos, de forma que duas alternativas foram empiricamente analisadas. A primeira delas é considerar a TDF de uma janela de Hanning de $N_W = N$ amostras, tal qual utilizado no método original. A segunda considerou a TDF de uma janela de Hanning de $N_W = 5N$ amostras, mais estreita no domínio da frequência, de forma a aproveitar melhor o aumento da resolução espectral trazido pelos modelos AR.

5.3.2 Base de dados e testes realizados

Para a realização de testes, foram utilizados sons compostos da soma de amostras musicais compostas de notas tocadas isoladamente, disponíveis na base de dados Musical Instrument Samples, da Universidade de Iowa [67]. Para a escolha das amostras que compõem cada som de teste, um arquivo de amostras é inicialmente selecionado de forma aleatória e dividido em quadros de 4096 amostras, mantendo paralelismo com os testes realizados por Klapuri [46]. Um dos quadros com energia acima da média da energia dos quadros é selecionado (essa restrição busca evitar a seleção de quadros contendo silêncio). A frequência fundamental do quadro selecionado é estimada tanto pelo método Yin [54] quanto pela heurística proposta por Mitre [55], dois métodos bastante precisos para aplicações de áudio monofônico [68]. Caso a diferença de estimativa de ambos os métodos seja inferior a meio semi-tom, o quadro é normalizado para média zero e variância unitária e adicionado ao som de teste, exceto no caso da frequência fundamental detectada para o quadro ser igual a alguma frequência fundamental já existente no som de teste, considerando um erro de um semi-tom.

Foram gerados sons com níveis de polifonia² 1, 2, 4 e 6, sendo que 1000 sons de cada tipo foram utilizadas, num total de 4000 sons gerados de forma completamente aleatória, sob a única condição de que o intervalo mínimo entre duas frequências fundamentais de suas componentes seja de meio semi-tom. Para cada mistura, cada versão do detector de múltiplas frequências fundamentais recebeu como entrada um quadro de 4096 amostras e o número de frequências a detectar.

A ordem do preditor foi variada de modo a se obter os melhores resultados. Testes revelaram que a ordem ideal do preditor deve ser em torno de 2000 (ver Tabela 5.1).

²Número de frequências fundamentais presentes.

5.3.3 Resultados e discussão

A Tabela 5.1 mostra, para cada teste realizado, a fração de frequências fundamentais corretamente detectadas em cada nível de polifonia, considerando um erro aceitável de até meio semi-tom. Essa fração considera o universo de todas as frequências fundamentais presentes nas misturas de determinado nível de polifonia. Para o nível de polifonia 2, por exemplo, um índice de acerto de 75% significa que, das 2000 frequências fundamentais presentes no universo das 1000 misturas de dois sons, 1500 foram corretamente estimadas.

É importante perceber que, nesta tarefa, o número de sons da mistura é conhecido a priori, o que é uma informação que pode não estar disponível numa estimativa real. Como já visto na Seção 4.2.1, ao se requisitar que o sistema realize a estimativa de mais frequências fundamentais que as existentes, provavelmente será obtido um valor errôneo.

Tabela 5.1: Índices de acerto na detecção de múltiplas frequências fundamentais considerando diferentes formas de estimação espectral

Método	Ordem	N_W	1	2	4	6
Original	-	N	96,70%	90,25%	84,15%	72,96%
Expansão 5×	-	N	97,40%	90,95%	84,87%	76,03%
AR	100	N	60,10%	32,90%	10,22%	7,36%
AR	1000	N	97,10%	93,75%	84,65%	74,65%
AR	1500	N	97,50%	94,40%	87,42%	80,71%
AR	2000	N	97,40%	93,85%	88,50%	80,93%
AR	2500	N	97,20%	93,70%	88,30%	80,78%
AR	3000	N	97,20%	92,70%	87,57%	79,96%
AR	3500	N	97,00%	91,05%	86,30%	77,95%
AR	4000	N	97,00%	89,80%	84,97%	77,23%
AR	4095	N	97,00%	90,65%	85,32%	76,68%
AR	1500	5N	97,50%	90,95%	87,22%	80,85%
AR	2000	5N	97,40%	90,60%	87,77%	81,57%
AR	2500	5N	97,20%	88,00%	87,00%	81,23%
AR	3000	5N	97,20%	82,15%	85,35%	79,86%
AR	3500	5N	97,00%	77,60%	81,87%	78,88%
AR	4000	5N	97,00%	76,60%	80,35%	77,15%

Como pode ser verificado, os modelos AR de ordem 1500 e 2000 levaram aos melhores índices de acerto para a detecção de múltiplas frequências fundamentais. Verifica-se,

também, que o uso de modelos AR leva a melhores resultados que a simples expansão do sinal com zeros prévia ao cálculo da TDF, em especial para altos níveis de polifonia. O aumento excessivo na ordem do modelo, porém, leva à perda de desempenho de forma geral. Em especial, para a maior ordem possível (4095) verificam-se resultados semelhantes a aqueles obtidos sem o uso de predição.

Por fim, é interessante perceber que a diferença no tamanho da janela utilizada para estimativa do espectro cancelado leva a piores desempenhos para polifonia 2 e 4. Apesar disso, é notável o fato de que o modelo AR de ordem 2000 obteve melhores resultados para polifonia 6.

A seguir, o processo de extração de múltiplas frequências fundamentais baseado em modelos AR será utilizado na tentativa de melhorar o transcritor automático implementado conforme descrito no Capítulo 4. Uma vez que, potencialmente, mostrou-se haver ganho no processo de detecção de frequências, é possível que a qualidade do processo de transcrição, como um todo, também aumente.

Capítulo 6

Aplicação de predição linear na transcrição automática

No Capítulo 5, foi mostrado que o uso de modelos AR pode prover maior resolução em análises no domínio da frequência, de forma a melhorar o desempenho na tarefa de detecção de múltiplas frequências fundamentais. Neste capítulo, o transcritor automático implementado no Capítulo 4 será modificado de forma a incorporar a predição linear na etapa de processamento de sinais.

O processo de transcrição utilizado como base neste trabalho [34], tal qual implementado no Capítulo 4, pode ser analisado na forma de um diagrama de blocos, conforme a Figura 6.1.

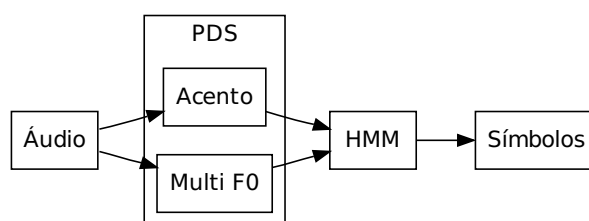


Figura 6.1: Diagrama de blocos mostrando a estrutura do sistema de transcrição.

O diagrama mostrado na Figura 6.1 evidencia que o sinal digitalizado de áudio segue dois caminhos isolados na etapa de processamento de sinais: o cálculo do acento é, em princípio, independente do cálculo das múltiplas frequências fundamentais presentes e suas respectivas saliências. Após, os dados desses dois caminhos são interpretados pelo HMM previamente treinado, que retorna uma seqüência de símbolos - a transcrição.

É de especial interesse que o bloco correspondente ao detector de múltiplas frequências fundamentais seja analisado de forma expandida, como mostra a Figura 6.2.

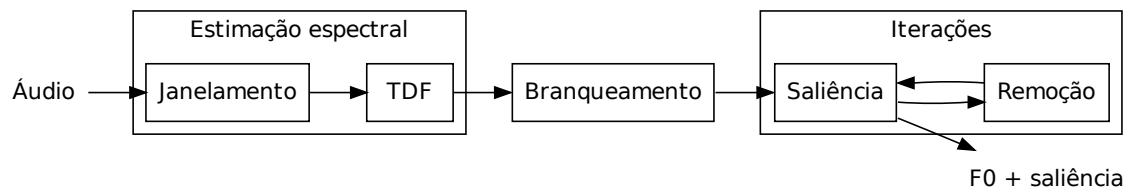


Figura 6.2: Diagrama de blocos do processo de detecção de múltiplas frequências fundamentais.

O processo de detecção de múltiplas frequências fundamentais inicia com um processo de estimação espectral que, no método original [46], é realizado através da multiplicação de um quadro de áudio por uma janela de Hanning seguida do cálculo da TDF, considerando a expansão do sinal com zeros até o dobro de seu comprimento original. A estimativa espectral passa por um processo de branqueamento, que busca equalizar a energia do sinal por todo espectro, e então entra num ciclo de iterações em que a F0 de maior saliência é estimada e removida da mistura sucessivamente.

Neste capítulo, serão propostas modificações ao funcionamento original do transcritor, tomando por base os resultados obtidos na Seção 5.3. Além das modificações ao método que usam técnicas específicas, foi também testada a simples substituição da expansão com zeros do sinal até duas vezes seu comprimento original (proposta original) pela expansão com zeros até cinco vezes o comprimento original, de forma a manter o paralelismo do número de amostras da representação espectral com o número de amostras utilizado na expansão de sinais pela predição linear.

Para cada uma das versões, o HMM é novamente treinado e executado, conforme descrito no Capítulo 4. Inicialmente, foram consideradas as versões do modelo sem a heurística de limite superior de F0 e com modelo musical uniforme. Tais modelos foram incorporados posteriormente.

A Tabela 6.1 traz uma comparação entre os resultados das duas formas de estimação espectral baseadas no formato original do transcritor (expansão com zeros até duas vezes e até cinco vezes o tamanho original).

Pode-se verificar que o simples aumento do número de pontos utilizados para cálculo da TDF representou um pequeno aumento dos índices de desempenho do sistema de transcrição. Apesar disso, o esforço computacional necessário foi, também, maior.

Tabela 6.1: Resultados do processo de transcrição para estimação espectral por TDF.

Método utilizado	Revocação	Precisão	Fator F
Original	49,40%	50,56%	48,90%
Expansão 5×	51,61 %	52,70 %	50,95%

6.1 Estimação espectral por modelo AR

A primeira modificação proposta é a mais simples de todas. Consiste em simplesmente substituir o processo de janelamento seguido da TDF do detector de múltiplas frequências fundamentais por um processo de estimação baseado em predição linear, conforme descrito na Seção 5.3.



Figura 6.3: Diagrama de blocos do processo de detecção de múltiplas frequências fundamentais, substituindo-se o processo de estimação baseado na TDF por um processo baseado em modelagem auto-regressiva.

Testes preliminares indicaram que o uso de $N_W = N$ é mais adequado para a tarefa. Os resultados de cada versão, junto aos resultados originais, são mostrados na Tabela 6.2.

Tabela 6.2: Resultados do processo de transcrição para estimação espectral por modelo AR.

Ordem de predição	Revocação	Precisão	Fator F
1500	48,52%	52,74%	49,08%
2000	49,87%	54,29%	50,58%
2500	50,09%	52,77%	50,56%
Método original	49,40%	50,56%	48,90%

Como pode ser verificado, o uso de modelos AR para a transcrição resulta em ganhos tanto na revocação quanto na precisão. A melhoria obtida, porém, é inferior à

observada na Seção 5.3.3, em que se considera somente a tarefa de detecção de múltiplas frequências fundamentais.

6.2 Modelo AR com pré-filtro

Como já foi discutido anteriormente, modelos AR tendem a apresentar resultados piores se a complexidade do sinal for muito grande. Uma vez que o baixo do sinal está, supostamente, localizado em faixas de frequência baixa, foi inserido um filtro passa-baixas FIR precedendo a estimação espectral, como mostrado na Figura 6.4.

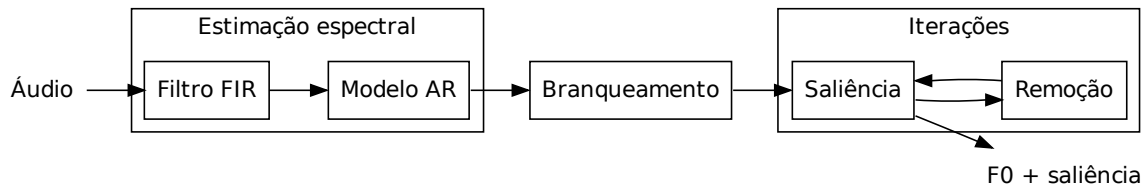


Figura 6.4: Diagrama de blocos do processo de detecção de múltiplas frequências fundamentais, substituindo-se o processo de estimação baseado na TDF por um processo baseado em modelagem auto-regressiva precedido de um filtro FIR.

Utilizando um modelo AR de ordem 2000, verificou-se os efeitos de dois diferentes filtros, um com frequência de corte 1000 Hz e o outro com frequência de corte 2000 Hz. Os filtros utilizados foram construídos com ordem 4096, e aplicados através da multiplicação do sinal por uma janela retangular no domínio da frequência. A Tabela 6.3 mostra os resultados dessa filtragem para a tarefa de transcrição, em comparação com os resultados do modelo AR sem filtro e o método original.

Tabela 6.3: Resultados do processo de transcrição para transcritores com pré-filtro FIR.

Frequência de corte	Revocação	Precisão	Fator F
1000 Hz	54,81%	58,77%	55,21%
2000 Hz	53,60	58,16%	54,37 %
Modelo AR sem filtro	49,87%	54,29%	50,58%
Método original sem filtro	49,40%	50,56%	48,90%

Como se pode verificar, a presença do filtro FIR representou um ganho sensível para o processo de transcrição, embora não tenha havido grande diferença entre as

duas frequências de corte testadas.

6.3 Pré-filtro aplicado ao sinal inteiro

A pré-filtragem FIR resultou em um ganho de desempenho decorrente da possibilidade de simplificação do sinal, em termos espectrais, sem eliminar significativamente as frequências relacionadas ao baixo. O sinal de acento, porém, não foi modificado. Neste novo experimento, um pré-filtro foi aplicado ao sinal inteiro, de forma que o sinal de acento também receberia um sinal cujo baixo é mais predominante. Assim, foi aplicado um filtro Butterworth passa-baixas de décima ordem, com frequência de corte de 1000 Hz, conforme o diagrama da Figura 6.5.

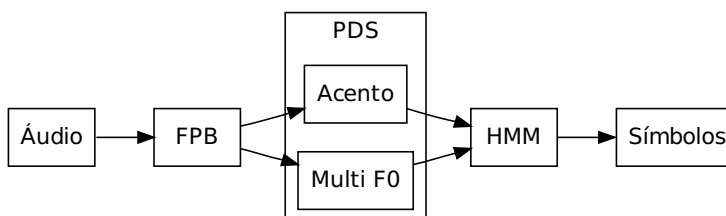


Figura 6.5: Diagrama de blocos mostrando a estrutura do sistema de transcrição considerando um pré-filtro passa-baixas.

Sobre o sinal filtrado, foram testadas duas formas de estimação espectral: através do método original, conforme a Figura 6.2, e através de um modelo AR de ordem 2000, conforme a Figura 6.3. A Tabela 6.4 mostra os resultados obtidos, em comparação com as versões sem pré-filtro.

Tabela 6.4: Resultados do processo de transcrição para transcritores com pré-filtro Butterworth.

Método	Frequência de corte	Revocação	Precisão	Fator F
Original	1000 Hz	51,81%	53,79%	51,45%
Modelo AR	1000 Hz	57,23%	58,69%	56,42%
Original	Sem filtro	49,40%	50,56%	48,90%
Modelo AR	Sem filtro	49,87%	54,29%	50,58%

Pode ser verificado que o pré-filtro levou à melhoria de desempenho em ambos os casos, embora o ganho para o caso do preditor AR tenha sido mais significativo.

A partir dos resultados decorrentes do uso do modelo original, pode-se inferir que a pré-filtragem, de fato, auxilia na exclusão de outros instrumentos, que não o baixo, do sinal analisado. Esse auxílio se reflete no ganho de desempenho conseguido.

No caso do modelo AR, além do isolamento, deve-se considerar que o sinal foi simplificado. Assim, a estimação do modelo AR foi realizada sobre um sinal mais simples, sobre o qual o baixo predomina mais fortemente, resultando numa estimação espectral mais acurada.

6.4 Melhorias no modelo

Após a incorporação de estimadores espectrais mais avançados, foram incorporados no modelo de transcrição o modelo musical, como descrito na Seção 4.3.4, e a heurística para limite superior de F0, como descrito na Seção 4.3.6. O processo de pré-filtragem foi utilizado na forma de um filtro Butterworth de ordem 10 e frequência de corte de 1 kHz, conforme o utilizado na Seção 6.3.

Durante a incorporação da heurística para limite superior de F0, observou-se que a normalização $\lceil y_q - c \rceil$, com $c = 4$, resulta em valores muito baixos na análise do sinal com pré-filtro. Isso decorre do fato de que o processo de pré-filtragem reduz, implicitamente, a média ponderada das múltiplas frequências fundamentais estimadas. Melhores resultados foram conseguidos para $c = -8$.

A Tabela 6.5 traz os resultados obtidos nessa execução, em comparação com os anteriores.

Tabela 6.5: Resultados do processo de transcrição para transcritores com heurística de máxima F0 e modelo musical.

Método	Pré-filtro	Máxima F0/Modelo musical	Revocação	Precisão	Fator F
Original	Nenhum	Não	49,40%	50,56%	48,90%
Modelo AR	1000 Hz	Não	57,23%	58,69%	56,42%
Original	Nenhum	Sim	51,79%	54,96%	52,29%
Modelo AR	1000 Hz	Sim	57,12%	60,12%	57,09%

Tais resultados mostram que melhorias no modelo cognitivo não foram capaz de melhorar significativamente os resultados obtidos para o melhor estimador espectral obtido. Também, é importante observar que a melhoria na estimação espectral foi

capaz de melhorar os resultados obtidos de forma mais significativa que o uso de um modelo cognitivo mais avançado.

Ao final do processo de implementação de modificações ao método original, verifica-se na Tabela 6.5 que a melhor versão obtida é aquela com pré-filtro Butterworth com frequência de corte de 1 kHz, modelo AR de ordem 2000 e incorporando o uso da heurística de máxima F0 e o modelo musical. Essa versão superou a versão original em todos os índices de desempenho, com destaque especial para o fator F , que evidencia uma melhoria geral de 5%.

Capítulo 7

Discussões

Durante a realização deste trabalho, surgiram diversas questões ligadas aos diferentes aspectos presentes no algoritmo aqui descrito. Muitas dessas questões não possuem respostas apropriadas na literatura, de modo que sua inclusão nesta dissertação pode representar uma contribuição adicional e um fechamento adequado ao projeto de pesquisa realizado. Assim, este capítulo tem como caráter principal fomentar a discussão em torno dessas questões em aberto, provendo terreno fértil para a criação e desenvolvimento de novas linhas de pesquisa.

7.1 Categorias de transcritores de música

Como já discutido no Capítulo 1, um transcritor automático de música é um dispositivo ou programa de computador capaz de extrair informações de uma música de forma a descrevê-la através de um conjunto discreto de informações. Tal conjunto de informações é definido caso a caso, dependendo da aplicação a que se destina o transcritor. É possível, porém, classificar os transcritores automáticos de música em três categorias básicas, que serão descritas a seguir.

7.1.1 Conversão para partitura

A primeira categoria engloba transcritores que obtêm automaticamente a partitura da peça tocada por um músico a partir do sinal de áudio correspondente à sua execução, tal qual mostrado na Figura 7.1. Nesse caso, a transcrição obtida destina-se a ser lida por um músico. Existem aplicações comerciais lidando com esse tipo de transcrição, como o Audioscore [3].



Figura 7.1: Partitura de uma música.

É interessante perceber que a conversão para partitura envolve a utilização de determinadas regras que podem ser usadas no processo de transcrição [15], embora tais regras não sejam triviais.

Embora sejam importantes para fins de aprendizado, documentação e análise musical, as partituras ignoram determinadas marcas de expressão, como pequenas flutuações de tempo, altura e intensidade na execução de notas musicais. Tais marcas de expressão aparecerão em outras formas de transcrição, como será visto a seguir.

7.1.2 Obtenção de curva melódica

A segunda categoria de transcritores automáticos engloba sistemas que desejam simplesmente inferir qual é a altura (*pitch*) da melodia de uma música a cada instante de tempo. Um exemplo é mostrado na Figura 7.2. Assim, pode-se obter pontos de uma relação na qual a frequência fundamental da melodia é considerada uma função do tempo. Essa relação determina a curva melódica de uma música [29, 30]. A melodia, em geral, é a parte da música mais marcante para o ouvinte. Assim, a obtenção da curva melódica está fortemente ligada ao reconhecimento de músicas tanto em seres humanos quanto em computadores.

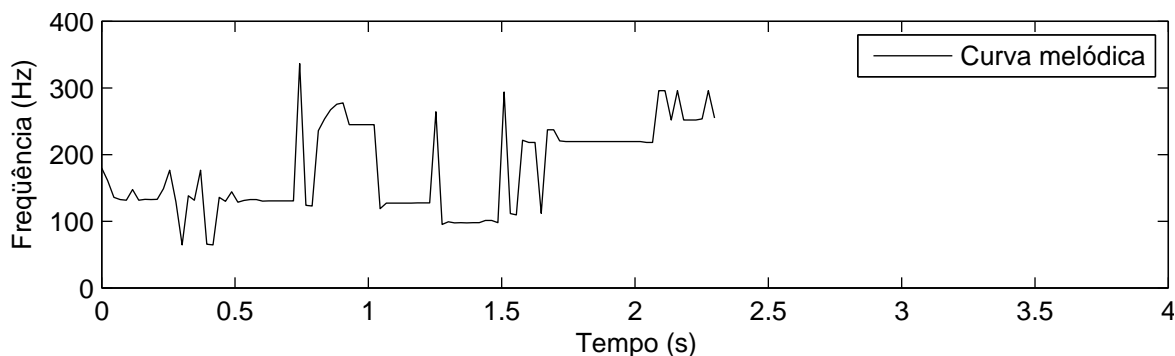


Figura 7.2: Curva melódica de uma música.

No caso de áudio monofônico, essa tarefa é, hoje, trivial, uma vez que já existem métodos bastante confiáveis para detectar a frequência fundamental de sinais harmônicos simples [54, 55]. No caso de áudio polifônico, descobrir a curva melódica significa

descobrir a frequência fundamental emitida pelo instrumento principal - uma guitarra solo ou a voz principal, por exemplo - a cada instante de tempo.

Esse tipo de transcrição, como observado por Poliner [30], tem a vantagem de preservar, no processo transcritivo, detalhes da execução de marcas de expressão como o *vibrato* e o *glissando*, ou mesmo desafinações características do estilo musical. Ao mesmo tempo, trata-se de uma notação que ignora eventos como o início e o final de notas, de forma que a característica rítmica da música é ignorada.

7.1.3 Conversão wave-MIDI

A terceira categoria de transcritores automáticos visa a detecção de notas musicais presentes na música analisada, especificando seu tempo de início (*onset*), tempo de final (*offset*), altura (*pitch*) e, opcionalmente, a técnica musical utilizada. A notação MIDI permite analisar flutuações de tempo e ritmo da música, embora os detalhes da execução de certas técnicas sejam ignoradas. A Figura 7.3 utiliza uma representação do tipo *piano-roll* para demonstrar correspondência entre um sinal de áudio e sua notação MIDI.

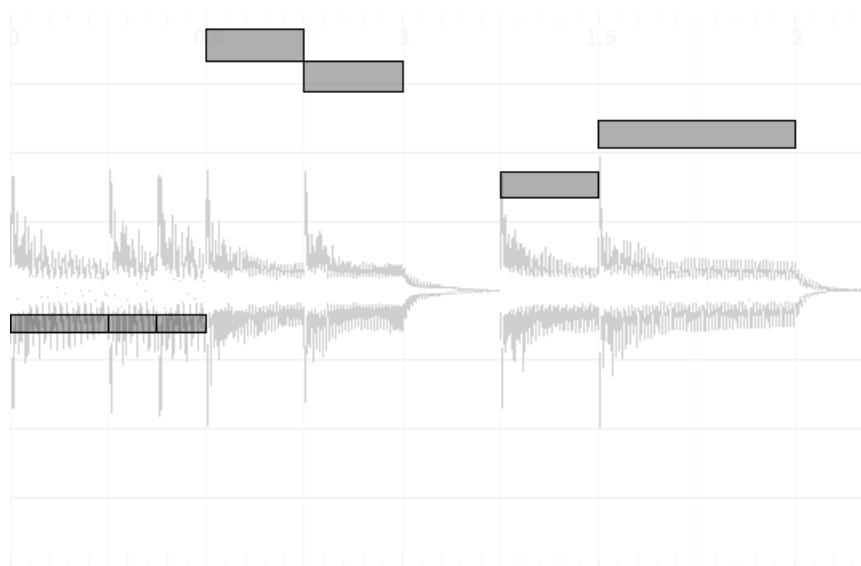


Figura 7.3: Representação *piano-roll* de uma música sobreposta ao sinal musical correspondente.

Além de estar presente em diversas aplicações comerciais, como o Solo Explorer [1], trata-se da categoria de transcritores mais abordada na literatura. Excluindo-se os trabalhos presentes em [15], [29] e [30], todos os trabalhos citados na Seção 1.2 abordam transcritores desta categoria. Este trabalho, também, aborda a construção de um

conversor wave-MIDI, de forma que as discussões que se seguem levam em consideração esta categoria de transcritores.

Embora trate-se de um problema definido de forma simples, os conceitos de início e final de uma nota são nebulosos, como será discutido a seguir.

7.2 O início de uma nota

O início de uma nota é marcado por uma transformação perceptual do comportamento do sinal musical que permite que um ouvinte treinado caracterize um determinado evento: uma tecla do piano foi pressionada, uma corda do violão foi dedilhada, um tambor foi tocado, e assim por diante. Essa definição perceptual é utilizada por Toh [69] em seu sistema de transcrição.

É importante perceber que o início de uma nota envolve uma série de eventos diferentes: o músico tem a intenção de tocar uma nota, executa ações em seu instrumento, o instrumento responde a essas ações de uma certa forma e emite o som. Ao mesmo tempo que há trabalhos como o de Toh [69] e Ryynanen [34] que buscam determinar o comportamento das ações executadas sobre o instrumento, verifica-se que programas como o Solo Explorer [1] tentam prever a intenção do músico, de forma a melhorar a qualidade da transcrição para o usuário final.

Essa diferença entre conceitos de início de nota levou Bello [70] a diferenciar três eventos que ocorrem no início de uma nota, a saber, Ataque, Transiente e *Onset*, de forma objetiva:

- *Ataque* é o intervalo de tempo em que a envoltória do sinal acústico aumenta sua intensidade;
- *Transiente* é o intervalo de tempo em que a onda musical apresenta comportamento não-estacionário;
- *Onset* é um instante que marca o início do transiente inicial de uma nota.

A correlação dessas três definições é grande a tal ponto que diversos métodos de transcrição simplesmente os confundem, sem que isso necessariamente deteriore o desempenho do sistema [69, 34, 33]. Apesar disso, é importante perceber que o objetivo de um conversor para notação MIDI é encontrar os instantes de tempo em que as instruções para iniciar a nota são executadas - ou seja, os *onsets*.

É importante atentar para o fato de existirem diversos tipos de *onsets* na música:

- Percussivo, harmônico (*Pitched Percussive* - PP): sons cujo ataque é percussivo, embora o som do instrumento seja geralmente harmônico. É o caso do piano e do violão, por exemplo.
- Percussivo, não-harmônico (*Non-pitched Percussive* - NPP): sons puramente percussivos, como tambores e pratos.
- Não percussivo, harmônico (*Pitched Non-Percussive* - PNP): tratam-se de sons puramente harmônicos, com entrada suave, como é o caso de violinos e outros instrumentos de arco e de instrumentos de sopro como a flauta e o saxofone.
- Complexos: uma mistura dos outros tipos, como é o caso de instrumentos como a marimba.

Como observado por Bello [70], cada um desses tipos de *onset* é melhor detectado por um processo diferente. Isso decorre do fato de que a frequência é um aspecto que só pode ser observado com um período relativamente longo de observação, de forma que é impossível detectar, sem o uso de informações adicionais, o instante exato em que um sinal senoidal começou a ser emitido. Pode-se, porém, através da observação do comportamento do sinal, inferir o instante de início de sua emissão.

7.3 O fim de uma nota

Embora o conceito de início de uma nota tenha sido amplamente estudado, o conceito de fim de uma nota é, usualmente, negligenciado. Isso ocorre porque, embora seja simples ouvir o início de uma nota, em especial no caso de *onsets* percussivos, perceber o fim da nota não é uma tarefa tão simples.

Numa definição análoga à proposta por Bello [70] quanto aos eventos que marcam o início de uma nota, é possível definir eventos que marcam o fim dessa mesma nota:

- *Soltura* é o intervalo de tempo em que a envoltória do sinal acústico diminui drasticamente sua intensidade;
- *Transiente* é o intervalo de tempo em que a onda musical apresenta comportamento não-estacionário. A soltura faz parte do transiente final de uma nota.
- *Offset* é um instante que marca o início do transiente final de uma nota.

Assim, o *offset* é o momento em que as cordas do violão começam a ser abafadas ou que a tecla de um piano deixa de ser pressionada.

Tratam-se de eventos mais dificilmente perceptíveis, por características próprias de instrumentos musicais. Uma corda vibrante, por exemplo, deixará de soar após algum tempo, mesmo que o músico não faça nada para isso. O ato de abafar a vibração da corda, quando a vibração ainda está forte, é percebido de forma diferente do mesmo gesto se realizado quando a vibração da corda já está naturalmente mais fraca.

Tendo por base essas informações, pode-se discutir o processo de avaliação de transcritores automáticos de música.

7.4 Avaliação de transcritores

A dificuldade de avaliação da conversão para notação MIDI começa na determinação de um conjunto de testes apropriado. Trata-se de uma tarefa bastante difícil determinar uma transcrição de referência para o uso na avaliação, uma vez que o início e o final de uma determinada nota musical nem sempre podem ser detectados facilmente. Esse problema é parcialmente resolvido pela base de dados RWC [59], que é composta de gravações que seguem estritamente arquivos MIDI pré-preparados.

Apesar de toda a precisão no processo de construção da base de dados, Bay [71] observa que obter correspondência absoluta entre uma gravação e um arquivo MIDI é praticamente impossível, uma vez que mesmo a síntese de instrumentos diretamente a partir de instruções MIDI não leva em consideração reverberações naturais do instrumento, sendo, portanto, imprecisa. É importante perceber que, apesar de ser difícil determinar exatamente o tempo de início e final de uma nota, um músico suficientemente habilidoso poderia reproduzir de maneira bastante fiel a execução musical contida numa gravação.

A partir dessa necessidade de avaliação objetiva, surgiu em 2005 o *Music Information Retrieval EXchange - MIREX* [72], uma competição anual na qual algoritmos ligados a tarefas de cognição em música - entre elas, a transcrição automática - são testados frente a um conjunto de dados comum, com métricas discutidas amplamente entre os participantes.

Inicialmente, foram discutidos os limites dentro dos quais é possível gerar uma transcrição de referência precisa. Nas discussões do MIREX-2009 [73], foi estabelecido que a transcrição de referência pode apresentar desvios de até 100ms em relação à marcação verdadeira devido aos limites humanos de percepção. O limite de 100 ms foi adotado

no MIREX por simples consenso, embora tenha sido observado por Loscos [74] que, em transcrições realizadas manualmente, encontram-se desvios próximos a 150 ms entre os resultados gerados por diferentes pessoas sobre a mesma música. Tal resultado é especialmente válido para transcrições referentes à voz humana, na qual é difícil encontrar precisamente o instante de início de uma nota.

Há, ainda, a discussão sobre o que significa uma nota ser corretamente transcrita. É imediata a percepção de que uma nota corretamente transcrita deve apontar para a mesma nota musical da referência, mas ainda há muita discussão quanto aos limites, no tempo, dentro dos quais uma nota ainda pode ser considerada corretamente transcrita.

Também por consenso, o MIREX-2009 adotou um erro de transcrição de 50ms como aceitável. Esse erro está ligado tanto ao fato dos quadros utilizados para análise de áudio através da TDF geralmente assumirem durações de 46ms ou 93ms, o que traz ao sistema de transcrição um erro inerente na ordem de 50ms. Somando-se o erro inerente aos sistemas e o erro inerente à confecção da referência, chegou-se a um limite de erro de 150ms.

Em edições anteriores do MIREX, observou-se que o desempenho dos sistemas avaliados cai significativamente quando se considera tanto o início quanto o final da nota dentro do erro aceitável. Por esse motivo, é possível encontrar na literatura sistemas de transcrição de áudio cuja avaliação leva em consideração apenas o início da nota [34]. No MIREX 2009, uma solução distinta foi adotada: o início da nota deve estar dentro de um limite de 150ms do início da referência, e o final da nota deve ser tal que a duração da nota transcrita desvie no máximo 20% da duração da nota de referência.

É importante perceber que as métricas adotadas no MIREX não estão ligadas a bases científicas sólidas sobre o que significa a transcrição correta de uma nota, mas sim a consensos que envolvem os limites dentro dos quais os transcritores de áudio atuais são capazes de atuar.

Capítulo 8

Conclusão

Neste trabalho, foi abordado o problema da transcrição automática de música, em especial a conversão wave-MIDI do baixo em músicas populares.

A transcrição automática de música depende, essencialmente, de algoritmos para detecção de frequências fundamentais, usualmente dependentes da transformada discreta de Fourier (TDF) de quadros de curta duração. Como discutido, o uso da TDF implica num compromisso entre a resolução da análise nos domínios do tempo e da frequência. A proposta deste trabalho foi utilizar a predição linear para expandir artificialmente cada um dos quadros analisados, de forma a aumentar a resolução da análise no domínio da frequência sem prejudicar a resolução no domínio do tempo. Com isso, buscou-se melhorar o processo de detecção de frequências fundamentais e, assim, melhorar o processo de transcrição como um todo.

O trabalho iniciou-se com o estudo dos conceitos fundamentais de música e psicoacústica que seriam utilizados posteriormente. Esse estudo também envolveu uma revisão da literatura científica relacionada à transcrição automática, que serviu como base para a escolha de um método de transcrição automática para implementação.

O método escolhido foi o desenvolvido por Ryyanen [34] com o intuito de transcrever o baixo em músicas populares. Tal método se baseia em, inicialmente, aplicar algoritmos de processamento digital - entre eles, um detector de frequências fundamentais - a um sinal de áudio digital e, depois, utilizar um Modelo Oculto de Markov (*Hidden Markov Model* - HMM) para inferir a sequência de notas que mais provavelmente foi tocada para gerar esse mesmo sinal.

O método foi implementado conforme descrito pelo autor original [34]. Para tal, foi necessária a codificação de diversos pacotes computacionais, relacionados tanto ao processamento digital de sinais quanto ao HMM. O código original não pôde ser utili-

zados por questões de direitos autorais. Verificou-se no processo de implementação que determinados detalhes do método de transcrição não foram descritos e tiveram que ser inferidos, uma vez que o sistema de transcrição não funcionaria corretamente sem eles. Assim, foi necessário o estudo profundo de todos os algoritmos utilizados no processo de transcrição, conforme registrado nesta dissertação. Tal estudo, embora tenha prolongado o tempo de realização do trabalho, resultou também em grandes contribuições ao conhecimento sobre as bases teóricas do método implementado.

Após a implementação, o método de detecção de frequências fundamentais foi modificado de forma a incorporar a predição linear, visando assim melhorar a resolução da análise no domínio da frequência. Inicialmente, as vantagens dessa melhoria foram investigadas em sinais de teste, compostos de sons conhecidos somados artificialmente. Depois disso, o método de detecção de frequências modificado foi incorporado ao transcritor previamente implementado.

Verificou-se que as modificações propostas trouxeram ganhos consideráveis no desempenho do transcritor. Tais ganhos foram ainda mais pronunciados quando foi adicionado ao sistema um pré-filtro passa-baixas, com a função de tornar o baixo mais evidente na música como um todo. Embora tenham havido ganhos de desempenho, há espaço para melhorias significativas e ainda não é possível assumir o problema da transcrição do baixo em músicas populares como é resolvido.

Futuramente, os conhecimentos adquiridos na realização deste trabalho serão utilizados na construção de um transcritor automático de música capaz de utilizar conhecimento musical de forma explícita, através de um algoritmo especializado, ao invés de um algoritmo genérico como o HMM. Ao se incorporar conhecimento especializado, espera-se obter um transcritor mais eficaz.

Referências Bibliográficas

- [1] Recognisoft, “Solo explorer - wav to midi conversion software,” <http://www.recognisoft.com/>, 2002.
- [2] IMS, “Intelliscore,” <http://www.intelliscore.net/>.
- [3] Neuratron, “AudioScore Ultimate 6,” <http://www.neuratron.com/audioscore.htm>, 2008.
- [4] Wei, J., Boo, J., Wang, Y., and Loscos, A., “A violin music transcriber for personalized learning,” in *2006 IEEE International Conference on Multimedia and Expo*, Toronto, Canada, Julho 2006, pp. 2081–2084.
- [5] Franzblau, C. A., “Computer-aided learning system employing a pitch tracking line,” US patent application number 11/853,062, Setembro. 2007.
- [6] Ghias, A., Logan, J., Chamberlin, D., and Smith, B. C., “Query by humming: musical information retrieval in an audio database,” in *MULTIMEDIA '95: Proceedings of the third ACM international conference on Multimedia*, 1995, pp. 231–236.
- [7] Kline, R. L. and Glinert E. P., “Approximate matching algorithms for music information retrieval using vocal input,” in *MULTIMEDIA '03: Proceedings of the Eleventh ACM International Conference on Multimedia*, Berkeley, CA, USA, Novembro 2003, pp. 130–139.
- [8] Durey, A. S. and Clements, M. A., “Melody spotting using hidden markov models,” in *Proc. 2nd International Conference on Music Information Retrieval*, Bloomington, USA, Outubro 2001.
- [9] Celemony, “Direct Note Access,” <http://www.celemony.com/>.
- [10] Ryyänänen, M., Virtanen, T., Paulus, J., and Klapuri, A., “Accompaniment separation and karaoke application based on automatic melody transcription,” in

- Proc. 2008 IEEE International Conference on Multimedia & Expo*, Hannover, Germany, Junho 2008, pp. 1417–1420.
- [11] Goto, M., Fujihara, H., and Okuno, H., “Automatic system for temporal alignment of music audio signal with lyrics,” US patent application number 11/834,778, Agosto 2007.
- [12] Tavares, T. F., Barbedo, J. G. A., and Lopes, A., “Transcrição Automática de Sinais de Áudio Monofônico Baseada em Quadros de Tamanho Variável,” in *Anais do V Congresso de Engenharia de Áudio*, Maio 2007, pp. 47–51.
- [13] Ryyänänen, M., “Probabilistic modelling of note events in the transcription of monophonic melodies,” Dissertação de Mestrado, Tampere University of Technology, Março 2004.
- [14] Hajimolahoseini, H., Taban, M. R., and Abutalebi, H. R., “Automatic transcription of music signal using harmonic elimination method,” in *Telecommunications, 2008. IST 2008. International Symposium on*, Tehran, Iran, Agosto 2008, pp. 559–563.
- [15] Simões, G., Freitas, A., and Souza, H., “Desenvolvimento de um sistema computacional de transcrição de melodias monofônicas para partitura,” in *Anais do XXVI Simpósio Brasileiro de Computação*, Julho 2006, pp. 22–27.
- [16] Trevilatto, N., Barbedo, J. G. A., and Lopes, A., “Transcrição Automática de Sinais de Áudio Monofônico,” in *Anais do 10 Simpósio Brasileiro de Computação Musical*, 2005, vol. 1, pp. 291–294.
- [17] Martin, K. D., “Automatic transcription of simple polyphonic music,” *The Journal of the Acoustical Society of America*, vol. 100, pp. 2813–2817, Outubro 1996.
- [18] Martin, K. D., “Automatic transcription of simple polyphonic music,” M.I.T. Media Laboratory Perceptual Computing Session Technical Report No. 385.
- [19] Marolt, M., “A connectionist approach to automatic transcription of polyphonic piano music,” *IEEE trans. multimedia*, pp. 439–449, 2004.
- [20] Reis, G. F. and Ferndandez, F. N., “Genetic algorithm approach to polyphonic music transcription,” in *Signal Processing and Information Technology, 2008. ISSPIT 2008. IEEE International Symposium on*, Outubro 2007, pp. 1–6.

- [21] Reis, G. F., Fonseca, N., Ferndandez, F. N., and Ferreira, A., “Genetic algorithm approach with harmonic structure evolution for polyphonic music transcription,” in *Proc. IEEE International Symposium on Intelligent Signal Processing, 2008.*, Outubro 2008, pp. 491–496.
- [22] Raphael, C., “Automatic transcription of piano music,” in *Proc. 3rd International Conference on Music Information Retrieval*, Paris, France, Outubro 2002.
- [23] Niedermayer, B., “Non-negative matrix division for the automatic transcription of polyphonic music,” in *Proc. 9th International Conference on Music Information Retrieval*, Philadelphia, USA, Setembro. 2008, pp. 544–549.
- [24] Lee, D. D. and Seung, H. S., “Algorithms for non-negative matrix factorization,” *Neural Information Processing Systems*, pp. 556–562, 2000.
- [25] D’Accord, “ichords,” ”<http://www.daccord.com.br/>”.
- [26] Lee, K. and Slaney, M., “Automatic chord recognition from audio using an hmm with supervised learning,” in *Proc. 7th International Conference on Music Information Retrieval*, Victoria, Canada, Outubro 2006.
- [27] Ryyänänen, M. and Klapuri, A., “Automatic transcription of melody, bass line, and chords in polyphonic music,” *Computer Music Journal*, vol. 32, no. 3, pp. 72–86, 2008.
- [28] Mauch, M., Noland, K., and Dixon, S., “Using musical structure to enhance automatic chord transcription,” in *Proc. 10th International Conference on Music Information Retrieval*, Kobe, Japan, Outubro 2009, pp. 231–236.
- [29] Goto, M., “A robust predominant-f0 estimation method for real-time detection of melody and bass lines in cd recordings,” in *Proc. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, Junho 2000, vol. 2, pp. II757–II760.
- [30] Poliner, G. E., Ellis, D. P. W., Ehmann, A. F., Gomez, E., Streich, S., and Beesuan O., “Melody transcription from music audio: Approaches and evaluation,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1247–1256, Maio 2007.

- [31] Poliner, G. E. and Ellis, D. P. W., “A classification approach to melody transcription,” in *Proc. 6th International Conference on Music Information Retrieval*, London, England, Outubro 2005.
- [32] Paiva, R. P., Mendes, T., and Cardoso, A., “On the detection of melody notes in polyphonic audio,” in *Proc. 6th International Conference on Music Information Retrieval*, London, England, Outubro 2005.
- [33] Ryyänänen, M. and Klapuri, A., “Transcription of the singing melody in polyphonic music,” in *Proc. 7th International Conference on Music Information Retrieval*, Victoria, BC, Canada, Outubro 2006, pp. 222–227.
- [34] Ryyänänen, M. and Klapuri, A., “Automatic bass line transcription from streaming polyphonic audio,” in *Proc. 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawai’i, USA, Abril 2007, pp. 1437–1440.
- [35] Ryyänänen, M., *Automatic Transcription of Pitch Content in Music and Selected Applications*, Tese de Doutorado, Tampere University of Technology, Dezembro 2008.
- [36] Oppenheim, A. V., Schafer, R. W., and Buck, J. R., *Discrete-time signal processing*, Prentice Hall Inc., 2 edition, 1999.
- [37] Helmholtz, H., *On the Sensation of Tone*, Dover Publications Inc., 4 edition, 1885.
- [38] Olson, H. F., *Music, Physics and Engineering*, Dover Publications Inc., 2 edition, 1967.
- [39] Deutsch, D., *The Psychology of Music*, Academic Press, 1 edition, 1982.
- [40] Schulter, M., “Pythagorean Tuning and Medieval Polyphony,” ”<http://www.medieval.org/emfaq/harmony/pyth.html>”, 1998.
- [41] MIDI Manufacturers Association, “MIDI Specifications,” ”<http://www.midi.org/>”, 1996.
- [42] Kennan, K., *Counterpoint*, Prentice Hall Inc., 4 edition, 1999.
- [43] Barbedo, J. G. A., *Avaliação Objetiva de Qualidade de Sinais de Áudio e Voz*, Tese de Doutorado, Universidade Estadual de Campinas, 2004.

- [44] Zwicker, E. and Feldkeller, R., *Das Ohr als Nachrichtenempfänger*, Hirzel Verlag, 1967.
- [45] Robert, G., “Critical Bands,” ”<http://www.music.gla.ac.uk/george/audio/psy/psy.html>”, 1996.
- [46] Klapuri, A., “Multiple fundamental frequency estimation by summing harmonic amplitudes,” in *Proc. 7th International Conference on Music Information Retrieval*, Victoria, BC, Canada, Outubro 2006, pp. 1–2.
- [47] Klapuri, A., Eronen, J., and Astola, T., “Analysis of the meter of acoustic musical signals,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 242–355, 2006.
- [48] Pappoulis, A. and Pillai, S. U., *Probability, Random Variables and Stochastic Processes*, McGraw Hill, 4 edition, 2002.
- [49] Tomasi, S., “Estimating gaussian mixture densities with em - a tutorial,” ”<http://www.cs.duke.edu/courses/spring04/cps196.1/handouts/EM/tomasiEM.pdf>”, 2004.
- [50] Rabiner, L. R., “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, Fevereiro 1989.
- [51] Murphy, K., “Hidden markov model toolbox for matlab,” ”<http://www.cs.ubc.ca/murphyk/Software/HMM/hmm.html>”, 2005.
- [52] Cambridge University Engineering Department, “Hidden markov model toolkit - v341,” ”<http://htk.eng.cam.ac.uk/>”, 2003.
- [53] Haykin, S., *Neural Networks: A Comprehensive Foundation*, Prentice Hall Inc., 2 edition, 1998.
- [54] Cheveigné, A. and Kawahara, H., “YIN, a fundamental frequency estimator for speech and music,” *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Abril 2002.
- [55] Mitre, A., Queiroz, M., and Faria, R. R. A., “Accurate and Efficient Fundamental Frequency Determination from Precise Partial Estimates,” in *Proceedings of the 4th AES Brazil Conference*, Maio 2006, pp. 113–118.

- [56] Ryyänänen, M. and Klapuri, A., “Modelling of note events for singing transcription,” in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio*, Jeju, Korea, Outubro 2004.
- [57] Ryyänänen, M. and Klapuri, A., “Polyphonic music transcription using note event modeling,” in *Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, Outubro 2005, pp. 319–322.
- [58] Krumhansl, C., *Cognitive Foundations of Musical Pitch*, Oxford University Press, 1 edition, 1990.
- [59] Goto, M., Hashigushi, H., Nishimura, T., and Oka, R., “RWC Music Database: Popular, classical, and jazz music databases,” in *Proc. of the 3rd International Conference on Music Information Retrieval*, Paris, France, Outubro 2002, pp. 281–288.
- [60] Cannam, C., “Sonic Visualizer,” ”<http://www.sonicvisualiser.org/>”, 2009.
- [61] Tavares, T. F., Barbedo, J. G. A., and Lopes, A., “Towards the evaluation of automatic transcription of music,” in *Anais do VI Congresso de Engenharia de Áudio*.
- [62] Kay, S. M., *Modern Spectral Estimation*, Prentice Hall Inc., 1 edition, 1988.
- [63] Makhoul, J., “Linear prediction: A tutorial review,” *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, Abril 1975.
- [64] Atal, B. S. and Hanauer, S. L., “Speech analysis and synthesis by linear prediction of the speech wave,” *The Journal of the Acoustical Society of America*, vol. 50, no. 2, pp. 637–655, Abril 1971.
- [65] Lansky, P. and Steiglitz, K., “Synthesis of timbral families by warped linear prediction,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Abril 1981, vol. 6, pp. 576–578.
- [66] Härmä, A. and Laine, U. K., “Speech analysis and synthesis by linear prediction of the speech wave,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 579–588, Julho 12001.
- [67] University of Iowa, “Musical Instrument Samples,” ”<http://theremin.music.uiowa.edu/MIS.html>”, 2005.

- [68] Tavares, T. F., Barbedo, J. G. A., and Lopes, A., “Performance evaluation of fundamental frequency estimation algorithms,” in *Proceedings of the International Workshop on Telecommunications - 2009*, Fevereiro 2009, pp. 94–97.
- [69] Toh, C. C., Zhang, B., and Wang, Y., “Multiple-feature fusion based onset detection for solo singing voice,” in *Proc. of the 9th International Society for Music Information Retrieval Conference*, Philadelphia, USA, Outubro 2008, pp. 515–520.
- [70] Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M.B., “A tutorial on onset detection in music signals,” *Speech and Audio Processing, IEEE Transactions on*, vol. 13, pp. 1035–1047, Setembro. 2005.
- [71] Bay, M., Ehmann, A. F., and Downie, J. S., “Evaluation of multiple-f0 estimation and tracking systems,” in *Proc. of the 10th International Society for Music Information Retrieval Conference*, Kobe, Japan, Outubro 2009, pp. 315–320.
- [72] Downie, J. S., “The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research,” *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, Fevereiro 2008.
- [73] MIREX 2009, “Discussion wiki,” ”<http://www.music-ir.org/mirex/2009/>”, 2009.
- [74] Loscos, A., *Spectral Processing of the Singing Voice*, Tese de Doutorado, Pompeu Fabra University, 2007.