DETERMINAÇÃO DA ENVOLTÓRIA DE NOTAS MUSICAIS NO DOMÍNIO DO TEMPO

Rafael George Amado

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

Orientador: Luiz Wagner Pereira Biscainho

Rio de Janeiro
Junho de 2012

DETERMINAÇÃO DA ENVOLTÓRIA DE NOTAS MUSICAIS NO DOMÍNIO DO TEMPO

Rafael George Amado

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

Examinada por:

Prof. Luiz Wagner Pereira Biscainho, D.Sc.

Prof. Eduardo Antônio Barros da Silva, Ph.D.

Prof. Tadeu Nagashima Ferreira, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
JUNHO DE 2012

Amado, Rafael George

Determinação da Envoltória de Notas Musicais no Domínio do Tempo/Rafael George Amado. – Rio de Janeiro: UFRJ/COPPE, 2012.

XVIII, 110 p.: il.; 29, 7cm.

Orientador: Luiz Wagner Pereira Biscainho

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2012.

Referências Bibliográficas: p. 98 – 102.

1. Áudio. 2. Envoltória. 3. Notas Musicais. I. Biscainho, Luiz Wagner Pereira. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Dedicado à minha família.

Agradecimentos

Agradeço de coração a todos os que me ajudaram nessa etapa.

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

DETERMINAÇÃO DA ENVOLTÓRIA DE NOTAS MUSICAIS NO DOMÍNIO DO TEMPO

Rafael George Amado

Junho/2012

Orientador: Luiz Wagner Pereira Biscainho

Programa: Engenharia Elétrica

A extração de parâmetros descritivos de sinais de música pré-gravados compreende, entre outras análises, a obtenção de uma envoltória temporal de sua amplitude (ou, alternativamente, sua potência). Uma possível abordagem é analisar cada nota musical (ou emissão) individualmente, o que pressupõe algum método para separá-las no caso polifônico. Esta dissertação se divide em duas partes principais. Na primeira, propõe-se um algoritmo para estimação da envoltória temporal de amplitude de notas musicais isoladas calcado em Morfologia Matemática, juntamente com um critério perceptivo que permite determinar automaticamente seus parâmetros de operação. Testes com sinais contendo notas musicais de diversas alturas gerados por instrumentos musicais de diferentes famílias mostraram o bom desempenho do método proposto quanto à suavidade e acurácia das envoltórias obtidas. Na segunda parte do trabalho, investigam-se as dificuldades associadas ao caso polifônico. Elegendo a NMF (*Non-Negative Matrix Factorization*) como o método de separação de fontes sonoras associado, examinaram-se combinações de notas musicais sequenciais sem e com sobreposição quanto à qualidade das envoltórias obteníveis associando-se a matriz $\mathbf{H}$ de ganhos resultante da separação com o método de extração de envoltórias proposto. No sentido inverso, fizeram-se experimentos sobre a possibilidade de melhorar o desempenho da separação introduzindo informação de padrões de envoltória previamente extraídos. Os resultados em ambos os casos não foram positiva ou negativamente impactantes, indicando a necessidade de investigação adicional.

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

ENVELOPE ESTIMATION OF MUSICAL NOTES IN TIME-DOMAIN

Rafael George Amado

June/2012

Advisor: Luiz Wagner Pereira Biscainho

Department: Electrical Engineering

The extraction of descriptive parameters from previously recorded music signals encompasses among other analises the obtention of temporal magnitude (or power) envelopes. A possible approach is to analyze each individual musical note (or emission), which implies the use of some separation procedure in the polyphonic case. This dissertation is structured in two main parts. In the first one, an algorithm for estimation of the temporal magnitude envelope of individual musical notes based on Mathematical Morphology is proposed, along with a perceptual criterion to automatically set its operation parameters. Tests with signals composed of musical notes of several pitches emitted by musical instruments of different families show the good performance of the proposed method as to both smoothness and accuracy of the obtained envelopes. In the second part of the work, some issues inherent to the polyphonic case are examined. After choosing the NMF (Non-Negative Matrix Factorization) as the sound source separation method to be applied, the situation when two musical notes are sequentially combined with and without superposition is investigated as to the quality of envelopes attainable by combination of the gain matrix $\mathbf{H}$ provided by the NMF with the proposed method for envelope extraction. Reversely, some experiments assess the possibility of ameliorate the separation performance by including information from a previously obtained envelope template. In both cases the results have not been clearly bad or good, thus indicating that further careful investigation is needed.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AM	*Amplitude Modulation*, p. 16
BSS	*Blind Source Separation*, p. 69
DE	Distância Euclidiana, p. 106
DFT	*Discrete Fourier Transform*, p. 26
DKL	Divergência de Kullback-Leibler, p. 66, 106
FDLP	*Frequency-Domain Linear Prediction*, p. 23
FIR	*Finite Impulse Response*, p. 16
G&L	*Algoritmo de Griffin e Lim*, p. 108
ICA	*Independent Component Analysis*, p. 65
IIR	*Infinite Impulse Response*, p. 16
MM	Morfologia Matemática, p. 31
MSTFTM	*Modified Short-Time Fourier Transform Magnitude*, p. 108
MSTFT	*Modified Short-Time Fourier Transform*, p. 108
NMF	*Non-negative Matrix Factorization*, p. 65
PCM	*Pulse Code Modulation*, p. 15
RMS	*Root Mean Square*, p. 20
RTISI-LA	*Real-Time Interactive Spectrogram Inversion with Look-Ahead*, p. 68
RTISI	*Real-time Iterative Spectrogram Inversion*, p. 109
SAR	*Source-to-Artifacts Ratio*, p. 70
SDR	*Source-to-Distortion Ratio*, p. 70

SIR *Source-to-Interferences Ratio*, p. 70

STFT *Short-Time Fourier Transform*, p. 108

TAE *True Amplitude Envelope*, p. 26

Capítulo 1

Introdução

O campo de Processamento Digital de Áudio é muito amplo, e suas técnicas são úteis em áreas tão diversas quanto Telecomunicações (telefonia, *VoIP*), Entretenimento (jogos eletrônicos, *players* portáteis de música e vídeo) e Música. Dentre os muitos problemas por ele abordados, podem-se enumerar compressão, codificação, análise, síntese etc.

Alguns exemplos de aplicações são sistemas de transcrição musical automática, em que se procura, a partir de uma gravação, obter a partitura que melhor represente tal gravação; sistemas de restauração de gravações degradadas, que buscam retirar ruídos indesejados e para tal necessitam diferenciar partes do sinal de entrada que sejam informação útil de indesejada; e extração de parâmetros para posterior ressíntese com finalidades diversas, entre elas a "reexecução" do que o instrumentista tocou no momento da gravação como forma de realizar a restauração por ressíntese.

Pode-se enquadrar o presente trabalho na área de análise de sinais musicais, em que se busca extrair parâmetros e características dos sinais a fim de processá-los, ou ainda para obter informações úteis e significativas para algum contexto.

Os parâmetros mais intuitivos de um sinal musical são o *pitch*, relacionado com a altura (em Hz) do sinal em determinado momento; o *timbre*, que é determinado pelas diversas componentes presentes no sinal, resultado das interações entre instrumentista (com sua técnica), a construção do instrumento e o ambiente em que o mesmo está inserido; e, finalmente, a *envoltória*, que pode ser entendida como sendo a evolução da intensidade do sinal ao longo do tempo.

1.1 Aplicações de Processamento Digital de Sinais em Música

Em geral, sinais musicais são bastante complexos, pois vários "sons" podem ser tocados simultaneamente. Os sinais em que isso ocorre são chamados sinais *polifônicos*,

em contraste com sinais *monofônicos*, em que apenas um "som" é tocado a cada vez [4]. Claramente, "som" é uma ideia demasiadamente vaga, que será mais bem definida posteriormente, porém é intuitiva o suficiente para permitir as explanações que se seguem. Por ora, pode-se entender o "som" como sendo cada célula básica que compõe o sinal musical.

No caso da análise/síntese de sinais musicais, existe uma cadeia típica de operações:

$$\text{Separar} \rightarrow \text{Modificar} \rightarrow \text{Ressintetizar}$$

Estamos falando em *separar* os "sons", extraindo parâmetros ou não, de modo a ter entidades ou partes significativas; *modificar* essas partes, entidades ou "sons", realizando seu processamento, modificando, retirando ou adicionando partes etc; e *ressintetizar* as partes desejadas (não necessariamente todas as extraídas), de modo a obter o efeito desejado no sinal musical resultante.

Esse efeito desejado poderia ser, por exemplo, retirar um instrumento, retirar alguma sequência de notas, retirar ruído (quando este pode ser diferenciado do sinal de interesse) ou algo mais simples, como mudar o padrão frequencial de algum instrumento, etc.

1.1.1 Um exemplo desafiador: análise para ressíntese

Um tema que vem ganhando importância é a Análise para Ressíntese, na qual um sinal musical é analisado e todas as suas informações relevantes são armazenadas de alguma forma, para posterior ressíntese.

No caso tradicional em que se pretende modificar características do sinal ("pitch", timbre etc), a fidelidade não é a meta: pode-se, então, modificar os instrumentos, timbres, ambientação, posição dos microfones etc; essa abordagem abre um mundo de possibilidades, pois em teoria é possível conseguir-se qualquer combinação timbre/ambiente.

Por sua vez, no caso da ressíntese pura, busca-se uma reprodução fiel do sinal original, de modo que a precisão no detalhamento de cada nota executada, bem como dos demais elementos presentes (reverberação, técnica do instrumentista etc) devem ser levados em conta. Esse processo pode ser útil em sistemas de restauração de gravações como em [5] e [6], muitas vezes degradadas a um grau tão extremo que as técnicas usuais de eliminação do ruído não são capazes de limpar o sinal, sendo uma alternativa mais viável a extração da informação de execução e a ressíntese do sinal sem o ruído. Um exemplo dessa aplicação em específico são as "reexecuções" criadas pelo Zenph Studios® [7], cuja ideia é exatamente extrair as informações

das notas na forma de um MIDI[1] [8] elaborado que, por sua vez, controla a reexecução por um piano acústico que, finalmente, é regravado. Nesse caso, busca-se a maior fidelidade possível ao sinal original, atentando-se para todas as características timbrísticas do instrumento e detalhes gerais da gravação, como posicionamento e tipo dos microfones, reverberação do ambiente etc.

Grande parte da informação de execução/interpretação reside diretamente na forma da envoltória do sinal. Desta forma, cabe uma discussão um pouco mais aprofundada das diversas maneiras de definir, extrair e interpretar uma envoltória.

1.2 Envoltória temporal de potência/magnitude

Para muitas das aplicações em análise/síntese, uma representação interpretável e modificável da envoltória do sinal é necessária.

Ao se observar uma forma de onda de um sinal qualquer, simples ou complexo, monofônico ou polifônico, pode-se visualmente "desenhar" a envoltória do sinal (lembrando que a envoltória está ligada à evolução da intensidade do sinal ao longo do tempo). Entretanto, na maioria dos casos, essa forma de onda é resultado da soma de diversas partes, sons ou instrumentos tocados simultaneamente, e isso faz com que ela carregue relativamente pouca informação. Feita diretamente, tal análise estaria restrita à determinação da energia (ou da potência) do sinal completo.

Para que seja possível uma melhor interpretação, pode-se dividir o sinal em partes e extrair a envoltória de tais partes, de modo que seja possível reconstruí-lo a partir de tais partes. A interpretação de "parte" é muito subjetiva e, no contexto do trabalho, poderia ser entendida como sendo uma *Fonte Sonora*.

O conceito de fonte sonora ainda precisa ser especificado, pois depende do contexto em que se procura realizar essa separação em partes. O mais intuitivo seria considerar uma fonte sonora como sendo um instrumento emitindo algum som, porém existem contra-exemplos: a bateria, que é um instrumento formado por diversos outros sub-instrumentos, ou um naipe de violinos ou de metais, em que vários instrumentos iguais ou parecidos por vezes tocam em uníssono a mesma melodia — nesse caso, o público em geral tende a considerar esse naipe como sendo um instrumento apenas.

Outra possibilidade de definição seria considerar como fonte sonora cada elemento físico que gera algum tipo de vibração. Desta forma, um violão seria dividido

[1]Sigla em inglês para Interface Digital para Instrumentos Musicais, é basicamente um padrão de mensagens que são enviadas a um sintetizador para que uma sequência de notas seja executada de uma certa maneira. Um arquivo MIDI não armazena o som, mas sim informações sobre a execução das notas, de forma que um equipamento possa interpretá-las e produzir o som que se deseja. Basicamente, o MIDI armazena os tempos de início e fim da nota, a intensidade com que deve ser tocada e a sua altura, entre outros parâmetros.

em 6 fontes, uma para cada corda. Mas existem instrumentos que, por construção geram mais de uma vibração ao mesmo tempo para uma nota; temos como exemplo o piano, no qual algumas teclas fazem com que o martelo golpeie até três cordas ao mesmo tempo, o que gera vibrações de cada corda, da interação entre elas e da ressonância do corpo do piano.

Uma terceira opção de definição seria considerar como fonte sonora aquilo que se percebe auditivamente como uma única fonte, mas isso ainda pode gerar discussão, pois se o público escuta o naipe de violinos como uma unidade, um maestro pode ser capaz de diferenciar um instrumento em separado.

Ao longo desta dissertação serão claramente especificadas as partes (elementos ou componentes) das quais se deseja encontrar a envoltória.

1.3 Representação de música

Sabendo que este trabalho busca conseguir representações detalhadas das envoltórias de sinais musicais, neste ponto cabe um pequeno resumo de como se representa música, em geral.

Em termos perceptivos, o som musical costuma ser caracterizado por três elementos principais: *pitch*, relativo à altura (em Hz) do som, *loudness*, relativo à intensidade sonora, e *timbre*, resultado da sua composição frequencial ao longo do tempo e que permite lhe dar uma "cor" própria que o identifica. Quando esses elementos são limitados por características temporais como *onset* (definido a seguir) e duração, o resultado é uma *Nota Musical* [9].

Em termos gerais, uma nota é um som, com altura definida, inserido num contexto temporal.

A representação musical mais comum é a partitura, que é uma forma simbólica de descrição da música. Cada símbolo define uma altura e uma duração para a nota. Existem outros símbolos de marcação de tempo e ritmo, que fogem ao escopo do trabalho.

Nota-se que em uma partitura não é possível representar a evolução temporal da nota, ou seja, cada símbolo informa quanto a nota deve durar, mas não especifica a maneira como ela deve evoluir ao longo desse tempo.

Supondo uma nota tocada isoladamente, uma representação que mostra sua evolução ao longo do tempo é a forma de onda dessa nota. Contudo, a forma de onda é algo de difícil interpretação, uma vez que não possui informação frequencial direta e embute os parâmetros em uma única dimensão.

Muitas vezes é necessária uma representação individual parametrizada das notas, que não seja simbólica como uma partitura nem de difícil interpretação como uma forma de onda. Uma descrição precisa do perfil de energia de cada nota pode ser

parte dessa representação, que pode ser mais ou menos detalhada, dependendo da finalidade. Essa representação consiste na chamada "envoltória".

1.3.1 Envoltória de uma nota musical

A literatura [4] define um modelo[2] (Figura 1.1) para a execução de uma nota musical, composto de 4 partes (supondo que a nota seja emitida em um meio silencioso):

- Ataque: região em que a amplitude da envoltória aumenta;
- Decaimento ou transitório: tempo para estabilização da execução da nota;
- Sustentação: período em que a amplitude se mantém aproximadamente constante (envolve a interferência do músico);
- Relaxamento: período de extinção da nota.

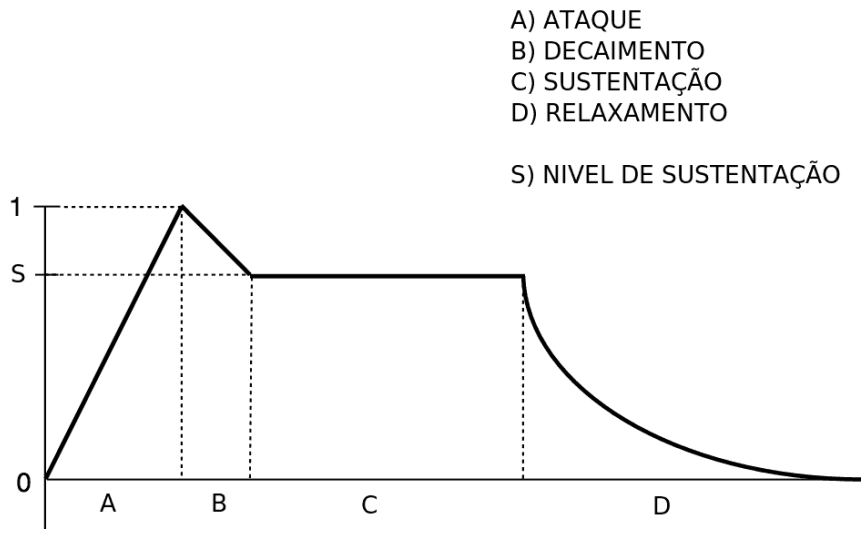


Figura 1.1: Envoltória de uma nota executada (extraída de [1])

A forma apresentada na Figura 1.1 nem sempre é representativa dos casos reais (de fato, na maioria dos casos não o é). É uma ideia advinda dos modelos de síntese sonora, e pode sofrer variações significativas em todas as suas partes. Dependendo do instrumento ou da técnica utilizada, a envoltória da nota pode variar significativamente; um arco tangenciando as cordas de um violino, por exemplo, pode criar um padrão de ataque abrupto ou suave, dependendo da velocidade e força que o instrumentista utiliza durante a execução da nota.

[2]Esse modelo é generalista, pois cada instrumento possui um padrão distinto; somente o Ataque e o Relaxamento estão presentes em qualquer envoltória.

Outro problema intrínseco, ainda no caso monofônico, é a sucessão de notas, pois deve-se encontrar um limite entre elas, o que nem sempre é trivial. Notas tocadas em sequência podem ser ligadas (*legato*), podem ser claramente isoladas (*staccato*), ou ainda diversas misturas das formas anteriores, que dependerão da técnica utilizada pelo instrumentista e das condições em que foi realizada a gravação. Muitas vezes uma nota começa a ser executada enquanto a nota anterior ainda não se extinguiu completamente, fazendo com que as duas soem simultaneamente. Esse é um caso que será tratado mais adiante no trabalho.

Uma pequena discussão sobre o início e a extinção de uma nota musical é realizada a seguir:

I) *Onset* - surgimento de uma nota musical

O *onset* é o instante de tempo marcado como sendo o início da execução da nota, ou seja, o instante durante o ataque a partir do qual se assume que a nota está presente. A Figura 1.2 ilustra um exemplo de *onset* marcado.

No grupo das notas emitidas de maneira isolada e espaçada a marcação dos *onsets* é óbvia, uma vez que é claro o momento em que se inicia uma nova emissão. Contudo, no caso de emissões sequenciais e ligadas essa marcação não é tão trivial. Deve-se convencionar o critério a ser adotado para se afirmar que uma nova emissão começou.

O fato de a energia não variar significativamente entre uma emissão e outra pode ser contornado com outros métodos que levam em conta, por exemplo, variações frequenciais [10] e modelos psicoacústicos [11], entre outros.

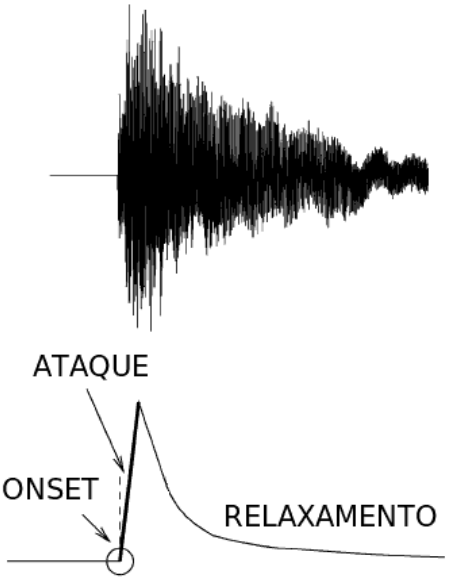


Figura 1.2: Exemplo de marcação de *onset* (extraído de [2])

II) *Offset* - Extinção de uma nota musical

É polêmica a definição de qual o momento em que se pode dizer que ocorreu a completa extinção de uma nota executada. No caso monofônico pode-se atentar apenas à energia do sinal; em notas tocadas em sequência, dependendo da aplicação, a execução de uma nota pode ser dada como terminada quando começa a próxima, de forma que a informação do momento da extinção da nota é dada pelo próximo *onset*. Por sua vez, em sinais polifônicos a presença de várias notas simultâneas impede que a informação de energia sozinha seja suficiente, sendo necessária a utilização de técnicas mais avançadas, envolvendo características frequenciais, por exemplo.

Muitas vezes o final de uma nota se torna indefinido, pois dependendo da gravação o relaxamento da nota pode ser ocultado: uma sustentação proporcionada pela reverberação do ambiente pode ser confundida com a manutenção proposital da nota pelo intérprete.

A extinção de uma nota é, obviamente, dependente da construção do instrumento, do modo de excitação do mesmo (através de um impulso, como um martelo, ou através de movimentação de ar, beliscando-se ou raspando-se a corda). Além disso, alguns instrumentos permitem a manutenção da energia da nota ao longo da sua existência, como por exemplo instrumentos em que é possível manter a excitação constante (um sopro constante ou um movimento contínuo de arco). Outros instrumentos não permitem esse controle; uma vez que se emite a nota, tem-se pouco ou nenhum controle sobre a evolução da mesma. Alguns exemplos são o piano e o violão, em que se excitam (cada um à sua maneira) as cordas, e estas vibram livremente.

1.4 Fidelidade versus processamento

Num sistema que visa a obter a representação detalhada da envoltória podem ser desejáveis duas configurações distintas: representar com a finalidade de manter a fidelidade ao sinal original ou obter-se uma representação que possibilite um posterior processamento das variáveis envolvidas.

1.4.1 Representação buscando fidelidade

Se o objetivo é manter a fidelidade ao sinal original, a representação deve ser capaz de garantir a manutenção das características do sinal original no caso de uma reprodução posterior. Esse contexto demanda uma série de assunções, tais como:

- O ambiente é parte integrante da representação, não se separam os efeitos próprios do instrumento dos gerados pelo ambiente;

- As condições nas quais foi realizada a gravação devem ser mantidas (posição dos microfones, perspectiva para o ouvinte etc.);
- Todas as características timbrísticas do instrumento e da execução devem ser preservados (técnica utilizada, construção do instrumento etc.);
- Se algum pós-processamento foi realizado após a gravação, ele deve ser reproduzido também.

1.4.2 Representação para posterior processamento

Por outro lado, pode ser desejável conseguir uma representação que possibilite a variação dos parâmetros envolvidos obtendo-se, na ressíntese, um sinal diferente do original. Esse sinal conteria as mesmas notas, porém admite-se outra ambientação, outra perspectiva para o ouvinte, diferentes posições de microfones, outra forma de timbrar o instrumento etc.

Um bom exemplo de uso desse tipo de representação são as "reexecuções" criadas pelo Zenph Studios® [7], em que as gravações originais são analisadas em detalhes e, a partir dessa representação, faz-se uma regravação das peças musicais com outros instrumentos, em um ambiente diferente do original.

Podem existir diversas maneiras de se representar a informação e ter o total controle da mesma. Numa situação extrema, as notas seriam representadas anecoicamente e o ambiente seria reinserido posteriormente. Para tal, seria possível criar diversas situações e combinações diferentes variando alguns parâmetros, tais como:

- Geometria, materiais das paredes e reverberação da sala (onde foi gerada a gravação?);
- Tipos, modelos, marcas dos microfones e como estão posicionados (como foi gerada a gravação?);
- Pós-processamento realizado (reverberação artificial, mixagem, masterização etc.).

Neste trabalho, o foco está na representação descrita na Seção 1.4.1.

1.5 Possível classificação de instrumentos/notas musicais

Uma vez que o objetivo é conseguir informações de envoltória de notas musicais (seja para posterior processamento ou para ressíntese fidedigna), é interessante esclarecer

como são geradas notas nos instrumentos; essa informação é muito útil, uma vez que o sistema precisa estimar uma envoltória fiel à forma com que o instrumento (ou naipe de instrumentos) gera suas notas.

Existem alguns modelos de classificação de instrumentos musicais na literatura. O mais conhecido é o apresentado por Hornbostel e Sachs [12], que os classifica de acordo com a natureza do material que produz o som (coluna de ar, membrana, corda etc.) e com o corpo do instrumento (forma, material de construção etc.). Existem classificações que utilizam outros critérios, como por exemplo a dinâmica [13], mas a maioria tem como base o método de Hornbostel e Sachs.

Resumidamente, a classificação usual divide os instrumentos em cinco grandes grupos:

- **Idiofones:** grupo de instrumentos musicais em que o som é provocado pela vibração do seu próprio corpo, sem a necessidade de nenhuma tensão. Este grupo engloba a maior parte dos instrumentos acionados por atrito (como o reco-reco), por agitação (como o chocalho, o caxixi e o ganzá), assim como muitos instrumentos de percussão melódica, como os xilofones. Os blocos sonoros, claves e pratos são exemplos de idiofones percutidos sem intenção melódica.

- **Membranofones:** são instrumentos de percussão, que produzem som através da vibração de membranas sob tensão. Neste grupo estão os tambores percutidos, como os tímpanos, e os tambores friccionados, como a cuíca, entre outros.

- **Cordofones:** grupos de instrumentos cuja fonte primária de som é a vibração de uma corda tensionada quando beliscada, percutida ou friccionada. Todos os instrumentos pertencentes a esse grupo podem ser executados de qualquer uma das três formas citadas, porém cada um possui uma maneira mais usual: cordas beliscadas ou tangidas por plectros, unhas, palhetas ou dedos são a maneira usual de se produzir som em violões, harpas e liras, grupo que também inclui instrumentos de teclado e plectro como cravos e clavicórdios; cordas percutidas como o berimbau e o piano; e cordas friccionadas, caso dos instrumentos de arco, como a família dos violinos.

- **Aerofones:** neste grupo, o som é produzido principalmente pela vibração do ar sem a presença de membranas ou cordas e sem que a própria vibração do corpo do instrumento tenha influência significativa no som produzido. Inclui todas as flautas, metais (como o trompete e o trombone), instrumentos de palhetas simples (algumas gaitas-de-fole, clarinete, saxofone etc.) e palhetas

duplas (como o oboé, algumas gaitas-de-fole e o fagote). Podem ser incluídos nesta categoria todos os tipos de órgão, com exceção dos elétricos.

- **Eletrofones:** originalmente não presente na classificação de Hornbostel e Sachs, este grupo foi incluído com o aparecimento dos instrumentos em que o som é produzido com a intervenção de corrente elétrica. Começou com o teremim, como primeira experiência, e hoje inclui todos os tipos de sintetizadores analógicos e digitais, órgãos e pianos elétricos, guitarras e baixos elétricos, entre outros.

Neste trabalho, o foco são as notas musicais, mais especificamente a maneira como essas notas "aparecem" nos instrumentos musicais. Dessa forma, é útil um modelo de classificação de notas.

A maneira como a nota "aparece" depende do tipo de instrumento (e, portanto, a classificação de instrumentos é parte desse modelo), do tipo de excitação aplicada a esse instrumento (contínua, impulsiva etc.), ou ainda da sua construção ou da técnica utilizada pelo instrumentista.

Como visto anteriormente, vários instrumentos possuem mais de um modo de execução; por exemplo, o violino pode ser tocado com a fricção de um arco ou o tanger dos dedos. Dessa forma, nem sempre a classificação do instrumento com base na sua construção é suficiente. Propõe-se aqui uma divisão diferente. Uma possível classificação quanto ao tipo de emissão das notas musicais é a seguinte:

1.5.1 Emissão de altura fixa

Quando a tecla de um piano é acionada, o martelo choca-se contra uma ou mais cordas associadas a esta tecla e as vibrações destas cordas são transmitidas ao corpo do piano, gerando o som que se ouve. Se o instrumentista toca a mesma tecla, a altura da nota emitida será sempre a mesma e, a menos dos pedais de abafamento ou de sustentação, não há controle algum sobre a evolução da nota até sua extinção completa.

Instrumentos de teclas em geral, como cravos e pianos, apresentam a mesma característica, pois cada tecla está associada a uma corda ou conjunto de corda, apenas. Órgãos em geral também se enquadram nesse grupo, pois cada tecla está associada a um ou um conjunto de tubos, que também fazem com que a coluna de ar sempre emita um mesmo conjunto de frequências, de altura definida.

A maioria dos instrumentos de percussão, como tambores e a maioria dos membranofones, também não permite um controle da nota emitida. Apenas o choque da mão (ou de algum objeto utilizado como acionador) com a membrana é realizado e a nota é emitida; a membrana vibra livremente até a extinção e, novamente, não

se pode alterar a altura da nota durante sua emissão. Alguns tipos de tambores possuem controle da emissão como é o caso do tímpano, que possui um pedal de controle de altura da nota emitida.

A grande maioria dos instrumentos de corda (com exceção de pianos, cravos e afins, onde a corda é acionada por uma tecla), pertencem ao outro grupo, detalhado a seguir.

1.5.2 Emissão de altura variável

Contrapondo-se aos exemplos apresentados na seção anterior, existem as emissões de notas com altura variável.

Ao friccionar um arco contra a corda de um violino, a corda vibra e transmite sua vibração através da ponte ao corpo do instrumento, gerando o som audível. Enquanto o instrumentista desliza o arco por sobre as codas, existe geração de som e o executor possui quase total controle sobre como a nota evolui, pois a mesma existe apenas enquanto o arco possui movimento. Imaginando uma situação hipotética de um arco circular, movido por uma máquina, a nota poderia perdurar indefinidamente.

Enquanto o arco fricciona a corda, o instrumentista tem o controle da altura da nota emitida conforme o local em que seu dedo pressiona a corda no braço do instrumento, ou seja, a altura da nota pode ser variada durante sua emissão e evolução temporal.

Os instrumentos de corda acionados por beliscões ou tangidos normalmente permitem algum controle sobre a altura durante a existência da nota. Num violão, o instrumentista pode esticar mais ou menos a corda e mudar de casa (espaço entre duas marcações no braço do instrumento) alterando o comprimento da corda, antes que a nota seja extinta, quando a corda para de vibrar.

Além dos exemplos acima citados, os instrumentos de sopro em geral são também parte desse grupo, uma vez que enquanto houver movimento de ar, existe emissão de nota e a altura é passível de mudança. Mesmo quando não há controle do comprimento do tubo (como têm os trombones e trompetes, cada qual com seu mecanismo) ou da saída intermediária de ar (como têm os saxofones, clarinetes etc), o instrumentista pode mudar a maneira como sopra, e a vibração de palhetas ou lábios para alterar a altura das notas emitidas. Caso se utilize uma geração constante de ar, como na respiração circular, a nota também pode perdurar indefinidamente.

O presente trabalho utiliza sinais extraídos da base de dados RWC [14], que consiste em um conjunto de notas gravadas individualmente de diversos instrumentos musicais, amostradas em 44,1kHz e 16 bits.

Os sinais utilizados no trabalho foram escolhidos de maneira a formar um con-

junto que compreende diversos tipos de instrumentos, com diferentes características de emissão de nota. Ao longo dos exemplos e testes, serão apresentados formas de onda e comentários sobre sinais advindos dos seguintes instrumentos:

- **Piano**
  - Tipo cordofone;
  - Emissão de altura fixa;
  - Sem controle sobre a altura da nota após o ataque;
- **Violoncelo**
  - Tipo cordofone;
  - Emissão de altura variável;
  - Possibilita controle sobre a altura da nota após o ataque;
- **Flauta**
  - Tipo aerofone;
  - Emissão de altura variável;
  - Possibilita controle sobre a altura da nota após o ataque;
  - Normalmente é tocada com tremolo e vibrato, e proporciona um perfil de energia oscilatório, interessante para os testes;
- **Clarineta**
  - Tipo aerofone;
  - Emissão de altura variável;
  - Possibilita controle sobre a altura da nota após o ataque;
  - Diferentemente da Flauta, a Clarineta quase não possui vibrato, sendo escolhida por proporcionar notas com um comportamento frequencial constante ao longo da existência da nota (mantendo a altura fixa).

1.6 Modelo geral da emissão de notas

Todas as emissões de nota com altura fixa possuem uma característica em comum: a energia da nota é sempre decrescente, pois o instrumentista não possui controle sobre a evolução da mesma. Em contrapartida, quando o executor controla a evolução da nota, ele pode até fazer com que a energia da nota seja crescente ao longo de certo intervalo de tempo.

Essas diferenças de evolução da nota e classificação levam a um modelo simplificado que, em princípio, pode modelar a geração de notas por diversos instrumentos.

O modelo baseia-se na ideia de uma excitação por parte do instrumentista, filtrada pelo corpo do instrumento em dada configuração. Por exemplo, se o ar é injetado de maneira constante através da entrada de uma flauta, as notas são variadas alterando-se a configuração dos furos tapados ou abertos. No caso de se manter a configuração de furos abertos/fechados e se alterar a maneira de injeção de ar, altera-se a emissão das notas. Obviamente, se os furos abertos/fechados são mantidos, a altura da nota emitida é constante; entretanto, a evolução temporal da mesma fica totalmente dependente do sopro do flautista. Esse modelo pode ser ilustrado na Figura 1.3 a seguir:

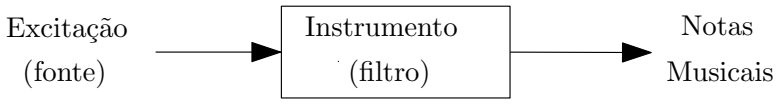


Figura 1.3: Modelo fonte-filtro

1.7 Organização do trabalho

Este trabalho buscará trabalhar com as diversas partes, elementos ou parâmetros que compõem o sinal de música, de modo a fornecer informações detalhadas. Esse elemento a ser trabalhado é a nota musical, definido anteriormente.

Entretanto, um trecho de música normalmente é formado por notas sucessivas e sobrepostas. Assim sendo, o desenvolvimento do trabalho se inicia, no Capítulo 2, com a abordagem um elemento apenas: a nota musical isolada. Essa nota pode ter sido gerada isoladamente ou separada de um sinal mais complexo, de alguma maneira. Nessa etapa são expostos diversos métodos presentes na literatura, sendo apresentados exemplos de aplicação dos mesmos; em seguida, será detalhado o algoritmo de estimação de envoltória proposto, desenvolvido durante o trabalho, discutindo todas as nuances sobre escolha de parâmetros.

No Capítulo 3 será realizada uma discussão de como um algoritmo de separação de fontes se comporta diante do problema da separação de sinais formados por notas sobrepostas e também serão detalhados estudos sobre as possíveis aplicações para as informações extraídas das notas separadas. Nesta etapa será detalhado o método escolhido para separação de fontes, bem como serão expostos os elementos e informações deste visando a atacar o problema da estimação de envoltória num contexto polifônico simples.

Finalmente, o Capítulo 4 tece as considerações finais sobre o trabalho, ressaltando as contribuições desta dissertação e sugerindo os caminhos a serem seguidos a partir dela.

Capítulo 2

Envoltória de uma nota isolada

O objetivo desta etapa é criar uma forma sistemática de obter a envoltória de uma nota isolada que apresente um equilíbrio satisfatório entre suavidade e detalhe.

Conforme exposto anteriormente, uma nota é uma célula sonora individual inserida num contexto temporal, e que pode ter sido gerada por um único instrumento ou por um conjunto de instrumentos, dependendo do que se considera Fonte Sonora em dada situação.

Um algoritmo de estimação de envoltória, não importando o método utilizado, terá em sua saída um levantamento da evolução da intensidade do sinal ao longo do tempo. Um bom método pode ser entendido como o que fornece compromisso entre suavidade e detalhe: detalhes suficientes para capturar as variações de intensidade perceptíveis e suavidade tal que a envoltória não possua descontinuidades de intensidade perceptíveis auditivamente.

O problema da estimação da envoltória vem sendo estudado há algum tempo e já motivou algumas soluções, calcadas em diversas técnicas e abordagens. A seguir são expostos alguns métodos de detecção de envoltória presentes na literatura [15], seguidos de uma nova proposta para solução desse problema.

2.1 Métodos de estimação de envoltória

Dentre muitos métodos existentes, destacam-se a seguir alguns dos mais relevantes e que apresentam os melhores resultados, para que se possa compará-los com o método proposto, que será detalhado mais adiante no capítulo.

Os sinais utilizados em análises como esta normalmente são notas musicais gravadas em formato de arquivos .wav PCM a 44,1kHz e 16 bits. Entretanto, dependendo do método em questão, diversos tipos de pré-processamento podem ser realizados.

Os métodos apresentados neste trabalho são encontrados na literatura utilizando como entrada a versão retificada (de onda completa) do sinal original, o que será mantido.

Comparar envoltórias estimadas através de métodos diferentes é uma tarefa difícil, uma vez que a escolha dos parâmetros deve ser equivalente de modo a permitir tal comparação. Inicialmente, serão apresentados alguns resultados dos métodos, com parâmetros escolhidos manualmente apenas para fins ilustrativos. Em seguida será feita uma comparação mais justa entre eles, considerando um critério comum para avaliação do seu desempenho.

2.1.1 Filtragem Passa-baixas

O meio mais intuitivo de se obter um sinal suave que siga a evolução temporal da forma de onda original é realizar uma filtragem passa-baixas. A ideia é a mesma da demodulação clássica de sinais modulados em amplitude (AM, do inglês *amplitude modulation*) [2], em que a informação desejada encontra-se na amplitude do sinal. No caso da envoltória, deseja-se remover as componentes que carregam em si a altura percebida, deixando apenas a evolução temporal de longo prazo. Neste caso, a informação com altura é a portadora.

Uma vez que se busca remover componentes de altura percebida, as componentes de baixa frequência (não-percebidas como *pitch*) podem ser entendidas como uma visualização da parte "lenta" do sinal. Para obtê-la, filtra-se o sinal passando-o por um filtro passa-baixas, cuja saída é, então, a envoltória do sinal filtrado.

Apesar de ser uma ideia simples, existem muitas variáveis envolvidas, pois vários parâmetros afetam diretamente o resultado final: o tipo de filtro utilizado, sua ordem, sua frequência de corte etc.

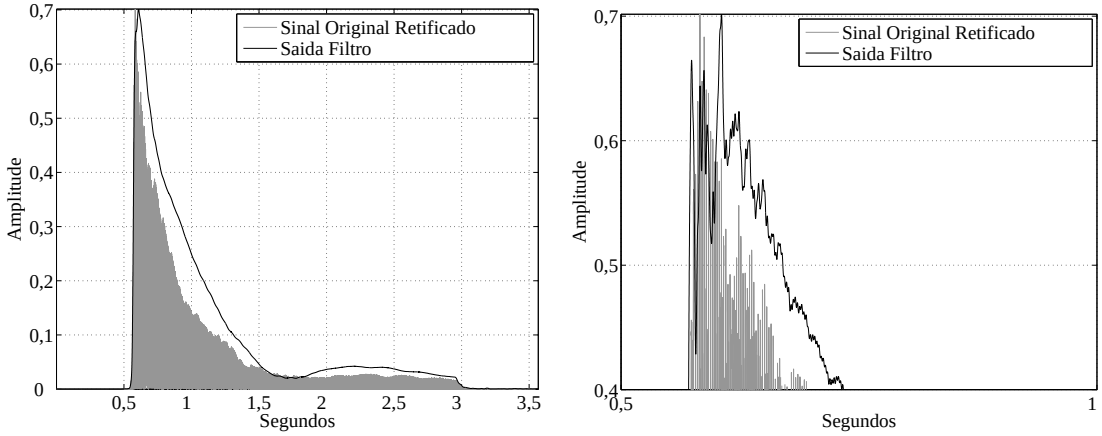
O tipo de filtro escolhido afeta diretamente o resultado, pois cada tipo possui resposta em frequência com *ripples* inerentes na faixa de passagem e/ou rejeição e respostas ao impulso finita (FIR, do inglês *Finite Impulse Response*) ou infinita (IIR, do inglês *Infinite Impulse Response*); diferentes ordens do filtro produzem larguras de faixa de transição diferentes; por fim, a escolha de uma frequência de corte elevada implica a obtenção de uma saída com informações não-desejadas (parte da portadora, como se diria no contexto de comunicações, por exemplo), enquanto uma frequência de corte baixa produz uma saída excessivamente suave, não acompanhando mudanças importantes na amplitude. Essa multiplicidade de opções na escolha dos parâmetros não torna as coisas mais fáceis, se não houver uma forma robusta de escolhê-los para garantir um bom desempenho.

A fim de ilustrar os efeitos discutidos acima, foram projetados dois filtros, sendo um deles com FIR (pelo método de Parks-McClellan) e outro tipo IIR (Chebyshev tipo II) [16], com as especificações abaixo:

- Frequência de amostragem: 44,1kHz
- Frequência do final da faixa de passagem: 10Hz

- Frequência do início da faixa de rejeição: 80Hz
- Máxima atenuação na faixa de passagem: 1dB
- Mínima atenuação na faixa de rejeição: 80dB

Duas envoltórias estimadas com estes filtros são mostradas como exemplo nas Figuras 2.1 e 2.3. Todas as análises foram realizadas utilizando os filtros projetados de modo a se ter ganho de 0dB na faixa de passagem.



(a) Envoltória completa. (b) Detalhe do ataque.

Figura 2.1: Nota Lá4 ($f_0 = 440$Hz) de um piano. Envoltória estimada com filtragem passa-baixas, tipo FIR. Ordem do filtro: 1604.

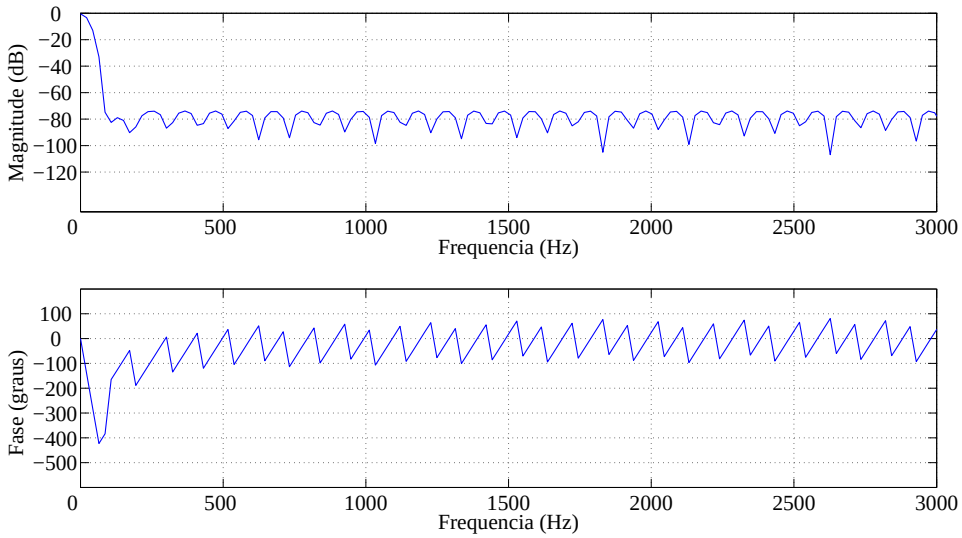
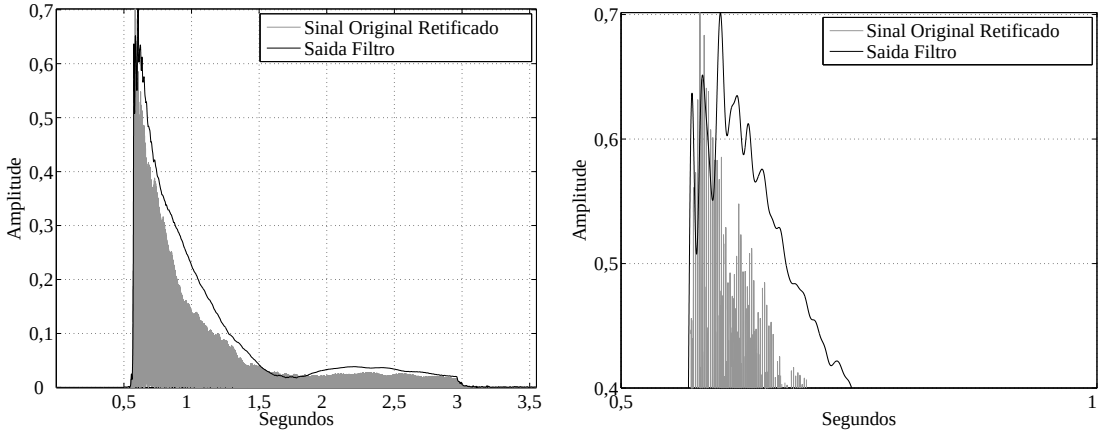


Figura 2.2: Respostas em magnitude e fase do filtro FIR, de ordem 1604, utilizado na Figura 2.1.

Para que se possa observar a variação da suavidade da saída, abaixo seguem os mesmos sinais, porém agora filtrados com frequências de corte maiores, ou seja,



(a) Envoltória completa.

(b) Detalhe do ataque.

Figura 2.3: Nota Lá4 ($f_0 = 440$Hz) de um piano. Envoltória estimada com filtragem passa-baixas, IIR. Ordem do filtro: 5.

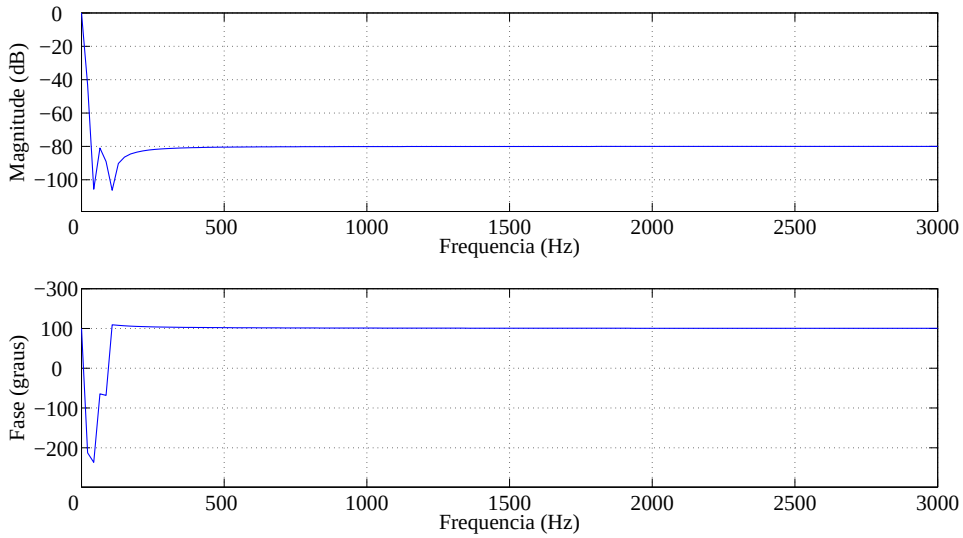


Figura 2.4: Respostas em magnitude e fase do filtro IIR, de ordem 5, utilizado na Figura 2.3.

conteúdo espectral mais amplo estará presente na envoltória resultante. Abaixo seguem as novas especificações dos filtros, mudando apenas seus limites frequenciais:

- Frequência de amostragem: 44,1kHz
- Frequência do final da faixa de passagem: 200Hz
- Frequência do início da faixa de rejeição: 500Hz
- Máxima atenuação na faixa de passagem: 1dB
- Mínima atenuação na faixa de rejeição: 80dB

As envoltórias estimadas com estes filtros são mostradas nas Figuras 2.5 e 2.7. Dois aspectos ligados ao projeto do filtro devem ser considerados:

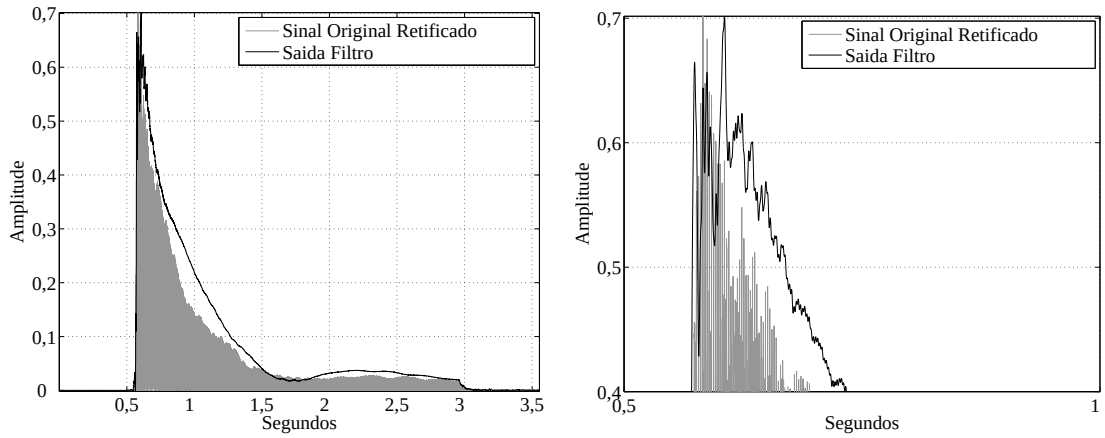
1. A ordem do filtro: quanto maior a ordem, mais lentamente ele responderá a modificações no sinal. Isso pode ser observado na Figura 2.1.

2. O atraso de grupo variável com a frequência, que pode ser observado na Figura 2.3.

Uma vez que a ideia principal da filtragem é eliminar frequências altas (responsáveis por oscilações indesejadas na envoltória), essa escolha é dependente da frequência fundamental do sinal analisado (f_0).

Ao se escolher uma frequência de corte em torno de 20Hz, todas as componentes tonais audíveis seriam eliminadas, sobrando apenas a parcela mais lenta e imperceptível como nota musical; entretanto, como se pode observar da Figura 2.1, por exemplo, detalhes são perdidos, pois a envoltória apresenta-se demasiadamente suave. Para evitar isso, um critério possível seria eliminar apenas as componentes abaixo da f_0, o que demanda seu conhecimento. Não se pode requerer o pré-conhecimento de f_0 ao longo de todo o processamento.

O método de estimação de envoltória deveria ser robusto o suficiente para não depender de conhecimento prévio sobre o sinal a ser analisado; nesse sentido, a filtragem passa-baixas apresenta uma dificuldade.



(a) Envoltória completa.

(b) Detalhe do ataque.

Figura 2.5: Nota Lá4 ($f_0 = 440$Hz) de um piano. Envoltória estimada com filtragem-passa baixas, tipo FIR. Ordem do filtro: 382.

Analisando as envoltórias estimadas através de filtragem passa-baixas, nota-se que, nos casos em que a envoltória consegue "acompanhar" a subida rápida no momento do *onset*, no transitório da nota (o ponto mais alto nas Figuras 2.5 e 2.7) a envoltória estimada é extremamente ruidosa, não apresentando grau de suavidade adequado para descrever as regiões de decaimento da nota.

Em contrapartida, nos casos em que um grau elevado de suavidade aparentemente constante aparece ao longo da duração de toda a nota, a envoltória estimada não acompanha variações rápidas na transição (como nas Figuras 2.1 e 2.3).

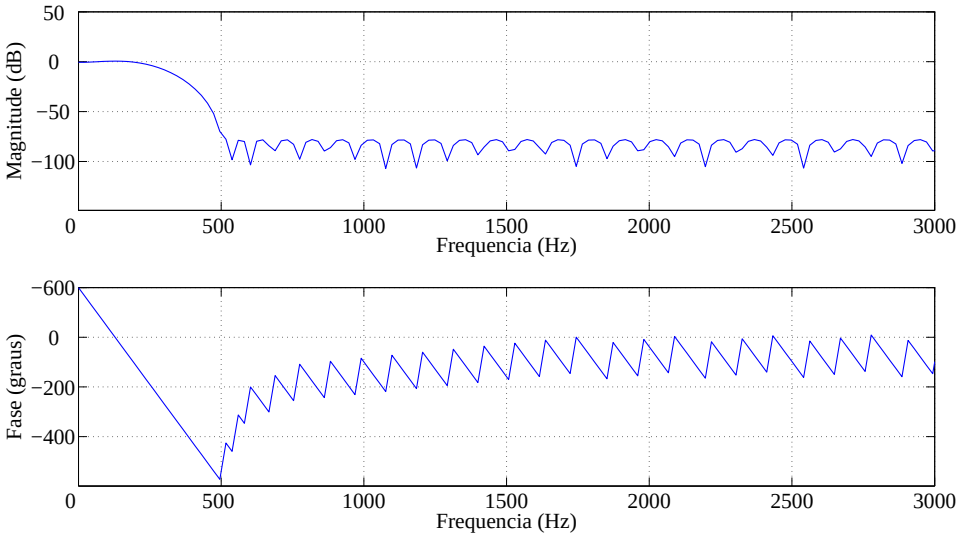
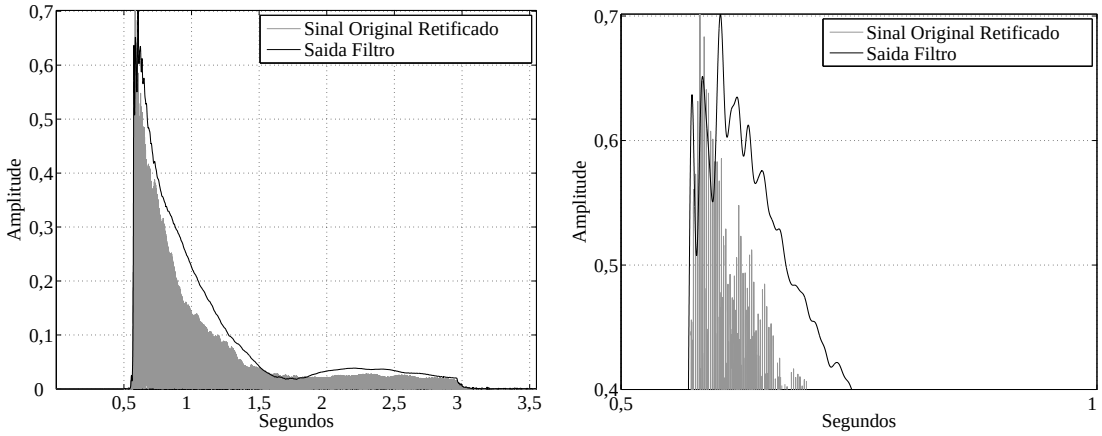


Figura 2.6: Respostas em frequência e fase do filtro FIR, de ordem 382, utilizado na Figura 2.5.



(a) Envoltória completa.

(b) Detalhe do ataque.

Figura 2.7: Nota Lá4 ($f_0 = 440$Hz) de um piano. Envoltória estimada com filtragem-passa baixas, IIR. Ordem do filtro: 8.

Vale ressaltar que a ordem deste segundo filtro é maior que a do filtro IIR anterior apesar de, em Hz, possuir faixa de transição maior. Na verdade, o que importa no caso de filtros de Chebyshev é a razão entre os limites superior (F_{stop}) e inferior (F_{pass}) da faixa de transição, já que sua ordem é proporcional ao inverso do $\cosh^{-1}\left(\frac{F_{stop}}{F_{pass}}\right)$ [16].

2.1.2 Valor Quadrático Médio da Energia (RMS)

O valor quadrático médio (RMS, do inglês *root mean square*) é, possivelmente, o método mais popular [17] para se estimar a evolução temporal da energia de um sinal. Ele pode ser obtido através da aplicação sucessiva da Equação (2.1).

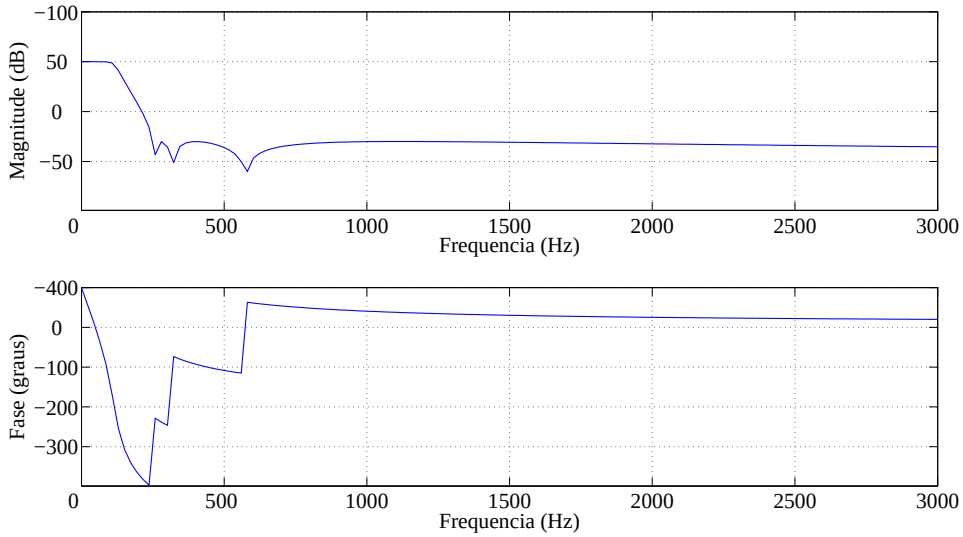


Figura 2.8: Respostas em frequência e fase do filtro IIR, de ordem 8, utilizado na Figure 2.7.

$$RMS(n) = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} s^2\left(n - \frac{N-1}{2} + i\right)} \tag{2.1}$$

onde $s(n)$ é o sinal original após retificação de onda completa, cuja potência média é calculada dentro de uma janela deslizante de comprimento N (ímpar) antes de sofrer a aplicação de uma raiz quadrada. Para tal cálculo, qualquer tipo de janela pode ser usado [18], embora a mais comum seja a retangular. Se for desejado manter a taxa de amostragem do sinal, a janela pode obedecer a um deslizamento de uma amostra apenas por vez.

A ideia principal é rastrear a variação lenta da potência média local do sinal, como um estimador da envoltória.

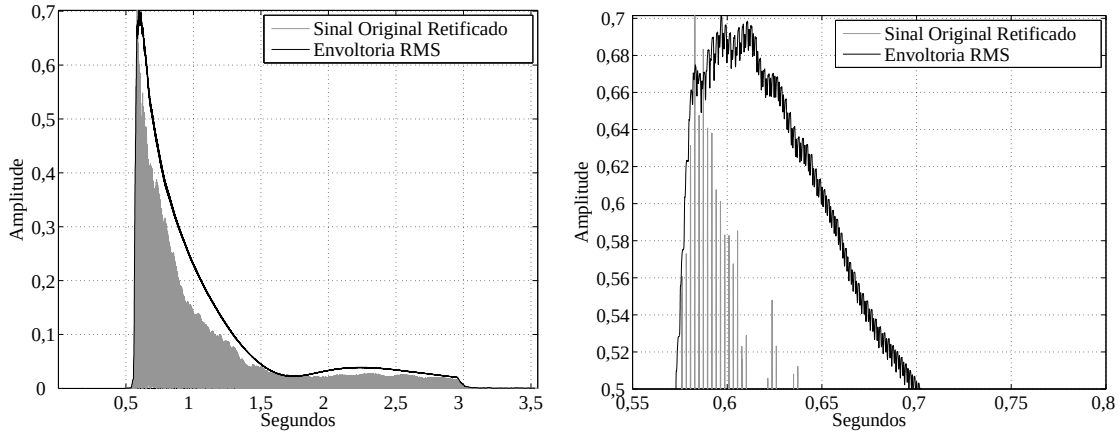
O cálculo do valor RMS atua como uma espécie de filtro passa-baixas (no domínio da potência) que suaviza o sinal, portanto é uma filtragem não-linear de $s(n)$. Como num filtro passa-baixas de fato, o tamanho da janela de cálculo afeta diretamente a suavidade do resultado final. Uma janela pequena produz um resultado que "acompanha" mais de perto as variações do sinal, porém carrega informações não-desejadas; por sua vez, uma janela excessivamente grande produz uma envoltória suave, que porém pode possuir pouca relação com o sinal original.

Por construção, o resultado do método depende da frequência fundamental f_0 do sinal analisado, uma vez que o parâmetro de ajuste do método afeta sua frequência de corte. Portanto é necessária informação prévia do sinal analisado (ou da estimação automática da f_0).

Para a visualização da influência do parâmetro acima exposto, duas envoltórias foram calculadas, empregando janelas de comprimento diferente, deslizando esta

janela amostra a amostra.

A Figura 2.9 ilustra a envoltória calculada com janela de comprimento 20ms.

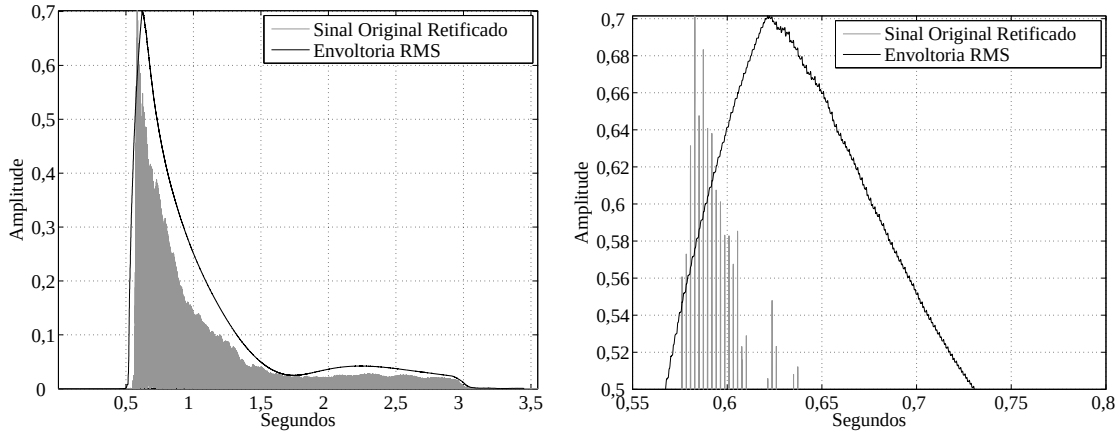


(a) Envoltória completa.

(b) Detalhe do ataque.

Figura 2.9: Nota Lá4 ($f_0 = 440$Hz) de um piano. Envoltória estimada com cálculo do valor RMS, utilizando janela de comprimento 20ms.

A Figura 2.10 ilustra a envoltória calculada com janela de comprimento 100ms.



(a) Envoltória completa.

(b) Detalhe do ataque.

Figura 2.10: Nota Lá4 ($f_0 = 440$Hz) de um piano. Envoltória estimada com cálculo do valor RMS, utilizando janela de comprimento 100ms.

Conforme pode ser visto nas Figuras 2.9 e 2.10, quanto maior o tamanho da janela, maior a suavidade da envoltória resultante; entretanto, essa suavidade é conseguida perdendo-se (ainda mais) o acompanhamento do contorno do sinal. Um valor RMS calculado com poucas amostras "desenham" melhor a envoltória, porém

esta apresenta-se mais ruidosa, enquanto que mais amostras no cálculo implicam uma envoltória disforme.

Assim sendo, todos os problemas reportados na análise do método de filtragem passa-baixas se aplicam aqui: desde o problema com os transitórios da nota, gerando envoltórias ruidosas (como pode ser observado na Figura 2.9) ou que não acompanham as variações abruptas no momento do *onset* até a "lentidão" da envoltória estimada em acompanhar a variação da amplitude do sinal analisado.

2.1.3 Predição Linear no Domínio da Frequência (FDLP)

A predição linear tradicional [19] estima a envoltória espectral a partir do sinal no domínio do tempo. A ideia básica do *Frequency-Domain Linear Prediction* (FDLP) [20] é explorar a dualidade tempo-frequência para extrair a amplitude temporal a partir da aplicação da predição linear sobre a representação espectral do sinal de entrada.

Nesse caso em particular, é adequada a utilização de uma representação espectral que possui apenas valores reais. A fim de satisfazer essa condição, o método emprega a Transformada Discreta de Cossenos (DCT) [21] em quadros longos e aplica a predição linear sobre a saída da DCT. A envoltória de fato é a resposta em frequência determinada pelos polos obtidos através do modelo de predição linear; sendo assim, a quantidade de polos do modelo (que deve ser previamente informada) afeta diretamente a suavidade da envoltória obtida, ou seja, uma quantidade excessiva de polos produz uma envoltória com muitas oscilações (seguindo o *pitch*), e um número baixo de polos produz um resultado demasiadamente suave, deixando para trás grande parte da característica temporal do sinal.

O FDLP é um método desenvolvido para aplicações em processamento de fala, com banda reduzida, e inclui-se tal método para fins de comparação, apenas. O método foi desenvolvido para ser aplicado quadro-a-quadro, e resultados de sua aplicação são mostrados nas Figuras 2.11 e 2.13.

Para efeito de ilustração, aplicou-se o método FDLP à nota Dó3 de uma Flauta, com 4 e 16 polos, respectivamente nas Figuras 2.11, 2.12 e 2.13. Em ambos os casos a janela utilizada foi a retangular, de comprimento 20ms e sobreposição de 50%.

Observando as Figuras 2.11 e 2.13 nota-se que a envoltória estimada pelo método possui um formato semelhante à variação de amplitude do sinal original, porém mostra-se demasiadamente ruidosa (ver Figura 2.12), o que é indesejável para estimação de envoltória.

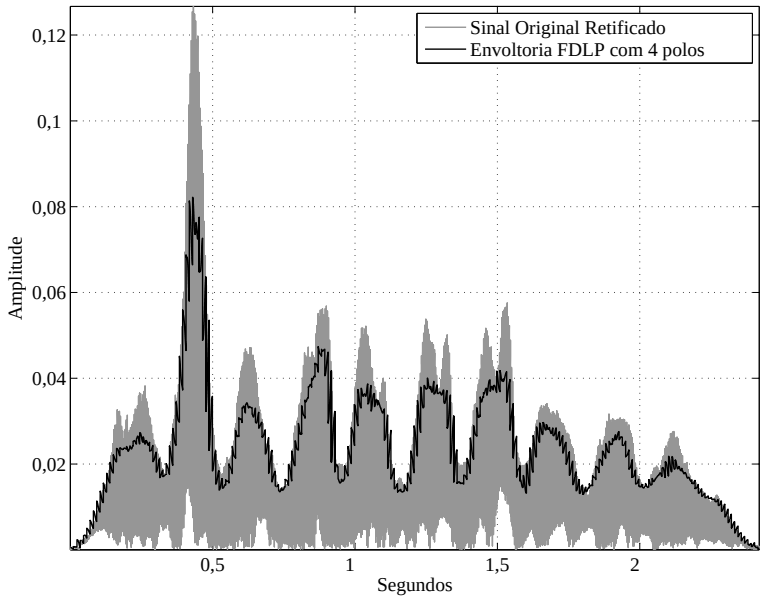


Figura 2.11: Nota Dó3 ($f_0 = 138,81$Hz) de uma flauta. Envoltória estimada através do método FDLP com 4 polos.

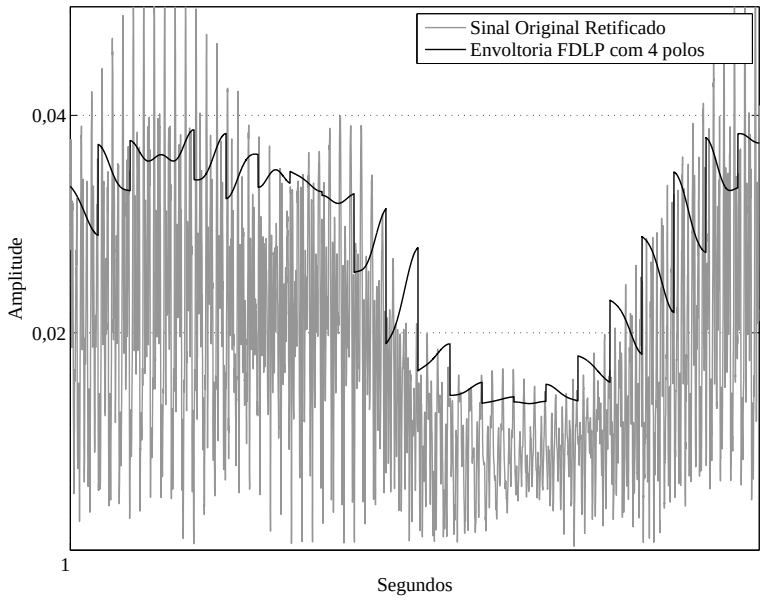


Figura 2.12: Nota Dó3 ($f_0 = 138,81$Hz) de uma flauta. Detalhe das descontinuidades da envoltória estimada.

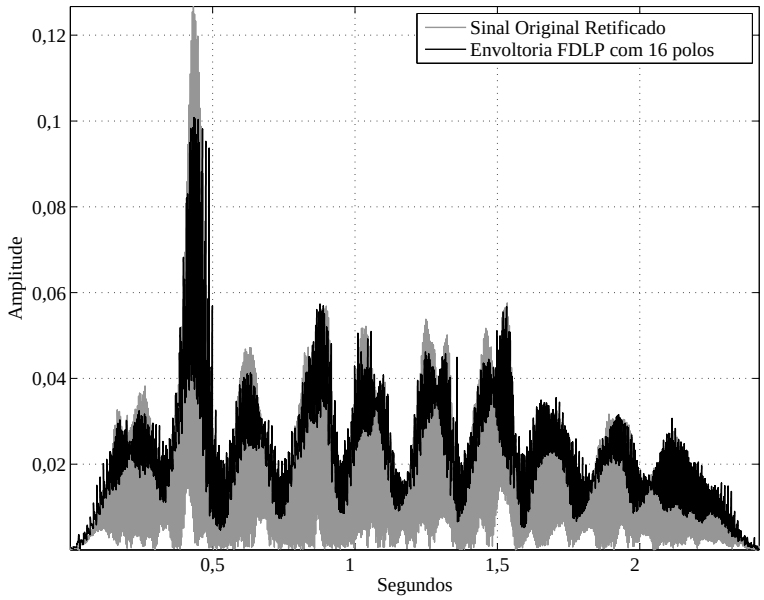


Figura 2.13: Nota Dó3 ($f_0 = 138,81$Hz) de uma flauta. Envoltória estimada através do método FDLP com 16 polos.

2.1.4 *True Amplitude Envelope* (TAE)

True Envelope [22] é um método desenvolvido para estimação da envoltória espectral de um sinal que mostrou um desempenho superior ao da predição linear [19] ou dos métodos *cepstrais* tais como *discrete cepstrum* [23].

O *cepstrum* é uma operação matemática que é definida como a Transformada de Fourier Inversa do logaritmo do espectro do sinal. O nome *cepstrum* é uma inversão da ordem das primeiras quatro letras de *spectrum*. Existem diversos tipos de *cepstrum*; no caso do *True Envelope* é empregado o *real cepstrum*, que utiliza a função logarítmica aplicada sobre o espectro de magnitude do sinal.

O método consiste em, iterativamente, calcular o *cepstrum* [24], que será a primeira estimação da envoltória, e suavizá-lo utilizando uma técnica chamada *cepstral smoothing*, eliminando algumas das suas componentes. Um diagrama de blocos da operação é mostrado na Figura 2.14.

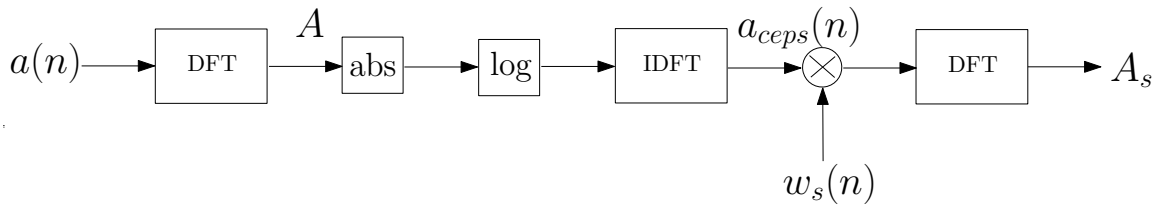


Figura 2.14: Diagrama de blocos do *cepstral smoothing*.

Sendo $a(n)$ um sinal qualquer, no domínio do tempo, A sua versão no domínio da frequência, e $a_{ceps}(n)$ seu *cepstrum*, a envoltória espectral obtida pelo *cepstral smoothing* pode ser definida como:

$$A_s = DFT[w_s(n) \times DFT^{-1} \log_{10}(|A|)] \tag{2.2}$$

onde w_s é uma janela que atua como filtro passa-baixas no domínio do *cepstrum*. Essa janela em muitos casos é quadrada, como mostrado pela Equação 2.3 abaixo:

$$w_s(n) = \begin{cases} 1 & |n| < n_c \\ 0,5 & |n| = n_c \\ 0 & |n| > n_c \end{cases} \tag{2.3}$$

onde n_c é o número de componentes que se deseja eliminar. Quanto mais componentes são retiradas, maior a suavidade do sinal resultante.

Essa suavização fará com que a estimação da envoltória "preencha" os vales do sinal original, criando um sinal suavizado. Repete-se esse processo até que se atinja um grau de suavização desejado para a envoltória espectral.

Uma forma de determinar a envoltória temporal foi proposta em [15] e consiste em empregar o *dual* do *True Envelope*, ou seja: em lugar de aplicar o método sobre

um sinal no domínio da frequência, um sinal no domínio temporal é processado pelo algoritmo. Dessa forma, a envoltória calculada pelo método será a temporal, e não a espectral.

Basicamente, realiza-se um pré-processamento do sinal original, deixando-o com a aparência da magnitude de um espectro e, sobre esse sinal, aplica-se o *True Envelope*. Esse pré-processamento é descrito a seguir:

Denotando $x(n)$, de comprimento M, como sendo o sinal original a ser processado, os passos do pré-processamento são os seguintes:

- Primeiramente cria-se uma versão auxiliar do sinal passando-o por um retificador de onda completa:

$$s(n) = |x(n)| \tag{2.4}$$

- Completa-se o sinal $s(n)$ com zeros (*zero-padding*) até que seu comprimento seja uma potência de 2 (a mais próxima possível):

$$s_{zp}(n) = \begin{cases} s(n) & t \leq M \\ 0 & M < t \leq 2^{\lceil \log_2 M \rceil} - M \end{cases} \tag{2.5}$$

- O novo sinal $s_{zp}(n)$, que agora possui comprimento $N = 2^{\lceil \log_2 M \rceil}$, sofre finalmente uma extensão simétrica modo a imitar as frequências negativas. Essa operação, obviamente, dobra o tamanho do sinal $s_{zp}(n)$:

$$s_{tr}(n) = \begin{cases} s_{zp}(n) & n \leq N \\ s_{zp}(2N - n) & N < n \leq 2N - 1 \end{cases} \tag{2.6}$$

  O sinal $s_{tr}(n)$ é então utilizado como entrada do algoritmo *True Envelope* original, conforme descrito em [15].

De modo a ilustrar esse processo, a Figura 2.15 mostra a forma de onda de uma nota Lá 4 ($f_0 = 440$Hz) de piano e a Figura 2.16 exibe o mesmo sinal após sofrer o pré-processamento descrito acima.

O desafio do TAE é encontrar a ordem ótima para atingir a relação suavidade/detalhe desejada.

O autor de [15] recomenda o uso da ordem

$$O = \alpha \frac{f_0}{f_s} \times N, \tag{2.7}$$

onde f_s é a frequência de amostragem N é o número de amostras (comprimento) do sinal a ser considerado. O fator $0 < \alpha \leq 1$ limita o número de oscilações por segundo a αf_0.

As Figuras 2.17 e 2.18 ilustram uma envoltória estimada através deste método, empregando $\alpha = 1$; para fins de comparação, as Figuras 2.19 e 2.20 mostram a mesma nota, porém com $\alpha = 1/4$.

Uma característica importante do método TAE é que a escolha da ordem dita diretamente a quantidade de "ondulações" que a envoltória final poderá ter. Conforme observado nas figuras acima, quanto maior a ordem, mais "ondulações" estarão pre-

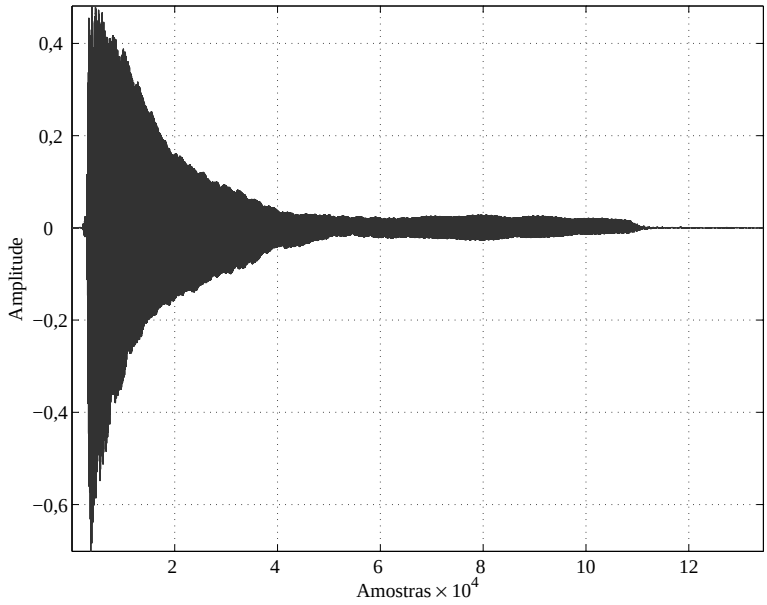


Figura 2.15: $x(n)$ – Forma de onda da nota Lá4 de um piano.

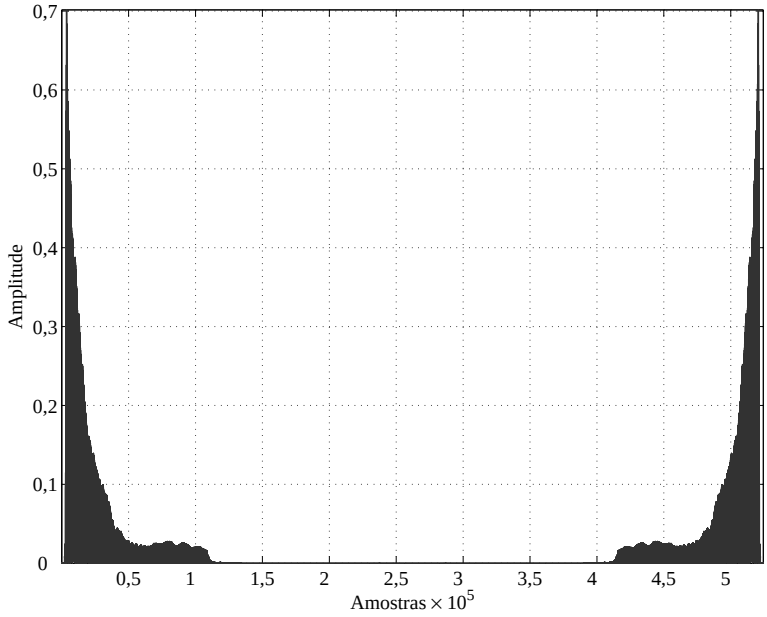


Figura 2.16: $s_{tr}(n)$ – Forma de onda da nota Lá4 de um piano após o pré-processamento.

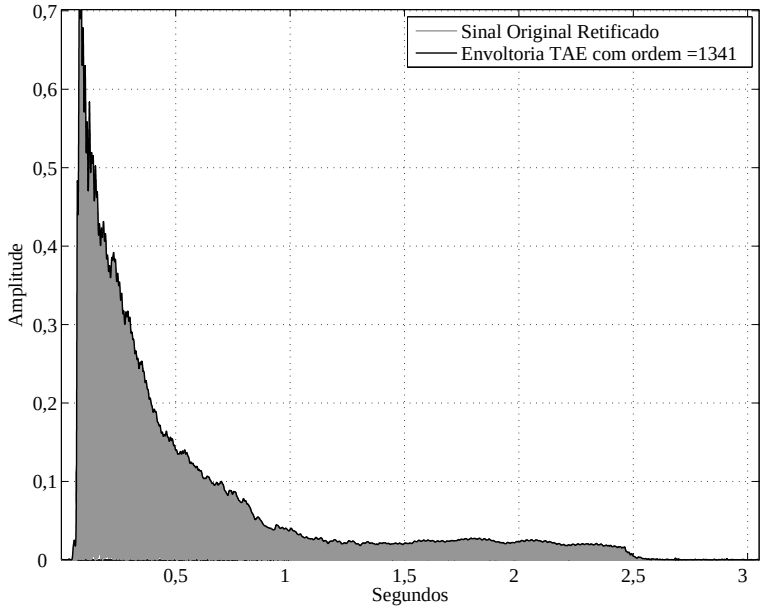


Figura 2.17: Nota Lá4 ($f_0 = 440$Hz) de um piano. Envoltória estimada através do método TAE com ordem proporcional à frequência fundamental do sinal.

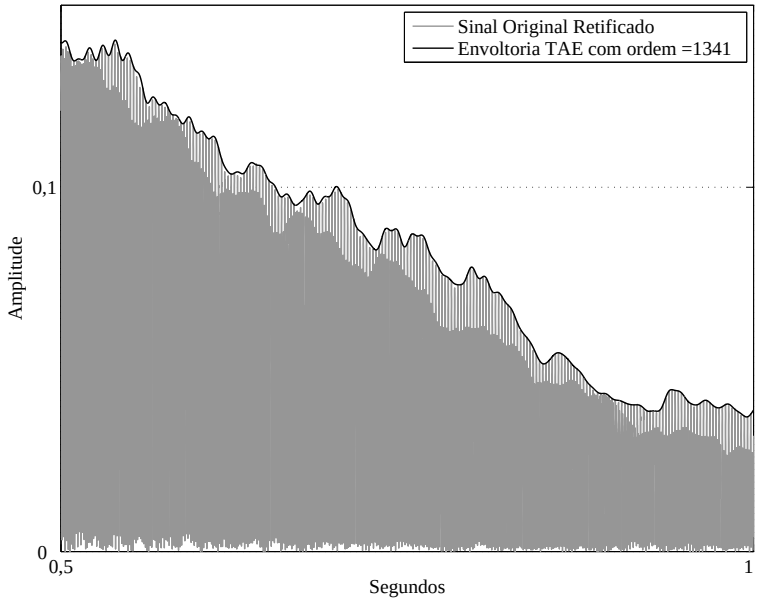


Figura 2.18: Detalhe da envoltória da nota Lá4 de um piano. Envoltória estimada através do método TAE com ordem proporcional à frequência fundamental do sinal.

sentes e mais "acidentes" serão descritos pela envoltória pois, para uma dada ordem, o número de ondulações é sempre fixo. Se se escolhe uma ordem excessivamente elevada, a envoltória será ruidosa; em caso contrário, a envoltória estará toda acima da forma de onda e ainda apresentando ondulações que claramente não pertencem à envoltória do sinal.

Por exemplo, ao analisar as Figuras 2.17 e 2.18 nota-se que a envoltória estimada apresenta um grau de suavidade visivelmente adequado, estando bem apoiada sobre

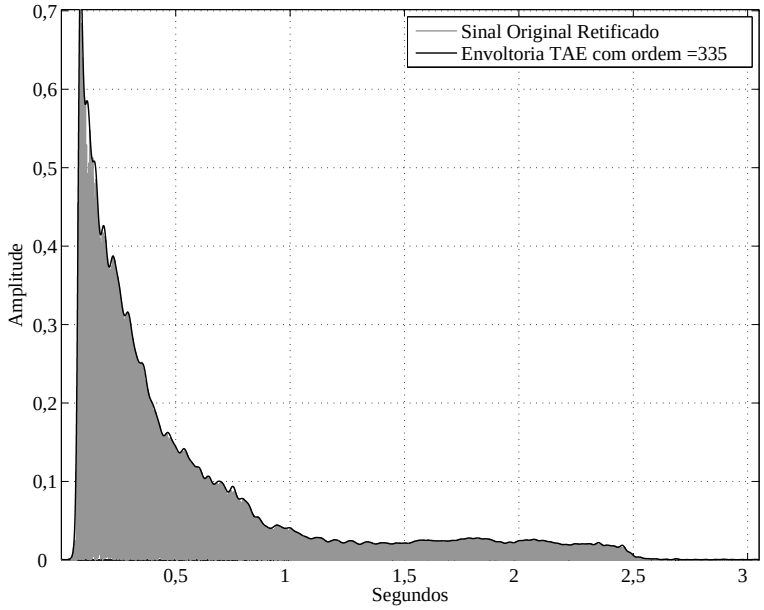


Figura 2.19: Nota Lá4 ($f_0 = 440$Hz) de um piano. Envoltória estimada através do método TAE com ordem proporcional a 1/4 da frequência fundamental do sinal.

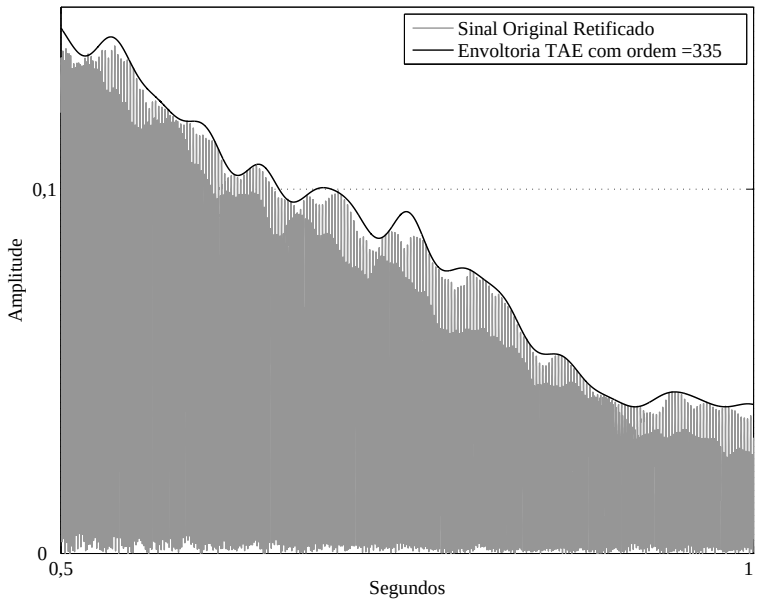


Figura 2.20: Detalhe da envoltória da nota Lá4 de um piano. Envoltória estimada através do método TAE com ordem proporcional a 1/4 da frequência fundamental do sinal.

a forma de onda do sinal. Entretanto, as Figuras 2.19 e 2.20 mostram o caso em que a ordem foi escolhida erroneamente.

Dentre os métodos apresentados até aqui, o TAE é o que apresenta os melhores resultados, sendo assim o mais indicado para ser comparado ao método que será proposto posteriormente. A maior dificuldade do método é sua dependência da f_0, que o deixa dependente de informação prévia sobre o sinal analisado ou demandando

uma estimação automática da f_0, que pode ser imprecisa.

2.2 Abordagem proposta neste trabalho

Após a exposição de alguns métodos de estimação de envoltória presentes na literatura, esta seção se destina a detalhar a abordagem proposta neste trabalho para o problema da estimação de envoltória. Em todos os casos, considera-se uma nota isolada e faz-se um pré-processamento a fim de criar um sinal auxiliar, que é utilizado como entrada do algoritmo de estimação da envoltória.

Nesse trabalho, o sinal auxiliar é calculado fazendo-se a retificação de onda completa no sinal de entrada. Adota-se esse procedimento para possibilitar a comparação com outros métodos da literatura.

A base do método proposto é a Morfologia Matemática (MM) [25], uma teoria utilizada em processamento de imagens que se mostrou adequada para o problema em questão. A fim de apresentar o método, faz-se uma breve explanação sobre algumas ferramentas dessa família; em seguida, é detalhado o método utilizado na estimação da envoltória de uma nota musical isolada.

2.2.1 Morfologia Matemática

Morfologia Matemática (MM) pode ser definida como uma técnica para análise de estruturas geométricas. É chamada morfologia porque reside na análise da forma dos objetos. É matemática porque é baseada em teoria dos conjuntos, geometria integral e em reticulados *lattice* [25]. A MM não é apenas uma teoria, mas também uma ferramenta largamente utilizada em análise de imagens.

A base da morfologia consiste em extrair as informações relativas à geometria e à topologia de um conjunto desconhecido pela transformação através de outro conjunto bem-definido, chamado *elemento estruturante*.

O conjunto desconhecido poderia ser uma imagem (conforme os exemplos de operações que serão explanados na seção a seguir), uma forma de onda retificada (como será o caso da aplicação no presente trabalho) etc.

As operações são ilustradas com exemplos encontrados na literatura onde o conjunto desconhecido é bidimensional.

Neste trabalho será utilizada uma ferramenta específica da MM: uma operação chamada fechamento, que é a composição de duas operações básicas, a erosão e a dilatação. Estas duas operações básicas serão detalhadas a seguir.

2.2.2 Operações básicas em Morfologia Matemática

A partir da definição do tamanho e da forma de um chamado elemento estruturante, podem ser realizadas diversas operações. Destacamos as mais importantes a fim de introduzir a operação escolhida como base do método proposto. Denotamos $\mathbf{X}$ (que posteriormente será definido a partir da versão retificada do sinal original) o conjunto onde serão aplicadas as operações, e $\mathbf{B}$ o elemento estruturante nela envolvido, ou seja, o conjunto definido que introduz a forma e o tamanho do operador. Segue uma pequena explanação intuitiva sobre algumas destas operações.

Erosão

Define-se a operação de erosão como:

$$\mathbf{E}(\mathbf{X}) = \{p_i \mid \mathbf{B}(p_i) \subseteq \mathbf{X}\}, \tag{2.8}$$

onde $\mathbf{E}(\mathbf{X})$ é o conjunto resultante da erosão do conjunto $\mathbf{X}$ pelo elemento estruturante $\mathbf{B}(p_i)$, centrado em p_i. Uma notação simplificada também pode ser utilizada:

$$E = X \ominus B. \tag{2.9}$$

Intuitivamente, a erosão de $\mathbf{X}$ por $\mathbf{B}$ é o conjunto de todos os pontos alcançados pelo centro de $\mathbf{B}$ quando $\mathbf{B}$ se move no interior de $\mathbf{X}$, sem sair dele. Isso leva a uma diminuição no seu tamanho original. Dessa característica vem o nome erosão. A Figura 2.21 abaixo ilustra esse processo:

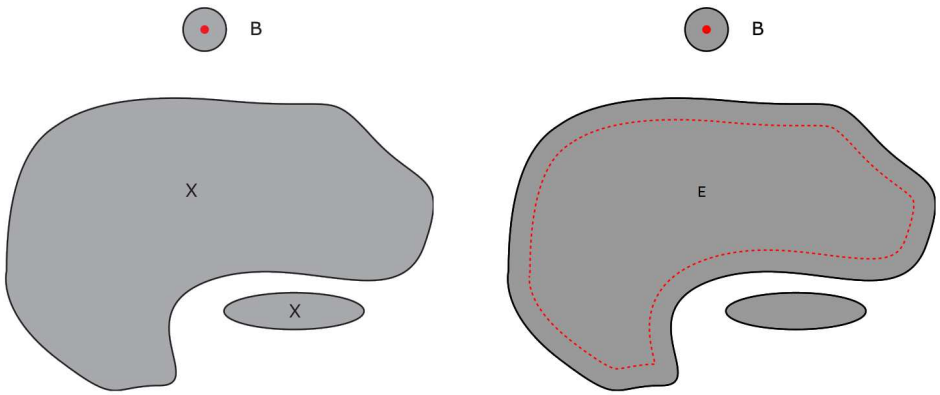


Figura 2.21: Exemplo de erosão (extraído de [3]). A forma final é o conjunto cinza interior à linha pontilhada vermelha

Dilatação

Define-se a operação de dilatação como:

$$\mathbf{D}\,(X) = \cup\{\mathbf{B}\,(p_i) \mid p_i \in \mathbf{X}\} \tag{2.10}$$

onde $\mathbf{D}\,(X)$ é o conjunto resultante da dilatação do conjunto $\mathbf{X}$ pelo elemento estruturante $\mathbf{B}\,(p_i)$, centrado em p_i. Utilizando uma notação simplificada:

$$D = X \oplus B. \tag{2.11}$$

A dilatação de **X** por **B** pode ser entendida como sendo o conjunto dos pontos delimitados pelo centro de B quando B se move sobre o exterior de X, interceptando X. Isso faz com que o conjunto original aumente de tamanho, como mostra a Figura 2.22. Dessa característica vem o nome de dilatação.

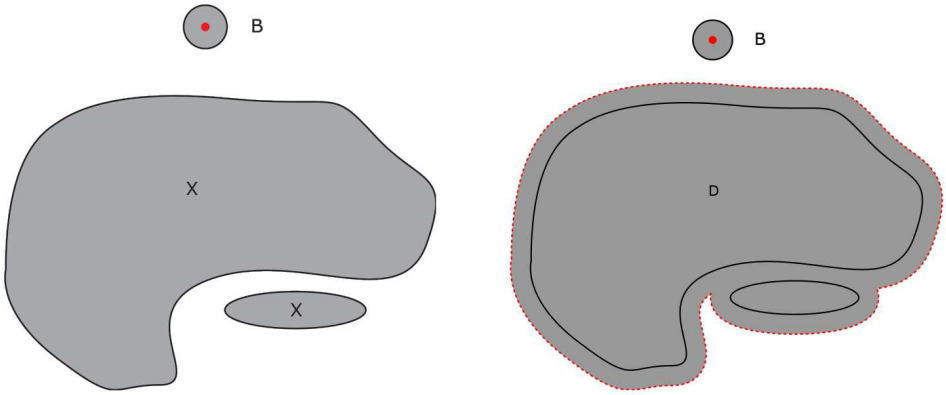


Figura 2.22: Exemplo de dilatação (extraído de [3]). A forma final é o conjunto cinza delimitado pela linha pontilhada vermelha.

Abertura

Denota-se a operação abertura como:

$$D = X \circ B = (X \ominus B) \oplus B \tag{2.12}$$

A abertura é uma operação derivada das outras duas, uma vez que é feita uma erosão seguida de uma dilatação. Intuitivamente, o elemento estruturante $\boldsymbol{B}$ "varre" o interior de $\boldsymbol{X}$, sem cruzar a fronteira, moldando a borda de $\boldsymbol{X}$ ao formato da borda de $\boldsymbol{B}$. Essa operação é ilustrada pela Figura 2.23:

Fechamento

Denota-se a operação fechamento como:

$$D = X \bullet B = (X \oplus B) \ominus B \tag{2.13}$$

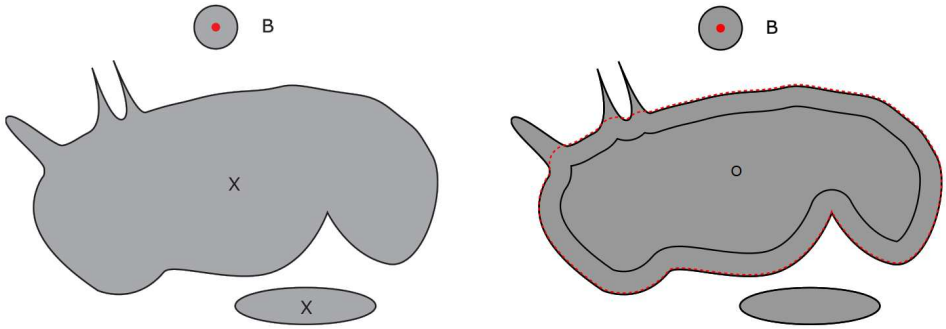


Figura 2.23: Exemplo de abertura (extraído de [3]). A forma final é a região limitada pela linha vermelha pontilhada.

Analogamente à abertura, o fechamento é uma sequência de operações, pois é feita uma dilatação seguida de uma erosão. A estrutura **B** tangencia as bordas de **X** de modo que, quanto menor o elemento estruturante **B**, mais próximo do formato da borda de **X** estará a nova fronteira. De outra forma, o elemento estruturante **B** molda a fronteira de **X**, porém através do seu exterior. A Figura 2.24 ilustra esse processo:

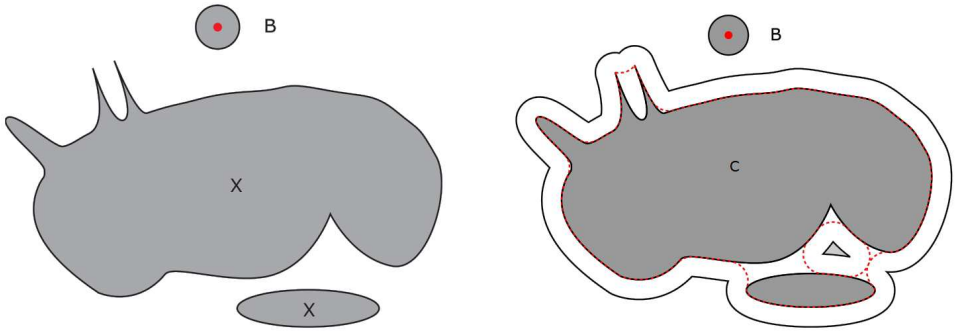


Figura 2.24: Exemplo de fechamento (extraído de [3]). A forma final é a região limitada pela linha vermelha pontilhada.

2.3 Método Proposto

A abordagem proposta no trabalho utiliza a operação de fechamento a fim de encontrar a envoltória temporal de notas musicais, aproveitando a característica dessa operação de contornar o exterior do conjunto que sofre o fechamento.

Visão Geral do Algoritmo

1. É gerado um sinal auxiliar a partir do sinal original a ser processado, para servir como entrada do sistema.

2. Determina-se o elemento estruturante, de acordo com algum critério.

3. Realiza-se o fechamento do elemento estruturante sobre o sinal de entrada.

4. Efetua-se um pós-processamento sobre o resultado dos fechamentos para atingir o desejado grau de suavidade.

O conjunto a ser considerado poderia ser a própria forma de onda da nota musical, sua versão retificada ou mesmo um sinal analítico real proveniente da mesma.

Especificamente nos exemplos mostrados neste trabalho, o conjunto processado foi um sinal auxiliar gerado através da retificação de onda completa do sinal original para possibilitar a comparação com os métodos mais comuns na literatura.

No caso do método proposto, em se tratando de um sinal unidimensional (que é o caso de formas de onda de sinais musicais), o elemento estruturante escolhido foi uma linha paralela ao eixo X. A operação de abertura é calculada, então, utilizando-se simplesmente um Filtro Unidimensional [26] definido a seguir.

Dada uma sequência $x_0, ..., x_{n-1}$, e um inteiro $p > 1$, o resultado do fechamento é a sequência

$$y_i = \max_{0 \leq j \leq p} x_{i+j} \tag{2.14}$$

para $i = 0, ..., n - p$, onde p é o comprimento do trecho da sequência a ser analisada (ou seja, o comprimento do elemento estruturante).

O sinal resultante do fechamento é uma estimativa inicial da envoltória, ainda constante por partes.

No pós-processamento, utilizam-se os pontos onde essa estimativa inicial toca a forma de onda do sinal de entrada como "âncoras" para uma interpolação. O tipo de interpolação adotado no algoritmo proposto é o *Piecewise Cubic Hermite Interpolating Polynomial* (PCHIP) [27], cujo sinal resultante preserva a forma e a monotonicidade do sinal original.

O resultado dessa interpolação é a estimação final da envoltória.

2.3.1 Comprimento do Elemento Estruturante

A estrutura escolhida pode, a princípio, possuir qualquer forma ou tamanho. O caso do presente trabalho envolve apenas um conjunto unidimensional (forma de onda retificada da nota musical), de modo que é aceitável aplicar um elemento estruturante em forma de *linha*. Assim sendo, o único parâmetro a ser definido é o *comprimento* da linha.

O comprimento do elemento estruturante é um parâmetro que influencia diretamente na forma final da envoltória a ser calculada. Uma vez que a abordagem proposta utiliza o fechamento, o elemento estruturante se deslocará na "superfície" da forma de onda e se encaixará (ou não) em vales da forma de onda conforme o

comprimento da estrutura. Isso resulta em um efeito interessante: caso um elemento estruturante excessivamente curto seja utilizado, o resultado poderá ser um contorno "acidentado", pois o elemento estruturante "se encaixará" em mais vales; por sua vez, um elemento estruturante muito longo resulta em um contorno diferente da forma original da forma de onda, uma vez que apenas os vales mais largos serão atingidos pelo elemento estruturante. A Figura 2.25 ilustra essa diferença.

Assim sendo, o próximo passo é definir qual o comprimento do elemento estruturante a ser utilizado em cada caso, em cada nota da qual se deseja extrair a envoltória.

A ideia mais simples e direta seria fixar um comprimento único para todas as notas analisadas. Nesse caso, notas com alturas diferentes (consequentemente, formadas por f_0 diferentes) seriam tratadas da mesma maneira. Se ajustássemos o comprimento para obter uma certa suavização para notas graves, fatalmente as notas agudas seriam excessivamente suavizadas; caso ajustássemos a suavização para notas agudas, as notas graves teriam uma envoltória demasiadamente ruidosa. As Figuras a ilustram esse comportamento.

Essas características são consequência do conteúdo espectral de cada nota, ou seja, notas mais graves possuem componentes de mais baixa frequência, possuindo períodos de maior duração. Com um tamanho fixo, o elemento estruturante "se encaixa" em mais vales nas formas de onda dessas notas mais graves do que nas formas de onda das notas agudas. Uma discussão mais profunda acerca do critério de escolha desse comprimento é realizada na Seção 2.4, uma vez que um comprimento fixo de estrutura não se mostrou adequado.

O sinal de saída obrigatoriamente é constante por partes, já que o elemento estruturante é uma linha. Esse fato explica a característica visual do sinal resultante do fechamento, que são as mudanças abruptas de nível em forma de "escada". Claramente essa estimativa de envoltória não é desejável, o que demanda um pós-processamento visando a tornar esse sinal resultante uma envoltória "aceitável". Mais adiante será discutido o que se poderia considerar uma envoltória "aceitável".

2.3.2 Efeito do pós-processamento da saída da operação morfológica

A aplicação pura e simples do elemento estruturante unidimensional não se mostra muito adequada pois não é suave o suficiente, conforme se pode observar nas figuras já apresentadas. A etapa de pós-processamento (Passo 4) do algoritmo tem a função de suavizar o sinal resultante do Fechamento.

O algoritmo proposto é sequencial por construção, a menos da determinação do comprimento do elemento estruturante. Caso o comprimento ótimo deste já seja

conhecido, não há iterações.

Nas Figuras 2.29 e 2.30, nota-se claramente a suavização da envoltória resultante e da interpolação incluída no algoritmo proposto.

2.4 Compromisso entre suavidade e detalhe

A fim de atingir o equilíbrio entre uma envoltória suave sem perder os detalhes relevantes da evolução da nota, um critério de suavidade pode ser considerado. Evidentemente cada método possui características e parâmetros próprios que possibilitam controlar essa relação; entretanto o critério deve ser o mesmo, a fim de possibilitar uma comparação entre os diversos métodos.

O maior desafio é encontrar um critério de suavidade que reflita o que se espera de uma envoltória, pois é difícil dizer se uma estimação de envoltória está boa, mas é fácil detectar uma envoltória mal estimada.

Nas seções seguintes, realiza-se uma discussão sobre algumas maneiras de se resolver o problema da suavidade.

2.4.1 Critério de suavidade associado ao *pitch*

Uma vez que cada nota musical possui um *pitch* definido, é intuitivo pensar em sua frequência fundamental f_0 ou em seu período fundamental $\frac{1}{f_0}$. A solução direta para a escolha do comprimento do elemento estruturante é associá-lo a esse período fundamental; desta forma, o elemento estruturante estaria "apoiado" sobre os picos das senoides de maior período, traçando a envoltória de maneira satisfatória.

Assim sendo, o comprimento L do elemento estruturante (em amostras), para uma nota com frequência fundamental f_0, amostrada a uma frequência f_s pode ser expresso como:

$$L = \frac{f_s}{f_0} \tag{2.15}$$

Desta abordagem surge a necessidade de se possuir a informação da f_0. Esse dado pode ser conhecido previamente ou obtido através de diversas técnicas como as apresentadas em [28], [29], [30], [31], [32], dentre outras.

Evidentemente, a introdução de um estimador de f_0 introduz mais imprecisões na cadeia de operações. Num contexto em que não se tem informação alguma acerca dos sinais a serem processados, certamente essa etapa de estimação é necessária. Entretanto, para os testes realizados nesta etapa do trabalho, os possíveis erros introduzidos pela estimação imprecisa desse parâmetro dificultariam a avaliação do método de estimação de envoltória proposto. Assim, nos testes realizados nesta etapa, assume-se que a f_0 de cada nota analisada é previamente conhecida.

Comprimento igual ao período fundamental da nota

Em muitos casos, devido principalmente a ruídos originados no momento da gravação ou mesmo ressonâncias do instrumento, usar um critério de suavidade diretamente relacionado à f_0 pode gerar resultados pouco suaves, principalmente em notas mais agudas. No caso analisado o comprimento do elemento estruturante é igual ao período fundamental da nota em questão. Exemplos desta abordagem são as Figuras 2.31 a 2.34. Nota-se que para as notas mais agudas a envoltória encontrada é extremamente ruidosa.

Uma alternativa é ajustar o parâmetro de suavidade com um valor proporcional a f_0, escalando-o por um fator na tentativa de suavizar o resultado. Neste caso, o comprimento do elemento estruturante será um múltiplo do período fundamental.

Comprimento igual a um múltiplo do período fundamental da nota

O problema dessa abordagem é que o mesmo fator multiplicativo não pode ser utilizado em todas as notas, ou seja, quando se ajusta a suavidade visualmente para uma dada nota, esse fator multiplicativo nem sempre é adequado para outras notas, ainda que nos mesmos instrumentos.

A fim de ilustrar essa característica, as figuras seguintes mostram as mesmas notas, porém agora com um fator multiplicativo de 10 sobre o parâmetro de ajuste. Nota-se que, para a nota mais aguda, esse valor é adequado para que a envoltória tenha uma suavidade tal que não seja ruidosa nem perca os detalhes; entretanto, para a nota mais grave, os vales não foram contornados corretamente. As Figuras 2.35 a 2.38 ilustram a característica descrita.

Conforme visto nas comparações feitas até este ponto, nota-se que não é possível estabelecer um valor único de parâmetro, ou mesmo um único fator multiplicativo para "amarrar" esse parâmetro à f_0. Assim, surge a necessidade de ajustar o parâmetro individualmente para cada nota o que resulta em outro desafio: avaliar objetivamente uma envoltória com "suavidade aceitável" e automatizar esse ajuste.

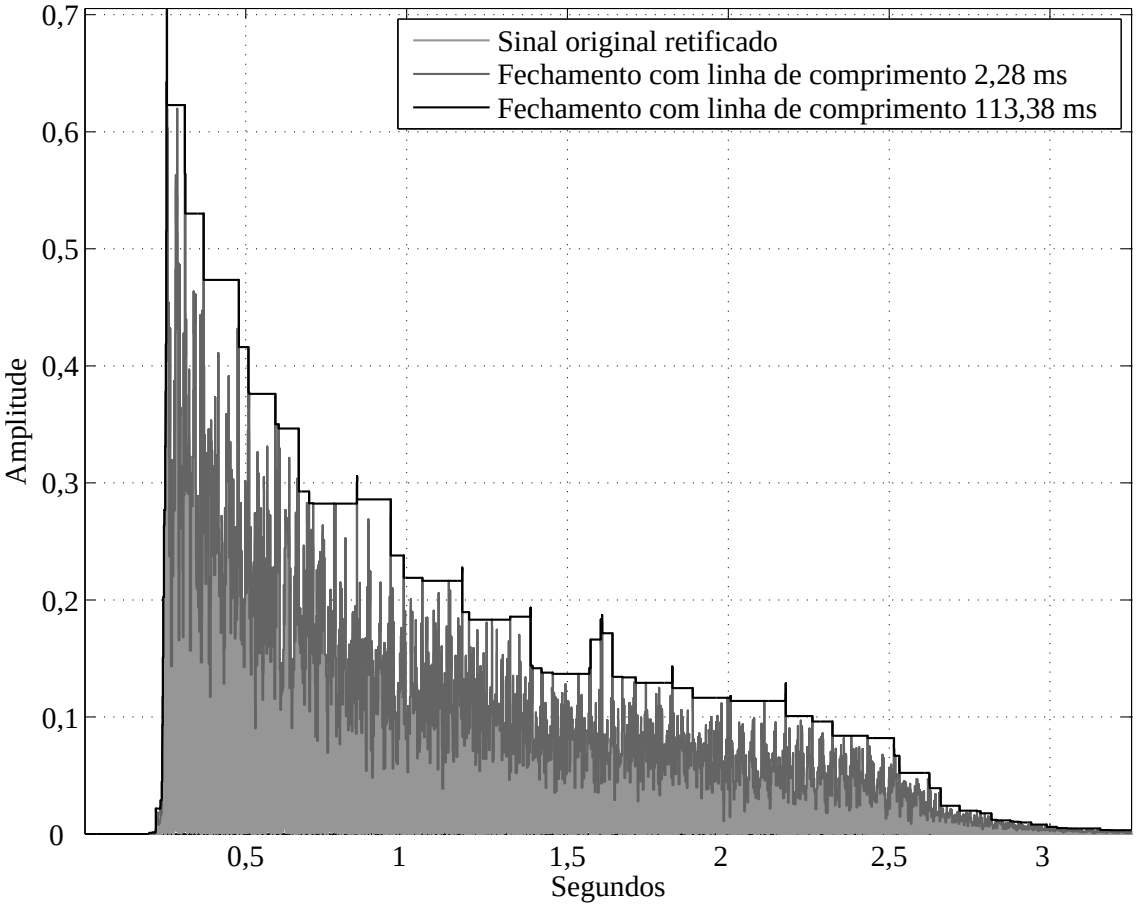


Figura 2.25: Comparação entre comprimentos de linha. A nota utilizada para a ilustração é a mesma Lá 4 ($f_0 = 440$Hz), de um piano, utilizada nos testes anteriores

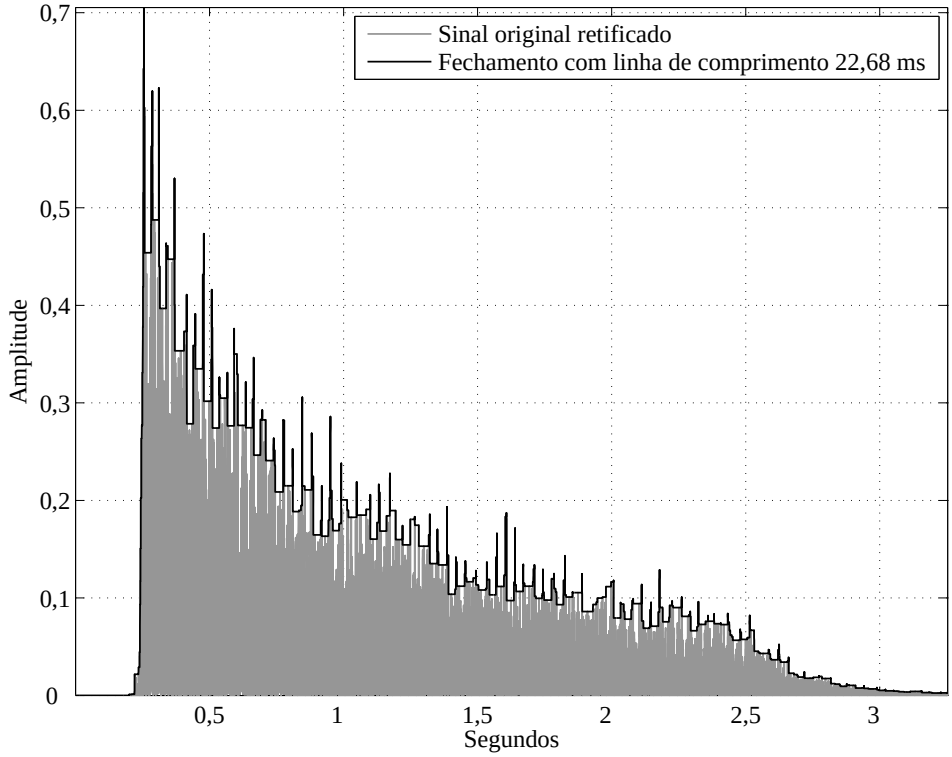


Figura 2.26: Nota Lá 1 ($f_0 = 55$Hz), pianoforte, sendo a estrutura uma linha de comprimento 22,68 ms

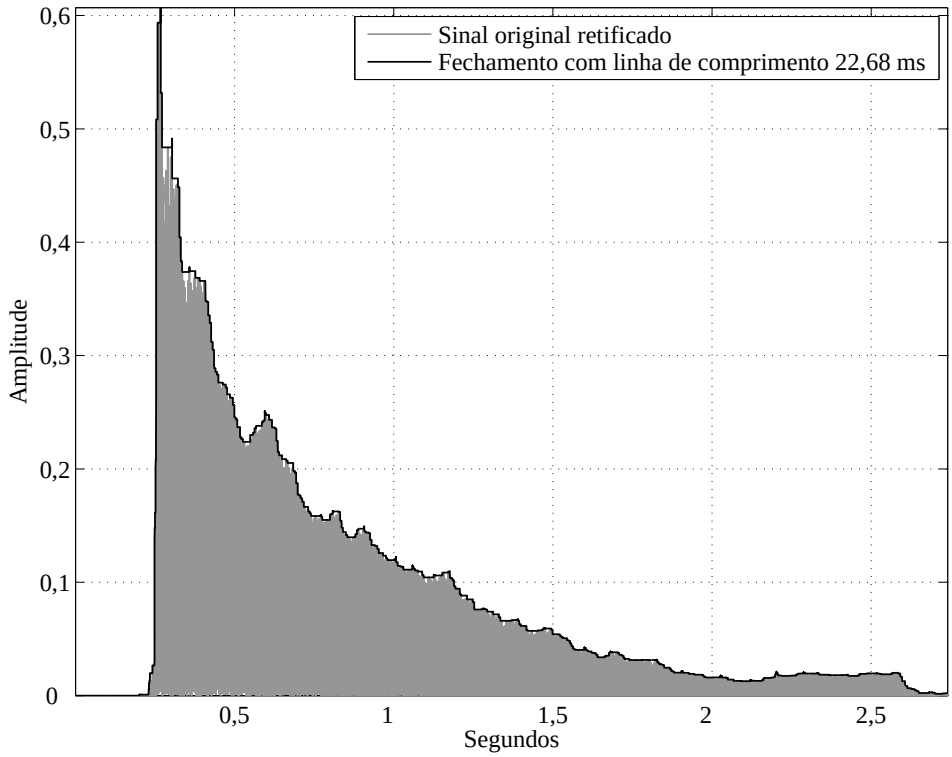


Figura 2.27: Nota Lá 4 ($f_0 = 440$Hz), piano, sendo a estrutura uma linha de comprimento 22,68 ms

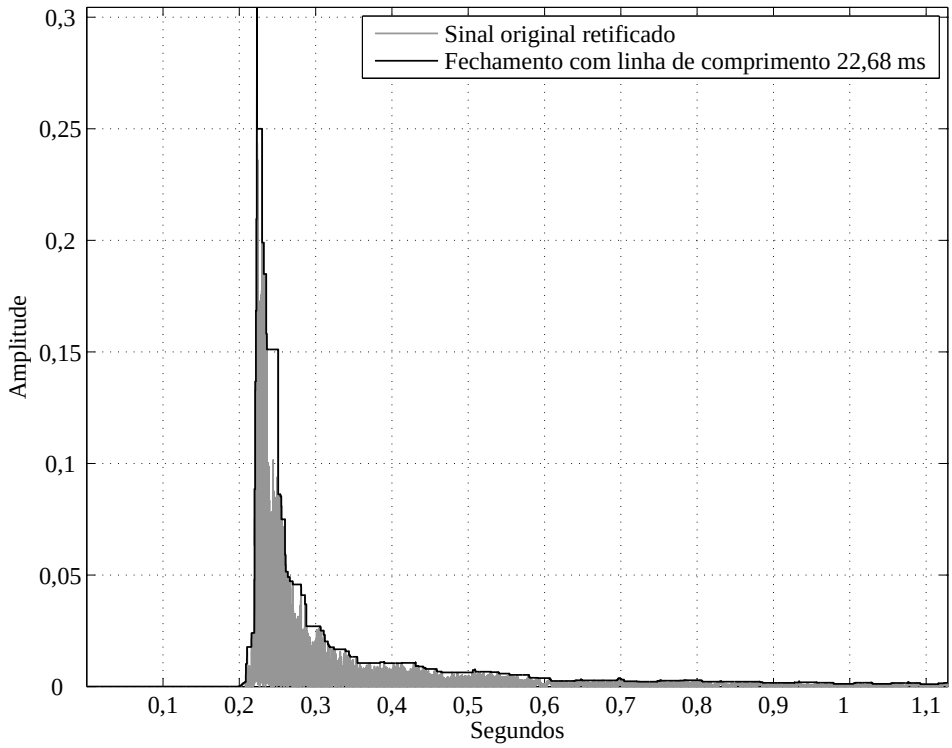


Figura 2.28: Nota Lá 7 (f_0 = 3520Hz), piano, sendo a estrutura uma linha de comprimento 22,68 ms

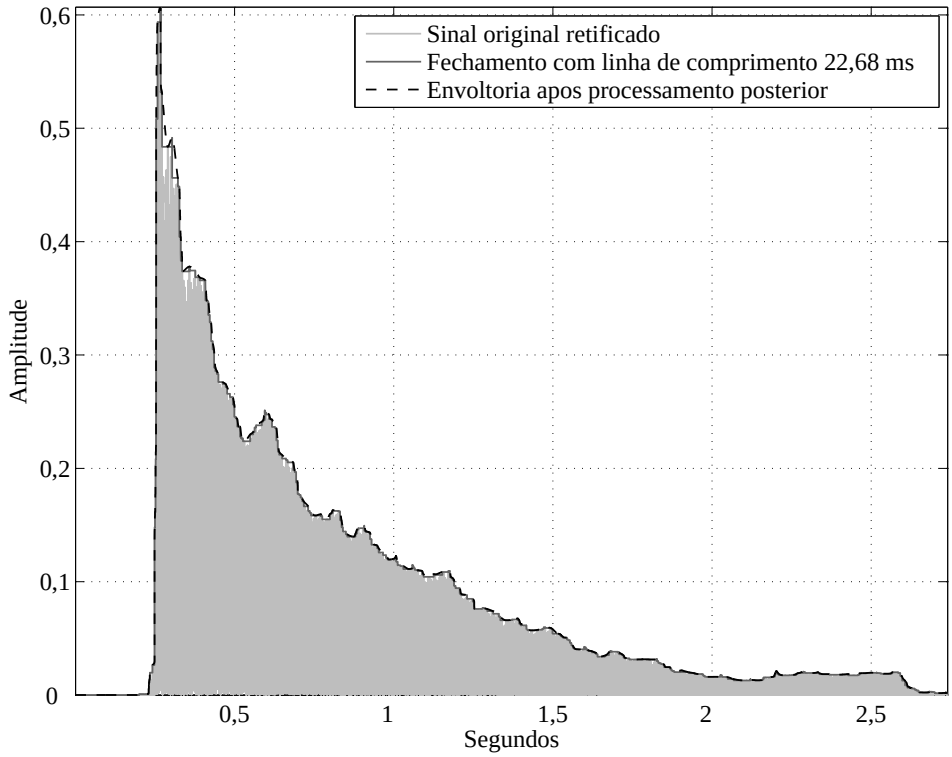


Figura 2.29: Nota Lá 4 ($f_0 = 440$Hz), piano, comparação entre a envoltória antes e após o pós-processamento

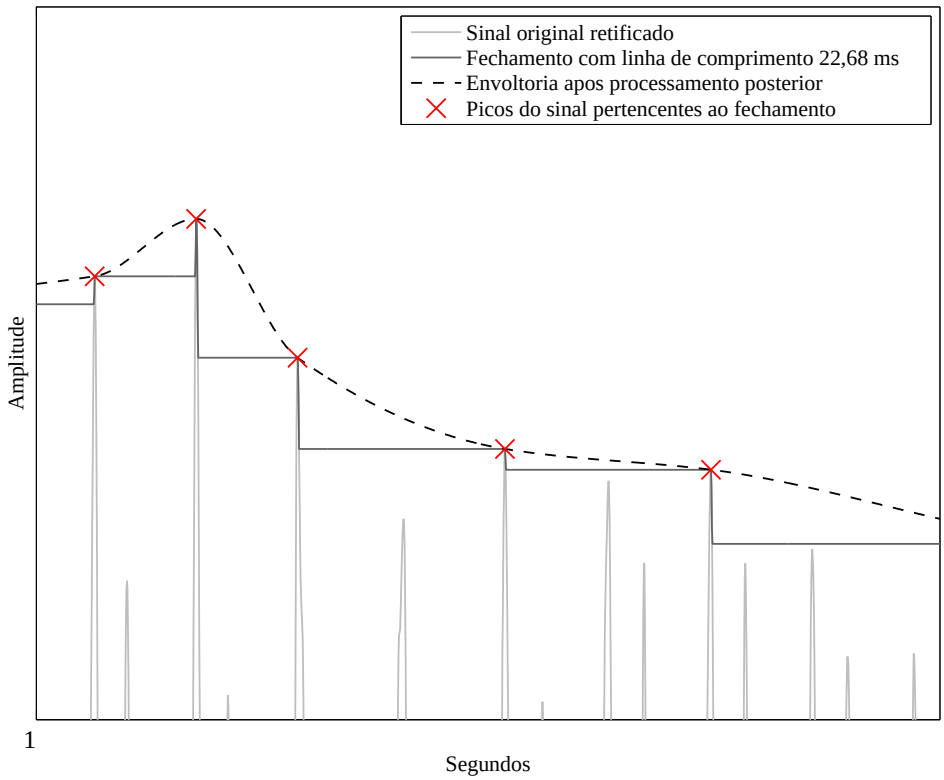


Figura 2.30: Detalhe da comparação entre a envoltória antes e após o pós-processamento

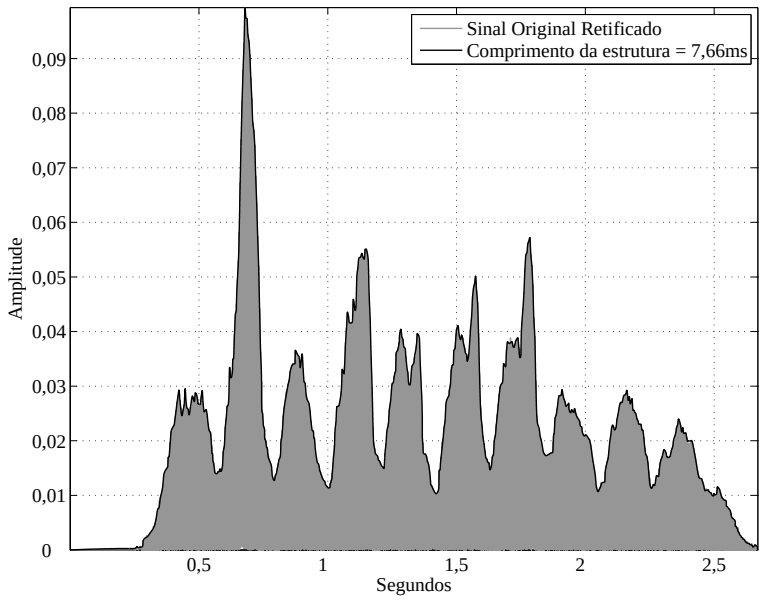


Figura 2.31: Nota Dó3 ($f_0 = 130,81$Hz) de uma flauta doce. Comprimento do elemento estruturante igual ao período fundamental da mesma.

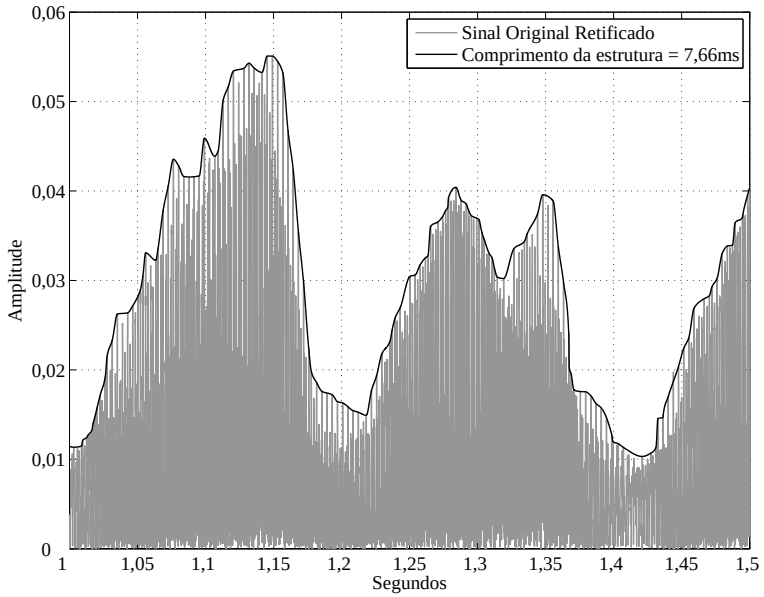


Figura 2.32: Detalhe da envoltória da nota Dó3 ($f_0 = 130,81$Hz) de uma flauta doce. Comprimento do elemento estruturante igual ao período fundamental da mesma.

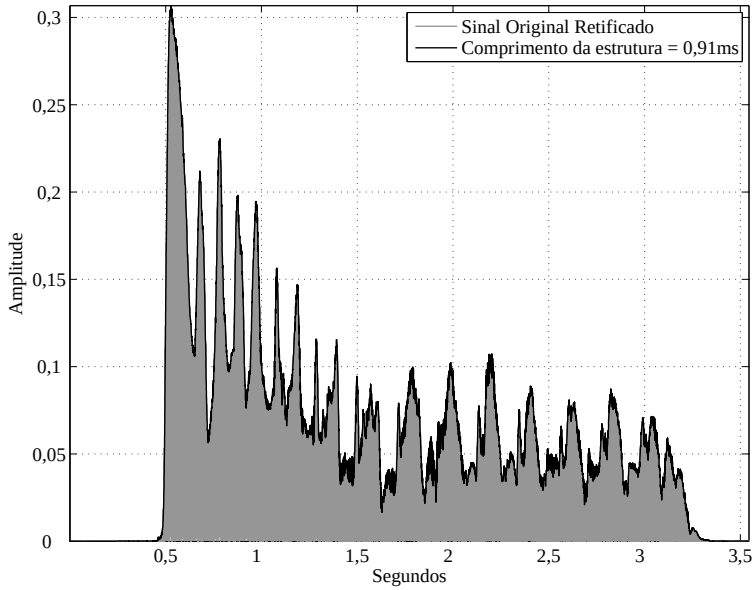


Figura 2.33: Nota Dó#6 ($f_0 = 1108,73$Hz) de uma flauta doce. Comprimento do elemento estruturante igual ao período fundamental da mesma.

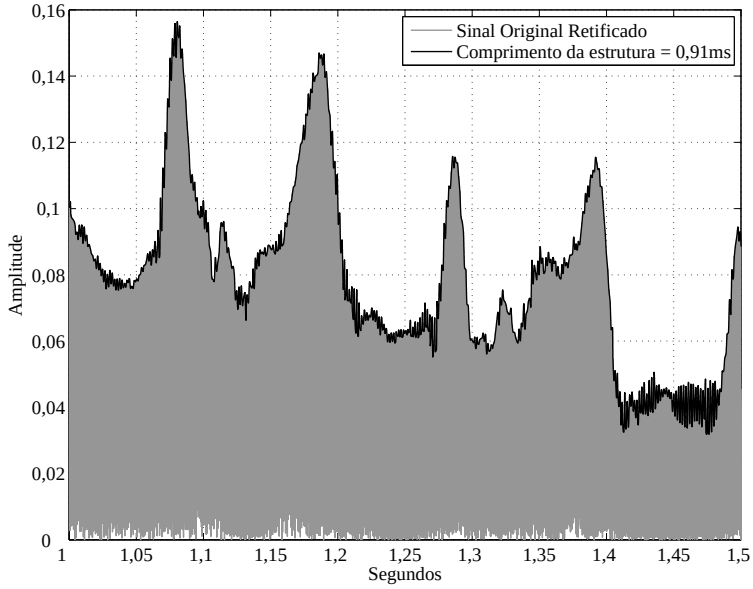


Figura 2.34: Detalhe da envoltória da nota Dó#6 ($f_0 = 1108,73$Hz) de uma flauta doce. Comprimento do elemento estruturante igual ao período fundamental da mesma.

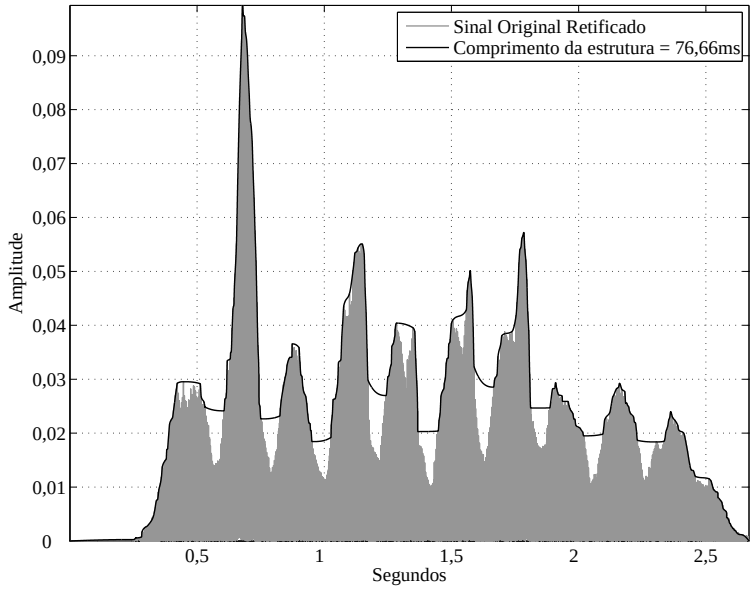


Figura 2.35: Nota Dó3 ($f_0 = 130,81$Hz) de uma flauta doce. Comprimento do elemento estruturante igual a 10 vezes o período fundamental da mesma.

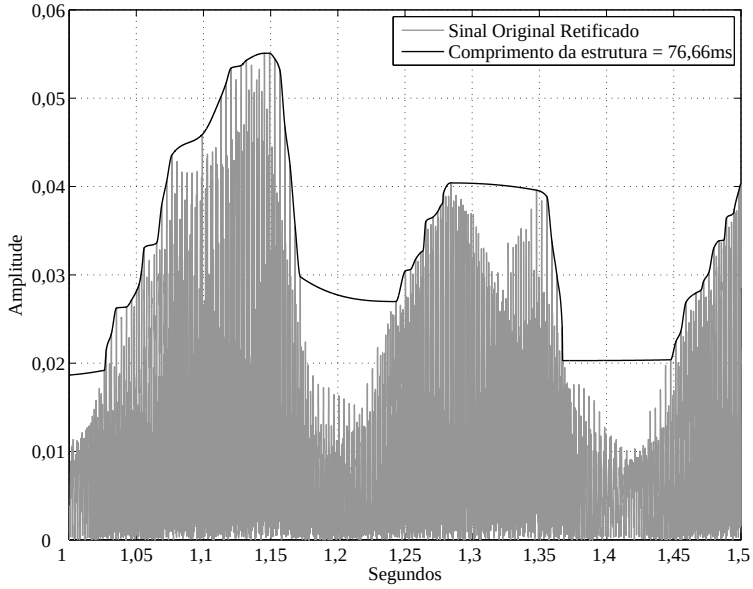


Figura 2.36: Detalhe da envoltória da nota Dó3 ($f_0 = 130,81$Hz) de uma flauta doce. Comprimento do elemento estruturante igual a 10 vezes o período fundamental da mesma.

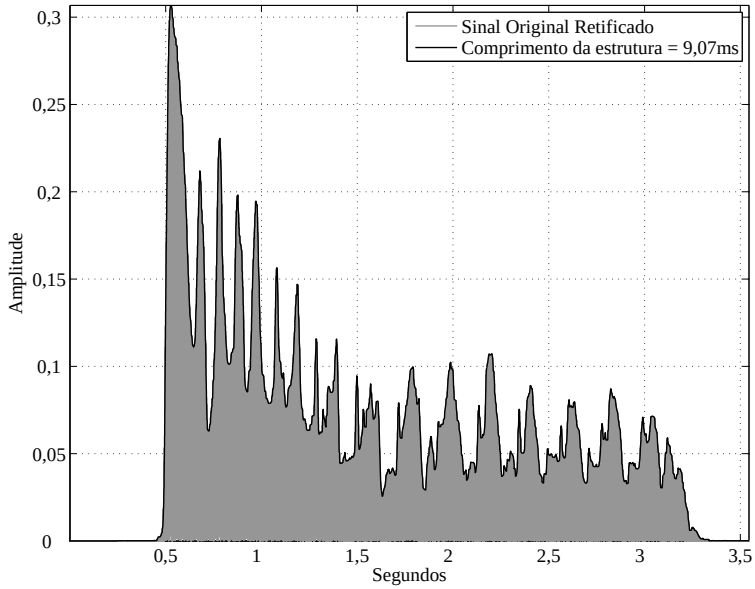


Figura 2.37: Nota Dó#6 ($f_0 = 1108,73$Hz) de uma flauta doce. Comprimento do elemento estruturante igual a 10 vezes o período fundamental da mesma.

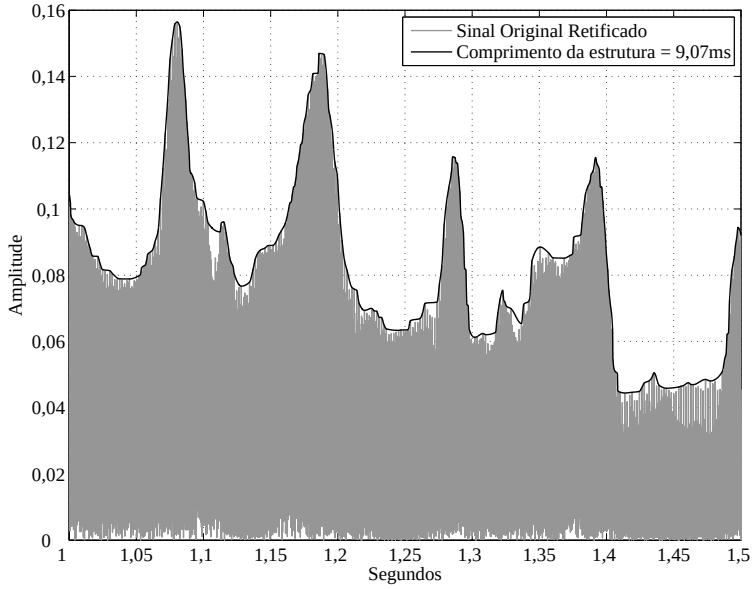


Figura 2.38: Detalhe da envoltória da nota Dó#6 ($f_0 = 1108,73$Hz) de uma flauta doce. Comprimento do elemento estruturante igual a 10 vezes o período fundamental da mesma.

2.4.2 Suavidade associada a um critério perceptivo

Diversas tentativas foram realizadas, sempre deixando perceber a dificuldade de avaliar objetivamente a qualidade da estimação de envoltórias. Associar a suavidade com a f_0 não é uma boa alternativa, como pôde ser observado na seção anterior.

Uma vez que a envoltória é uma evolução temporal da intensidade da nota, vale a analogia com um efeito conhecido: o tremolo, que é uma oscilação periódica de amplitude [4]. Sendo uma oscilação de amplitude audível, sua frequência não deve superar os 20Hz pois, conforme essa frequência aumenta, o ouvido tende a integrar essas oscilações e tais variações não são mais perceptíveis em separado, mas sim como um tom de intensidade constante [33]. A melhor maneira de exemplificar esse pensamento é ouvir um tremolo e ir gradativamente aumentando sua frequência de oscilação; a partir de determinada frequência o ouvido não mais distingue tais oscilações e a intensidade é percebida como sendo constante.

Seria interessante encontrar uma característica parecida nas envoltórias, que deveriam idealmente descrever variações de intensidade do sinal que pudessem ser percebidas como tal.

Considerando essa característica desejável, foram calculadas as taxas de picos das envoltórias entregues pelos métodos nos diversos testes realizados e observou-se que a maioria das envoltórias "aceitáveis" apresentavam uma taxa de picos por amostra similares, em torno de 5×10^{-4}. Esse valor de taxa de picos, à frequência de amostragem dos sinais (44,1kHz), leva a envoltórias com uma frequência de oscilação em torno de 20Hz, conforme esperado. Vale ressaltar que as taxas de picos das envoltórias foram calculadas considerando amostras centrais das notas, compreendendo entre 20% e 80% de sua energia, de forma a excluir regiões de transitórios e de baixíssima energia – como o final do decaimento, onde o chão de ruído se aproxima do sinal de interesse e não reflete corretamente as características do sinal.

Dessa forma, esse critério da taxa de picos da envoltória é um critério simples, porém eficiente e reflete o que perceptivamente se espera de uma envoltória com suavidade "aceitável".

Aplicação do critério perceptivo ao método proposto

A fim de automatizar a escolha do comprimento da estrutura do método proposto, o critério perceptivo detalhado acima será utilizado.

A etapa de escolha do comprimento da estrutura do algoritmo proposto na Seção 2.3 pode ser automática, seguindo a seguinte sequência de passos:

1. Uma estimativa inicial do comprimento é realizada localizando-se os picos do sinal completo e escolhendo a maior distância entre picos como sendo essa primeira estimativa.

2. Em seguida, realiza-se o Fechamento sobre o sinal (utilizando o comprimento inicial estimado) e calcula-se a taxa de picos através dos critérios descritos na seção anterior.

3. Sabendo-se a porção do sinal a ser utilizada para o cálculo da taxa de picos (vale lembrar que essa taxa é calculada numa região entre 20% e 80% da energia do sinal), calcula-se a quantidade de picos necessária para se atingir o valor ótimo (equivalente a 20Hz).

4. Com essa estimativa, sabe-se qual a razão entre a quantidade de picos da primeira estimativa e a ideal. Essa razão é utilizada como multiplicador do comprimento da estrutura e um novo fechamento, agora com uma estrutura de comprimento ajustado, é realizado.

5. Caso a taxa de picos da envoltória após esse novo fechamento seja próxima o suficiente do ideal (em torno de 15% é o bastante[1]), realiza-se a interpolação, conforme descrito no algoritmo, e a envoltória está calculada.

6. Se a taxa de picos calculada no item anterior não estiver próxima o suficiente, faz-se uma busca utilizando algum algoritmo de minimização para determinar o comprimento ótimo. A função a ser minimizada é a distância absoluta entre a taxa de picos ótima e a calculada para um dado comprimento de estrutura.

Seguindo o algoritmo descrito acima, um conjunto de teste de 664 sinais foi utilizado, sendo cada um dos sinais uma nota musical isolada (extraída da base RWC [14]), contendo diversos instrumentos, tais como violoncelo, clarineta, violão, baixo elétrico, flauta, harmônica, harpa, marimba, órgão, piano, trombone, trompete, tuba e vibrafone.

As envoltórias de todos os sinais foram calculadas e alguma estatística foi extraída, de forma a criar um panorama do desempenho do método, segundo os critérios descritos anteriormente. O algoritmo de minimização utilizado foi o Método da Bisseção. Neste método, divide-se um intervalo sucessivamente em subintervalos dentro dos quais, espera-se, está o mínimo da função.

O histograma da Figura 2.39 ilustra a diferença percentual absoluta entre a taxa de picos ótima (correspondente aos 20Hz) e a obtida pelo método da Bisseção, para cada um dos sinais acima citados. Nota-se que, para grande parte dos sinais, o comprimento desejado da estrutura foi obtido corretamente (em torno de 70% dos

[1] O valor de 15% para a diferença máxima entre o valor tido como ótimo e o encontrado pelo algoritmo foi definido experimentalmente. Notou-se que o critério era robusto o suficiente para se utilizar uma margem ampla (algo como 50% ou mais em alguns casos) e ainda produzir boas envoltórias. Desta forma, optou-se por um valor que possibilita uma convergência mais rápida e ainda assim mantém a precisão do método.

casos). Em princípio essa proporção não seria satisfatória, uma vez que o ideal é obter um comprimento que leve à taxa de picos por amostra mais próxima possível do valor desejado.

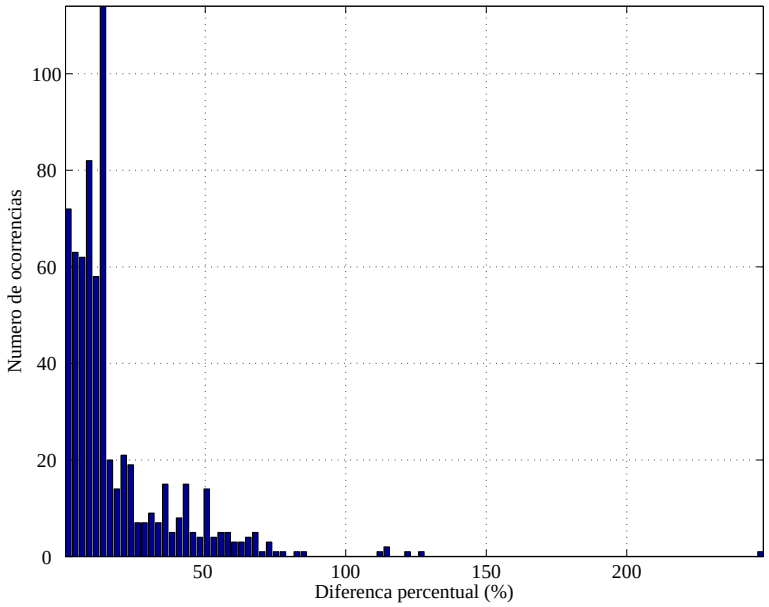


Figura 2.39: Diferença percentual absoluta na convergência do método da Bisseção.

Vale ressaltar que a função a ser minimizada possui grandes "platôs" indesejados para comprimentos de estrutura maiores que a faixa dos que satisfazem a condição, apesar de o mínimo da função também ser um platô, já que existem diversos comprimentos de estrutura que satisfazem o critério de suavidade desejado. Além disso, vários sinais não possuem um comprimento ótimo, ou seja, a taxa de picos nunca será próxima o suficiente do valor de referência correspondente aos 20Hz.

Um exemplo de sinal que não atinge o critério de convergência é a nota Fá#5 de uma marimba. A fim de ilustrar o comportamento desse sinal, as Figuras 2.40 e 2.41 mostram a curva do erro de estimação da envoltória em função do comprimento da estrutura utilizada. Nota-se claramente que não é possível atingir-se a taxa ideal (nem mesmo a tolerância dos 15%), por maior que seja o comprimento da estrutura.

Comparativamente, a nota seguinte da mesma marimba, Sol5, atinge a convergência, conforme pode ser visto nas Figuras 2.42 e 2.43.

A fim de se obter uma visão global das características dos sinais, realizou-se um teste de convergência em que diferentes comprimentos de estrutura foram utilizados para o fechamento e as taxas de picos correspondentes foram calculadas.

Teste de Convergência

A estrutura utilizada para o teste teve seu comprimento aumentado de 0,02ms (uma amostra), em passos de também uma amostra, até 100ms, e foram calculadas as

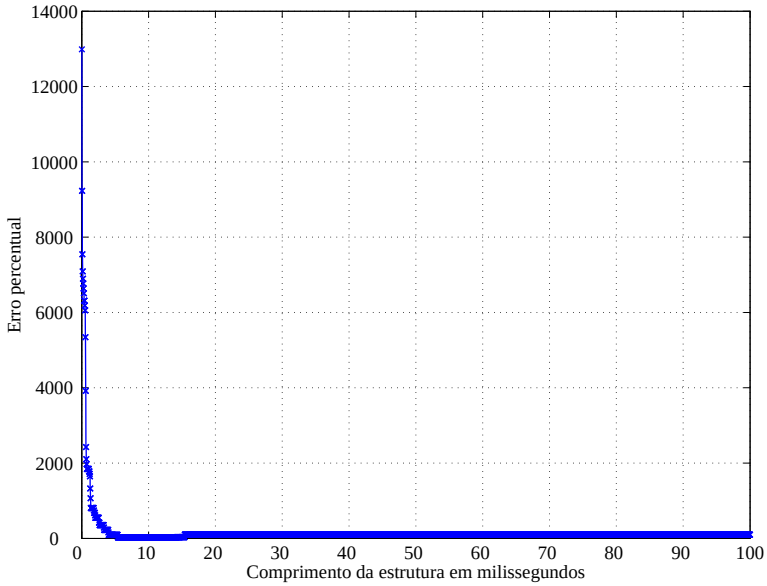


Figura 2.40: Curva de convergência para uma nota Fá#5 de marimba.

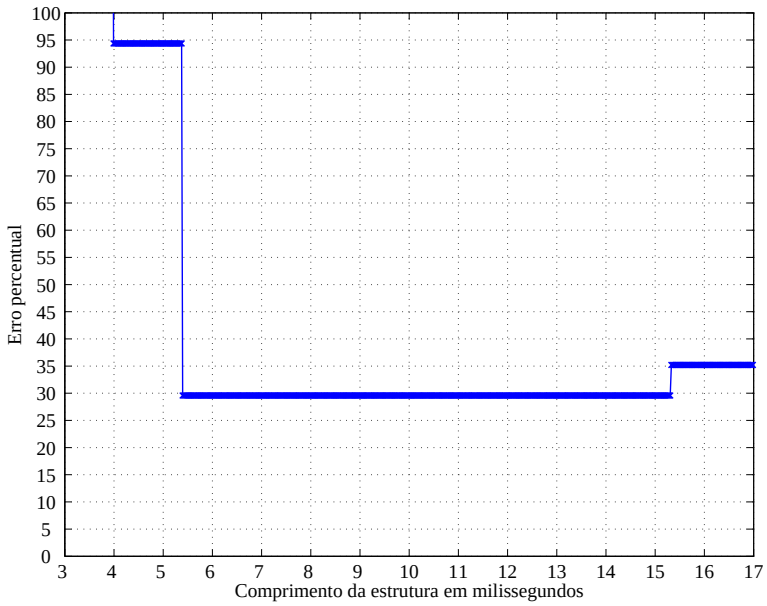


Figura 2.41: Detalhe da curva de convergência para uma nota Fá#5 de marimba.

taxas de picos correspondentes em todos os casos. A partir desses dados montou-se a Figura 2.44, que mostra o histograma das menores diferenças possíveis entre a taxa de picos ótima (correspondente aos 20Hz) e as taxas de picos obtidas para os os sinais do conjunto de testes. Pode-se observar que existe um percentual de sinais que não possibilitam encontrar um comprimento de estrutura que leve a uma taxa de picos por amostra adequada. Para o conjunto em questão essa proporção é de 14,5% (96 sinais).

Comparando-se os valores de comprimento de estrutura obtidos com o método da Bisseção com os obtidos para os menores erros possíveis, os valores de taxas de

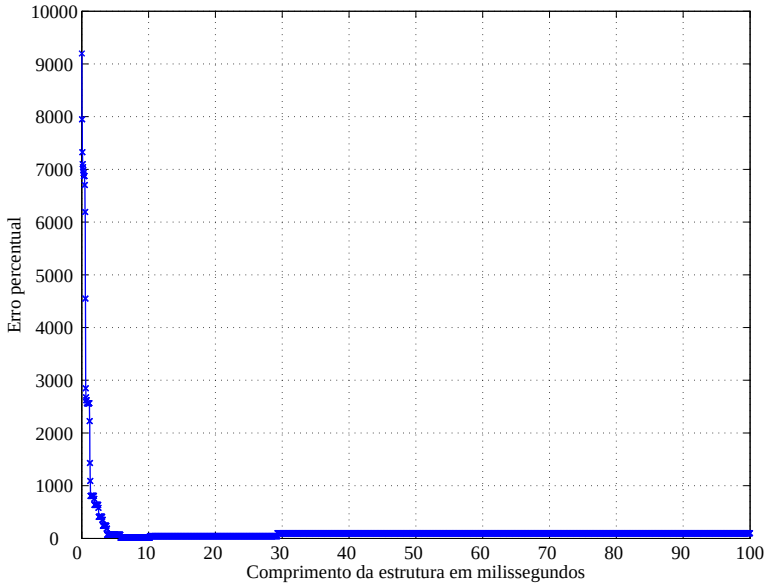


Figura 2.42: Curva de convergência para uma nota Sol5 de marimba.

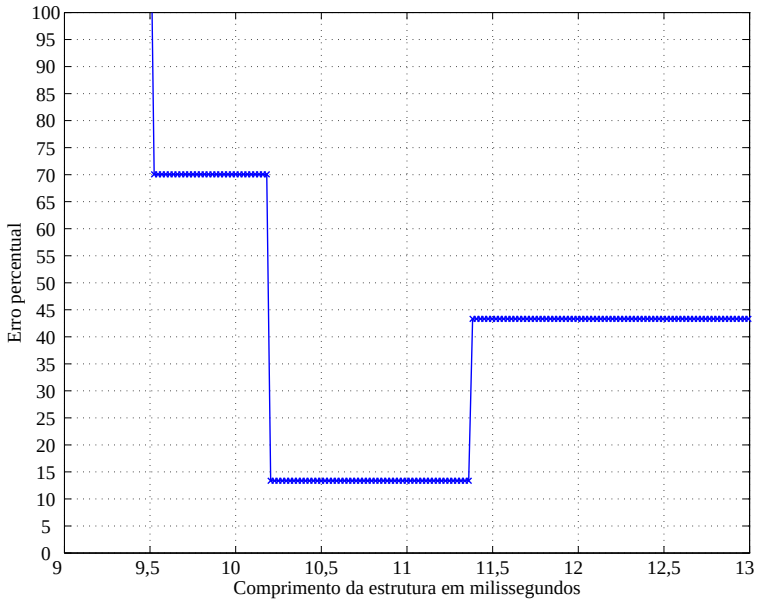


Figura 2.43: Detalhe da curva de convergência Sol5 para uma nota de marimba.

picos foram ordenados e a Figura 2.45 mostra tal relação. Nota-se claramente que o erro da Bisseção sempre está acima do erro mínimo, porém isso se deve ao fato de que o algoritmo busca o ponto em que o erro cruza a fronteira dos 15%, e não o erro mínimo para cada sinal.

Levando-se em conta apenas o número de sinais que realmente possibilitam a determinação do comprimento ótimo, o método da Bisseção possui uma taxa de sucesso de 81,4% (462 sinais).

O método de minimização adotado foi escolhido pela sua boa relação entre desempenho e simplicidade. Obviamente um algoritmo que consiga superar os desafios

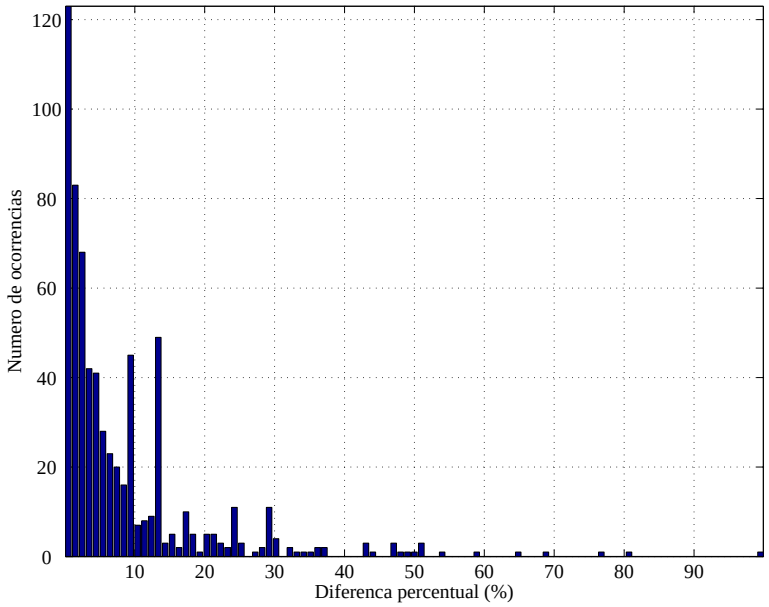


Figura 2.44: Diferença percentual absoluta mínima possível.

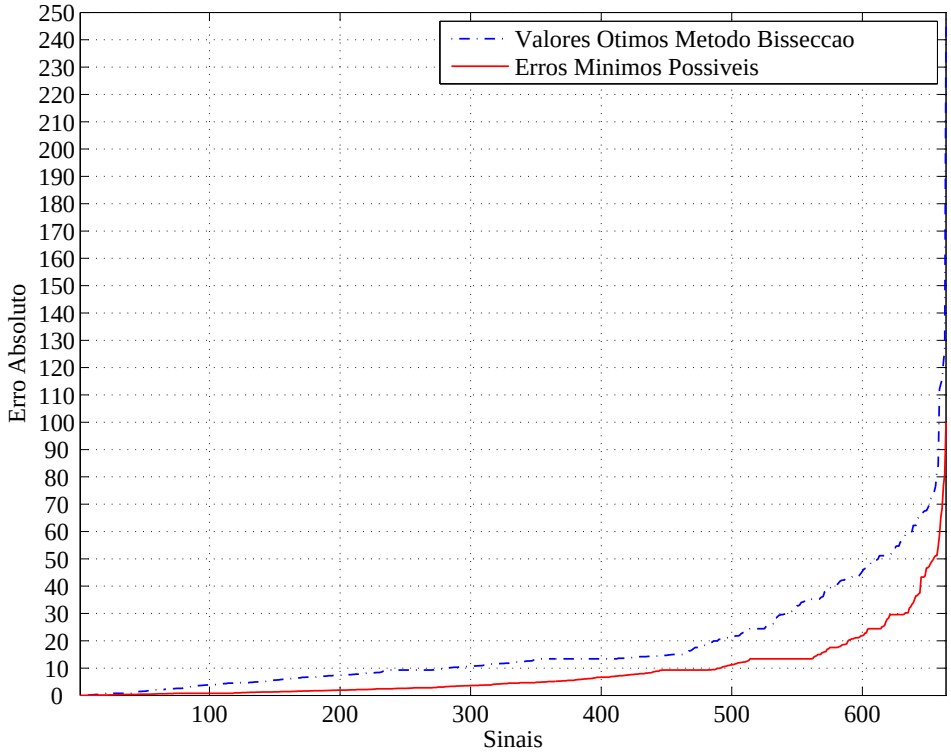


Figura 2.45: Comparação entre os resultados do método de minimização e os menores erros possíveis.

dos "platôs" tenderá a obter melhores resultados, até o limite mostrado no teste de convergência

Outra característica interessante é a robustez do método perceptivo, pois, mesmo nos casos em que não é possível alcançar-se a taxa de picos ótima, a envoltória

estimada é visualmente adequada.

Toma-se como exemplo a mesma nota Fá#5 de marimba anteriormente citada que, como visto, não permite a escolha de um comprimento que leve à taxa de picos ótima. A Figura 2.46 mostra a forma de onda retificada da nota Fá#5 de marimba e envoltória calculada com o método da Bisseção.

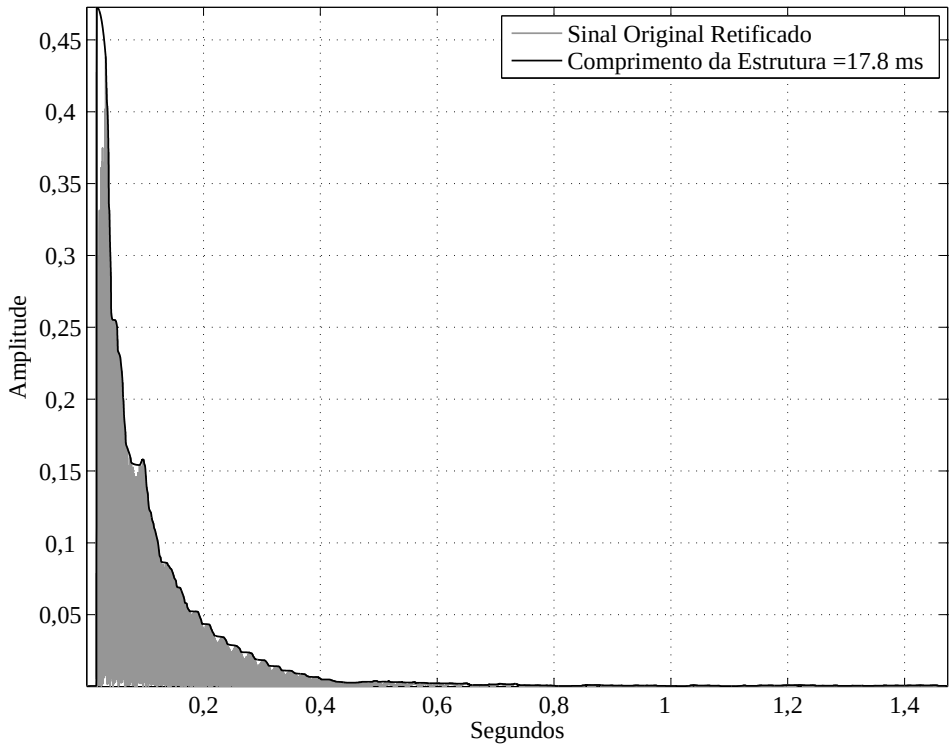


Figura 2.46: Nota Fá#5 ($f_0 = 739,98$Hz) de uma marimba. Envoltória estimada com o método proposto, minimizado com Bisseção.

Nota-se que, apesar de não permitir a definição do comprimento ótimo, a envoltória apresentada consegue acompanhar aceitavelmente o contorno da forma de onda retificada.

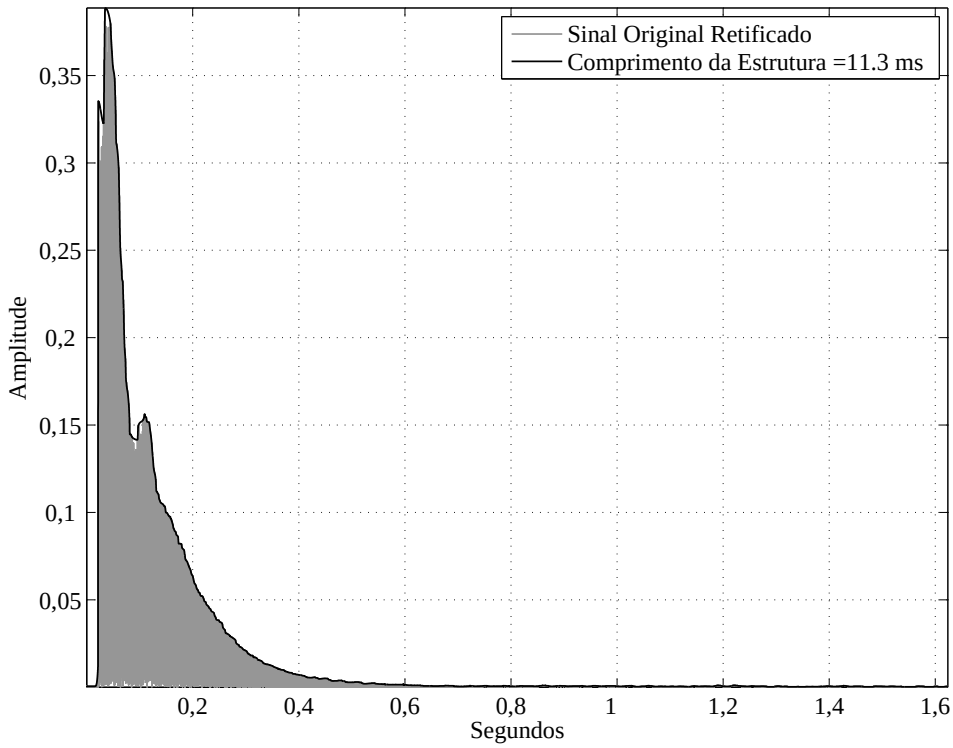


Figura 2.47: Nota Sol5 ($f_0 = 783,99$Hz) de uma marimba. Envoltória estimada com o método proposto, minimizado com Bisseção.

2.5 Complexidade computacional

Finalizando este capítulo, realiza-se uma análise da complexidade computacional dos métodos descritos e comparados até esta etapa do trabalho. A análise aqui realizada leva em conta apenas as operações realizadas pelos métodos segundo a literatura descreve.

2.5.1 Valor Médio Quadrático

O método por RMS é simples e eficiente. Denota-se N o número de amostras em que o valor RMS será computado (comprimento da janela deslizante) e M o número da amostras total do sinal a ser analisado.

O número de operações realizadas é listado a seguir:

- Multiplicações: $N + 3M + 1$
- Divisões: 1
- Somas: $N + 2M$
- Operações Lógicas: 0

O fato de a janela deslizante ter um deslocamento de apenas uma amostra entre as janelas possibilita otimizar os cálculos: após a primeira janela, basta retirar da média a energia da primeira amostra da janela anterior e acrescentar a energia da nova amostra à média.

Como exemplificação, a nota Lá 4 de um piano utilizada nos testes (extraída da base RWC [14]), que possui 3 segundos de duração, amostrados a 44,1kHz (o que leva a $M = 132300$ amostras) foi calculada, na Figura 2.9, com $N = 882$ amostras.

2.5.2 *True Amplitude Envelope*

O método TAE realiza Transformadas de Fourier iterativamente para calcular o *cepstrum*, o que o torna excessivamente complexo computacionalmente.

Sendo $\bar{N}$ o número de amostras do sinal auxiliar gerado após o *zero-padding* a reversão no tempo (para fins de comparação, no melhor caso, $N = \bar{N}$); e k o número de iterações realizadas durante a execução, segue uma análise de operações realizadas.

- Multiplicações: $2(k+2)[\frac{\bar{N}}{2}\log_2 \bar{N} - \frac{3\bar{N}}{2} + 2] + \bar{N}$
- Divisões: $\bar{N}$
- Somas: $k\bar{N}[2\log_2 \bar{N} + 1] + 4\bar{N}\log_2 \bar{N}$

- Operações Lógicas: $(4k + 3)\bar{N}$

A dependência no número de iterações embute uma lentidão considerável ao método. Por exemplo, a envoltória da mesma nota Lá 4 de piano é calculada com $k = 3192$.

Conforme exposto anteriormente, o *zero-padding* e a reversão no tempo promovem um aumento significativo na complexidade computacional do método TAE, uma vez que tais operações aumentam muito o número de amostras a serem processadas.

2.5.3 Morfologia Matemática

Denotando por N o número de amostras do sinal sob análise e k o número de iterações realizadas durante a execução, o número de operações realizadas é listado abaixo.

- Multiplicações: $4N$
- Divisões: $8N + k + 2$
- Somas: $(2N + 1)k + 18N + 1$
- Operações Lógicas: $(3N + 2)k + 9N + 2$

O método proposto, incluindo a determinação do comprimento da estrutura, é comparativamente mais eficiente, embora seja iterativo também. Para fins de comparação, a mesma nota Lá 4 tem sua envoltória estimada pelo método proposto com $k = 1$.

Para os resultados apresentados utilizando-se o método da Bisseção, o número máximo de iterações foi fixado em $k = 20$, pois notou-se que era o suficiente para os piores casos: se o algoritmo segue iterando então ele não é capaz de diminuir o erro, mesmo que k seja alto.

Mais uma vez, esse paradigma é passível de melhoria por algum outro método de otimização que consiga transpor os "platôs" e atingir o menor valor possível com um número de iterações mais baixo.

A fim de ilustrar o acima exposto, a Figura 2.48 mostra um histograma do número de iterações realizado na análise do conjunto de testes descrito anteriormente. O grande número de ocorrências para $k = 20$ descreve os casos em que o número de iterações foi truncado buscando acelerar o processo de otimização.

Por sua vez, a Figura 2.49 ilustra o erro absoluto na primeira iteração do método, o que mostra que a estimativa do número de picos ótimo foi suficiente para a convergência em muitos casos. Uma porcentagem dos sinais (17,6% ou 117 sinais)

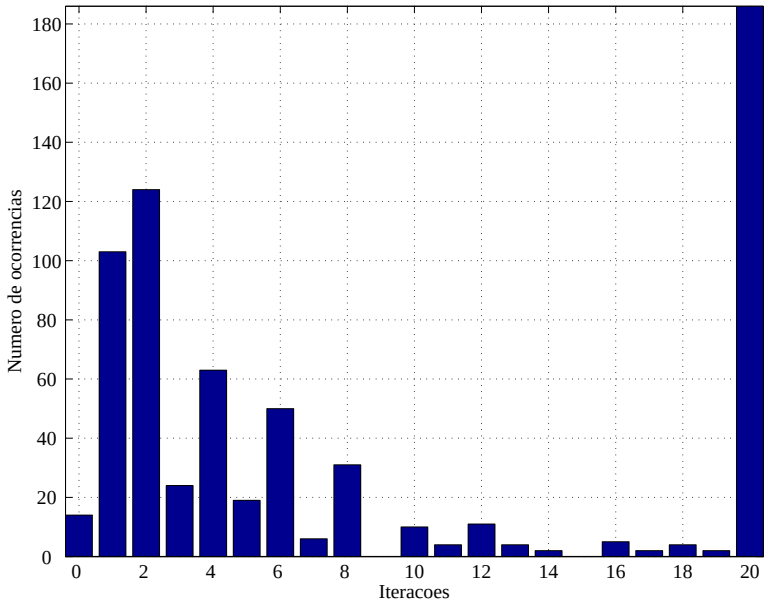


Figura 2.48: Número de iterações realizadas até a convergência.

convergiram em uma iteração ou nenhuma (entenda-se por nenhuma quando o comprimento ótimo da estrutura é a maior distância entre picos consecutivos do sinal de entrada - Passo 1 do algoritmo proposto, descrito em 2.4.2).

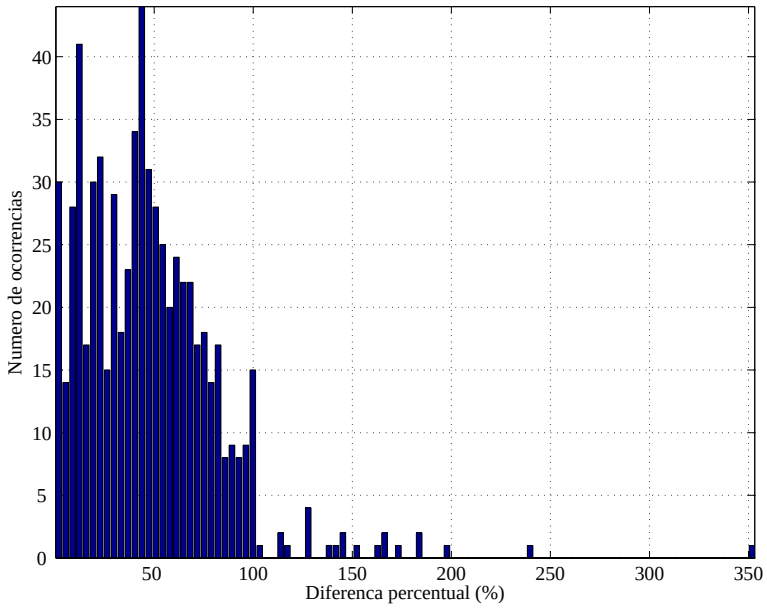


Figura 2.49: Erro percentual absoluto na primeira iteração.

2.6 Testes Subjetivos

Adicionalmente às análises previamente detalhadas, realizou-se um teste subjetivo informal para comparar as envoltórias produzidas pelo método proposto com as produzidas por dois métodos concorrentes.

2.6.1 Metodologia do Teste

A metodologia é muito simples: a cada etapa do teste, uma gravação de um instrumento musical emitindo uma única nota foi utilizada como referência e comparada com três sinais sintetizados, com evoluções temporais (envoltória) que procuram se assemelhar à do sinal original.

A geração dos sinais sintéticos foi realizada seguindo os seguintes passos:

1. Foram selecionados 3 tipos diferentes de instrumentos: flauta, piano e violoncelo, de modo a abranger três tipos (sopro, percussão e arco, respectivamente);

2. Selecionadas algumas notas ao longo da tessitura de cada instrumento, as 20 primeiras parciais de cada nota foram detectadas e trilhas senoidais foram extraídas (utilizando a *toolbox* de modelagem senoidal FlexSM [34]) a partir de tais parciais;

3. Escolhendo uma região da nota em que as trilhas apresentam pouca variação de frequência, armazenou-se a informação das frequências e relação de amplitudes entre as senoides presentes em cada uma das trilhas selecionadas.

4. Em seguida, as mesmas notas utilizadas para extração de trilhas e síntese com as senoides tiveram suas envoltórias estimadas por 3 métodos:

   - Morfologia Matemática
   - Valor Médio Quadrático
   - *True Amplitude Envelope*

5. Por fim, cada nota previamente analisada foi sintetizada a partir das informações de frequência e relação de amplitudes das senoides, juntamente com a envoltória estimada pelos 3 métodos, de forma a apresentar uma evolução temporal estimada pelos métodos e características de regime permanente das notas originais.

Os avaliadores atribuíram notas semelhantes aos conjuntos de sinais, o que indica que as envoltórias geradas pelos métodos envolvidos são perceptivamente comparáveis. Assim sendo, a comparação entre os métodos deve considerar outras medidas, tais como automatização, robustez e velocidade de convergência.

Conforme visto na Seção 2.5, o método proposto é totalmente automático e independente de informações prévias acerca dos sinais a serem analisados. Em contrapartida, o método TAE depende do conhecimento da f_0 e é demasiadamente lento devido, principalmente, à sua característica iterativa e às DFTs e DFTs inversas que são calculadas a cada iteração.

2.7 Comparação final

Por fim, faz-se uma comparação final entre os método proposto (descrito na Seção 2.3) e o método TAE (descrito na Seção 2.1.4), este com a ordem definida pela Equação 2.7 com $\alpha = 1$.

Para fins de comparação, a Tabela 2.1 mostra alguns resultados relevantes. Além dos parâmetros utilizados pelos métodos (ordem, no caso do TAE e comprimento da estrutura, para o método proposto), mostram-se as taxas de picos por amostra das envoltórias resultantes, que foram calculadas através do critério apresentado na Seção 2.4.2.

Tabela 2.1: Comparação Final. Taxas de Picos ($\times 10^{-4}$)

	TAE		Morfologia Matemática	
	Ordem	Taxa de Picos	Comp. (ms)	Taxa de Picos
Flauta Dó3	316	3,56	7,5	5,25
Violoncelo Dó2	135	4,47	23,8	4,47
Piano Lá4	1341	25,0	12,2	5,04
Piano Dó8	5883	306,0	20,8	5,67
Harmônica Fá#4	1077	32,0	21,0	5,35

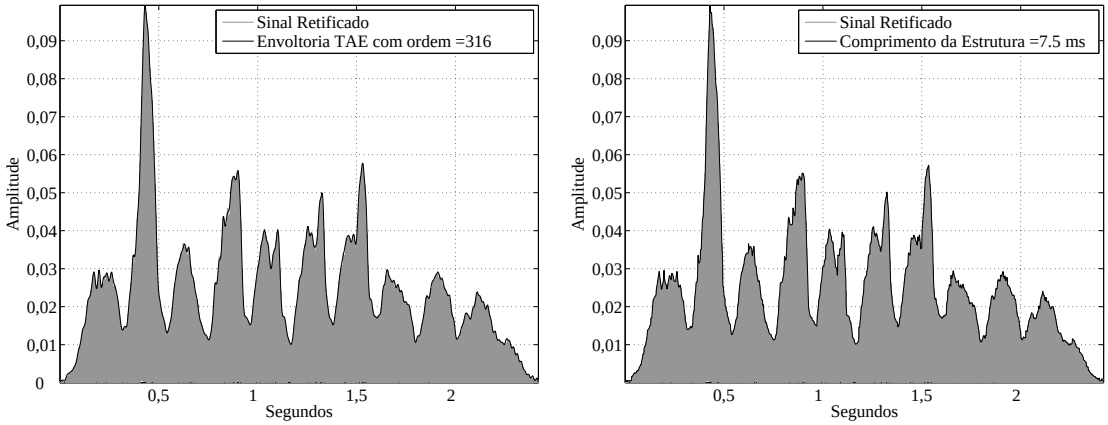
As Figuras 2.50 e 2.51 mostram, respectivamente, notas de Flauta e Violoncelo cujas envoltórias foram corretamente estimadas e, além disso, possuem um aspecto visual muito próximo. Nota-se, da Tabela 2.1, que possuem taxas de picos semelhantes e próximas ao valor descrito pelo critério perceptivo exposto na Seção 2.4.2.

A Figura 2.52 exemplifica um caso em que o resultado obtido com o TAE não é tão suave quanto os resultados anteriores, vide valor da taxa de picos mostrado na Tabela 2.1, cinco vezes maior que a obtida pelo método proposto. Para fins de comparação, mostram-se as envoltórias em detalhe de modo a explicitar a diferença entre elas.

Finalmente, as Figuras 2.53 e 2.54 ilustram exemplos nos quais o método TAE não foi capaz de estimar a envoltória de maneira suave. Isso pode ser comprovado observando-se as taxas de picos de tais envoltórias, que são significativamente maiores (no caso do Piano Dó8, 60 vezes) que as obtidas pelo método proposto e indicam

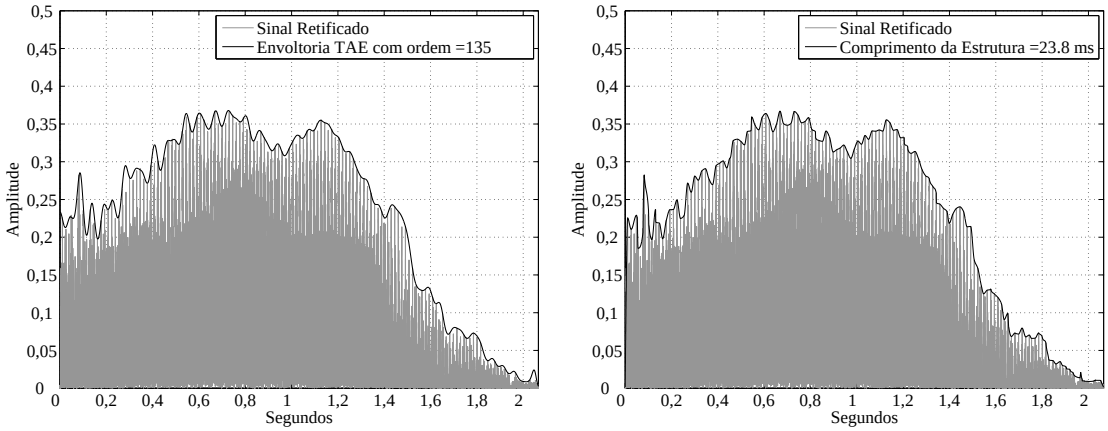
a característica ruidosa visível nas envoltórias. Também para tais sinais, detalhes das envoltórias são mostrados, para melhor visualização e comparação.

Pode-se concluir, dos resultados obtidos, que nem sempre utilizando a ordem recomendada pelo autor do método, o TAE é capaz de estimar envoltórias com um grau de suavidade "aceitável", enquanto o método proposto consegue atingir uma suavidade coerente para diversos tipos de notas/instrumentos envolvidos, e o faz de maneira automática e sem conhecimento prévio algum do sinal a ser analisado.



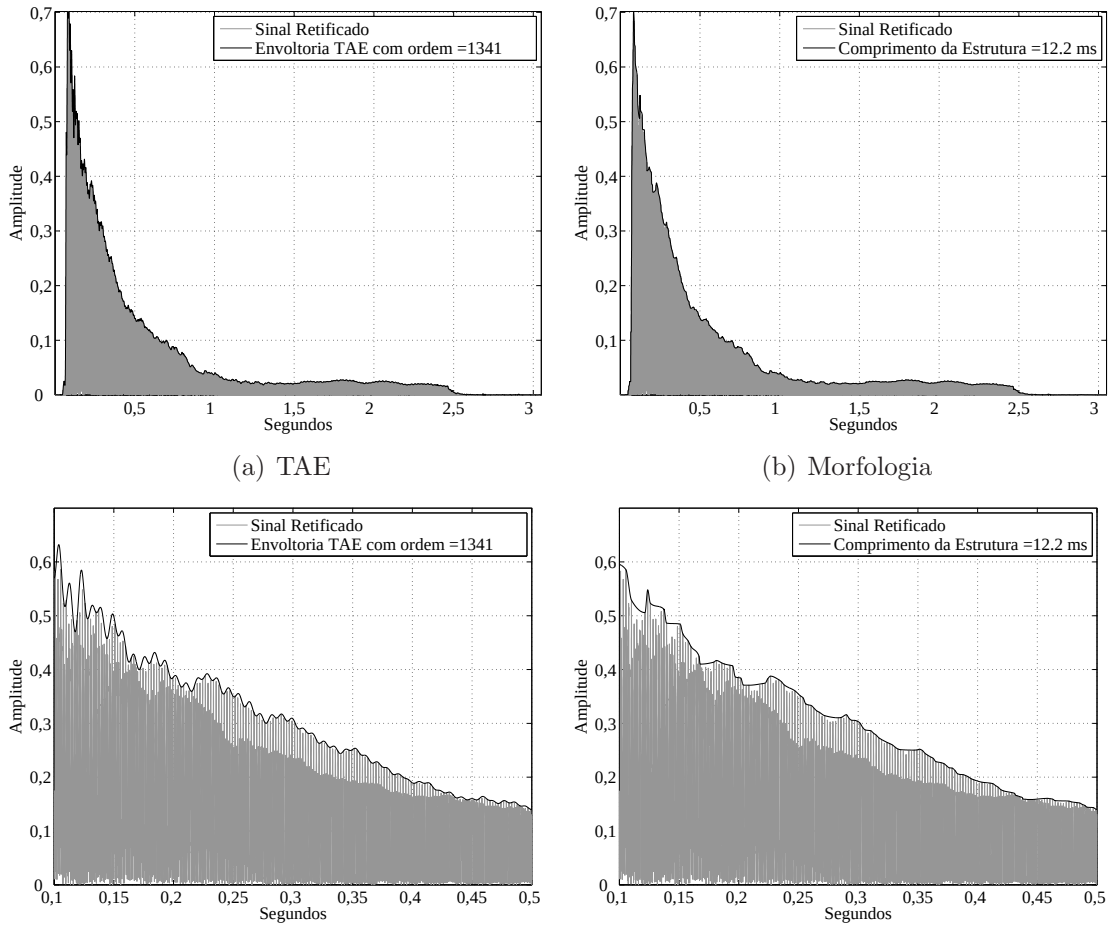
(a) TAE (b) Morfologia

Figura 2.50: Envoltórias da nota Dó3 ($f_0 = 130,81$Hz) de uma Flauta.



(a) TAE (b) Morfologia

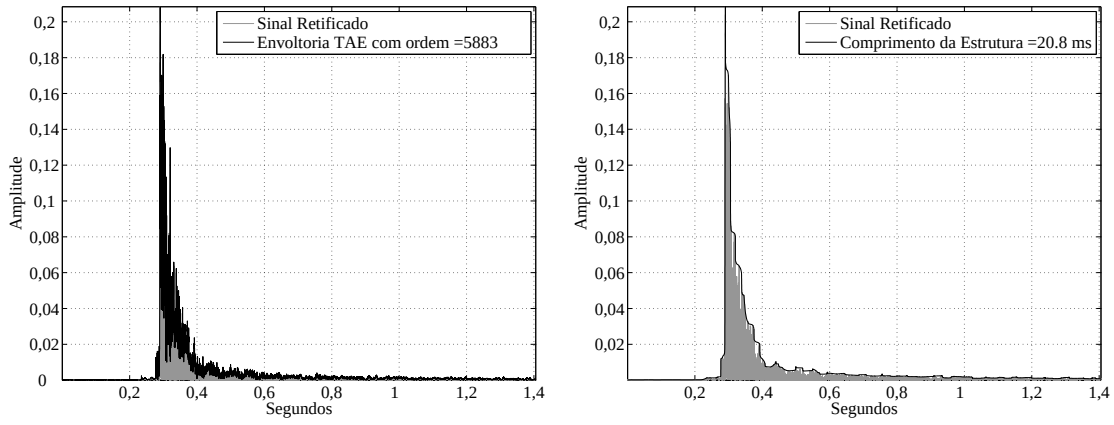
Figura 2.51: Envoltórias da nota Dó2 ($f_0 = 65,41$Hz) de um Violoncelo.



(a) TAE (b) Morfologia

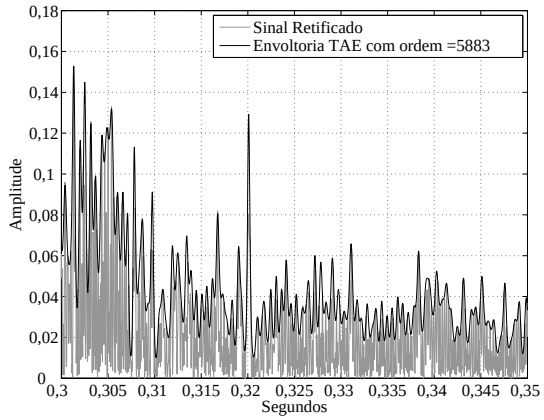
(c) TAE (Detalhe) (d) Morfologia (Detalhe)

Figura 2.52: Envoltórias da nota Lá4 ($f_0 = 440$Hz) de um Piano

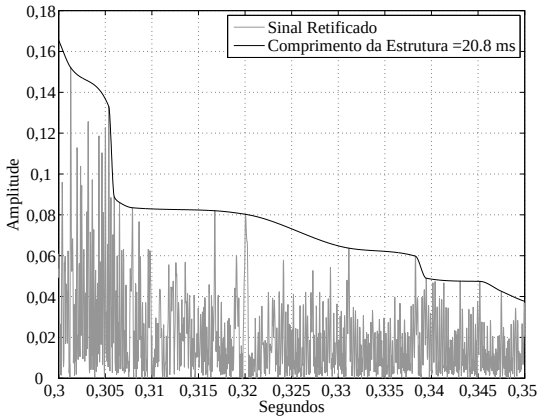


(a) TAE

(b) Morfologia

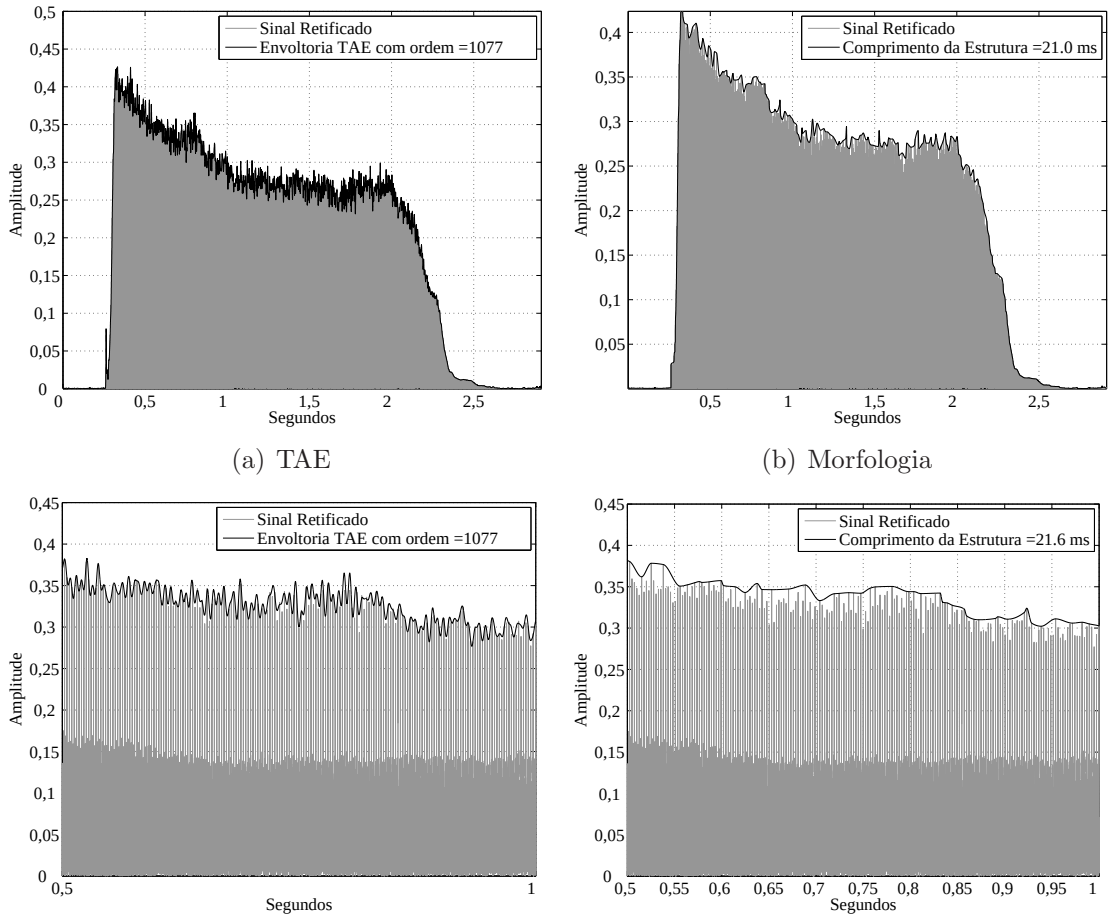


(c) TAE (Detalhe)



(d) Morfologia (Detalhe)

Figura 2.53: Envoltórias da nota Dó8 ($f_0 = 4186,01$Hz) de um Piano



(a) TAE

(b) Morfologia

(c) TAE (Detalhe)

(d) Morfologia (Detalhe)

Figura 2.54: Envoltórias da nota Fá#4 ($f_0 = 369,99$Hz) de uma Harmônica

Capítulo 3

Envoltória de notas sequenciais

Conforme visto na Seção 1.2, sinais musicais em geral são polifônicos, contendo notas tocadas simultaneamente e, em diversas aplicações de análise/ressíntese, uma representação da envoltória do sinal é necessária para fornecer informações da evolução do sinal ao longo do tempo. Entretanto, a forma de onda de sinais polifônicos é resultado da soma de diversos sons simultâneos (mesmo desconsiderando a reverberação), cujas interferências construtivas e destrutivas podem impossibilitar a identificação individual das formas de onda, e consequentemente das envoltórias das notas presentes. Isso faz com que a envoltória desse sinal composto por diversos outros carregue relativamente pouca informação.

Para fins de nomenclatura, este sinal composto será denominado *sinal de mistura* ou, simplesmente, *mistura*. Os sinais que compõem a mistura serão denominados *sinais das fontes originais* ou, simplesmente, *fontes originais*.

Para que seja possível uma melhor interpretação das características temporais de sinais polifônicos, pode-se dividir o sinal em partes e extrair a envoltória de cada uma delas. Cada uma dessas partes será denominada *sinal separado* ou *fonte estimada*.

Uma opção para se conseguir uma melhor representação foi mencionada na Seção 1.2: separar o sinal polifônico em *Fontes Sonoras*. Essa divisão é possível através de um algoritmo de Separação de Fontes, que, idealmente, entrega o sinal proveniente de cada uma das fontes originais do sinal de mistura. Desta forma, pode-se estimar a envoltória de tais fontes isoladamente, utilizando o método proposto no capítulo anterior.

Através da envoltória das fontes é possível se obter informação acerca da evolução temporal de cada nota. Com isso, seria possível inferir qual a técnica ou forma que o instrumentista executou tal nota, se é uma nota em *staccato* ou se a sustentação da nota foi mantida; é possível também que se obtenham informações sobre o ambiente em que a nota está inserida, dentre outras possibilidades.

Lidar com um sinal polifônico, como é o caso de um sinal musical genérico, pode ser extremamente complexo conforme o número de notas que possam ser emitidas

simultaneamente. Uma possível simplificação para permitir estudos de caso seria considerar o caso em que apenas duas notas musicais ocorrem no sinal, podendo ocorrer sobreposição temporal e frequencial entre as notas. Este capítulo se limita a realizar estudos envolvendo este caso simplificado. Por ser um caso intermediário, permite interpretar e julgar mais facilmente as ideias propostas antes de se buscar soluções para o caso mais geral. Neste capítulo é realizada uma discussão de como um algoritmo de separação de fontes se comporta diante do problema dos sinais formados por notas sobrepostas e são feitos estudos sobre as possíveis aplicações para as informações extraídas das notas separadas.

3.1 Escolha do algoritmo de separação

Esta seção se destina a apresentar o algoritmo escolhido para a realização das discussões que serão detalhadas mais adiante no capítulo.

Uma técnica que já foi muito utilizada para separação de fontes é a ICA (*Independent Component Analysis*) [35]; porém, ela assume independência estatística entre as fontes originais e demanda ao menos N sinais de mistura para efetuar a separação de N fontes. Atualmente, a técnica mais largamente utilizada é a *Non-negative Matrix Factorization* (NMF) [36], principalmente porque demanda apenas um sinal de mistura para qualquer número de fontes separadas [4].

Dentro do escopo deste trabalho, utiliza-se a chamada separação *monaural*, que a partir de apenas um sinal que mistura duas fontes sonoras originais deve ser capaz de entregá-las separadas em sua saída.

Assim sendo, a ferramenta de separação escolhida foi a NMF (*Non-Negative Matrix Factorization*) [36].

3.1.1 Non-negative Matrix Factorization (NMF)

A Fatoração de Matrizes Não-negativas, primeiramente apresentada por [36], é um método que permite decompor uma matriz de elementos não-negativos em duas outras, também contendo apenas elementos não-negativos:

$$\mathbf{V} \approx \mathbf{\Lambda} = \mathbf{WH} \tag{3.1}$$

onde $\mathbf{V} \in \mathbb{R}_+^{N\times M}$, $\mathbf{W} \in \mathbb{R}_+^{N\times D}$ e $\mathbf{H} \in \mathbb{R}_+^{D\times M}$ são todas matrizes não-negativas, e $D = \min\{N, M\}$.

O produto $\mathbf{WH}$ é chamado fatoração não-negativa de $\mathbf{V}$, porém $\mathbf{V}$ não é necessariamente igual a $\mathbf{WH}$. Na prática, sua aproximação $\mathbf{\Lambda} \in \mathbb{R}_+^{N\times M}$ é o que realmente

é calculado, sendo necessária a utilização de métodos de otimização para o cálculo dos fatores **W** e **H**.

No presente trabalho será adotada a versão clássica da NMF, que atribui padrões espectrais fixos para cada uma das fontes. Mais detalhes do método de otimização e da função-custo, bem como da medida de distância utilizadas (Divergência de Kullback-Leibler) são encontradas no Apêndice A.

Para o caso da aplicação em áudio, vamos considerar que **V** representa a magnitude (ou valor absoluto) de um espectrograma proveniente de uma STFT, *Short-time Fourier Transform* [37] (uma representação tempo-frequência em que para cada quadro uma DFT é calculada). Cada coluna da matriz **V** representa a DFT de um quadro do sinal e cada linha, a evolução temporal de uma raia da DFT.

A matriz **W** pode ser compreendida como um padrão espectral que se repete ao longo dos quadros do espectrograma representado. Cada coluna dessa matriz contém o padrão de uma fonte. A matriz **H** descreve a intensidade com que cada padrão espectral ocorre em cada quadro [38], o que leva à ideia de evolução temporal ou envoltória. Cada linha dessa matriz contém o padrão temporal para uma fonte.

De forma a ilustrar essa interessante característica das matrizes resultantes da fatoração de um espectrograma, uma fatoração resultando em uma única fonte estimada foi realizada sobre os sinais de notas individuais (fontes originais). Com isso, busca-se uma referência para as matrizes **H** e **W**, de modo a se ter uma ideia do comportamento da NMF no cenário de mais baixa dificuldade.

A Tabela 3.1 mostra os parâmetros utilizados para obtenção do espectrograma utilizado na fatoração de referência. Os valores foram escolhidos buscando um compromisso entre resolução temporal e espectral para a DFT e a sobreposição de 75% permite reconstrução perfeita para uma possível ressíntese, já que a janela utilizada foi a de *Hanning* [16] — mais detalhes sobre o janelamento podem ser encontrados no Apêndice B. Esses parâmetros serão empregados em todos os experimentos apresentados neste capítulo.

Tabela 3.1: Parâmetros utilizados na fatoração de referência

Comprimento da DFT	4096 pontos
Comprimento da janela de análise	20ms
Sobreposição entre janelas adjacentes	75%

Uma representação gráfica bastante intuitiva da saída da NMF pode ser vista na Figura 3.1, obtida para um sinal contendo a emissão da nota Lá 4 ($f_0 = 440$Hz) por uma Flauta. Na figura, o quadro da esquerda representa a matriz **W** (que nesse exemplo é apenas um vetor, já que há apenas uma fonte), ou seja, as raias da representação frequencial; o quadro superior representa a matriz **H** (novamente,

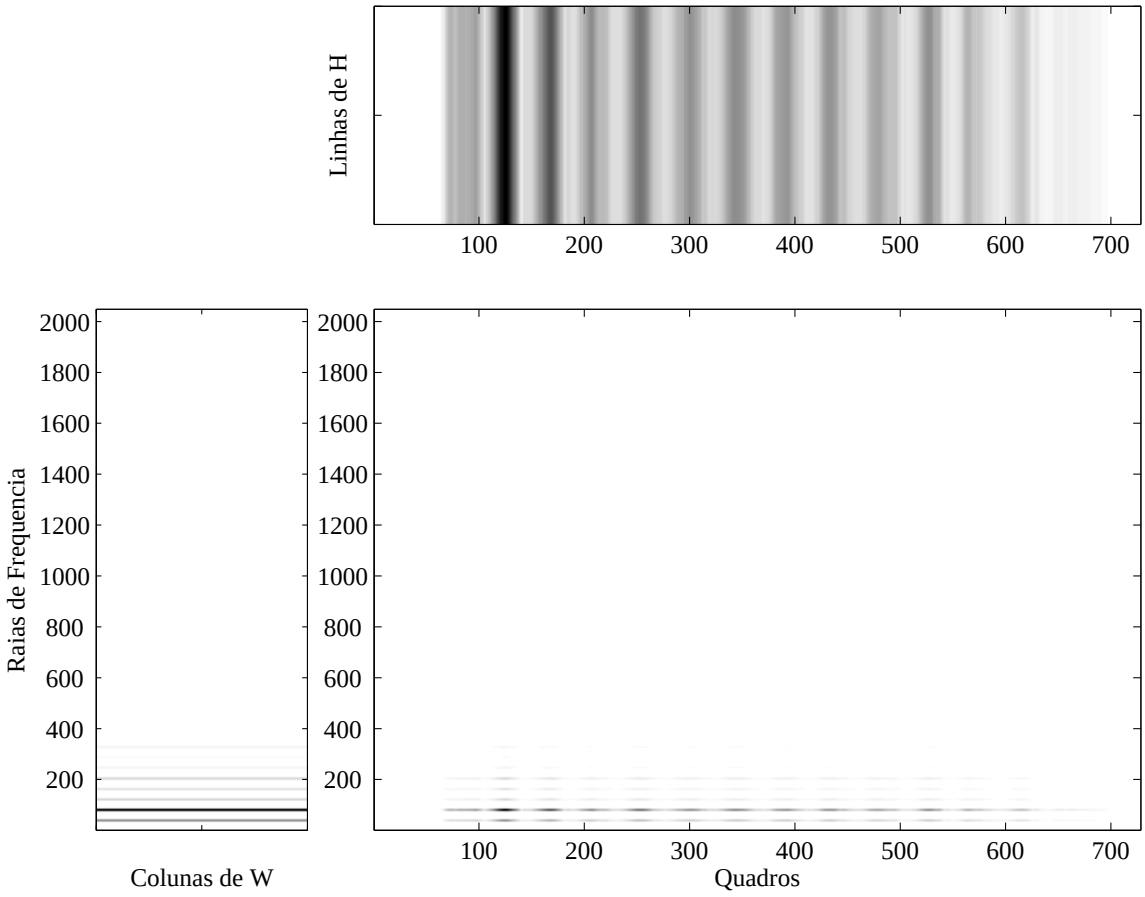


Figura 3.1: Representação gráfica do resultado da fatoração de uma nota Lá 4 de uma Flauta.

apenas um vetor), ou seja, as intensidades da parcela temporal da NMF; e o quadro central representa o espectrograma (matriz $\mathbf{\Lambda}$) resultante da multiplicação de $\mathbf{W}$ por $\mathbf{H}$, conforme a Equação (3.1).

Conforme pode ser visto na Figura 3.2, que representa a forma de onda retificada da nota da Flauta e o vetor $\mathbf{H}$ reamostrado na taxa do sinal, o vetor $\mathbf{H}$ contém informação bastante correlacionada com a envoltória a ser estimada.

Outro exemplo da saída da NMF para um sinal contendo a emissão da nota Sol#2 ($f_0 = 103,83$Hz) por um Piano pode ser visto na Figura 3.3. A Figura 3.4 mostra a forma de onda retificada juntamente com o vetor $\mathbf{H}$ resultante da fatoração.

Por construção, sendo uma nota isolada fatorada como uma fonte única, a NMF deveria fornecer uma representação temporal e frequencial completa. De fato, após realizada a fatoração conforme descrito anteriormente, o sinal foi ressintetizado utilizando a técnica descrita na Seção 3.2 e obteve-se um sinal que, auditivamente, é idêntico ao original.

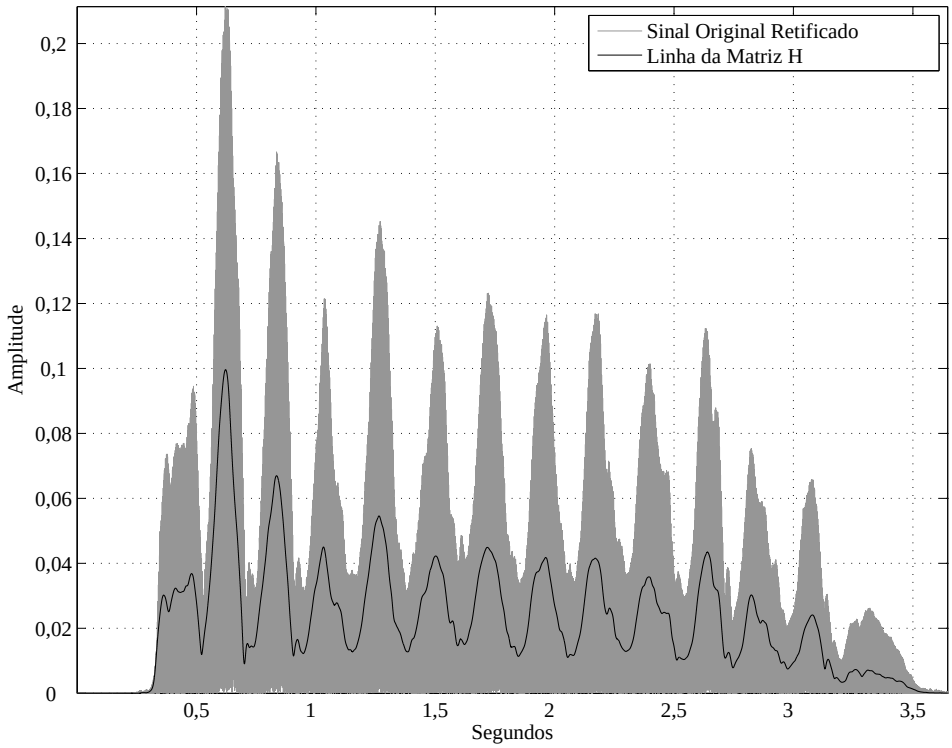


Figura 3.2: Nota Lá 4 de uma Flauta - Saída da NMF - Representação do vetor **H** sobre sinal de entrada.

3.2 Ressíntese das fontes

De acordo com a cadeia descrita na Seção 1.1, após uma separação, pode ser desejável realizar a ressíntese da(s) fonte(s) resultante(s). Devido ao fato de a NMF utilizar apenas a informação de magnitude do espectrograma, esta é também a única informação conhecida sobre os sinais separados. Com isso, faz-se necessário a estimação da fase de cada um desses sinais separados a fim de se obter um sinal real no domínio do tempo.

Para a realização desta estimação de fase foi escolhido o algoritmo RTISI-LA (*Real-Time Interactive Spectrogram Inversion with Look-Ahead*) [39], que realiza a estimação da fase de forma iterativa. Utilizando o espectro de magnitude de uma fonte separada entregue pela NMF e a janela empregada no momento da análise do sinal de mistura, o algoritmo RTISI-LA realiza iterativamente o cálculo de um sinal temporal cujo espectrograma de magnitude seja o mais próximo possível do que foi entregue pela NMF, para a fonte separada em questão. O algoritmo reconstrói quadro a quadro, calculando-os sequencialmente e utilizando também informação dos quadros posteriores no processo iterativo de estimação do sinal no domínio do tempo. Uma exposição mais detalhada sobre alguns algoritmos de síntese pode ser encontrada no Apêndice B.

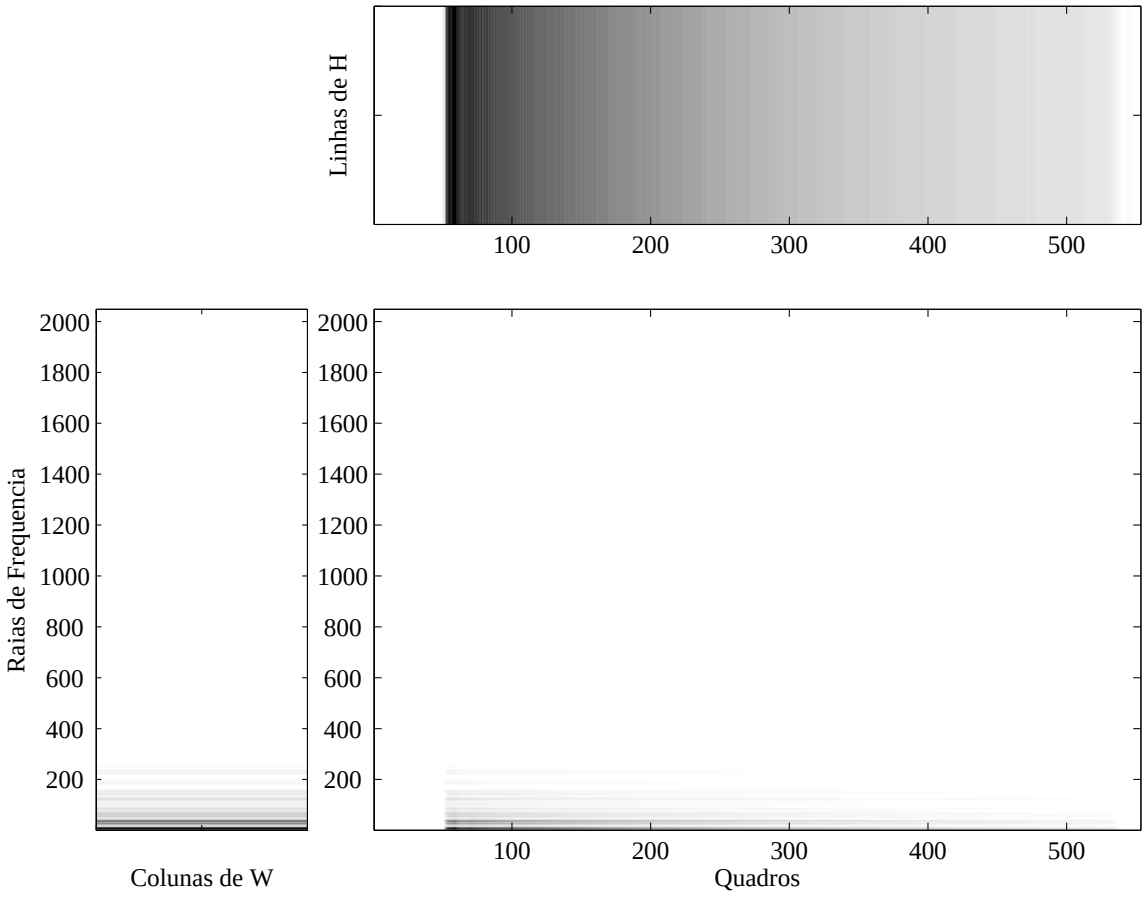


Figura 3.3: Representação gráfica do resultado da fatoração de uma nota Sol#2 de um Piano.

3.3 Metodologia de avaliação

A fim de possibilitar uma avaliação objetiva dos estudos de caso que serão apresentados nas próximas seções, são descritas algumas figuras de mérito apresentadas em [40] especificamente para o problema de separação de fontes.

A fim de facilitar a descrição das métricas, o sinal separado é modelado como em [40]:

$$s_d = s_{\text{alvo}} + e_{\text{inter}} + e_{\text{artef}} + e_{\text{ruído}}, \tag{3.2}$$

onde s_d é a fonte separada, s_{alvo} é a fonte original, e_{inter} é a interferência causada por outras fontes e e_{artef} são os defeitos possivelmente inseridos pelo processo de separação. O termo $e_{\text{ruído}}$ é utilizado caso haja presença de ruído na mistura.

A partir deste modelo, podem ser definidas quatro medidas de qualidade:

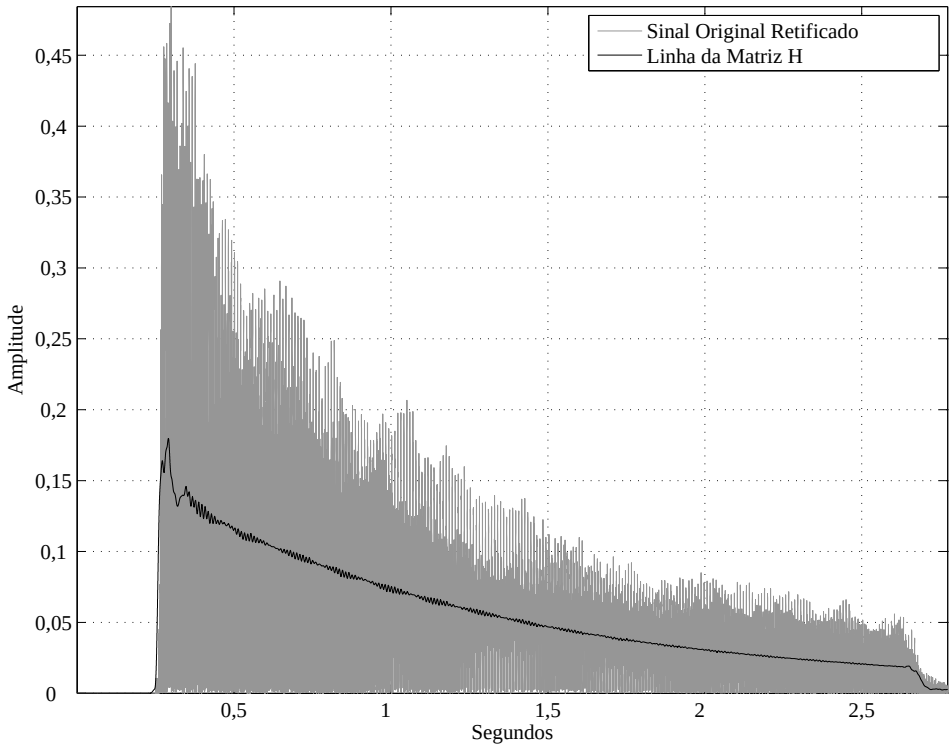


Figura 3.4: Nota Sol#2 de um Piano - Saída da NMF - Representação da matriz **H** sobre sinal de entrada.

***Source-to-Distortion Ratio* (SDR):** A razão fonte-distorção fornece uma medida de qualidade geral da separação. É calculada como

$$\mathrm{SDR} = 10\log_{10}\frac{||s_{\mathrm{alvo}}||^2}{||e_{\mathrm{inter}} + e_{\mathrm{artef}} + e_{\mathrm{ruído}}||^2}. \tag{3.3}$$

***Source-to-Interferences Ratio* (SIR):** A razão fonte-interferência fornece uma medida da potência de sinal das outras fontes que foi inserida na fonte de interesse. É uma medida de qualidade da separação em si, e é calculada como

$$\mathrm{SIR} = 10\log_{10}\frac{||s_{\mathrm{alvo}}||^2}{||e_{\mathrm{inter}}||^2}. \tag{3.4}$$

***Sources-to-Artifacts Ratio* (SAR):** A razão fontes-defeitos fornece uma medida da quantidade de defeitos que foram inseridos no processo de separação, ou seja, a inserção de elementos que não estavam presentes na mistura original. A SAR é calculada como

$$\mathrm{SAR} = 10\log_{10}\frac{||s_{\mathrm{alvo}} + e_{\mathrm{inter}} + e_{\mathrm{ruído}}||^2}{||e_{\mathrm{artef}}||^2}. \tag{3.5}$$

***Sources-to-Noise Ratio* (SNR):** Nos casos em que a mistura contém ruído, a razão fontes-ruído fornece uma medida da quantidade de ruído que restou junto às fontes.

$$\mathrm{SNR} = 10\log_{10}\frac{||s_{\text{alvo}} + e_{\text{inter}}||^2}{||e_{\text{ruído}}||^2}. \tag{3.6}$$

As medidas são invariantes ao ganho de uma fonte em relação à outra, e à ordenação dos sinais. Isto significa que variações de ganho não são penalizadas, e que cada estimativa de fonte separada é comparada com todas as fontes originais, e aquela que possuir maior SDR é considerada a fonte correta. Todas as medidas podem ser calculadas utilizando-se o pacote disponível em [41].

3.4 Escolha dos sinais para os testes

Nesta seção, serão descritos os sinais a serem utilizados nos testes que empregam sinais com contribuições de duas fontes sonoras. As notas isoladas foram extraídas dos instrumentos citados no início do trabalho (Piano, Violoncelo, Flauta e Clarineta), tendo como representantes notas musicais dispostas em intervalos de quarta e de sétima aumentada, por razões que serão expostas ao longo do texto.

Na Tabela 3.2 pode ser visto um resumo dos sinais de mistura formados por notas com sobreposição. As misturas foram obtidas através da soma dos sinais emitidos por cada instrumento. Na Tabela 3.2, o *Onset* 1, correspondente ao início da primeira nota, sempre está em 0 (zero); o *Offset* 1 corresponde ao final da primeira nota; e assim por diante. Para algumas explanações serão utilizados sinais de mistura contendo notas não-sobrepostas. Estes são mostrados na Tabela 3.3.

O Piano foi escolhido por ser um instrumento de percussão com emissão de altura fixa e decaimento livre (quando o instrumentista deixa o pedal de sustentação acionado) e extensão ampla; a Flauta foi escolhida por possuir componentes ruidosas de alta frequência, devido ao sopro do instrumentista; e a Clarineta foi escolhida por apresentar um padrão espectral relativamente constante ao longo da emissão das notas.

Tabela 3.2: Misturas utilizadas nos testes (valores de *Onset* e *Offset* em segundos)

Misturas	Clarineta Lá4 + Clarineta Ré5	Flauta Lá4 + Piano Sol#2	Piano Sol#2 + Flauta Lá4
Sobreposição (%)	53	44	44
Offset 1	4,08	3,40	3,10
Onset 2	1,08	1,40	1,10
Offset 2	5,63	4,49	4,49
Razão frequencial	4:3	15:8	15:8

Tabela 3.3: Misturas (notas não sobrepostas) utilizadas nos testes (valores de *Onset* e *Offset* em segundos)

Misturas sem sobreposição	Clarineta Lá4 + Clarineta Ré5	Flauta Lá4 + Piano Sol#2	Piano Sol#2 + Flauta Lá4
Sobreposição (%)	0	0	0
Offset 1	4,08	3,40	3,10
Onset 2	5,08	4,40	4,10
Offset 2	9,63	7,50	7,50
Razão frequencial	4:3	15:8	15:8

3.5 Aplicação da NMF para extração de envoltória

Conforme exposto no início do capítulo, busca-se aqui uma forma de estimação de envoltória para notas sobrepostas. Foi adotado o algoritmo de NMF para tentar separar o sinais e extrair a envoltória deles. Embora a técnica proposta para atacar o problema de notas sobrepostas seja um método de separação de fontes, uma boa "separação de envoltórias" é o objetivo final.

Existem três abordagens possíveis para estimar envoltórias a partir da saída da NMF:

1. Interpolar a matriz **H** até a taxa de amostragem do sinal (conforme mostrado na Seção 3.1.1);

2. Ressintetizar cada uma das fontes e estimar a envoltória desses sinais ressintetizados;

3. Lançar mão de um pós-processamento sobre a saída do algoritmo, com a finalidade de extrair alguma informação do decaimento de cada nota [42].

Contudo, antes de atacar diretamente o problema da envoltória para mais de uma fonte, é necessária uma análise de como a NMF se comporta ao ser aplicada sobre o tipo de sinal sub análise. Analogamente ao realizado na seção anterior, porém agora com sinais envolvendo notas musicais sobrepostas, analisa-se a saída da NMF para alguns casos.

Após a fatoração de diversos sinais, foram escolhidas as misturas mostradas na Tabela 3.2 para ilustrar os resultados. A princípio, serão mostrados resultados das misturas formadas a partir de uma nota Sol#2 (f_0 = 103,83Hz) de um Piano e uma nota Lá4 (f_0 = 440Hz) de uma Flauta. Esta escolha justifica-se por conter dois tipos diferentes de instrumentos, sendo um percussivo (com emissão de altura fixa) e um de sopro (com emissão de altura variável). Além disso, o piano contém

várias ressonâncias inerentes à construção do corpo do instrumento e a flauta tem como complicador o sopro do instrumentista. A combinação ainda é interessante por ser um intervalo de sétima aumentada, com 11 semitons de diferença — as duas oitavas de distância entre as notas busca abranger uma maior faixa de frequências no espectro. Este intervalo é considerado o mais dissonante possível (depois do intervalo de segunda menor, que é seu inverso), pois não há coincidência de harmônicos entre as notas. Dessa forma, tem-se uma mistura que permite analisar o desempenho do método de separação.

A próxima seção mostra resultados da fatoração de tais misturas em dois sinais separados.

3.6 Fatoração em Duas Fontes

Além da representação gráfica já apresentada nas fatorações com uma única fonte, uma representação alternativa será também utilizada na representação das saídas das fatorações com duas ou mais fontes: curvas representando cada linha da matriz **H** e cada coluna da matriz **W**. A ordem das fontes é sempre lida **de baixo para cima**: as saídas para a primeira fonte estão nas curvas inferiores e assim por diante.

As Figuras 3.5 e 3.6 mostram um exemplo de fatoração em duas fontes, tendo como entrada o sinal Piano Sol#2 + Flauta Lá4, nesta ordem, conforme descrito na última coluna da Tabela 3.2.

Em seguida, gerou-se o sinal de mistura composto pelas notas Lá4 de Flauta + Sol#2 de Piano, conforme descrito na terceira coluna da Tabela 3.2 e realizou-se a fatoração com os mesmos parâmetros; e a saída da NMF é mostrada nas Figuras 3.9 e 3.10.

3.6.1 Análise da fatoração

Observando as saídas das fatorações, nas Figuras 3.6 até 3.12, são feitas algumas observações.

A ordem das notas no sinal de entrada altera o resultado; ou seja, apesar de a fatoração ser calculada diretamente sobre todos os quadros do espectrograma utilizado como entrada do algoritmo de separação, o resultado da fatoração é dependente da ordem das notas. A fatoração resultante do sinal Piano Sol#2 + Flauta Lá4 é diferente da fatoração do sinal Flauta Lá4 + Piano Sol#2, conforme pode ser visto, nas matrizes **H** e **W**, representadas nas Figuras 3.5 e 3.9, respectivamente.

Outro ponto interessante a notar é a presença de um "pré-eco" na matriz **H** da Figura 3.5. Observando as curvas da representação de H, supõe-se que a curva inferior corresponda à emissão da Flauta devido às suas ondulações significativas

(comparar com o vetor **H** oriundo da fatoração da mesma nota de Flauta, na Figura 3.2), e a linha superior, que corresponda à emissão do Piano. Entretanto, apesar de a emissão da Flauta ser iniciada após a emissão do Piano, o valor da linha de **H** correspondente à Flauta possui valores não-nulos na região temporal em que só o Piano está presente. Esse pré-eco possui um formato muito similar ao *onset* da nota do Piano, como pode ser visto na Figura 3.7.

Ao escutar as duas fontes ressintetizadas percebe-se uma emissão parecida à do Piano em termos temporais, porém com o conteúdo frequencial da Flauta. Isto pode ser explicado pelo fato de o conteúdo espectral da nota da Flauta estar contido nas frequências presentes na nota do Piano — que é um Sol#2, portanto mais grave que a Flauta (Lá4). A NMF busca padrões temporais e espectrais e, encontrando uma parte do espectro da Flauta na nota do Piano, reúne o que "parece" semelhante.

O fato interessante aqui é que na fatoração envolvendo as mesmas notas, porém invertendo a ordem das mesmas no sinal de entrada (Flauta + Piano), esse efeito é menos aparente, vide Figuras 3.11 e 3.12. Inclusive, este pré-eco não foi percebido em testes informais de audição. Analogamente à explanação anterior, a parte do conteúdo espectral do Piano semelhante à Flauta é agregado à fonte correspondente à Flauta pela NMF; a diferença agora é que, pelo fato de a emissão do Piano ocorrer após a Flauta estar soando, essa agregação de informação frequencial não é percebida, justamente porque a Flauta e essa "parcela" do espectro do Piano se sobrepõem.

Em ambos os casos, na fonte correspondente ao Piano, praticamente não se percebe a presença de elementos pertencentes à Flauta. Isto ocorre devido ao fato de que a parcela da Flauta presente na nota de Piano é somada às parciais mais agudas da nota do Piano, pois a maior energia da Flauta está na sua f_0 — 440Hz (Lá4) — que é mais que duas oitavas mais aguda que a f_0 da nota do Piano — 103,83Hz (Sol#2).

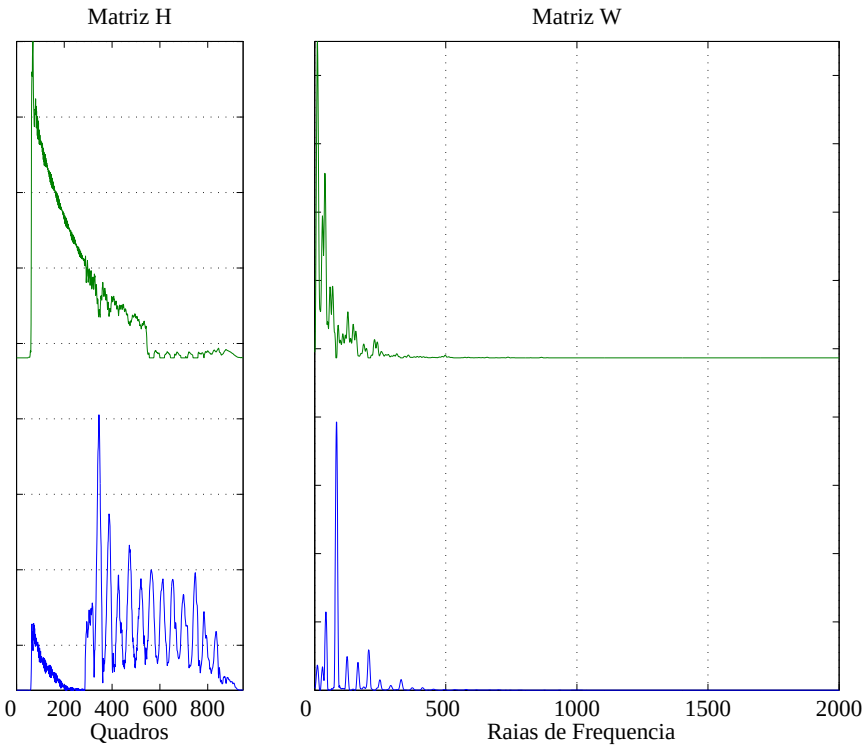


Figura 3.5: Matrizes **H** e **W** - Piano Sol#2 e Flauta Lá4

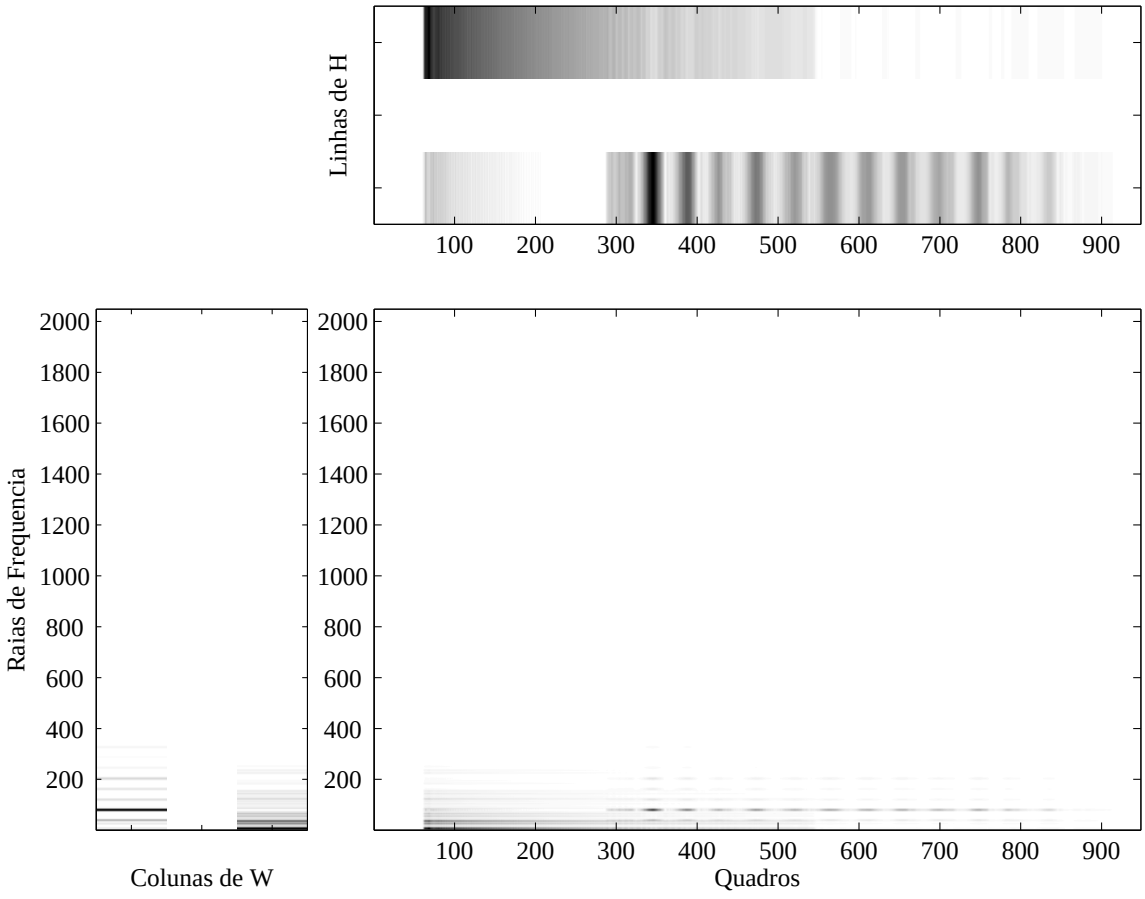


Figura 3.6: Representação gráfica do resultado da fatoração - Piano Sol#2 e Flauta Lá4

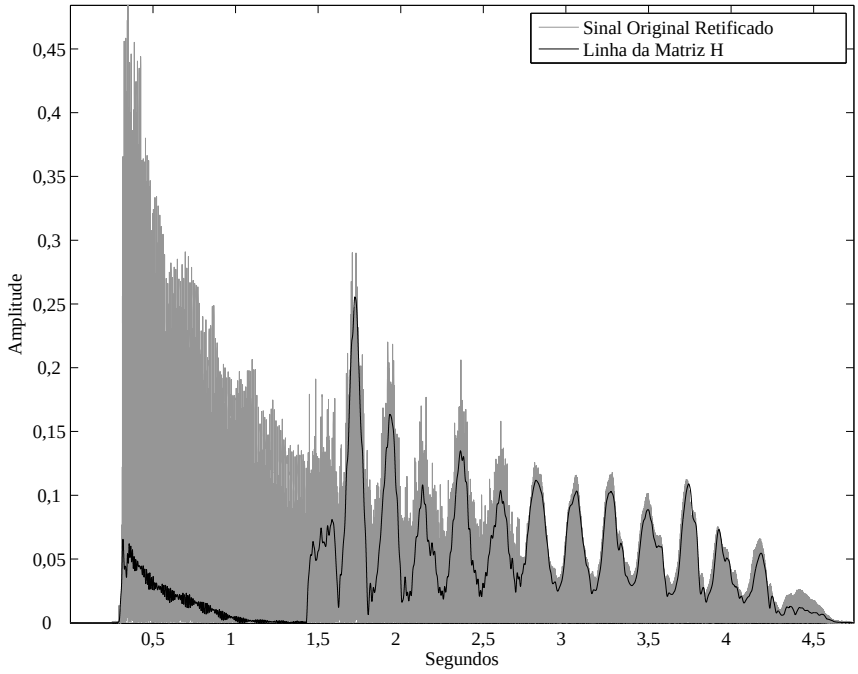


Figura 3.7: Linha 1 da matriz **H** sobre sinal original retificado - Piano Sol#2 e Flauta Lá4

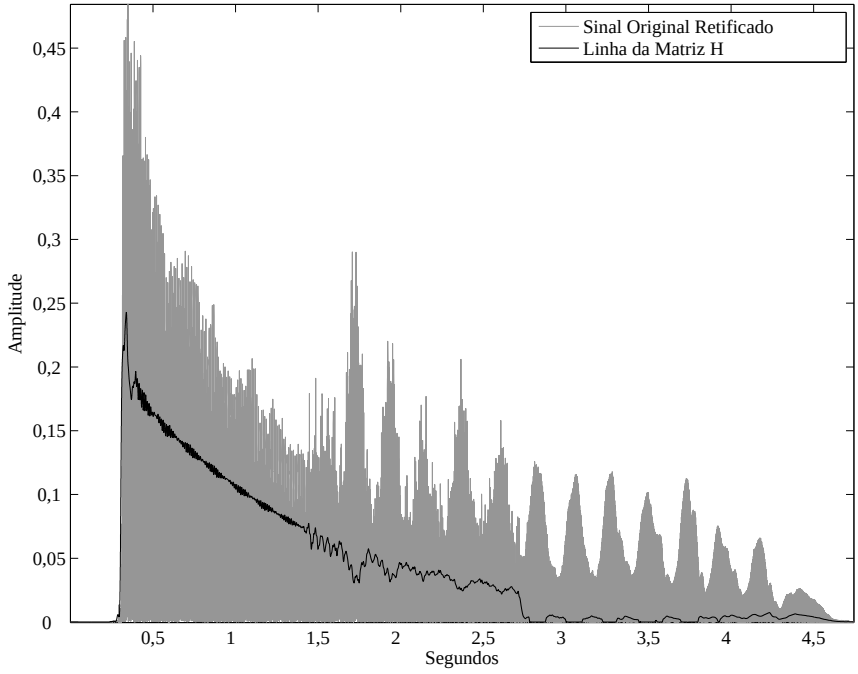


Figura 3.8: Linha 2 da matriz **H** sobre sinal original retificado - Piano Sol#2 e Flauta Lá4

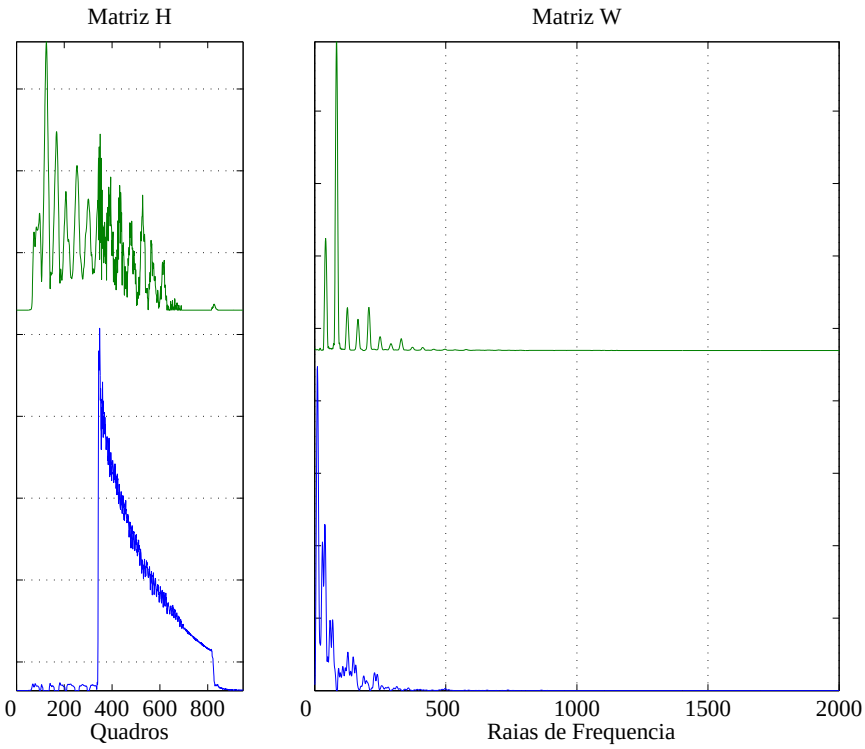


Figura 3.9: Matrizes **H** e **W** - Flauta Lá4 e Piano Sol#2

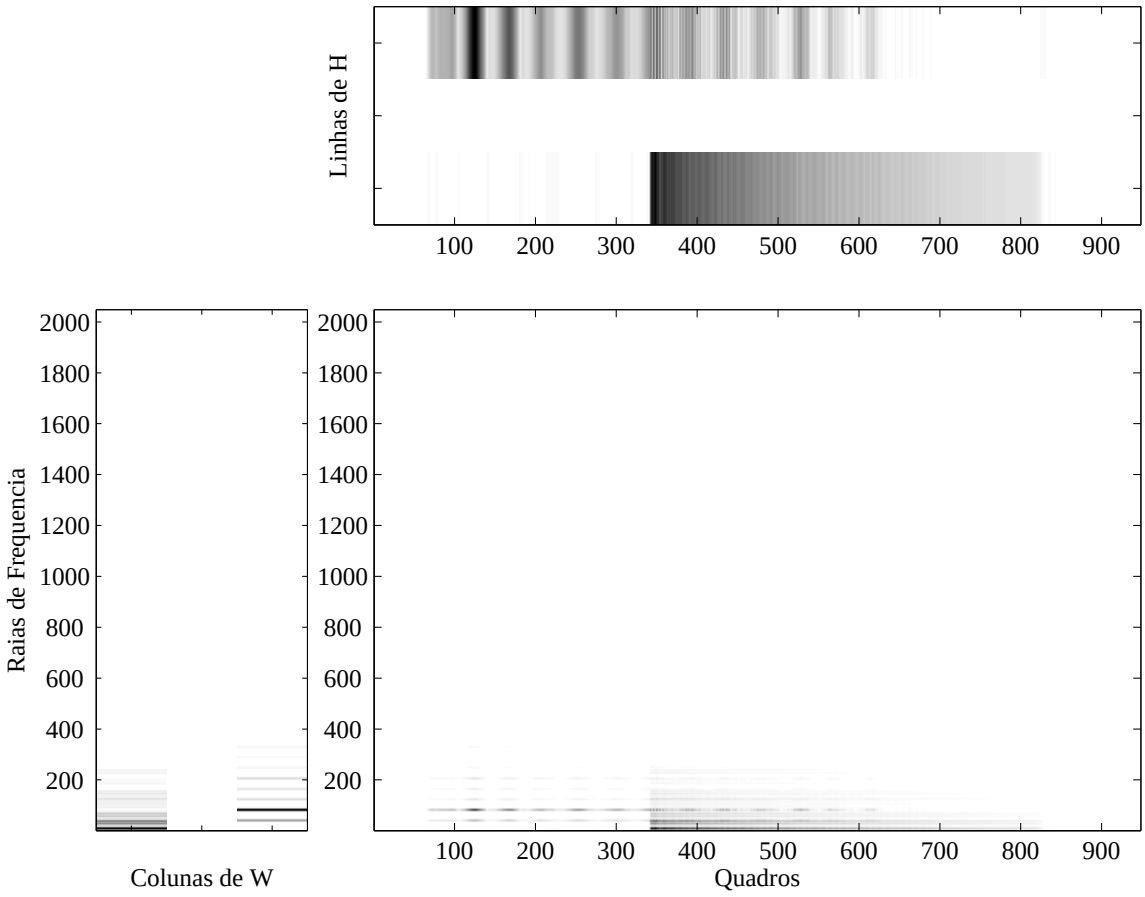


Figura 3.10: Representação gráfica do resultado da fatoração - Flauta Lá4 e Piano G#2

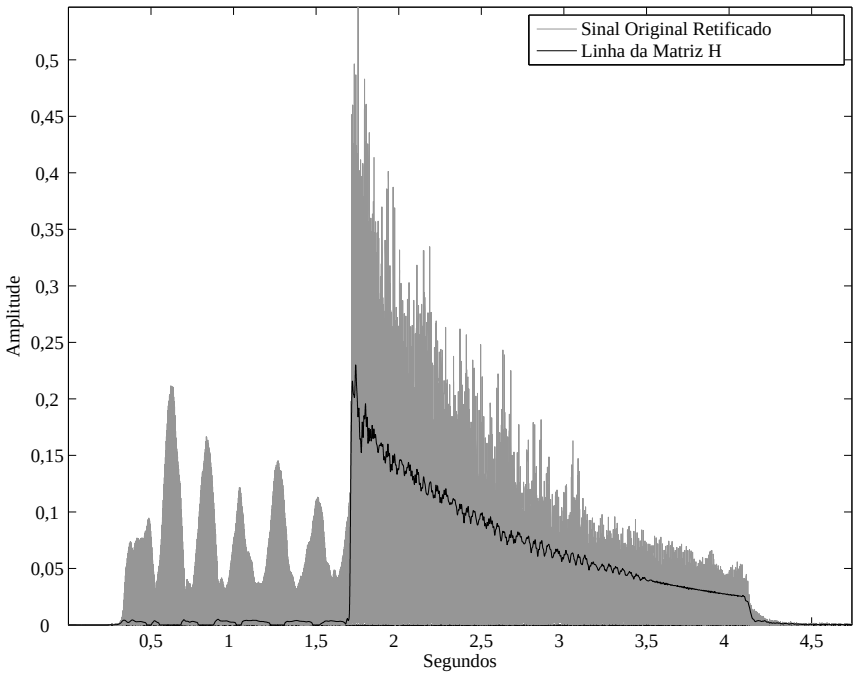


Figura 3.11: Linha 1 da matriz **H** sobre sinal original retificado - Flauta Lá4 e Piano Sol#2

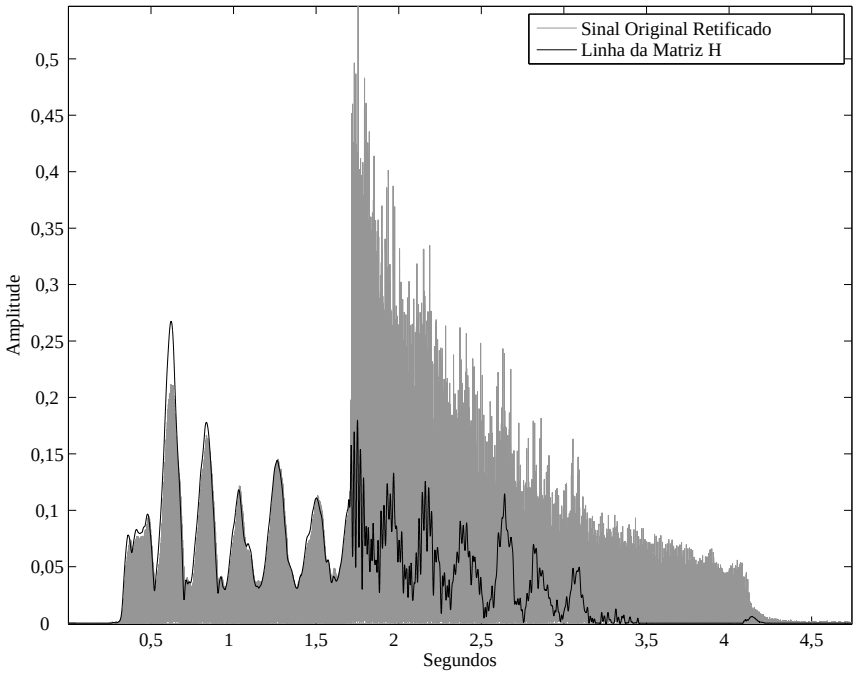


Figura 3.12: Linha 2 da matriz **H** sobre sinal original retificado - Flauta Lá4 e Piano Sol#2

3.6.2 Análise do comportamento da NMF para notas não-sobrepostas

A seção anterior tratou de exemplificar alguns resultados obtidos com a fatoração via NMF de sinais contendo notas sobrepostas. Algumas características interessantes foram ressaltadas, dentre elas o fato de a NMF tender a juntar padrões espectrais parecidos de fontes distintas. A fim de ilustrar essa característica da NMF, as Figuras 3.13 a 3.18 apresentam os resultados da fatoração de sinais gerados a partir das mesmas notas de Piano e Flauta, porém com notas dispostas de maneira a não haver sobreposição entre elas. Estes sinais são detalhados na Tabela 3.3.

O fato de que ambas as notas são "apresentadas" à NMF simultaneamente — vale ressaltar que a fatoração se dá a partir do espectrograma da mistura, que possui informações das duas notas presentes na mistura — faz com que essa aglomeração de padrões espectrais semelhantes (aos olhos da NMF) seja realizada mesmo quando a mistura possui as duas notas sem sobreposição.

Analisando as figuras, as mesmas características dos resultados envolvendo notas sobrepostas são encontradas. Isto pode ser notado nas Figuras 3.13 e 3.16, em que parte do piano aparece na fonte que seria correspondente à flauta e vice versa.

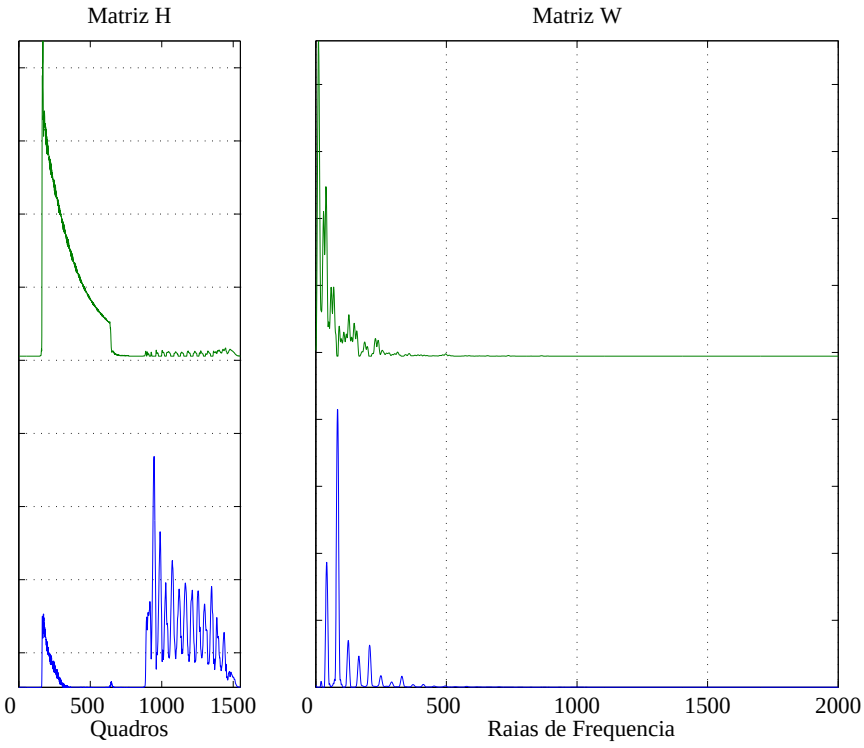


Figura 3.13: Matrizes **H** e **W** - Piano Sol#2 e Flauta Lá4

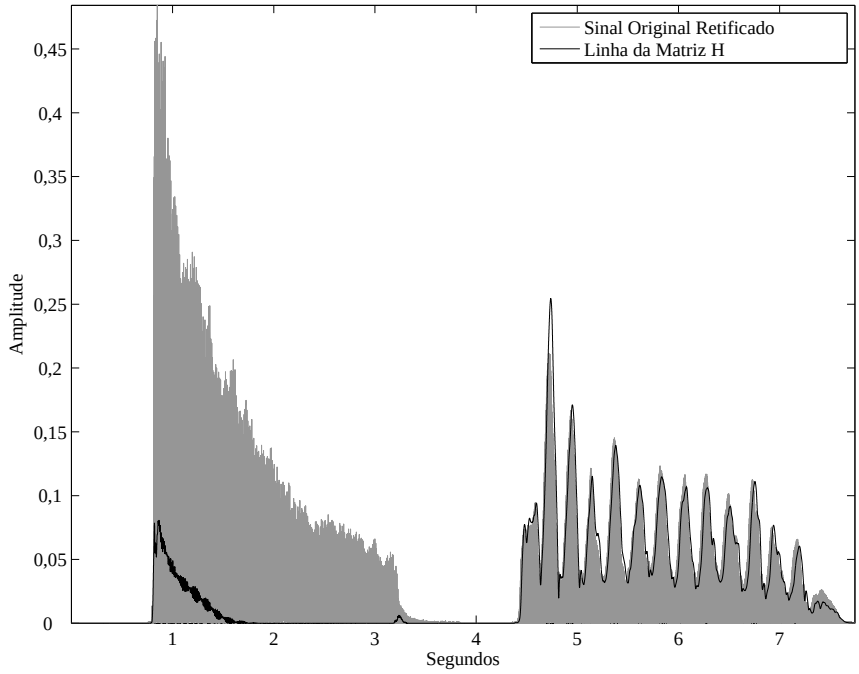


Figura 3.14: Linha 1 da matriz **H** sobre sinal original retificado - Piano Sol#2 e Flauta Lá4

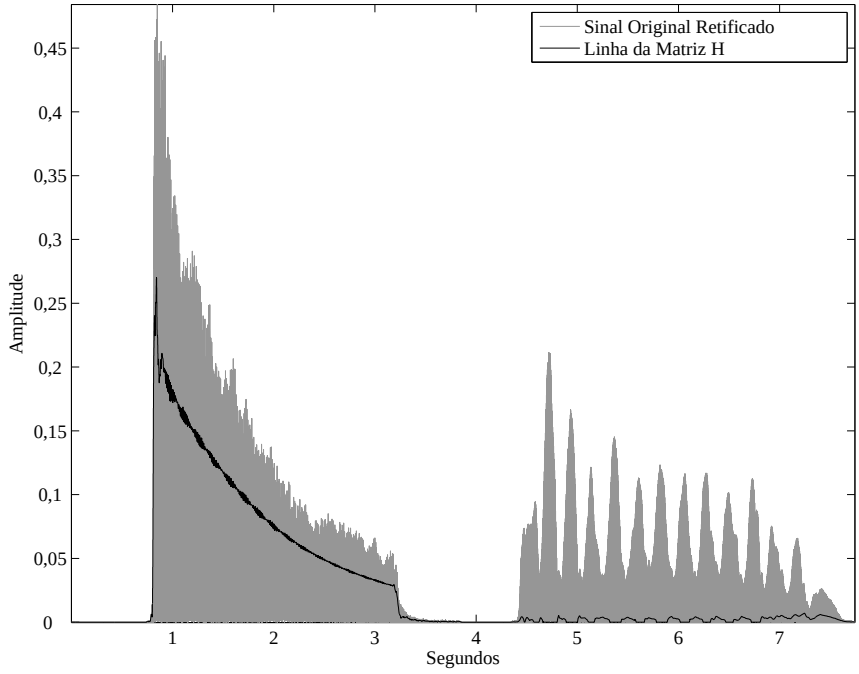


Figura 3.15: Linha 2 da matriz **H** sobre sinal original retificado - Piano Sol#2 e Flauta Lá4

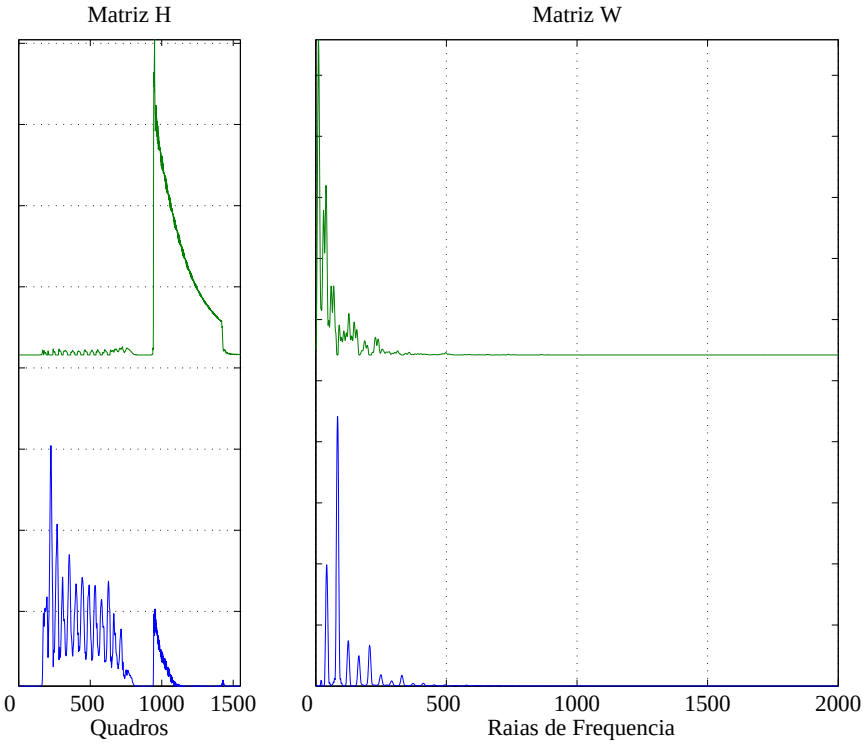


Figura 3.16: Matrizes **H** e **W** - Flauta Lá4 e Piano Sol#2

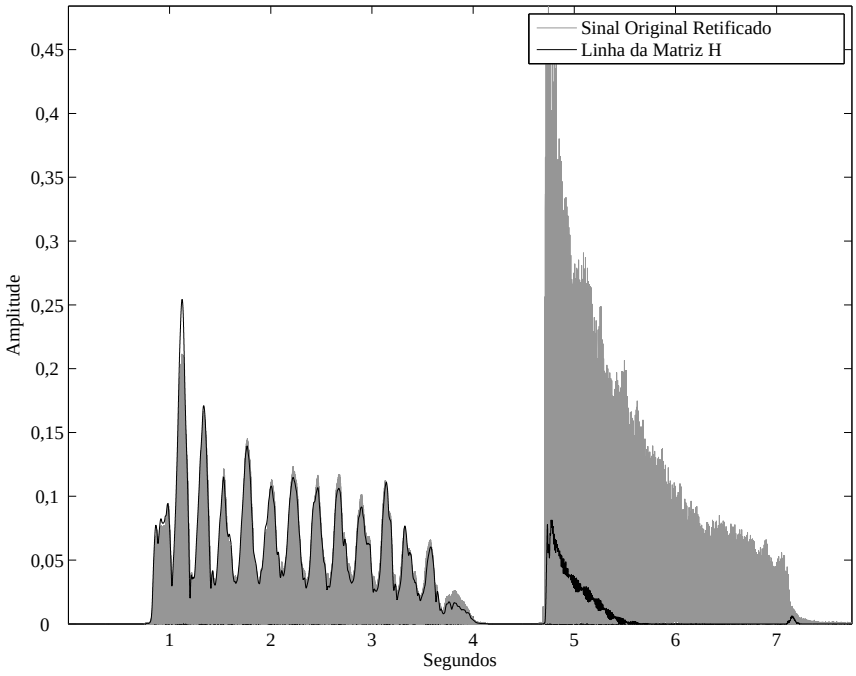


Figura 3.17: Linha 1 da matriz **H** sobre sinal original retificado - Flauta Lá4 e Piano Sol#2

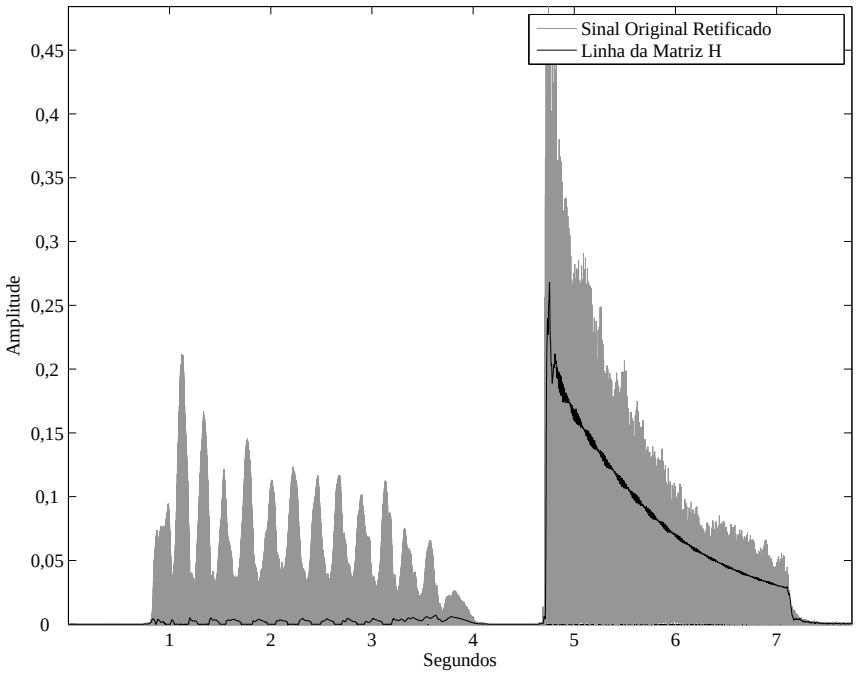


Figura 3.18: Linha 2 da matriz **H** sobre sinal original retificado - Flauta Lá4 e Piano Sol#2

3.7 Envoltória obtida diretamente da saída da NMF

Uma vez que se sabe que a matriz **H** possui informação temporal, pode-se pensar em usá-la diretamente para a obtenção da envoltória das fontes, bastando apenas interpolá-la até a taxa de amostragem do sinal de entrada.

Esse seria o método mais eficiente do ponto de vista computacional, porém é o mais dependente da eficácia da separação: a envoltória será tão boa quanto a descrição temporal das fontes provida pela matriz **H**.

Como parte da informação de uma fonte pode estar presente na estimativa da outra, talvez trabalhar diretamente sobre as fontes estimadas ressintetizadas forneça mais elementos para suas respectivas envoltórias; assim sendo, pode-se pensar em reconstruir ambas as fontes de modo a se obter cada um dos sinais de saída reconstruídos no domínio do tempo e estimar a envoltória de cada um deles.

A principal diferença entre simplesmente interpolar a matriz **H** e estimar a envoltória da respectiva fonte reconstruída e ressintetizada reside no fato de que a matriz **W** pode influenciar a forma de onda final e, consequentemente, a envoltória da fonte em questão. Assim sendo, a interpolação considera unicamente informação temporal — matriz **H** — e a estimação por ressíntese da fonte considera toda a informação fornecida pela NMF, tanto temporal quanto espectral — matriz **W**.

A Tabela 3.4 ilustra os valores de SDR, SIR e SAR para a separação das misturas [Piano Sol#2 + Flauta Lá4], [Flauta Lá4 + Piano Sol#2] e [Clarineta Lá4 + Clarineta Ré5] (todos descritos na Tabela 3.2) e a coluna "Ref." indica qual fonte original foi identificada como sendo mais parecida com a fonte estimada cujas avaliações são apresentadas na linha em questão. Tais medidas (descritas na Seção 3.3) serão utilizadas nas próximas seções a fim de possibilitar uma comparação objetiva entre os resultados obtidos com cada um dos estudos de caso que serão detalhados nas próximas seções.

A escolha da Clarineta, conforme já foi explicitado anteriormente, se deveu à sua característica de apresentar um padrão espectral constante ao longo da emissão das notas, e essa combinação (Lá4 e Ré5) é um intervalo de quarta justa (de razão 4:3) escolhido por ser um intervalo consonante e com um bom número de coincidências de harmônicos.

Notam-se, em todos os casos, valores de SDR e SAR negativos e valores de SIR positivos. Isso indica que há pouca potência de uma fonte inserida na outra e vice-versa (SIR positivos); além disso, indica a presença de uma grande quantidade de defeitos inseridos (SAR negativos), possivelmente devido ao processo de síntese.

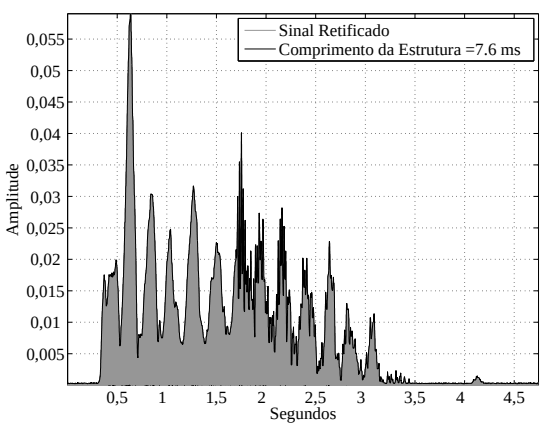
A Figura 3.19 ilustra as fontes estimadas pela NMF e as suas respectivas en-

Tabela 3.4: Figuras de mérito do resultado da separação. São mostrados resultados de sinais de mistura formados por notas isoladas (I) e sobrepostas (S) entre si.

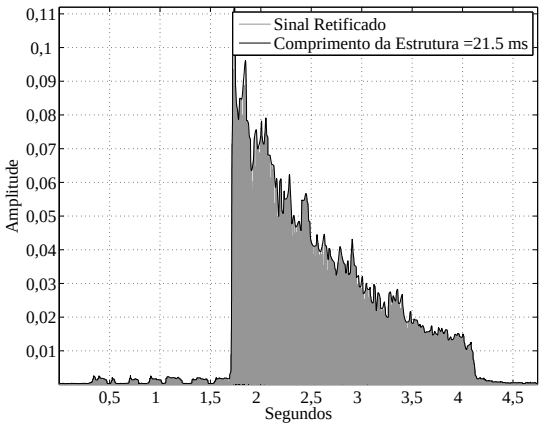
Sinal		SDR	SIR	SAR	Ref.
Piano Sol#2	I	-8,43	19,35	-8,37	Lá4
Flauta Lá4		-14,83	23,10	-14,81	Sol#2
	S	-11,96	15,85	-11,84	Lá4
		-24,02	13,32	-23,82	Sol#2
Flauta Lá4	I	-13,65	29,26	-13,65	Lá4
Piano Sol#2		-12,19	19,72	-12,14	Sol#2
	S	-12,61	29,95	-12,61	Lá4
		-8,82	23,69	-8,80	Sol#2
Clarineta Lá4	I	-6,9409	20,5941	-6,8955	Lá4
Clarineta Ré5		-3,8876	28,0708	-3,8781	Ré5
	S	-9,5585	16,2909	-9,4464	Ré5
		-3,7866	26,1187	-3,7716	Lá4

voltórias. Para fins de comparação, as linhas da matriz **H** resultantes da separação da mistura [Flauta Lá4 + Piano Sol#2, sobrepostas] foram interpoladas até a taxa de amostragem do sinal de mistura e dispostas juntas. Pode-se notar claramente na Figura 3.20 a região de sobreposição entre as notas e como a matriz **H** dispõe a informação temporal de cada uma das fontes estimadas.

De forma a conhecer as limitações do método de separação e explorar a NMF buscando melhorar a separação dos sinais e a estimação das envoltórias das notas presentes na mistura, foram realizados alguns estudos de caso envolvendo as matrizes entregues pela fatoração por NMF.



(a) Envoltória da Fonte 1.



(b) Envoltória da Fonte 2.

Figura 3.19: Exemplo de envoltórias das fontes resultantes da fatoração da mistura Flauta Lá4 + Piano Sol#2, notas sobrepostas.

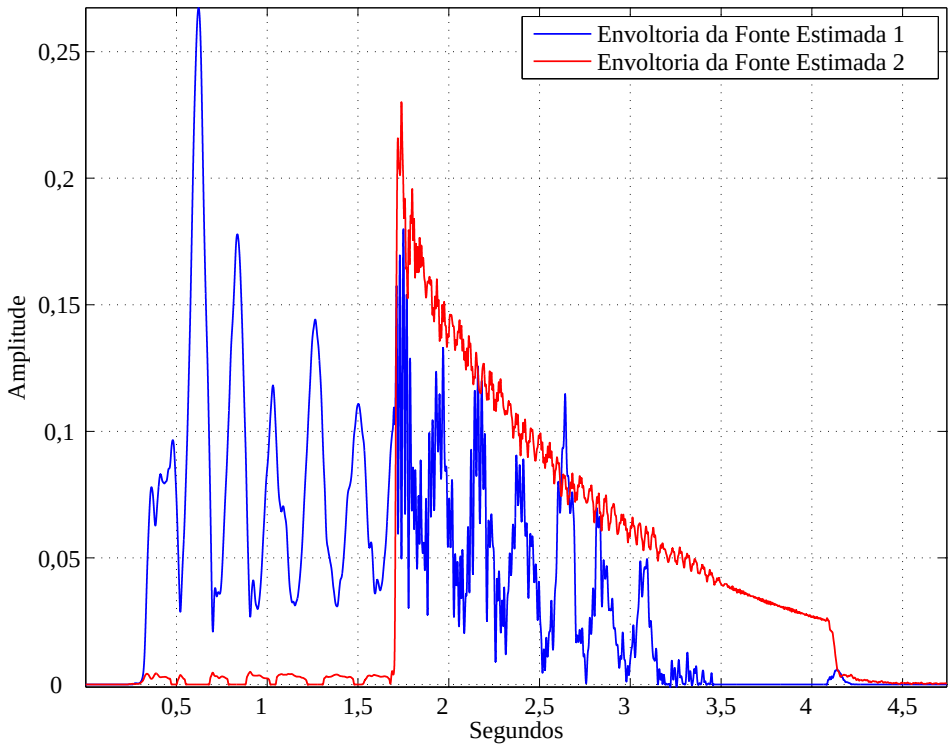


Figura 3.20: Envoltórias dos sinais estimados - Flauta Lá4 e Piano Sol#2

3.8 Caso 1: Envoltória a partir do processamento da envoltória da ressíntese

O processo detalhado a seguir busca tentar melhorar a envoltória obtida a partir do sinal separado fornecido pela NMF com informação das envoltórias obtidas a partir das estimativas das fontes separadas.

A hipótese do presente caso assume que não se possui nenhum conhecimento acerca dos sinais originais misturados; desta forma, resta apenas a informação oriunda da separação, as matrizes **W** e **H**. Uma vez que o foco é a envoltória, a matriz **W** não será modificada. Fazendo a matriz **H** com todos os elementos iguais a um e a matriz **W** inalterada, é gerado um sinal "sem envoltória" que posteriormente receberá a envoltória da fonte estimada (entregue pela NMF).

A reconstrução da fonte utilizando apenas a matriz **W** gera um sinal "sem envoltória", o que elimina o efeito da matriz **H** sobre o sinal ressintetizado no item 2 — efeito esse que pode ser nocivo à qualidade da separação (e consequentemente da envoltória) caso a matriz **H** não tenha sido bem estimada.

Todas as estimações de envoltórias foram realizadas utilizando o algoritmo desenvolvido no trabalho (Seção 2.3) e todas as reconstruções e ressínteses, quando citadas, utilizaram o algoritmo RTISI-LA, previamente descrito na Seção 3.2 e detalhado no Apêndice B.

Possuindo a informação das matrizes **W** e **H** e de cada uma das fontes resultantes da separação ressintetizadas, a sequência de passos seguidos para o estudo de caso 1 é detalhada a seguir:

1. Estima-se a envoltória do sinal resultante da ressíntese de uma fonte oriunda do algoritmo de NMF. Este sinal será denominado Sinal A.

2. Realiza-se a reconstrução da fonte, utilizando a matriz **W** e a matriz **H** feita unitária (todos os elementos iguais a 1), e realiza-se a ressíntese.

3. Aplica-se a envoltória previamente estimada sobre o sinal ressintetizado na etapa anterior. O sinal aqui gerado será denominado Sinal B.

4. Calcula-se a diferença das energias das envoltórias estimadas dos Sinais A e B. Caso essa diferença seja menor que 30% da energia da ressíntese, o processo se encerra e o Sinal B será a nova fonte ressintetizada.

5. Caso contrário, o Sinal B se torna o Sinal A e o processo recomeça.

A Tabela 3.5 ilustra os valores de SDR e SAR negativos, e SIR positivos. Tal resultado mostra que o processo não alterou a característica da separação (detalhada

na Seção 3.7) e, para fins de comparação, as três últimas colunas mostram a diferença entre os valores resultantes do processo do Caso 1 e a avaliação utilizando as fontes ressintetizadas diretamente após a reconstrução a partir da saída da NMF, presentes na Tabela 3.4 — que serão sempre os valores de referência já que são o que a NMF consegue entregar por si só. Valores positivos nas diferenças indicam melhoras e valores negativos, pioras.

A escolha do critério de parada de 30% da diferença de energias é empírico. Diversas maneiras de se interromper o processo foram testadas e, notou-se que a estimação iterativa da envoltória tende a acentuar os picos e vales presentes nas formas de onda dos sinais oriundos da NMF. Por exemplo, na Figura 3.21(a), a fonte estimada pela NMF apresenta vales em sua forma de onda. Após algumas iterações (resultado mostrado na Figura 3.21(b)), o processo tende a acentuar tais vales e atenuar os picos, degradando ainda mais as estimações. Isso pode ser observado comparando-se as Figuras 3.21 e 3.23(a), que é a forma de onda original da nota de Clarineta Lá4 utilizada na mistura. Tal degradação é comprovada observando-se os valores negativos de SDR e SAR na Tabela 3.5.

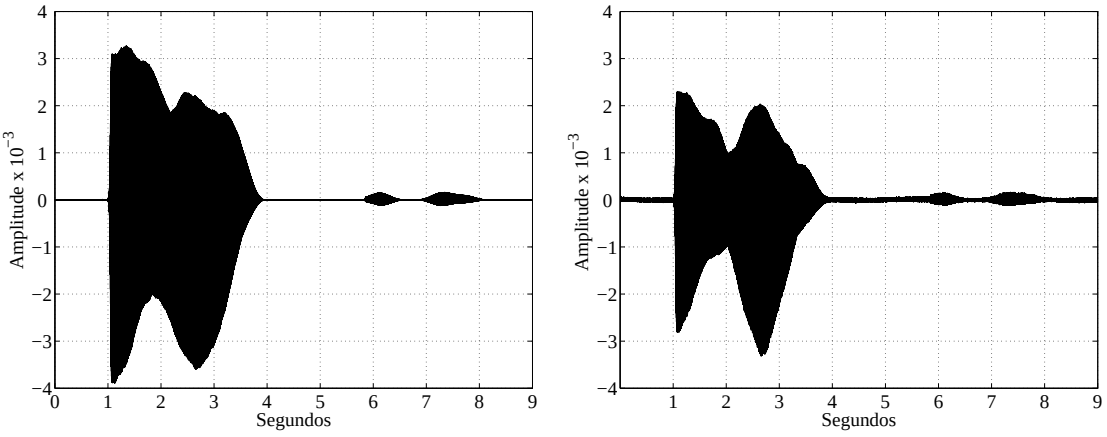
Para fins de comparação mostra-se, na Figura 3.22 a outra fonte entregue pela NMF e a processada pelo algoritmo definido no Caso 1. As envoltórias das fontes estimadas, antes e após o processo do Caso 1, são mostradas, respectivamente nas Figuras 3.24 e 3.25. Resultados semelhantes foram obtidos com diversos outros sinais testados, e os exemplos aqui mostrados visam a exemplificar resumidamente o que foi observado nos testes.

Observando os resultados na Tabela 3.5, não se pode afirmar que o método descrito nesta seção melhora o desempenho da separação. O mesmo resultado inconclusivo foi obtido ao se observar as envoltórias obtidas ao final do algoritmo. O resultado final ainda é altamente dependente do resultado da separação das fontes.

Uma forma de se contornar essa dependência do método de separação seria encontrar alguma maneira de incluir informações extras sobre as envoltórias das fontes. Por exemplo, poderia-se utilizar um *template* da emissão de nota dos instrumentos presentes no sinal em análise. Neste caso, seria necessário um banco de envoltórias contendo "carimbos" de emissões de notas advindas de diversos instrumentos.

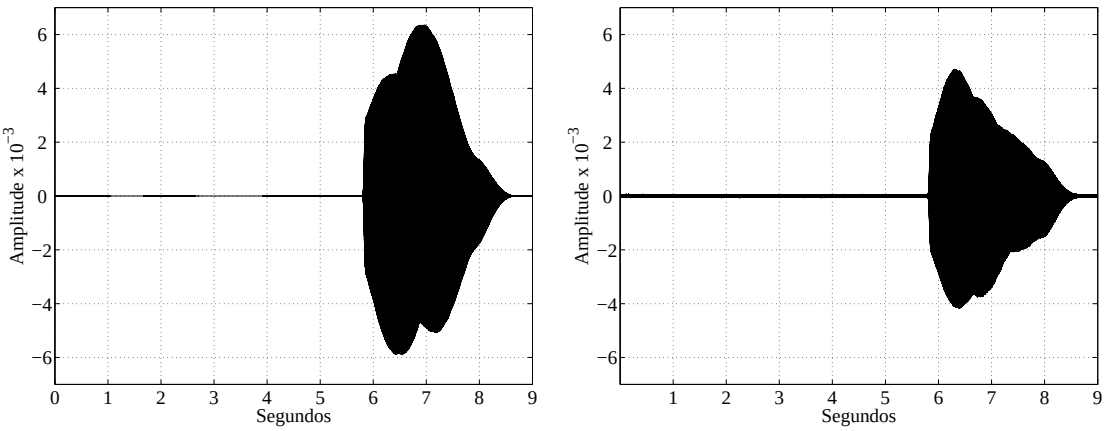
Tabela 3.5: Avaliação do caso de estudo 1. São mostrados resultados de sinais de mistura formados por notas isoladas (I) e sobrepostas (S) entre si.

Sinais		SDR	SIR	SAR	Ref.		Dif-SDR	Dif-SIR	Dif-SAR
Piano Sol#2	I	-8,9716	19,589	-8,9181	Lá4		-0,5416	0,2416	-0,5455
Flauta Lá4		-14,867	23,765	-14,8482	Sol#2		-0,0390	0,6649	-0,0421
	S	-12,105	15,3531	-11,9724	Lá4		-0,1410	-0,4989	-0,1270
		-23,741	11,0952	-23,4146	Sol#2		0,2763	-2,2219	0,4042
Flauta Lá4	I	-13,4924	28,2779	-13,4856	Lá4		0,1626	-0,9826	0,1641
Piano Sol#2		-12,2304	19,1429	-12,1746	Sol#2		-0,0426	-0,5807	-0,0356
	S	-12,9707	28,9007	-12,9648	Lá4		-0,3596	-1,0535	-0,3584
		-8,9102	23,4194	-8,888	Sol#2		-0,0851	-0,2688	-0,0839
Clarineta Lá4	I	-9,6220	18,2742	-9,5508	Lá4		-2,6811	-2,3199	-2,6553
Clarineta Ré5		-4,7326	28,7587	-4,7249	Ré5		-0,8450	0,6879	-0,8468
	S	-8,8380	16,1491	-8,7200	Ré5		0,7205	-0,1418	0,7264
		-3,6966	26,7339	-3,6835	Lá4		0,0900	0,6152	0,0881



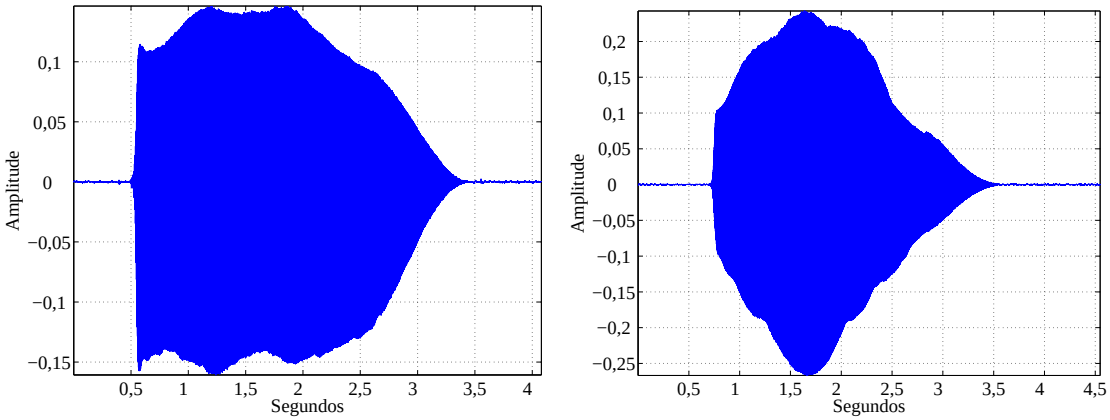
(a) Fonte 1 entregue pela NMF. (b) Mesma Fonte 1 após o processo do Caso 1.

Figura 3.21: Exemplo do efeito resultante do processo do Caso 1 sobre uma fonte resultante da NMF. Nota Lá4 de uma Clarineta, vindo de uma mistura não-sobreposta com uma nota Ré5 de Clarineta.



(a) Fonte 2 entregue pela NMF. (b) Mesma Fonte 2 após o processo do Caso 1.

Figura 3.22: Exemplo do efeito resultante do processo do Caso 1 sobre uma fonte resultante da NMF. Nota Ré5 de uma Clarineta, vindo de uma mistura não-sobreposta com uma nota Lá4 de Clarineta.



(a) Sinal Original: Clarineta Lá4. (b) Sinal Original: Clarineta Ré5.

Figura 3.23: Sinais originais utilizados nas misturas envolvendo as notas de Clarineta Lá4 ($f_0 = 440$Hz) e Ré5 ($f_0 = 587,33$Hz).

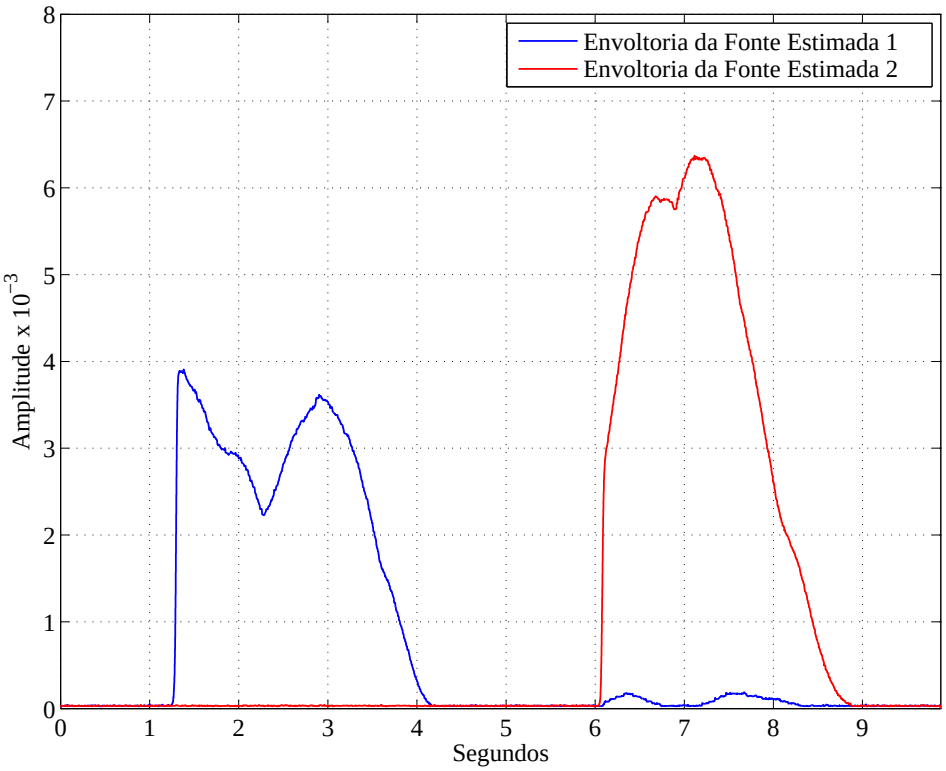


Figura 3.24: Envoltórias dos sinais estimados - Clarineta Lá4 e Clarineta Ré5. Entregues pela NMF.

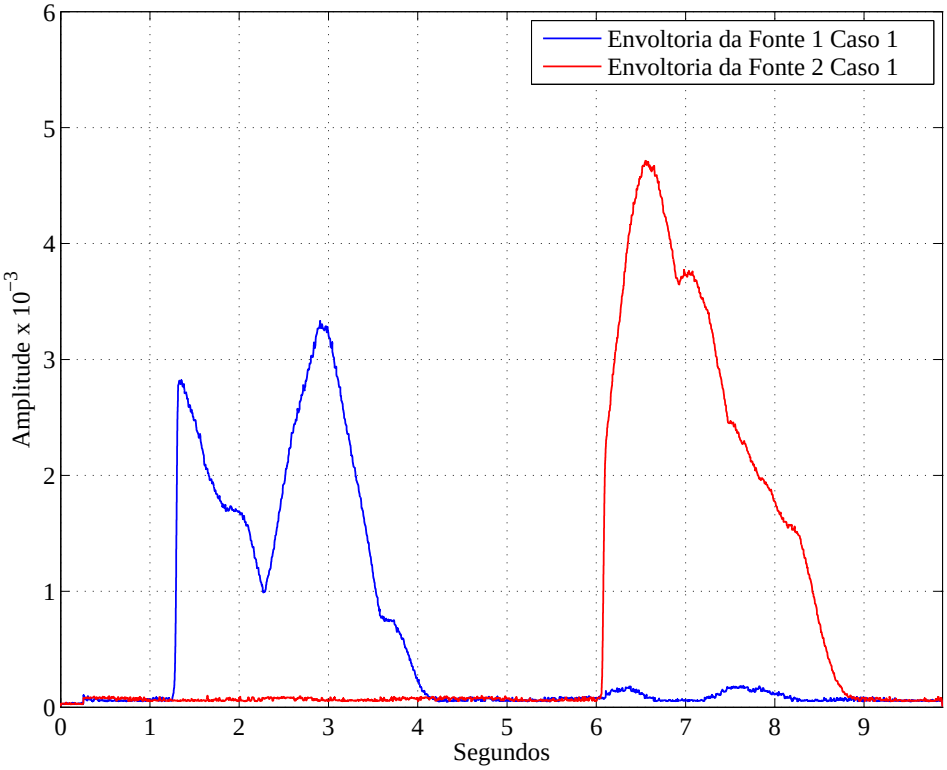


Figura 3.25: Envoltórias dos sinais estimados - Clarineta Lá4 e Clarineta Ré5. Após processo do Caso 1.

3.9 Caso 2: Melhorar a separação com informações de envoltória

Aqui, por um momento discute-se a relação entre algoritmo de separação e mecanismo de extração de envoltória na ordem inversa: de que forma envoltórias previamente extraídas poderiam ser utilizadas para tentar melhorar o desempenho de um método de separação de fontes?

O segundo caso de estudo pode ser ramificado em duas opções: substituir a matriz **H** pelo *template* da envoltória ou realizar o mesmo procedimento do Caso 1, porém aplicando o *template* sobre o sinal "sem envoltória".

Esse *template* poderia ser facilmente obtido utilizando notas reais e aplicando o algoritmo de estimação de envoltória sobre elas. De fato, assim foram criados os *templates* utilizados nos testes detalhados nas próximas seções: foram estimadas as envoltórias de algumas das notas da base RWC [14] e estas compuseram um banco de envoltórias que foi utilizado nos testes.

3.9.1 Substituição da matriz H pelo *template* de envoltória

O fato de a matriz **H** carregar informação temporal abre uma possibilidade de melhorar a estimação de cada uma das fontes na hora da reconstrução: substitui-se a linha da matriz **H** da fonte em análise por um *template* correspondente à emissão do instrumento que originou a nota. Com isso espera-se garantir maior tipicidade na envoltória da fonte estimada.

Abaixo são detalhados os passos para esse processo:

1. Primeiramente seleciona-se o *template* da envoltória com base num conhecimento prévio ou algum outro procedimento de identificação.

2. Uma decimação sobre o *template* é realizada, de modo a deixá-lo na taxa de amostragem da matriz **H**.

3. Realiza-se a reconstrução da fonte, utilizando a matriz **W**, oriunda da fatoração pela NMF e a matriz **H** construída pela decimação da envoltória *template*, e realiza-se a ressíntese.

4. Esse novo sinal gerado no item anterior será a nova fonte ressintetizada.

A escolha da envoltória a ser utilizada como *template* pode seguir critérios de similaridade diversos, automáticos ou não. Entretanto, uma vez que o foco do estudo é o efeito da inserção da informação da envoltória na qualidade separação, buscou-se utilizar o melhor caso possível: os *templates* utilizados são as envoltórias das notas

utilizadas na geração dos sinais de mistura. Este é o caso em que se pode conseguir a melhoria mais significativa na qualidade da separação, pois a informação temporal é a mais correta possível.

Tabela 3.6: Avaliação do caso de estudo 2.1. São mostrados resultados de sinais de mistura formados por notas isoladas (I) e sobrepostas (S) entre si.

Sinais		SDR	SIR	SAR	Ref.		Diff-SDR	Diff-SIR	Diff-SAR
Piano Sol#2	I	-6,1489	23,3071	-6,1238	Lá4		2,2811	3,9597	2,2488
Flauta Lá4		-12,6506	28,7113	-12,6444	Sol#2		2,1774	5,6112	2,1617
	S	-6,9625	30,847	-6,9582	Lá4		5,0015	14,995	4,8872
		-16,5176	18,7828	-16,4592	Sol#2		7,4997	5,4657	7,3596
Flauta Lá4	I	-12,6234	28,6881	-12,6172	Lá4		1,0316	-0,5724	1,0325
Piano Sol#2		-6,1582	23,4614	-6,134	Sol#2		6,0296	3,7378	6,005
	S	-10,4707	28,9952	-10,4647	Lá4		2,1404	-0,959	2,1417
		-5,6991	24,2407	-5,6784	Sol#2		3,126	0,5525	3,1257
Clarineta Lá4	I	-7,529	19,963	-7,4776	Lá4		-0,5881	-0,6311	-0,5821
Clarineta Ré5		-6,6149	22,7473	-6,5868	Ré5		-2,7273	-5,3235	-2,7087
	S	-9,5899	15,401	-9,4527	Ré5		-0,0314	-0,8899	-0,0063
		-4,331	25,2419	-4,3132	Lá4		-0,5444	-0,8768	-0,5416

Analisando a Tabela 3.6, pode-se observar que os valores de SDR e SAR continuam negativos, o que indica uma forte presença de defeitos possivelmente inseridos pelo processo de ressíntese e pouca interferência entre as fontes, pois os valores de SIR seguem positivos. Entretanto, a diferença entre os valores obtidos após a substituição da matriz **H** é, em quase todos os casos envolvendo a mistura [Piano + Flauta], positiva, o que indica uma melhora na separação.

Entretanto, ao observar os valores obtidos para a mistura de notas de Clarineta, nota-se que não houve melhoras na separação e, em alguns casos, o processo degradou os resultados em mais de 2dB. Essa dificuldade pode ser atribuída à má separação, também no âmbito espectral da mistura, como se pode observar na Figura 3.26 em que, na fonte 1, existem raias que pertencem à fonte 2. Isso se deve à grande coincidência de harmônicos nesse intervalo.

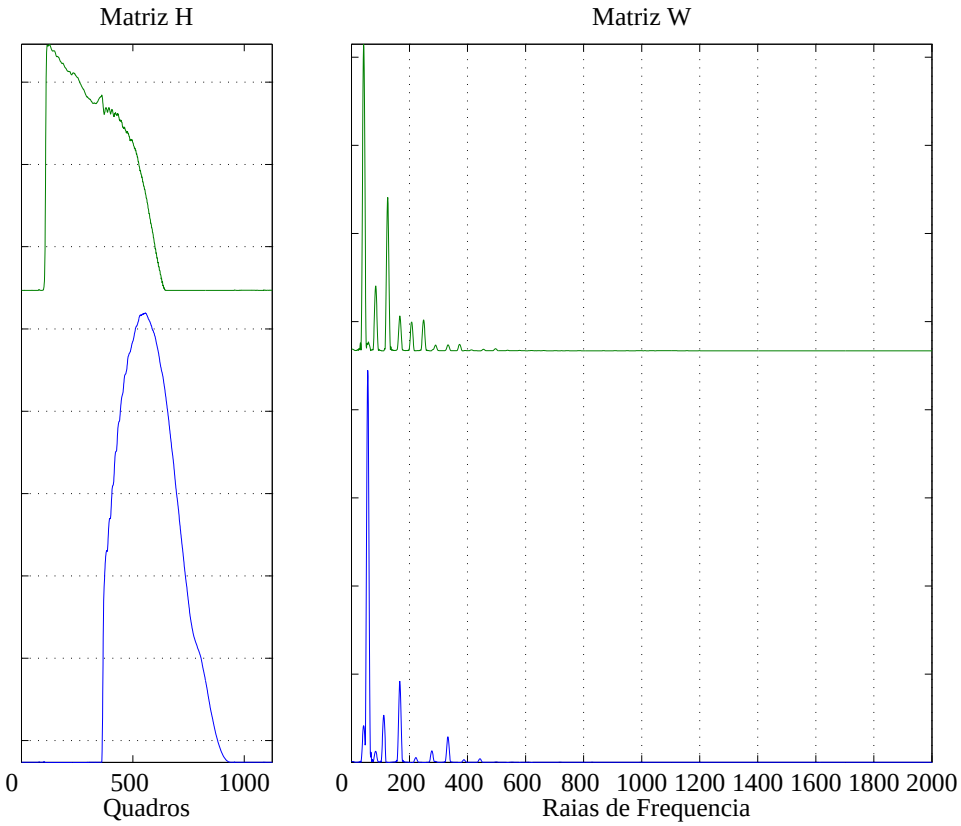


Figura 3.26: Matrizes **H** e **W** oriundas da NMF - [Clarineta Lá4 + Clarineta Ré5]

3.9.2 Aplicação do *template* de envoltória sobre a saída da NMF

A seção anterior mostrou o caso de se inserir informação temporal na separação utilizando um *template* como substituto da matriz **H** resultante da NMF. Uma segunda maneira de se utilizar o *template* é aplicá-lo sobre um sinal "sem envoltória", como no Caso 1. Em lugar de substituir a matriz **H**, a reconstrução do sinal é realizada fazendo todos os elementos desta matriz iguais a um e a matriz **W** oriunda da separação mantida inalterada. Reconstrói-se este sinal "sem envoltória" e aplica-se o *template* sobre ele.

Assim como no estudo anterior, foram escolhidas como *templates* as envoltórias das notas utilizadas na geração dos sinais de mistura, de modo a atingir a melhoria mais significativa possível na qualidade da separação.

A sequência de passos seguida no teste é detalhada a seguir:

1. Primeiramente seleciona-se o *template* da envoltória com base num conhecimento prévio ou algum outro procedimento de identificação.

2. Realiza-se a reconstrução da fonte, utilizando a matriz **W** e a matriz **H** feita unitária (todos os elementos iguais a 1), e realiza-se a ressíntese.

3. Aplica-se o *template* sobre o sinal ressintetizado na etapa anterior.

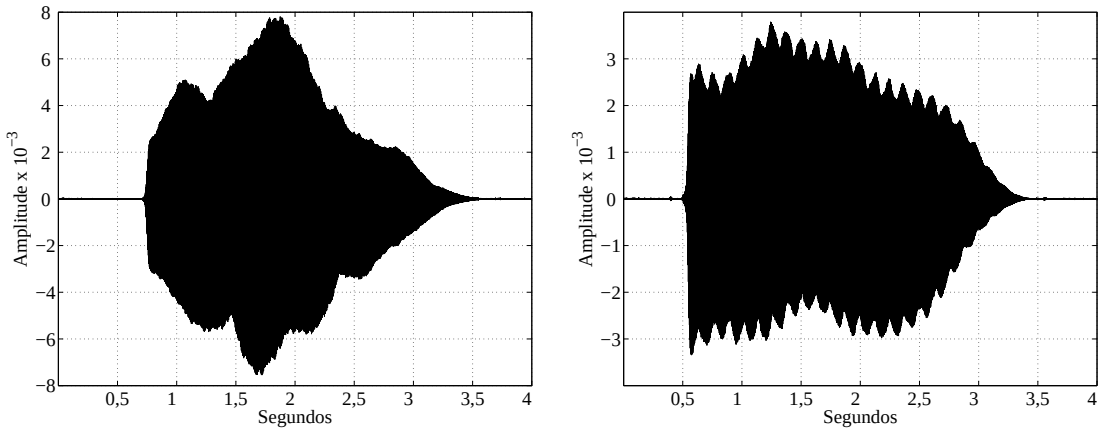
4. Esse novo sinal gerado no item anterior será a nova fonte ressintetizada.

Tabela 3.7: Avaliação do caso de estudo 2.2. São mostrados resultados de sinais de mistura formados por notas isoladas (I) e sobrepostas (S) entre si.

Sinais		SDR	SIR	SAR	Ref.		Dif-SDR	Dif-SIR	Dif-SAR
Piano Sol#2	I	-6.4707	23.1808	-6.4452	Lá4		1.9593	3.8334	1.9274
Flauta Lá4		-14.3629	27.0184	-14.3540	Sol#2		0.4651	3.9183	0.4521
	S	-8.3155	22.9340	-8.2902	Lá4		3.6485	7.0820	3.5552
		-14.0601	21.9941	-14.0316	Sol#2		9.9572	8.6770	9.7872
Flauta Lá4	I	-14.3742	27.0995	-14.3654	Lá4		-0.7192	-2.1610	-0.7157
Piano Sol#2		-6.4872	23.3815	-6.4629	Sol#2		5.7006	3.6579	5.6761
	S	-10.8846	27.8078	-10.8768	Lá4		1.7265	-2.1464	1.7296
		-6.4326	24.9481	-6.4156	Sol#2		2.3925	1.2599	2.3885
Clarineta Lá4	I	-6.2728	21.1959	-6.2321	Lá4		0.6681	0.6018	0.6634
Clarineta Ré5		-4.8126	24.6748	-4.7929	Ré5		-0.9250	-3.3960	-0.9148
	S	-11.7984	14.1450	-11.6232	Ré5		-2.2399	-2.1459	-2.1768
		-4.0196	26.3434	-4.0055	Lá4		-0.2330	0.2247	-0.2339

Analisando a Tabela 3.7 pode-ser notar uma melhora na separação, principalmente para o Piano. A diferença entre as duas abordagens que utilizam *template* é que, nesse segundo caso, o *template* é empregado na taxa original do sinal, ou seja, não precisa ser decimado para adequar-se à matriz **H**. Essa abordagem tem sua vantagem no fato de utilizar toda a informação do *template*, uma vez que este não é decimado, porém, o uso de uma matriz H contendo apenas uns dificulta o processo de estimação de fase.

Analogamente ao exposto anteriormente, a mistura de Clarinetas apresenta uma separação de componentes frequenciais inadequada, conforme se pode observar na Figura 3.26, dificultando a melhora da estimação. Os resultados são ilustrados na Figura 3.27 e pode-se compará-los às formas de onda dos sinais originais, apresentados na Figura 3.23, onde se pode observar que o processo de separação inverteu a ordem dos sinais originais na disposição das fontes estimadas.



(a) Fonte 1 após o processo.

(b) Fonte 2 após o processo.

Figura 3.27: [Clarineta Lá4 + Clarineta Ré5], originalmente misturados com sobreposição de 3 segundos.

Das análises realizadas ao longo do capítulo pode-se observar que, de um modo geral, a qualidade da envoltória das fontes estimadas através de um processo de separação por NMF é altamente dependente da qualidade dessa separação, principalmente da capacidade de separação das componentes frequenciais dos sinais das fontes originais. Nos casos em que as componentes espectrais são bem separadas é possível lançar mão de um processamento envolvendo um *template* a fim de melhorar a separação; entretanto, quando existem muitas componentes frequenciais mal-separadas (caso da mistura de Clarinetas anteriormente mostrado), esforços buscando melhorar a parcela temporal da saída do separador não surtem efeito sobre a qualidade da separação em geral.

Capítulo 4

Conclusões

Esta dissertação abordou a envoltória de sinais musicais no domínio do tempo. Foram detalhados alguns dos métodos mais comuns para a obtenção da envoltória de notas isoladas e proposta uma abordagem calcada em Morfologia Matemática. O método proposto mostrou-se rápido e eficiente; porém, sua maior vantagem sobre os demais é sua total automatização, não demandando nenhum conhecimento prévio do sinal de áudio a ser analisado.

O método é dependente de apenas um parâmetro: o comprimento da estrutura morfológica, que é estimado de maneira automática através de um critério perceptivo também proposto no trabalho. Tal critério mostrou-se robusto e adequado para diferentes tipos de sinais e notas, estimando o comprimento da estrutura de forma a obter-se uma envoltória que atende ao compromisso suavidade/detalhe.

Mais adiante no trabalho realizou-se um estudo sobre o emprego da NMF padrão como ferramenta na obtenção da envoltória de sinais formados por notas sequenciais sobrepostas e foram mostradas suas capacidades e limitações. Através desse estudo foi possível avaliar qualitativamente qual a influência da informação temporal, leia-se envoltória, sobre a qualidade da separação.

Mostrou-se ainda que, utilizando um *template* da envoltória é possível melhorar a qualidade da separação, exceto no caso em que a NMF não consegue separar as componentes espectrais de maneira adequada.

4.1 Trabalhos futuros

As sugestões para continuação deste trabalho são motivadas por desafios encontrados durante sua elaboração, e que não puderam ser resolvidos ou que deixam margem para melhorias, dentre os quais se pode citar:

Obtenção do comprimento ótimo: Foi possível notar que o problema de se encontrar o comprimento ótimo para a estrutura morfológica é um desafio devido à

forma da curva de convergência da função a ser minimizada. Possivelmente, através de um algoritmo de otimização mais eficiente, pode-se obter uma estimação do parâmetro próximo do ideal, apresentado na Figura 2.45.

Variantes da NMF: Através do uso de variantes da NMF pode ser possível obter-se uma melhor qualidade na separação, possibilitando a estimação mais robusta da envoltória de cada uma das fontes envolvidas.

Referências Bibliográficas

[1] MUSIC4C, http://ems.music.uiuc.edu/beaucham/software/m4c/m4c_intro_-html/M4C_intro.html, 2011, Último acesso em Fevereiro de 2011.

[2] BELLO, J. P., DAUDET, L., ABDALLAH, S., *et al.*, "A Tutorial on Onset Detection in Music Signals", *IEEE Transactions on Speech and Audio Processing*, v. 13, n. 5, pp. 1035–1047, Set. 2005.

[3] EIDHEIM, O. C., http://www.idi.ntnu.no/emner/tdt16/, 2011, Último acesso em Maio de 2012.

[4] KLAPURI, A., DAVY, M., *Signal Processing Methods for Music Transcription*. Nova Iorque, Springer, 2006.

[5] ESQUEF, P. A. A., VÄLIMÄKI, V., KARJALAINEN, M., "Restoration and Enhancement of Instrumental Recordings Based on Sound Source Modeling". In: *Proceedings of 110th Audio Engineering Society Convention*, Amsterdã, Holanda, Maio 2001.

[6] ESQUEF, P. A. A., VÄLIMÄKI, V., KARJALAINEN, M., "Restoration and Enhancement of Solo Guitar Recordings Based on Sound Source Modeling", *Journal of the Audio Engineering Society*, v. 50, n. 4, pp. 227–236, Abr. 2002.

[7] ZENPH, http://www.zenph.com, 2012, Útimo acesso em Fevereiro de 2012.

[8] MIDI, http://www.midi.org/, 2012, Último acesso em Fevereiro de 2012.

[9] LOY, G., *Musimathics: The Mathematical Foundations of Music, Volume 1*. The MIT Press, 2006.

[10] DIXON, S., "Onset Detection Revisited". In: *Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx'06)*, pp. 133–137, Montreal, Canada, Set. 2006.

[11] COLLINS, N., "A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions". In:

Proceedings of the AES 118th Convention, Barcelona, Espanha, p. 12, Barcelona, Spain, Maio 2005.

[12] HORNBOSTEL, E. M. V., SACHS, C., "Classification of Musical Instruments: Translated from the Original German by Anthony Baines and Klaus P. Wachsmann", *The Galpin Society Journal*, v. 14, pp. 3–29, 1961.

[13] WIECZORKOWSKA, A. A., "Multi-way Hierarchic Classification of Musical Instrument Sounds". In: *Proceedings of the IEEE CS International Conference on Multimedia and Ubiquitous Engineering (MUE 2007), in Seoul, Korea*, pp. 897–902, Abr. 2007.

[14] GOTO, M., NISHIMURA, T., "RWC Music Database: Music Genre Database and Musical Instrument Sound Database". In: *International Symposium on Music Information Retrieval (ISMIR)*, pp. 229–230, 2003.

[15] CAETANO, M., BURRED, J. J., RODET, X., "Automatic Segmentation of the Temporal Evolution of Isolated Acoustic Musical Instrument Sounds Using Spectro-Temporal Cues". In: *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx'10)*, Graz, Austria, Set. 2010.

[16] DINIZ, P., NETTO, S., SILVA, E. D., *Digital Signal Processing: System Analysis and Design*. New York, NY, USA, Cambridge University Press, 2002.

[17] HAJDA, J., "A New Model for Segmenting the Envelope of Musical Signals: The Relative Salience of Steady State Versus Attack, Revisited". In: *Audio Engineering Society Convention 101*, Nov. 1996.

[18] JENSEN, K., "Envelope Model Of Isolated Musical Sounds". In: *Proc. of the 2nd Int. Conference on Digital Audio Effects (DAFx'99)*, Trondheim, Norway, Dez. 1999.

[19] MAKHOUL, J., "Linear prediction: A tutorial review", *Proceedings of the IEEE*, v. 63, n. 4, pp. 561 – 580, Abr. 1975.

[20] ATHINEOS, M., ELLIS, D., "Frequency-domain linear prediction for temporal features". In: *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, pp. 261 – 266, Dez. 2003.

[21] AHMED, N., NATARAJAN, T., RAO, K., "Discrete Cosine Transfom", *IEEE Transactions on Computers*, v. C-23, n. 1, pp. 90 – 93, Jan. 1974.

[22] RöBEL, A., VILLAVICENCIO, F., RODET, X., "On Cepstral and All-Pole Based Spectral Envelope Modeling with Unknown Model Order", *Pattern Recognition Letters*, v. 28, pp. 1343–1350, Ago. 2007.

[23] GALAS, T., RODET, X., "An Improved Cepstral Method for Deconvolution of Source-Filter Systems with Discrete Spectra: Application to Musical Sound Signals". In: *Proceedings of the International Computer Music Conference (ICMC)*, Glasgow, Set. 1990.

[24] DELLER, JR., J. R., PROAKIS, J. G., HANSEN, J. H., *Discrete Time Processing of Speech Signals*. 1 ed. Upper Saddle River, NJ, USA, Prentice Hall PTR, 1993.

[25] SOILLE, P., *Morphological Image Analysis: Principles and Applications*. 2 ed. Secaucus, NJ, USA, Springer-Verlag New York, Inc., 2003.

[26] GIL, J., KIMMEL, R., "Efficient dilation, erosion, opening, and closing algorithms", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 24, n. 12, pp. 1606 – 1617, dec 2002.

[27] KAHANER, D., MOLER, C., NASH, S., *et al.*, *Numerical Methods and Software*, Prentice-Hall series in computational mathematics. Prentice Hall, 1988.

[28] KEDEM, B., "Spectral analysis and discrimination by zero-crossings", *Proceedings of the IEEE*, v. 74, n. 11, pp. 1477 – 1493, Nov. 1986.

[29] AMADO, R., FILHO, J., "Pitch detection algorithms based on zero-cross rate and autocorrelation function for musical notes". In: *International Conference on Audio, Language and Image Processing, 2008. ICALIP 2008.*, pp. 449–454, Shanghai, China, Jul. 2008.

[30] CUADRA, P. D. L., MASTER, A., "Efficient pitch detection techniques for interactive music". In: *In Proceedings of the 2001 International Computer Music Conference, La Habana*, 2001.

[31] GERHARD, D., *Pitch Extraction and Fundamental Frequency: History and Current Techniques*, Report, 2003.

[32] LAHAT, M., NIEDERJOHN, R., KRUBSACK, D., "A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech", *Acoustics, Speech and Signal Processing, IEEE Transactions on*, v. 35, n. 6, pp. 741 – 750, Jun. 1987.

[33] HOWARD, D. M., ANGUS, J., *Acoustics and Psychoacoustics*. 2 ed. Newton, MA, USA, Butterworth-Heinemann, 2000.

[34] NUNES, L. O., ESQUEF, P. A. A., BISCAINHO, L. W. P., "FlexSM: A Flexible Sinusoidal Modeling System", *Journal of The Audio Engineering Society*, v. 57, n. 12, pp. 1042–1056, 2009.

[35] COMON, P., "Independent component analysis, a new concept?", *Signal Process.*, v. 36, pp. 287–314, Abr. 1994.

[36] LEE, D. D., SEUNG, S. H., "Learning The Parts of Objects by Non-negative Matrix Factorization", *Nature*, v. 401, pp. 788–791, 1999.

[37] SMARAGDIS, P., BROWN, J. C., "Non-negative Matrix Factorization for Polyphonic Music Transcription", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177–180, 2003.

[38] TYGEL, A. F., *Método de Fatoração de Matrizes Não-Negativas para Separação de Sinais Musicais*. Dissertação M.Sc., PEE/COPPE;UFRJ, Rio de Janeiro, Brasil, Dez. 2009. Disponível em http://www.pee.ufrj.br/teses/index.php?Resumo=2009121701.

[39] ZHU, X., BEAUREGARD, G., WYSE, L., "Real-Time Iterative Spectrum Inversion with Look-Ahead". In: *Multimedia and Expo, 2006 IEEE International Conference on*, pp. 229–232, Jul. 2006.

[40] VINCENT, E., GRIBONVAL, R., FEVOTTE, C., "Performance measurement in blind audio source separation", *Audio, Speech, and Language Processing, IEEE Transactions on*, v. 14, n. 4, pp. 1462 –1469, jul 2006.

[41] VINCENT, http://bass-db.gforge.inria.fr/bss_eval/, 2012, Útimo acesso em Abril de 2012.

[42] SMARAGDIS, P., "Non-negative Matrix Factor Deconvolution; Extracation of Multiple Sound Sources from Monophonic Inputs", *5th International Congress on Independent Component Analysis and Blind Signal Separation (ICA)*, p. 8, Set. 2004.

[43] BERRY, M. W., BROWNE, M., LANGVILLE, A. N., *et al.*, "Algorithms and applications for approximate nonnegative matrix factorization", *Computational Statistics & Data Analysis*, v. 52, n. 1, pp. 155–173, September 2007.

[44] LEE, D. D., SEUNG, S. H., "Algorithms for Non-negative Matrix Factorization", *Neural Information Processing Systems*, pp. 556–562, 2000.

[45] ANTONIOU, A., LU, W.-S., *Practical Optimization: Algorithms and Engineering Applications*. Springer Publishing Company, Incorporated, 2007.

[46] VIRTANEN, T., "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria", *Audio, Speech, and Language Processing, IEEE Transactions on*, v. 15, n. 3, pp. 1066 –1074, mar 2007.

[47] GRIFFIN, D., LIM, J., "Signal estimation from modified short-time Fourier transform", *Acoustics, Speech and Signal Processing, IEEE Transactions on*, v. 32, n. 2, pp. 236–243, Abr. 1984.

[48] ZHU, X., BEAUREGARD, G., WYSE, L., "Real-Time Signal Estimation From Modified Short-Time Fourier Transform Magnitude Spectra", *Audio, Speech, and Language Processing, IEEE Transactions on*, v. 15, n. 5, pp. 1645–1653, Jul. 2007.

Apêndice A

Non-Negative Matrix Factorization (NMF)

Este apêndice reproduz as Seções 3.1 (subseções 3.1.1 e 3.1.2) da referência [38], e visa a detalhar algumas definições utilizadas ao longo do texto e explicitar o método de fatoração de matrizes não-negativas (NMF).

A.1 Definição do Problema

O problema da fatoração de matrizes não-negativas pode ser definido da seguinte maneira [43]:

Dada uma matriz não-negativa $\mathbf{V} \in \mathbb{R}_+^{N\times M}$ *e um inteiro positivo* $D < \min(N, M)$, *ache as matrizes não-negativas* $\mathbf{W} \in \mathbb{R}_+^{N\times D}$ *e* $\mathbf{H} \in \mathbb{R}_+^{D\times M}$ *que minimizem a função*

$$f(\mathbf{W}, \mathbf{H}) = \frac{1}{2}||\mathbf{V} - \mathbf{WH}||_F^2, \tag{A.1}$$

onde $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$ *denota a norma de Frobenius.*

O produto **WH** é chamado de fatoração não-negativa de **V**. Entretanto, **V** não é necessariamente igual a **WH**; em geral tem-se uma aproximação com posto no máximo igual a D. Ao longo do texto, a seguinte definição será utilizada:

$$\mathbf{V} \approx \mathbf{\Lambda} = \mathbf{WH}, \tag{A.2}$$

onde $\mathbf{\Lambda} \in \mathbb{R}_+^{N\times M}$ será a aproximação de **V**.

Os fatores **W** e **H** devem ser calculados através de algoritmos de otimização, de forma a solucionar o problema. A próxima subseção trata do algoritmo básico de otimização a ser usado.

A.1.1 Algoritmo de Otimização

Segundo [44], o problema da fatoração de matrizes não-negativas é convexo em $\mathbf{W}$ ou $\mathbf{H}$, mas não em ambos. Desta forma, utilizando o método de gradiente descendente [45] de forma alternada e com um passo suficientemente pequeno, garante-se que o erro $E = f(\mathbf{W}, \mathbf{H})$ sempre diminui. Definindo $W_{n,m}$ como um elemento de $\mathbf{W}$, a equação de atualização para $\mathbf{W}$ é dada por

$$W_{n,d} \leftarrow W_{n,d} - \mu_W \frac{\partial E}{\partial W_{n,d}}, \tag{A.3}$$

com

$$\frac{\partial E}{\partial W_{n,d}} = \sum_{n=1}^{N} \sum_{m=1}^{M} (V_{n,m} - \Lambda_{n,m}) \frac{\partial \mathbf{\Lambda}}{\partial W_{n,d}}. \tag{A.4}$$

A derivada de um elemento de $\mathbf{\Lambda}$ em relação a $W_{n,d}$ é

$$\frac{\partial \Lambda_{n',m'}}{\partial W_{n,d}} = \begin{cases} \frac{\partial}{\partial W_{n,d}} \sum_{k=1}^{D} W_{n,k} H_{k,m'} = H_{d,m'}, & \text{para} \quad n = n', \quad \text{e} \\ 0, & \text{para} \quad n \neq n', \end{cases} \tag{A.5}$$

e portanto

$$\frac{\partial E}{\partial W_{n,d}} = \sum_{m'=1}^{M} (V_{n,m'} - \Lambda_{n,m'}) H_{d,m'}. \tag{A.6}$$

Para todos os elementos de $\mathbf{W}$, tem-se:

$$\frac{\partial E}{\partial \mathbf{W}} = (\mathbf{V} - \mathbf{\Lambda})\mathbf{H}^T, \tag{A.7}$$

e com isso

$$\mathbf{W} \leftarrow \mathbf{W} + \mu_W (\mathbf{V} - \mathbf{\Lambda})\mathbf{H}^T,$$

ou

$$\mathbf{W} \leftarrow \mathbf{W} + \mu_W (\mathbf{V}\mathbf{H}^T - \mathbf{\Lambda}\mathbf{H}^T). \tag{A.8}$$

Esta equação de atualização não garante o atendimento à restrição de não-negatividade do problema. No entanto, caso as matrizes sejam inicializadas com valores não-negativos e a atualização seja multiplicativa por um fator não-negativo, garante-se automaticamente que os elementos nunca assumirão valores negativos. Então, deve-se escolher μ_W de forma que o valor seguinte de $\mathbf{W}$ seja ele próprio

multiplicado por um número não-negativo.

O passo μ_W, até aqui representado por um escalar, é substituído por uma matriz de dimensões $N \times D$. As operações de 'divisão entre matrizes' são realizadas ponto-a-ponto, e $\otimes$ denota o produto de Hadamard, onde os elementos são multiplicados também ponto-a-ponto. Fazendo

$$\mu_W = \frac{\mathbf{W}}{\mathbf{\Lambda H}^T}, \tag{A.9}$$

obtém-se

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \left(\mathbf{1} + \frac{\mathbf{V H}^T}{\mathbf{\Lambda H}^T} - \frac{\mathbf{\Lambda H}^T}{\mathbf{\Lambda H}^T} \right), \tag{A.10}$$

ou

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\mathbf{V H}^T}{\mathbf{W H H}^T}, \tag{A.11}$$

que é a regra de atualização desejada para a matriz $\mathbf{W}$. O símbolo $\mathbf{1}$ representa uma matriz $N \times D$ com todos os elementos iguais a 1.

Pode-se notar que para a decomposição de $\mathbf{V}^T$, as matrizes $\mathbf{W}$ e $\mathbf{H}$ seriam substituídas por $\mathbf{H}^T$ e $\mathbf{W}^T$, respectivamente. Portanto, todo o desenvolvimento feito para a matriz $\mathbf{W}$ pode ser estendido a $\mathbf{H}$, operando-se esta troca:

$$\mathbf{H} \leftarrow \left(\mathbf{H}^T \otimes \frac{\mathbf{V}^T \mathbf{W}}{\mathbf{H}^T \mathbf{W}^T \mathbf{W}} \right)^T, \tag{A.12}$$

ou

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T \mathbf{V}}{\mathbf{W}^T \mathbf{W H}}. \tag{A.13}$$

A prova de convergência do algoritmo pode ser encontrada em [44]. O Algoritmo a seguir mostra o procedimento básico da NMF que minimiza a distância euclidiana.

Entrada: Matriz não-negativa $\mathbf{V} \in \mathbb{R}_+^{N \times M}$ e número de fontes D.

1. Inicialize as matrizes $\mathbf{W} \in \mathbb{R}_+^{N \times D}$ e $\mathbf{H} \in \mathbb{R}_+^{D \times M}$ com valores aleatórios não-negativos distribuídos uniformemente entre 0 e 1;
2. Atualize $\mathbf{W}$ utilizando a equação (A.11);
3. Atualize $\mathbf{H}$ utilizando a equação (A.13);
4. Volte ao passo 2 até atingir a convergência ou um número máximo de iterações.

Saída: Matrizes $\mathbf{W} \in \mathbb{R}_+^{N \times D}$ e $\mathbf{H} \in \mathbb{R}_+^{D \times M}$.

A.1.2 Função-Custo

Na definição do problema foi utilizada como função-custo a distância euclidiana (DE), expressa na equação (A.1). Esta escolha traz consigo duas decisões de projeto: (1) o único objetivo da fatoração é a reconstrução, ou seja, a busca das matrizes $\mathbf{W}$ e $\mathbf{H}$ cujo produto seja o mais próximo possível de $\mathbf{V}$; e (2) a proximidade de $\mathbf{WH}$ em relação a $\mathbf{V}$ deve ser medida pela DE.

O desenvolvimento das equações de atualização foi mostrado utilizando a DE por simplicidade. Dependendo do problema, no entanto, pode ser favorável utilizar outras medidas para o cálculo da distância, e outros critérios, além da reconstrução.

Uma medida de distância comumente usada é inspirada na Divergência de Kullback-Leibler (DKL) [46]:

$$f_{\mathrm{KL}}(\mathbf{W}, \mathbf{H}) = \left\| \mathbf{V} \otimes \ln\left(\frac{\mathbf{V}}{\mathbf{WH}}\right) - \mathbf{V} + \mathbf{WH}) \right\|_F . \tag{A.14}$$

Rigorosamente, a medida só poderia ser chamada de divergência de Kullback-Leibler quando $\mathbf{V}$ e $\mathbf{\Lambda}$ representassem distribuições de probabilidades. No entanto, feita esta ressalva, adota-se este nome por simplicidade. A medida não representa uma distância, pois não é simétrica, mas tem seu mínimo em zero, que só é atingido quando $\mathbf{V} = \mathbf{\Lambda}$.

Além do critério de reconstrução, outros critérios podem ser inseridos na função-custo. Essa escolha também é dependente do problema, e pode, por exemplo, garantir algum tipo de estrutura para a matriz $\mathbf{\Lambda}$, para que ela possua sentido físico. Caso se tratasse de uma distribuição de probabilidades, por exemplo, um dos critérios de otimização seria a norma unitária.

Tanto uma mudança na medida de distância quanto no critério de otimização afetam diretamente as equações de atualização, que são o cerne do algoritmo. O desenvolvimento apresentado anteriormente, que culmina na equações (A.11) e (A.13), foi feito de acordo com o projeto que utiliza a reconstrução como único critério, e a distância euclidiana como medida de distância. Em [36], podem ser encontradas as equações de atualização referentes ao critério de reconstrução utilizando a divergência de Kullback-Leibler.

Apêndice B

Métodos de Síntese

No presente trabalho, em diversas situações, foi necessária uma ressíntese dos sinais envolvidos na separação de fontes realizada pela NMF. Conforme já exposto, a NMF entrega espectrogramas de magnitude das fontes estimadas e, por construção, a informação de fase é perdida. Em geral, este espectrograma modificado não é válido, no sentido de que é possível que nenhum sinal real possua tal espectrograma de magnitude [47]. A solução utilizada no presente trabalho é detalhada neste apêndice, que reproduz as Seções 6.1, 6.2 e 6.3 da referência [38].

B.1 STFT e MSTFT

Neste trabalho é utilizada a seguinte definição para o Transformada de Fourier de Tempo Curto (*Short-Time Fourier Transform*, STFT) para um sinal $x(k), k = 1, \ldots, K-1$:

$$X^m(n) = \sum_{k=0}^{N-1} x(k)w(k - mS)e^{-j\frac{2\pi n}{N}k}, \quad \text{para} \quad n = 0, \ldots, N-1, \tag{B.1}$$

onde n é o contador de raias, m é o contador de quadros, S é o avanço em amostras a cada quadro, N é o tamanho da janela (igual ao número de raias), e w é a janela de Hanning, definida como:

$$w(k) = \begin{cases} 2\frac{\sqrt{S/L}}{\sqrt{4a^2+2b^2}}\left(a + b\cos(\frac{2\pi k}{L} + \frac{\pi}{L})\right), \text{ para } 0 \leq n < L \\ 0, \text{ em outros casos,} \end{cases} \tag{B.2}$$

onde L é o tamanho da janela, $a = 0,5$ e $b = -0,5$.

Além disso, define-se a magnitude da transformada de Fourier de tempo curto (*Short-Time Fourier Transform Magnitude*, STFTM) como $|X^m(n)|$. Ao separar

um espectrograma de magnitude como a soma de espectrogramas gerados pelas fontes, criam-se as transformadas de Fourier de tempo curto modificadas (*Modified Short-Time Fourier Transform*, MSTFTMs).

B.2 Algoritmo de Griffin e Lim

Uma das primeiras soluções dadas na literatura para o problema da determinação de fase foi o algoritmo de Griffin e Lim (G&L) [47], cuja ideia é procurar um sinal no tempo cuja STFTM seja a mais próxima possível da MSTFTM desejada, no sentido de mínimos quadrados. Para isso, é utilizado um procedimento iterativo, no qual a estimativa da fase vai sendo aproximada enquanto o espectrograma de magnitude é fixado.

Os passos do algoritmo G&L estão descritos no algoritmo a seguir. Como entrada, o algoritmo recebe $|Y^m(n)|$, o MSTFTM alvo. Além disso, é necessário fornecer o tamanho da janela de análise L e o salto entre janelas S. Na saída, tem-se a estimativa $\hat{x}(k)$ do sinal no tempo.

Entrada: A MSTFTM $|Y^m(n)|$, o tamanho da janela de análise L e o salto entre janelas S.

1. Estimação inicial do sinal no tempo $\hat{x}(k)$, que pode ser feita com amostras de uma distribuição uniforme.

2. Geração de uma estimativa de STFT: $\hat{X}^m(n) = |Y^m(n)|e^{-j\angle\hat{x}(k)}$. Este passo é chamado de *Magnitude-Constrained*, pois gera-se uma STFT que possui a magnitude alvo $|Y^m(n)|$ e a fase da estimativa $\hat{x}(k)$.

3. Atualização de $\hat{x}(k)$ segundo a equação:

$$\hat{x}(k) = \frac{\sum_{m=-\infty}^{\infty} w(k - mS)\frac{1}{2\pi}\sum_{n=0}^{N-1}\hat{X}^m(n)e^{j\frac{2\pi n}{N}k}}{\sum_{m=-\infty}^{\infty} w(k - mS)^2} \tag{B.3}$$

   Esta equação busca a minimização do erro quadrático entre o alvo $|Y^m(n)|$ e a STFTM $|\hat{X}^m(n)|$ do sinal $\hat{x}(k)$ que está sendo estimado.

4. Volta ao passo 2 até a convergência: O passo 2 é executado novamente, desta vez com uma estimativa melhor do sinal no tempo e sua fase. A convergência pode ser medida pela diferença entre as estimativas $\hat{x}(k)$ a cada iteração.

Saída: Estimativa do sinal no tempo $\hat{x}(k)$.

B.3 Algoritmos *Real-time Iterative Spectrogram Inversion* (RTISI)

No algoritmo G&L, cada quadro utiliza informações de quadros passados e futuros, o que torna a sua utilização em tempo real inviável por definição. Além disso, o alto número de transformadas de Fourier torna o algoritmo custoso computacionalmente.

O algoritmo RTISI [48] propõe uma solução para ambos os problemas. Cada quadro só depende dos quadros anteriores, e a convergência é acelerada utilizando uma inicialização melhor.

A ideia principal do algoritmo é estimar um quadro por vez, ao contrário do algoritmo G&L, que estima o sinal inteiro. Considerando $L = 4S$, ou seja, uma sobreposição entre janelas de 75%, antes de se iniciar a estimação do quadro m, ele já possui 75% das amostras preenchidas pelos 3 quadro anteriores, e os 25% finais são nulos. Assim, em vez de se começar a estimativa da fase do quadro m com zeros, já se tem parte das amostras preenchidas, o que permite fazer uma inicialização mais próxima e coerente com a do quadro anterior. Em seguida, aplica-se esta fase à magnitude alvo do quadro e itera-se até a convergência. O método encontra-se sistematizado no algoritmo a seguir:

Entrada: A MSTFTM $|Y^m(n)|$, o tamanho da janela de análise L e o salto entre janelas S.

1. Estimativa inicial do sinal no tempo, $\hat{x}(k)$;
2. Para cada quadro m de $\hat{x}(k)$, definição do sinal $\hat{x}^m(k)$;
3. Até que a estimativa de $\hat{x}^m(k)$ convirja:
   (a) DFT de $\hat{x}^m(k)$, $X^m(n)$;
   (b) Geração de um quadro *Magnitude-Constrained*, $\hat{X}_m(n) = |Y^m(n)|\angle X^m(n)$;
   (c) iDFT de $\hat{X}^m(n)$, que resulta em $\hat{x}^m(k)$.
4. Após a convergência do quadro, *overlap-and-add* em $\hat{x}(k)$ e volta ao passo 2, com o quadro $m + 1$.

Saída: Estimativa do sinal no tempo $\hat{x}(k)$.

Este algoritmo possui uma versão avançada, também descrita em [48], denominada RTISI *Look Ahead* (RTISI-LA). Neste método, também são utilizados p quadros à frente na estimação do quadro m. Isso torna o algoritmo mais custoso

computacionalmente, além de impor um atraso estrutural de p quadros. Entretanto, a estimativa da fase tem melhora substancial.

No RTISI-LA, cada quadro m tem influência de amostras de quadros anteriores e posteriores, ao contrário do RTISI, em que apenas os quadros anteriores eram utilizados. No caso do RTISI-LA, após a estimação do quadro $m + p$, o quadro m é reestimado, desta vez levando em conta os p quadros posteriores que o influenciam.

Este é o método utilizado em todas as sínteses realizadas durante o trabalho.