ESTUDO SOBRE TRANSIENTES EM SINAIS DE FALA E MÚSICA

THIAGO DE ALMEIDA MAGALHÃES CAMPOLINA

ESTUDO SOBRE TRANSIENTES EM SINAIS DE FALA E MÚSICA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Engenharia Elétrica.

Orientadores: Hani Camille Yehia, Maurício Alves Loureiro.

Belo Horizonte

Agosto de 2012

Thiago de Almeida Magalhães Campolina
Estudo sobre transientes em sinais de fala e música / Thiago de Almeida Magalhães Campolina. — Belo Horizonte, 2012
xxiv, 66 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientadores: Hani Camille Yehia, Maurício Alves Loureiro.

1.Computação musical. 2.Modelagem de transientes. 3.Processamento de sinais. 4.Sons de fala plosivos. 5.Audição computacional. — Tese. I. Dissertação (Mestrado) — Escola de Engenharia Universidade Federal de Minas Gerais. II. Título.

CDU

"Estudo sobre Transientes em Sinais de Fala e Música"

Thiago de Almeida Magalhaes Campolina

Dissertação de Mestrado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Mestre em Engenharia Elétrica.

Aprovada em 10 de agosto de 2012.

Por:

Prof. Dr. Hani Camille Yehia
DELT (UFMG) - Orientador

Prof. Dr. Maurício Alves Loureiro
Escola de Música (UFMG) - Co-Orientador

Prof. Dr. Eduardo Mazoni Andrade Marçal Mendes
DELT (UFMG)

Dr. Leandro de Campos Teixeira Gomes
(Fundação CPqD)

Prof. Dr. Maurílio Nunes Vieira
DELT (UFMG)

Dedico esta dissertação a todas as pessoas que fazem parte da minha vida de forma positiva, sendo amadas por mim e me amando.

Agradecimentos

Agradeço aos meus orientadores Prof. Maurício Alves Loureiro e Prof. Hani Camille Yehia, pelos valiosos ensinamentos, conselhos e motivações. Ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e ao pessoal da Fundação CPqD (Centro de Pesquisa e Desenvolvimento em Telecomunicações), pelo suporte e contribuições à minha pesquisa. Agradeço imensamente aos meus pais e irmãos pelo apoio e carinho em todos os momentos, me motivando a sempre seguir o meu caminho. Aos meus amigos, que se preocupam comigo e estão sempre dispostos a me ajudar no que for preciso, e a dividir momentos de felicidade. Aos colegas do CEGeME (Centro de Estudos do Gesto Musical e Expressão), CEFALA (Centro de Estudos da Fala, Acústica, Linguagem e Música), Escola de Música e Escola de Engenharia pela amizade, contribuições e troca de conhecimentos. Agradeço também aos participantes dos testes realizados neste estudo.

"A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original."
(Albert Einstein)

Resumo

Atualmente, sistemas computacionais necessitam da modelagem de sinais de música e fala para vários propósitos: síntese, audição computacional, análise acústica, análise musical sistemática, transformação, composição, entre muitas outras. A modelagem por síntese de sinais de música e fala, com qualidade, é uma tarefa complexa, que ainda se apresenta como um desafio. Principalmente, quando se busca a representação com a menor quantidade de parâmetros possível, visando baixo custo computacional, procurando manter boa inteligibilidade e naturalidade.

Sinais de fala e música apresentam estados transitórios de vibração, que contêm transientes. Como é o caso de ataques de notas musicais e *bursts* em consoantes oclusivas. Por possuírem características impulsivas, transientes são como retratos instantâneos do trato vocal e instrumentos musicais, sendo importantes para a percepção de timbre e reconhecimento da fonte sonora. Muitas das técnicas de modelagem de música e fala não são eficientes para transientes. A modelagem por síntese dos transientes possibilita sua separação das partes determinísticas e estocásticas de sinais, representando maior flexibilidade para processamentos. Este estudo é dedicado à análise, modelagem, e percepção auditiva de transientes.

Na pesquisa desenvolvida, *Transient Modeling Synthesis* (TMS) é usado para modelar a componente transiente de sinais musicais e de fala. Em seguida, TMS é avaliado e comparado à tradicional modelagem senoidal *Spectral Modeling Synthesis* (SMS). Os resultados de um experimento de reconhecimento e qualidade MOS (Mean Opinion score) são usados para medir a importância de uma modelagem adequada para transientes como *bursts* de consoantes oclusivas.

Comparado à inteligibilidade de 98% obtida das elocuções originais, o TMS atingiu 95%, sendo significativamente maior que os 87% obtidos com o SMS. É também observado que a remoção dos *bursts* reduziu a inteligibilidade para 79%.

Finalmente, possibilitando a separação da componente transiente, o TMS permite a definição de um índice para medir a razão entre as energias do sinal original e sua componente transiente. Esse índice, chamado de Índice de Transiência, é avaliado. Os

valores obtidos para notas musicais de diferentes instrumentos são, então, comparados.

Palavras-chave: Computação musical, modelagem de transientes, processamento de sinais, sons de fala plosivos, audição computacional.

Abstract

Nowadays, computer systems require the modeling of musical and speech signals for several purposes: synthesis, computational hearing, acoustic analysis, systematic musicology, transformation, composition, and many others. The quality modeling synthesis of music and speech signals is a complex task, which is still presented as a challenge. This is particularly true when low dimensional parametric representation, low computational cost, intelligibility and naturalness are aimed.

Speech and musical signals have transient states of vibration, such as musical instrument note attacks and speech bursts in stop consonants. Due to its impulsive characteristics, transients are like snapshots of the vocal tract and musical instruments, being important for the perception of timbre and recognition of the sound source. Many of the modeling techniques of musical and speech signals are not efficient at transient parts. The separation of transients from deterministic and stochastic signal components represent a significant improvement in modeling synthesis flexibility. This study is devoted to the analysis, modeling, and auditory perception measurements of speech and music transients.

In the research carried out, Transient Modeling Synthesis (TMS) is used to model the transient components of speech and musical signals. Next, TMS is evaluated and compared to traditional sinusoidal based Spectral Modeling Synthesis (SMS). The results of a phoneme recognition experiment and a quality MOS (Mean Opinion Score) test are used to measure the importance of an adequate modeling of transients as bursts in stop consonants.

Compared to the intelligibility of 98% obtained for the original utterances, TMS modeling attained 95%, which are significantly higher than the 87% attained with SMS modeling. It was also observed that removing the transient component reduces the intelligibility to 79%.

Finally, by enabling the separation of transient components, TMS allows the definition of an index to measure the ratio between the energy of original and of transient signal components. This index, called Index of Transience, has been evaluated. The va-

lues obtained were then compared in a test and applied to different musical instrument notes.

Keywords: Computer Music, transient modeling, signal processing, speech bursts, computational hearing.

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas

TMS *Transient Modeling Synthesis* Síntese por modelagem de transientes.
SMS *Spectral Modeling Synthesis* Síntese por modelagem espectral.
MDCT *Modified Discret Cosine Transform* Transformada discreta em cossenos modificada.
MPEG *Moving Picture Experts Group* Grupo de Especialistas em Imagens com Movimento.
AAC *Advanced Audio Coding* Codificação avançada de áudio.
VPM *Voice Pulse Modeling* Modelagem do pulso de voz.
PSOLA *Pitch Synchronous Overlap Add* Soma sobreposta síncrona com a frequência fundamental.
DCT *Discret Cosine Transform* Transformada discreta em cossenos.
RTF *Radiodiffusion Télévision Française* Radiodifusão e televisão francesa.
bpm Batidas por minuto.
VOT *Voice Onset Time* Tempo de início do vozeamento.
IDCT *Inverse Discret Cosine Transform* Transformada discreta inversa em cossenos.
STFT *Short Time Fourier Transform* Transformada de Fourier de curto prazo.
DFT *Discret Fourier Transform* Transformada discreta de Fourier.
OLA *Overlap Add* Soma sobreposta.
RMS *Root Mean Square* Raiz da soma quadrática.
ITR Índice de Transiência Regional.
ITC Índice de Transiência Comparativo.
ITG Índice de Transiência Global.
FFT *Fast Fourier Transform* Transformada rápida de Fourier.
H0 Hipótese nula.
H1 Hipótese alternativa.
S1 a S9 Situação 1 a Situação 9.
MOS *Mean Opinion Score* Pontuação de opinião média.
ré bemol Nota musical ré bemol.

mi4 Nota musical correspondente a 330 Hz.

sol3 Nota musical correspondente a 196 Hz.

fá4 Nota musical correspondente a 349 Hz.

Sumário

Capítulo 1

Introdução

Música e fala são dois fenômenos de extrema importância para seres humanos. As pessoas precisam se comunicar, se expressar e apreciar expressões, e estes dois fenômenos são duas formas importantes de execução destas necessidades. Atualmente, sistemas computacionais necessitam da modelagem de sinais de música e fala para vários propósitos: síntese, audição computacional, análise acústica, análise musical sistemática, transformação, composição, entre muitas outras. A modelagem com qualidade de sinais de fala e música é uma tarefa complexa, que ainda se apresenta como um desafio, principalmente quando se busca a representação com a menor quantidade de parâmetros possível, visando baixo custo computacional, procurando manter boa inteligibilidade e naturalidade.

A produção do som no ar tem como origem processos complexos de vibração/oscilação de corpos com massa que, neste caso, são elementos formadores de instrumentos musicais acústicos e da voz. A movimentação destes corpos requer a aplicação de energia para retirá-los do estado de repouso, ou de menor energia, para estados estacionários de vibração. Na maioria dos casos, entre o repouso e o estado estacionário, existe um estado transitório de vibração, como é o caso de ataques de instrumentos musicais e inícios de fonemas como consoantes oclusivas (*bursts*). Esses eventos são importantes para a percepção de timbre e reconhecimento da fonte sonora.

Este estudo é dedicado à análise, modelagem, e percepção auditiva de transientes no contexto de música e fala. A modelagem por síntese possibilita a separação das componentes de um sinal, representando maior flexibilidade para processamentos. O TMS (*Transient Modeling Synthesis*) é uma técnica simples de modelagem por síntese que viabiliza a detecção e separação de transientes (Verma & Meng, 2000). Esta

técnica explora a dualidade tempo-frequência de senoides e impulsos, e considera que os transientes, por apresentarem característica impulsiva no domínio do tempo, são bem modelados por senoides no domínio da frequência. A modelagem de transientes em sinais musicais vem sendo bastante explorada (Daudet, 2006), mas o mesmo não acontece para sinais de fala. Neste estudo, foram realizados testes para avaliar, tanto a importância de uma modelagem adequada para transientes em sinais de fala quanto o desempenho do TMS para aplicações em fala e música.

1.1 Motivação

As regiões de transição de notas musicais, definidas como ataques (região de inicio da nota) e decaimentos (final da nota), apresentam transientes. Estas regiões são motivo de vários estudos e sua importância para a percepção é demonstrada por Luce (1963); Risset (1965); Grey (1977). Em Luce (1963), por exemplo, foi mostrado que a identificação do instrumento musical foi possível com apenas 60 ms do ataque, enquanto que, com a sustentação, eram necessários 150 ms. Em Loureiro et al. (2009) foram estudadas as transições entre notas musicais evidenciando diferentes características da qualidade das transições, sugerindo que tais características são determinantes na construção de uma interpretação musical. Os estudos de Loureiro et al. mostraram que a determinação dos instantes de final de ataque e início de decaimento, instantes delimitadores das regiões de transientes, ainda não está consolidado na literatura.

Estudos na área de fonética acústica mostram a existência de transientes no sinal de fala (Stevens et al., 1994). Transientes, no contexto da fala, são causados por constrições em algum trecho do trato vocal que passam a desempenhar função de fonte sonora, quando submetidas a um crescimento seguido de uma abrupta diminuição de pressão (Stevens, 2000). São encontrados, por exemplo, em consoantes oclusivas. Segundo Ladefoged & Maddieson (1996); Maia (1985) as consoantes oclusivas são encontradas em todas as línguas do mundo. Em Repp & Lin (1989), foi estudada a presença de transientes em consoantes oclusivas, e diversos experimentos perceptivos revelaram um alto grau de influência nas vogais seguintes. Foi observado desempenho semelhante no reconhecimento de oclusivas com a presença apenas dos transientes e com a presença de toda a parte antecedente às vogais. A explicação foi a presença de informação relevante nos transientes que, por sua característica impulsiva no domínio do tempo, são como retratos instantâneos do trato vocal ajudando, inclusive, na determinação do ponto de articulação.

Grande parte das técnicas de modelagem física de sinais de música e fala focam nas componentes estacionárias e estocásticas. Porém, ter um modelo dedicado aos transientes provou ser benéfico para sistemas de parametrização de sinais de áudio, pelo fato de modelagens senoidais e de ruído não serem eficientes para modelar estes importantes eventos para a percepção (Goodwin, 1996; Verma & Meng, 1998; Serra & Smith, 1990). O TMS é um método simples mas que apresenta uma abordagem convincente de modelagem de transientes. Apesar de ter sido uma das primeiras tentativas de modelagem de transientes, não foram encontradas na literatura análises esclarecedoras de aplicação deste método. De forma a cobrir esta lacuna, o TMS foi escolhido para ser avaliado nos testes. O TMS tem sido mais utilizado para sinais musicais, apesar de ter sido proposto para sinais de áudio em geral. A proposta deste trabalho é usá-lo também para a modelagem de fala, pelo fato de terem sido encontrados poucos estudos sobre esta abordagem.

1.2 Estrutura da dissertação

Esta dissertação está dividida em sete capítulos. O primeiro capítulo, **Introdução**, contém uma visão geral do assunto abordado e uma motivação para a escolha do tema e desenvolvimento da pesquisa. No segundo capítulo, **Fundamentação teórica**, inicialmente é apresentado um histórico sobre estudos relacionados a transientes em sinais de música e fala na literatura ao longo das últimas décadas. Em seguida, são mostrados os conceitos básicos do estado transitório de oscilador, fundamento importante para entender a origem dos transientes. É também apresentada neste capítulo uma descrição dos experimentos e resultados perceptivos de Pierre Schaeffer no **Solfejo dos Objetos Sonoros**. O capítulo é encerrado com duas seções contendo uma descrição de como os transientes estão presentes em sinais de fala e música. No terceiro capítulo, o TMS, método de modelagem e separação de transientes estudado e testado nos experimentos, é explicado com detalhes. São expostos seus princípios matemáticos e a descrição das etapas de processamento. O quarto capítulo apresenta métodos de detecção de regiões com presença de transientes, sendo o Fluxo Espectral o método utilizado para detectar regiões de transição de notas musicais. Formas de caracterização de transientes em sinais de fala e música são discutidas, e apresentadas as definições dos Índices de Transiência, propostos neste estudo. No capítulo seguinte, são apresentados os resultados de testes iniciais com o TMS, um experimento para medir a importância da presença de *bursts* para o reconhecimento de consoantes oclusivas e a importância da modelagem

adequada destes eventos. São apresentados também resultados de um teste para comparação dos Índices de Transiência para diferentes instrumentos musicais. O penúltimo capítulo contém a discussão dos resultados e, no último capítulo, são apresentadas as conclusões do estudo e propostas de trabalhos futuros.

Capítulo 2

Fundamentação teórica

2.1 Histórico

O ponto de partida deste histórico é o trabalho de Liberman et al. (1954), onde foram estudadas as pistas acústicas de consoantes oclusivas importantes para a percepção. As características espectrais dos *burts*, como a frequência central do espectro, foram avaliadas como importantes para diferenciar as consoantes. Em Repp & Lin (1989), a presença de transientes em consoantes oclusivas foi estudada e diversos experimentos perceptivos foram realizados. Nestes casos, os transientes ocorrem no início do sinal e um alto grau de influência nas vogais seguintes foi notado nos experimentos. Os sujeitos apresentaram desempenho semelhante no reconhecimento de diferentes consoantes oclusivas quando ouviram apenas os transientes e toda parte antecedente as vogais. Foi observado que os transientes ajudam, inclusive, na determinação do ponto de articulação.

Em Friedlander & Porat (1989) foi proposta a representação de Gabor para detecção de transientes em sinais, utilizando janelas exponenciais de um lado, alegando serem estas mais adequadas a este tipo de sinal. Mais tarde, em Hant et al. (1997) foi desenvolvido um modelo psicoacústico de predição de limiar de mascaramento de *bursts* de oclusivas não vozeadas, o qual pode ser aplicado em sistemas de síntese. Masri & Bateman (1996) usaram características de informações de altas frequências do espectro para detectar transientes, enquanto que Duxbury et al. (2001) propuseram a separação de transientes de sinais musicais através de técnicas de análise de multirresolução.

Levine & Smith (1998) desenvolveram um sistema de codificação de áudio de baixa taxa de transmissão que permite realizar transformações nos parâmetros. O

sistema separa o áudio de entrada em três partes: senoides, transientes e ruído. Cada parte pode assim ser quantizada separadamente, permitindo transformações temporais e espectrais com facilidade. O sistema utiliza *transform coding* para modelar os transientes, ou seja, utiliza informações a priori para realizar a modelagem. É baseado em uma simplificação do MPEG-AAC (*Advanced Audio Coding*) que utiliza MDCT (*Modified Discret Cosine Transform*).

Em 2003, foi proposto em Röbel (2003) uma abordagem para tratar processamento de transientes no *phase vocoder*. A abordagem parte do princípio de que existe uma imprevisibilidade de quadros em regiões de transientes e, por isso, o espectro de fase inevitavelmente deve ser deve reiniciado. A detecção de quadros transientes é feita através da comparação do cálculo do centro de gravidade do espectro a um limiar previamente ajustado. À medida que o quadro se desloca de regiões de transientes para regiões estacionárias, o valor do centro de gravidade se altera.

Em Molla & Torrésani (2004) foi proposta uma abordagem para determinar, o grau de tonalidade e o quão transiente (chamado pelo autor de *transientness*) é um sinal. O método se baseia no fato de ambos, sinais transientes e tonais, apresentarem expansão esparsa em bases *wavelet* e cossenoidais, respectivamente. Assumido isso, a tonalidade e a *transientness* são calculadas através de uma medida de entropia do quão esparso é o sinal nas duas bases.

Em meados da década de 2000, houve um aumento de estudos sobre separação e modelagem da componente transiente de sinais de fala. Szwoch et al. (2006) propuseram um algoritmo de detecção de transientes em sinais de fala, alegando que sistemas de codificação podem ser melhorados com a detecção e modelagem adequada dos transientes. O algoritmo é baseado na análise em multi-bandas de frequência. Rasetshwane et al. (2006) combinaram a abordagem de análise em sub-bandas com a transformada *wavelet* com taxa de amostragem variável para identificar e modificar transientes em sinais de fala.

Em 2007, foi desenvolvido por Bonada & Serra (2007) um sistema de síntese de voz cantada, que busca combinar os benefícios da fidelidade da síntese por concatenação e da flexibilidade da síntese por parametrização. O sistema utiliza um método denominado de VPM (*voice pulse modeling*) para modelar o pulso glotal filtrado isolado utilizando o espectro. O VPM faz isso centrando um pulso em uma janela no tempo, detectando picos harmônicos do espectro desta janela, e interpolando. Através

da ressíntese, o resíduo é extraído e, o ruído de aspiração é modelado usando um procedimento síncrono temporalmente com o *pitch*, o PSOLA (*Pitch Synchronous Overlap Add*). Finalmente, a parte transiente do resíduo é detectada e modelada pelo método descrito em Bonada & Serra (2007) que, pela integração da fase espectral, discrimina quais picos contribuem para a formação dos transientes.

Em Neto et al. (2012), o TMS foi utilizado para separar os transientes do resíduo da modelagem paramétrica baseada em estimação conjunta de fonte e filtro proposta. O resultado foi uma melhor modelagem do ruído sem a presença dos transientes.

Estudos relacionados à região de transientes em sinais musicais vêm sendo feitos desde a década de 1960. Em Luce (1963) foi mostrado que a identificação do instrumento musical foi possível com apenas 60 ms do ataque, enquanto que com a sustentação eram necessários 150 ms. Risset (1965) analisou sons de trompete e, após a obtenção de curvas individuais de evolução temporal das amplitudes e frequências para cada harmônico, conseguiu sintetizar sons de trompete a partir de aproximações, por segmentos lineares, das curvas obtidas na análise. Mostrou também que há diferença no tempo de início de cada harmônico no ataque. Grey (1977) mostrou que durante transições de notas ocorre uma mudança no equilíbrio entre os harmônicos devido a diferentes taxas de crescimento e decrescimento dos mesmos.

O TMS foi uma das primeiras tentativas de detecção e modelagem de transientes. Foi inicialmente proposto em Verma & Meng (1998) e refinado em Verma & Meng (2000). A técnica explora a dualidade tempo/frequência de senoides e impulsos e propõe a modelagem senoidal da transformada discreta em cossenos (DCT) do sinal. Sinais impulsivos no tempo passam a ser estacionários com a DCT, viabilizando a modelagem senoidal de transientes.

Em Daudet (2006), é apresentada uma revisão de alguns métodos de extração de transientes em sinais musicais, mencionando que podem ser aplicados a fala e outros sinais de áudio. Os métodos foram classificados de acordo com a natureza de suas saídas.

Em Loureiro et al. (2009) foram estudadas as transições entre notas musicais evidenciando diferentes características da qualidade das transições, que podem estar relacionadas à habilidade do músico, ao tempo de reverberação do ambiente ou a características acústicas do instrumento. Sugerindo que tais características são determinantes na construção de uma interpretação musical, o estudo buscou modelar estas

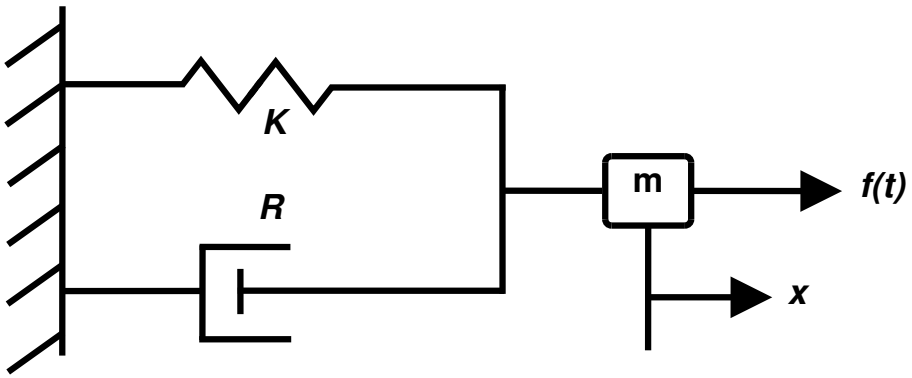

Figura 2.1. Representação de um oscilador forçado com amortecimento.

características a partir de descritores acústicos com a finalidade de construir um modelo de análise da expressividade musical. Os estudos de Loureiro et al. mostraram, ainda, que a determinação dos instantes de final de ataques e início de decaimento, instantes delimitadores das regiões de transição, ainda não está consolidada na literatura, e que a análise da presença de transientes nestas regiões pode auxiliar neste problema.

2.2 Estado transitório de um oscilador

A resposta transiente de um oscilador ocorre quando uma força externa é aplicada ao sistema com uma frequência f inicialmente diferente da frequência natural f_0. Quando a força externa é inicialmente aplicada ao oscilador, o movimento resultante pode ser um tanto quanto complicado (Fletcher & Rossing, 1998). Se o sistema for muito amortecido, a vibração transiente decai rapidamente. Se não houver um amortecimento grande, o oscilador pode permanecer no estado transiente por muitos ciclos de oscilação. Além disso, se f apresentar valor próximo a f_0, fortes batimentos podem ocorrer com a frequência resultante $|f - f_0|$. Se f for igual f_0, apenas uma alteração na amplitude ocorre. A representação de um oscilador amortecido com oscilação forçada é mostrada na Figura 2.1.

Dada uma força externa da forma

$$f_e(t) = F\cos(wt), \tag{2.1}$$

onde F é a amplitude da excitação gerada pela força externa e w é a frequência angular, a expressão matemática para o oscilador forçado amortecido é, então, uma equação

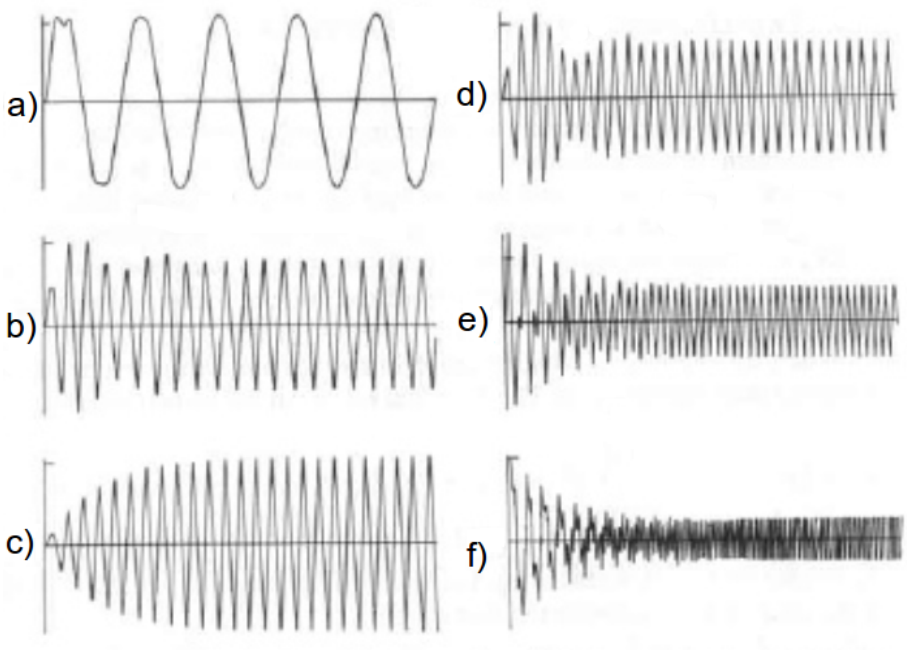

Figura 2.2. Respostas de um oscilador forçado com amortecimento: razão $\frac{f}{f_0}$ igual a $0,2$ em **a)**; $0,8$ em **b)**; $1,0$ em **c)**; $1,2$ em **d)**; $2,0$ em **e)**; $4,0$ em **f)**. Figura adaptada de Fletcher & Rossing (1998).

diferencial da forma

$$m\ddot{x} + R\dot{x} + Kx = F\cos(wt), \tag{2.2}$$

onde m a massa, R a resistência mecânica, e K a constante de mola. Resolvendo-se a equação diferencial descrita na Equação 2.2, chega-se à solução

$$x = Ae^{-t\alpha}\cos(w_d t + \phi) + \frac{F}{wZ}\sin(wt + \phi), \tag{2.3}$$

em que A e ϕ são constantes arbitrárias determinadas pelas condições iniciais, F é a amplitude da excitação gerada pela força externa, w é a frequência angular de $f_e(t)$,w_d é a frequência angular natural de amortecimento do oscilador e $\alpha = \frac{R}{2m}$ uma constante de amortecimento. A primeira parcela da soma da Equação 2.3 é referente ao caso de sub-amortecimento, em que $w_d = \sqrt{w_0^2 - \alpha^2}$. Se o amortecimento for muito pequeno, w_d pode ser substituído por w_0.

Na Figura 2.2, adaptada de Fletcher & Rossing (1998), estão presentes seis respostas de um oscilador simples a forças senoidais aplicadas repentinamente. Cada resposta é referente a uma razão $\frac{f}{f_0}$ diferente. Da letra **a)** a **f)**, os valores da razão $\frac{f}{f_0}$ variam de $0,2$ a $4,0$. Se w for igual a w_0, a amplitude da oscilação cresce exponencial-

mente, sem perturbações, até atingir o estado estacionário, como pode ser visto em **c)**. Note que, independentemente de como o movimento do oscilador é iniciado, o estado estacionário é sempre atingido.

2.3 Percepção de ataques em objetos sonoros

Pierre Henri Marie Schaeffer nasceu na França em 1910, tendo sido um importante compositor, engenheiro de telecomunicações e musicólogo de sua época. Desenvolveu estudos importantes sobre percepção auditiva, conduziu trabalhos pela *Radiodiffusion Télévision Française* (RTF) em Paris, e foi precursor da Música Concreta (Palombini, 2006). Um de seus estudos foi o **Solfejo dos Objetos Sonoros** (Schaeffer, 1967). O **Solfejo dos Objetos Sonoros** é uma sequência de áudios em que Schaeffer apresenta vários experimentos perceptivos. Guiados pela narração do autor, os áudios dos experimentos são apresentados, possibilitando aos ouvintes a verificação das conclusões com o próprio sistema auditivo. As considerações relativas ao ataque em objetos sonoros e limiares de audição serão apresentados nesta seção.

O conceito de objeto sonoro é abordado em detalhes em Palombini (2006). De uma forma resumida, objeto sonoro é o som em si, dissociado da fonte de produção e de contexto previamente envolvidos na produção e escuta dos sons. Segundo Schaeffer, o ataque é um dos eventos mais familiares na música. Schaeffer relata ainda que o conhecimento adquirido com a música concreta e técnicas disponíveis na época levaram às seguintes conclusões: O ataque é muito importante na criação de objetos musicais, e varia com a natureza do corpo em vibração e com a forma em que foi posto a vibrar. Varia também com a dinâmica (variação de energia) do início da nota, qualitativamente classificada como percussiva, explosiva, etc. A complexidade harmônica emitida pelo corpo ressonante é também uma variável importante.

2.3.1 Limiares de audição

Em Schaeffer (1967), são realizados experimentos com pulsos sintetizados. Até 24 pulsos por segundo, ouvem-se pulsos distintos. A partir de 29 pulsos por segundo, uma sensação de altura (percepção de frequência fundamental) é observada, e o som passa a ter uma granulosidade. À medida que a frequência dos pulsos é aumentada, uma sensação de rugosidade aparece. O mesmo foi feito a partir de um grão de uma nota grave de fagote. A sequência de sensações foi descrita como: inicialmente, choques regulares, seguidos de vestígios rítmicos, chamados de grão, paralelamente a

um crescente efeito de altura e, por último, a emergência de uma textura, colorindo a altura.

Outro experimento mostra que existe um limite para a acumulação temporal de objetos. De certa forma a música estabeleceu este limite como sendo semi-fusas a 60 batidas por minuto, o que corresponde a 62,5 ms ou $\frac{1}{16}$ s. Uma escala descendente de piano foi executada com durações iguais a semifusas ($\frac{1}{16}$ s) e, como esperado, é possível perceber a separação das notas. Não podendo o pianista executar mais rápido, foi usada a função acelerando do gravador e, com isso, as durações passaram a ser de $\frac{1}{32}$ s cada uma. Neste ponto, passa-se a não ter mais uma distinção clara das notas. O mesmo experimento foi realizado para palavras. Com sílabas de duração média de 40 ms, não foi mais possível reconhecer o que estava sendo dito. Ao aumentar a duração média para 80 ms, a inteligibilidade foi restaurada. A conclusão foi de que a barreira de separação é de 50 ms.

A constante de tempo, considerada o menor intervalo de tempo abaixo do qual o ouvido é indiferente à natureza dos estímulos, foi de 5 ms. Qualquer aparição entre 0 e 5 ms é percebida como ruído parasita, ou seja, uma pequena explosão.

Um novo experimento foi realizado com o intuito de medir o menor intervalo de tempo no qual o aparelho auditivo consegue reconhecer timbre e altura. Quatro séries de durações iguais a 3, 5, 10, 25, 50 e 250 ms, duas para clarineta e duas para trompete, foram executadas. Em aproximadamente 10 ms, começa a surgir a percepção de altura. A partir daí, começa a surgir a sensação de cor, timbre e, posteriormente, de reconhecimento de instrumento. Foi também demonstrado que, com durações abaixo de 6 ms, fragmentos de alturas diferentes concatenados fundem-se, criando uma melodia subliminar. A partir de 10 ms de duração, passaram a formar uma estrutura melódica.

2.3.2 Altura (*pitch*)

Schaeffer realizou uma filtragem em uma nota grave de piano, deixando apenas os três primeiros harmônicos. O timbre é afetado consideravelmente, porém, a altura ou *pitch* (percepção de frequência fundamental) e a intensidade não se alteram. Retirando apenas a fundamental, nenhuma diferença muito aparente foi constatada. Contudo, cortando-se a fundamental de uma nota de altura média, o timbre foi gravemente alterado. E por último, repetindo o procedimento para uma nota aguda, o observado é o oitavamento da nota. Portanto, a altura não está completamente associada com a

frequência fundamental.

Neste mesmo tema, Schaeffer realiza outro conjunto de experimentos. Filtragens com diferentes frequências centrais foram feitas no ruído branco, sucessivamente, resultando em uma percepção de sequência melódica. A mesma filtragem foi feita em um som complexo estruturado. Neste caso o timbre é alterado mas o som não evolui em tessitura.

2.3.3 Intensidade

Para sons sustentados, nos quais o conteúdo harmônico não varia consideravelmente no decorrer da nota, o ouvido é mais sensível à variação de energia no tempo. Em sons percussivos seguidos de ressonância, o ouvido é mais sensível a como a energia desaparece do que como ela aparece. Outra consideração é que sons do tipo percussão-ressonância têm a caracterização do timbre no momento do ataque. Já em sons sustentados, o ataque tem um papel secundário na caracterização do timbre. Nestes casos, o timbre é o resultado da combinação da percepção do ataque com a percepção do restante da nota.

Schaeffer mostra que suprimindo-se até um segundo do início de uma nota grave de piano, a nota permanece quase sem alteração no timbre. Paradoxalmente, quando o ataque de um som de sino soando é recortado, a alteração na percepção é bastante alterada. A supressão do ataque em alguns casos altera muito a percepção, e em outros casos não. Uma explicação é o fato da nota grave do piano ser formada por um único som, e uma nota de vibrafone ser formada por dois sons: um choque metálico muito breve, e uma ressonância que depende da construção do instrumento. Portanto, para o caso de um som único, a supressão do ataque não altera muito a percepção, enquanto que, para sons duplos, a retirada do ataque eliminaria uma das componentes do som. Outro experimento descarta a existência de uma correlação na forma de onda entre ataques parecidos. Oito stacatos de trompete foram gravados. Suas formas de onda foram comparadas com o intuito de achar algum padrão, sem nenhum sucesso. Porém, a diferença na forma da curva de evolução temporal de energia é um fator importante.

Os primeiros 50 ms foram cortados de um **ré bemol** de flauta e o observado foi uma diferença sutil para o original. A diferença segundo Schaeffer, é que os primeiros 50 ms fornecem uma espécie de ruído causado pelo sopro no instrumento. Ao realizar

a mesma experiência para o trompete, esta diferença sutil não foi encontrada. A explicação fornecida pelo autor é de que, no trompete, o ataque é principalmente influenciado pelo formato da curva de energia. Portanto, para reforçar seu argumento, Shaeffer produziu um ataque artificial em uma nota sustentada de trompete, cortando a fita magnética obliquamente, e o resultado foi bastante similar à nota original. O mesmo para o violino pôde ser constatado. Um ataque artificial foi realizado primeiramente com um corte reto e posteriormente com um corte oblíquo, e, ao serem comparados com o original, o corte oblíquo se mostrou bastante similar.

Finalmente, o autor apresenta dois experimentos, nos quais notas de instrumentos diferentes, piano e flauta, são transmutadas entre si apenas pela alteração da curva de energia. No primeiro caso, um **mi4** de piano teve seu ataque recortado na fita magnética e posta a soar simultaneamente com a nota **mi4** da flauta. As duas notas se mostraram bem parecidas. Outra manipulação foi feita, agora com um **fá4** de flauta, cuja curva de energia foi modelada por um modulador de envoltória de forma a ficar semelhante à de um **fá4** de piano. Mais uma vez, as notas foram comparadas e se mostraram semelhantes.

2.3.4 Duração

Shaeffer realizou um experimento em que sons complexos com ataques curtos e ressonâncias longas foram executados. Os mesmos sons foram executados de trás para frente e, curiosamente, na execução invertida, os ataques pareceram bem menores. O trajeto da escuta não é mais realizado nem com a mesma velocidade nem da mesma maneira. Os ataques foram cortados e executados separadamente, ficando claro que parecem muito menores isolados do que quando tocados no conjunto da nota. A conclusão deste conjunto de experimentos é que existe uma anamorfose tempo-duração. Há uma distorção entre a duração percebida e o tempo físico medido. Os elementos da forma perturbam consideravelmente os valores métricos.

2.3.5 Timbre

Shaeffer apresenta um conjunto de experimentos com o intuito de mostrar que o timbre não é dependente nem somente do espectro harmônico, nem somente da dinâmica (variação de energia), e sim de uma associação das duas coisas. Inicialmente oito sons provenientes de flauta, fagote, flauta, clarineta, oboé, trompete, e síntese, tiveram suas curvas de energia alteradas e, com isso, o reconhecimento da fonte produtora do

som foi bastante dificultado. Isto mostra que a curva de energia possui uma grande importância na composição do timbre. Para demonstrar a importância da composição harmônica, um tom puro (**sol3**), foi modulado com a mesma curva de energia de um **sol3** de piano, e o resultado foi uma nota parecida mas com uma cor diferente. Fazendo a mesma coisa mas agora modulando uma nota de flauta, que possui espectro harmônico bem mais próximo ao do piano, o resultado foi bem mais próximo ao original.

Schaeffer coloca a questão de como reconhecemos um timbre característico de um instrumento, como o piano por exemplo, se cada uma de suas notas possue timbre próprio. Foi constatado que a dinâmica das notas do piano se torna cada vez mais acentuada à medida que aumenta-se a altura. Schaeffer gravou 22 notas de piano sobre as sete oitavas consecutivamente, e podendo perceber claramente a mudança de dinâmica. Para constatar a mudança harmônica, foram gravadas notas graves transpostas pelo gravador para duas oitavas acima, e notas agudas transpostas pelo gravador para duas oitavas abaixo. As notas transpostas foram comparadas com as notas originais, percebendo-se claramente que notas mais graves possuem um espectro harmônico muito mais rico que notas mais agudas. Schaeffer define então o que ele chama de lei do piano. A lei do piano foi definida como uma lei de compensação entre timbre dinâmico e timbre harmônico. Em uma progressão do grave para o agudo, a inclinação da curva de energia cresce constantemente, enquanto que o timbre harmônico decresce proporcionalmente.

A causa é colocada como forte influente no discernimento das fontes emissoras. Um experimento foi feito com uma nota de trompete soando com um incidente no início. O incidente causa certa estranheza que o autor chamou de excesso de timbre. Em outro experimento, duas ressonâncias provenientes de uma excitação de chapa de metal e outra proveniente de uma simulação do piano foram comparados. Quando apenas a parte de ressonância é executada, é muito difícil dizer o que gerou ambos os sons. No momento em que o som por inteiro é tocado, o gerado pelo piano é imediatamente reconhecido. Segundo Schaeffer, quando o contexto de causalidade intervém, é inserido um novo fator: a psicologia da audição propriamente dita.

2.4 Transientes em sinais de fala

Focando na análise de sons de fala, estudos na área de fonética acústica mostram a existência de transientes na fala (Stevens et al., 1994; Stevens, 2000; Flanagan, 1972).

Os transientes ocorrem em situações da fala em que uma obstrução dos articuladores do trato vocal acontece em um intervalo de tempo, seguido por uma soltura repentina da corrente de ar. No momento da obstrução, há um súbito aumento de pressão na região obstrutora, seguido de uma diminuição abrupta. Esta variação rápida de pressão faz com que a região de articuladores que gerou a obstrução vibre, se comportando como uma fonte sonora independente das pregas vocais (Flanagan, 1972). Essa situação pode ser associada a um degrau de excitação aplicado a um oscilador. O resultado é um estado transitório de vibração, caracterizado pela presença de transientes.

As consoantes oclusivas [p], [t], [k], [b], [d], [g], apresentam uma região de transientes, importante em sua composição, denominada por *bursts* (Liberman et al., 1954; Repp & Lin, 1989). Segundo Kent & Read (2002), a região de *bursts* dura entre 10 e 30 milissegundos em média. As oclusivas não-vozeadas ([p], [t] e [k]), são formadas por uma região de silêncio, seguida por transientes ou *bursts*, e uma transição para o fonema seguinte. Na transição, podem aparecer componentes de africação e aspiração.

Na Figura 2.3, retirada de Stevens (2000), está ilustrada a sequência de eventos para a produção de oclusivas não vozeadas seguida de um fonema vozeado. A sequência da esquerda para a direita na Figura 2.3 é: transientes, africação, aspiração e vozeamento do próximo fonema.

Nas oclusivas vozeadas ([b], [d] e [g]), existe uma região temporal chamada de pré-sonora, anterior à região de *bursts*. Na região pré-sonora ocorre a vibração das pregas vocais simultaneamente ao aumento da pressão no ponto de obstrução.

Estudos na área de fonoaudiologia diferenciam as oclusivas do português brasileiro pelo ponto articulatório e pelo contraste de sonoridade (Melo et al., 2012). São procuradas pistas acústicas para caracterizar a sonoridade das oclusivas, com o intuito de compreender a causa de problemas na pronúncia correta. Algumas destas pistas são o *Voice Onset Time* (VOT) e amplitude dos *bursts*. O VOT é um parâmetro básico e fundamental para o estabelecimento do contraste de sonoridade das oclusivas (Bonatto, 2007). Esse parâmetro corresponde ao intervalo de tempo entre os *bursts* e o início do vozeamento.

Em Repp & Lin (1989) foi verificado que características espectrais dos *bursts*, como a frequência central do espectro, ajudam a diferenciar as consoantes. Os *bursts* ocorrem no início do sinal, mas um alto grau de influência nas vogais seguintes foi

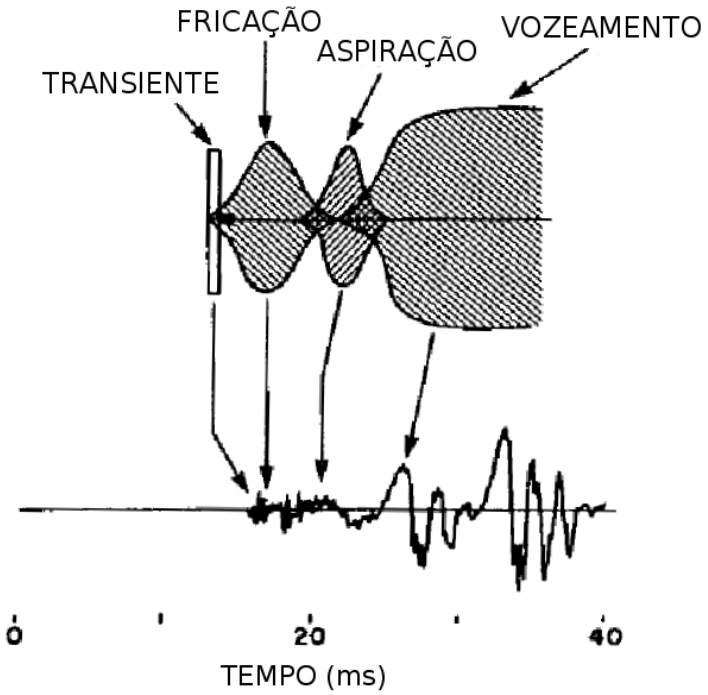

Figura 2.3. Representação da sequência de eventos para produção de oclusivas não-vozeadas. Figura adaptada de Stevens (2000).

notado nos experimentos. Os informantes apresentaram desempenho semelhante no reconhecimento de diferentes consoantes oclusivas quando ouviram apenas os transientes e toda a parte antecedente às vogais. A explicação para a presença de informação relevante nos transientes é que, por sua característica impulsiva no domínio do tempo, são como retratos instantâneos do trato vocal. Podem ajudar inclusive, na determinação do ponto de articulação.

2.5 Transientes em sinais musicais

Voltando a atenção a sinais musicais, as regiões de início e final das notas, definidas como ataques e decaimentos, são as regiões que apresentam maior quantidade de transientes. Estas regiões são também chamadas de regiões de transição da nota. A detecção e segmentação destas regiões é importante em análises. Em Loureiro et al. (2009) as transições entre notas musicais foram estudadas. Diferenças acústicas que podem estar relacionadas à habilidade do músico ou a características acústicas do instrumento foram evidenciadas, sugerindo que tais características são determinantes na construção de uma interpretação musical. Foi verificado também que, até 2009, a determinação dos instantes delimitadores das regiões de transição não estava

consolidada na literatura, e que a presença de transientes nestas regiões pode auxiliar no problema. Grey (1977) mostrou que durante as transições das notas, ocorre uma mudança no equilíbrio entre os harmônicos, devido a diferentes taxas de crescimento e decrescimento.

As transições são muito importantes para a percepção. Em Luce (1963) foi mostrado que a identificação do instrumento musical foi possível com apenas 60 ms do ataque. Para regiões de sustentação da nota, foram necessários em média, 150 ms.

Muitos modelos propostos para sinais musicais não são eficientes para regiões de transição. Pelo fato de serem importantes para a percepção, em muitos destes sistemas, o resíduo é somado integralmente em regiões de transição sem nenhuma modelagem. Por isso, ter um modelo dedicado aos transientes provou ser benéfico para sistemas de parametrização de sinais de áudio (Goodwin, 1996; Verma & Meng, 1998; Serra & Smith, 1990).

Neste capítulo, foram apresentados conceitos fundamentais necessários à compreensão do processo de formação e funções desempenhadas por transientes na música e na fala. A seguir, a atenção é focada na técnica de modelagem de transientes TMS.

Capítulo 3

Transient Modeling Synthesis (TMS)

O TMS (Verma & Meng, 1998, 2000) foi uma das primeiras propostas de detecção e modelagem de transientes. Foi descrito como uma extensão da modelagem senoidal SMS (*Spectral Modeling Synthesis*) (Serra & Smith, 1990). A modelagem senoidal, até então, foi muito utilizada para sinais de fala e música em aplicações como transformação, compressão, redução de ruído e análises. Porém, este tipo de modelagem é apropriada para sinais com componentes senoidais de variação lenta.

Os transientes são eventos de curta duração, que apresentam características impulsivas. Devido à dualidade tempo frequência de impulsos e senoides, sinais impulsivos no tempo apresentam espectro espalhado de variação lenta, de difícil detecção. Porém, com uma representação em frequência adicional, é possível utilizar a modelagem senoidal para parametrizar apenas os transientes do sinal previamente transformado. A representação em frequência adicional utilizada é a DCT (*Discrete Cosine Transform*).

O SMS modela bem a parte estacionária dos sinais, retornando uma saída do tipo Modelo + Resíduo. Este resíduo contém componentes de ruídos, transientes e erro de modelagem somados. O SMS apresenta também uma abordagem de modelagem da parte estocástica estacionária ou ruído, extraindo a envoltória do espectro do resíduo por decimação. Porém, como descrito em Serra & Smith (1990), a modelagem estocástica do resíduo não é eficiente para transientes. Por isso, a proposta do TMS foi utilizar o SMS para separar a parte determinística, e o próprio SMS para modelar os transientes através da DCT do resíduo. Feita a modelagem da DCT do resíduo,

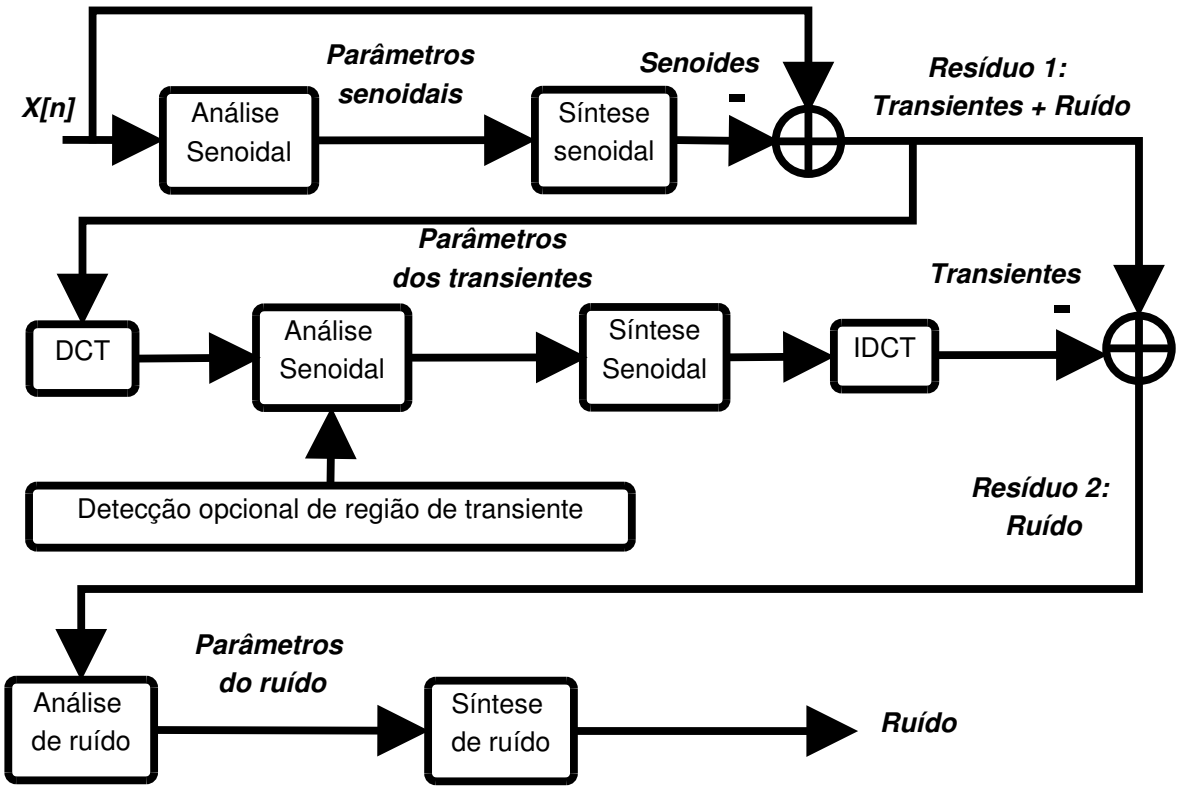

Figura 3.1. Diagrama de blocos representando as etapas de análise TMS.

os transientes são sintetizados com a aplicação da IDCT (*Inverse Discrete Cosine Transform*) na ressíntese da modelagem do SMS. Com o TMS, o sinal passa a ser modelado em três componentes: Modelo determinístico + Modelo estocástico + Modelo dos transientes.

Modelar e separar as componentes do sinal possibilita maior flexibilidade para processamentos, permitindo modificações ou análises isoladas em cada componente. Na Figura 3.1 pode ser visto um diagrama de blocos das etapas de análise TMS. $X[n]$ representa o sinal a ser processado.

Informações provindas de uma função de detecção de regiões de transientes podem ser utilizadas para que o TMS seja aplicado apenas nestas regiões, evitando processamentos computacionais desnecessários. Na Figura 3.2 pode ser visto um diagrama de blocos das etapas de síntese TMS. $X'[n]$ representa o sinal sintetizado final.

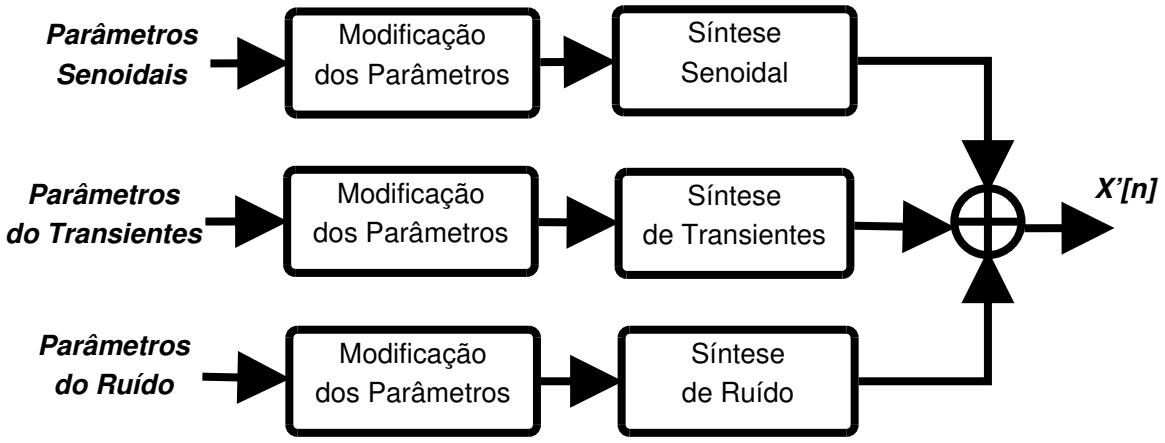

Figura 3.2. Diagrama de blocos representando as etapas de síntese do TMS.

3.1 Procedimentos de modelagem

Para o processamento do TMS, depois de extraída a componente estacionária por algum método, o sinal de entrada deve ser dividido temporalmente em blocos de análise, sem sobreposição. Verma & Meng (2000) sugerem utilizar um segundo de duração para cada bloco, ou um tamanho em que a largura dos picos dos transientes seja bem menor que os blocos. Para cada bloco, a DCT é calculada e analisada por SMS. Os parâmetros de análise do SMS influenciam drasticamente no resultado da modelagem. São sugeridos também de 30 a 60 senoides por blocos de análise para aplicações em que a síntese fiel é necessária e de 5 a 10 senoides para aplicações em que apenas um detector de transientes é desejado.

Os parâmetros são formados por uma tríplice de valores que representam as senoides do SMS, ficando da seguinte forma: $\{A^k_{l,m}, F^k_{l,m}, \phi^k_{l,m}\}$, em que $A^k_{l,m}$ é a amplitude, $F^k_{l,m}$ a frequência, e $\phi^k_{l,m}$ a fase da k-ésima senoide, l-ésimo quadro de análise, e m-ésimo bloco de DCT. Os parâmetros podem ser interpretados da seguinte forma: a frequência na DCT representa o instante temporal do transiente, a magnitude representa a intensidade, e a fase representa a direção (positiva ou negativa) de início do transiente.

A transformação nos parâmetros do TMS segue princípios análogos de transformação dos parâmetros senoidais. A modificação de escala de tempo sem mudança da altura (*pitch*) deve ser feita alterando-se a duração dos blocos de análise da DCT. Para modificações em amplitude, apenas um fator deve ser multiplicado a $A^k_{l,m}$ e, para modificações em altura, nada precisa ser feito nos parâmetros do TMS, devido às ca-

racterísticas impulsivas não-tonais dos transientes.

3.2 Modelagem senoidal

A modelagem senoidal tem sido utilizada para síntese tanto de sinais de fala como de música (McAulay & Quatieri, 1986; Serra & Smith, 1990). O princípio básico de modelagem senoidal é a representação do sinal através de uma soma de componentes senoidais. Considerando um sinal $s(t)$, sua representação senoidal é, portanto, da forma

$$s(t) = \sum_{n=1}^{N} A_n(t) \cos \theta_n(t), \tag{3.1}$$

onde N é o número de componentes senoidais, $A_n(t)$ e $\theta_n(t)$ são a amplitude instantânea e a fase instantânea das componentes senoidais. A fase instantânea é definida como

$$\theta_n(t) = \int_0^t \omega_n(\tau) d\tau + \theta_n(0) + \phi_n, \tag{3.2}$$

onde $\omega_n(\tau)$ é a frequência angular instantânea, $\theta_n(0)$ é o valor inicial da fase e ϕ_n é o deslocamento fixo de fase. Para a detecção dos valores dos parâmetros das componentes senoidais, são necessárias três etapas. A primeira é a análise em frequência dos quadros de curta duração do sinal, ou *Short Time Fourier Transform* (STFT). A segunda etapa é a detecção de picos do espectro, de acordo com algum critério. E, por último, a síntese utilizando os parâmetros extraídos.

3.2.1 *Short Time Fourier Transform* (STFT)

A modelagem senoidal é apropriada para regiões periódicas. Sinais de música e fala apresentam regiões de periodicidade alta. Porém, estas regiões não são perfeitamente periódicas. A análise do sinal através de janelamento permite contornar o problema da periodicidade, visto que em um tamanho suficientemente pequeno, o sinal quase periódico pode ser considerado periódico. O janelamento permite também a representação da evolução temporal do sinal. Em cada quadro de análise do janelamento, é aplicada a Transformada de Fourier. A Transformada de Fourier discreta (DFT) de um sinal $x(n)$ é definida por

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\omega_k n}, \tag{3.3}$$

onde $\omega_k = \frac{2k\Pi}{N}$ é a frequência angular, N é o número de amostras temporais, n é o índice da amostra temporal, $k = 0, 1, 2, ..., N-1$ é o índice do *bin* de frequência. A

transformada inversa é da forma

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j\omega_k n}. \tag{3.4}$$

Para a análise janelada do sinal, a Transformada de Fourier deve ser aplicada a cada quadro, ficando da forma

$$X_l(k) = \sum_{n=0}^{N-1} w(n) x(n + lh) e^{-j\omega_k n} \tag{3.5}$$

onde $w(n)$ é uma janela real que determina a região do quadro $l = 1, 2, 3...$ de análise, e h é o salto temporal da janela, podendo haver superposição.

A utilização de uma janela temporal diferente da retangular é essencial para melhorar a discriminação de picos no espectro de frequência. A janela retangular possui um espectro espalhado e, por isso, outras janelas de análise são usadas para melhorar a detecção de picos locais. Quando existe a necessidade de ressíntese do sinal, é comum utilizar a sobreposição de janelas não-retangulares para a aplicação do *Overlap Add* (OLA). A sobreposição de janelas não-retangulares na síntese melhora a suavização do espectro nas transições entre os quadros.

3.2.2 Detecção de picos

Em cada quadro de análise, o espectro complexo precisa ser convertido para coordenadas polares, a fim de detectar picos na magnitude. Picos são definidos como máximos locais na magnitude do espectro. Dado um espectro $X(b)$, o *bin* b será um máximo local se seguir o critério

$$|X(b-1)| \leq |X(b)| \geq |X(b+1)|. \tag{3.6}$$

Detectados os picos locais, um critério de seleção dos picos de interesse deve ser adotado. Os critérios de seleção podem ser simples, como selecionar por magnitudes acima de um limiar, ou selecionar os N picos de maior magnitude. Podem ser mais elaborados, como usar a detecção de frequência fundamental para selecionar apenas os possíveis harmônicos. Podem também, usar informação de outros quadros de análise, como o algoritmo *peak detection* descrito em Serra & Smith (1990).

3.2.3 Síntese

Para realizar a síntese a partir dos parâmetros do modelo, dois métodos são mais utilizados. O primeiro é a geração das senoides por bancos de osciladores McAulay & Quatieri (1986). O segundo é a recomposição do espectro a partir dos parâmetros e aplicação da Transformada Inversa de Fourier (Serra & Smith, 1990). Neste caso, o espectro é reconstruído a partir da resposta espectral da janela de análise utilizada em cada conjunto de parâmetros senoidais. Por último, é aplicada a superposição de quadros ou *Overlap Add* (OLA).

3.3 Transformada discreta em cossenos

A transformada discreta em cossenos ou *discrete cosine transform* (DCT) foi inicialmente apresentada por Ahmed et al. (1974). Esta primeira versão é hoje chamada de DCT II e sua inversa de DCT III. Pertence à classe de transformadas unitárias senoidais estudada em Jain (1979), as quais possuem bases ortogonais inversíveis.

A DCT é usada em processamento de sinais e imagens principalmente para compressão e descompressão. As versões II e III da DCT recebem uma atenção maior em processamento de sinais, pelo fato de a transformação ser real, ortogonal e separável. É utilizada por exemplo no padrão internacional de codificação de áudio *Moving Picture Experts Group* (MPEG) (Rao & Hwang, 1996).

A DCT é calculada como

$$C(k) = \beta(k) \sum_{n=0}^{N-1} x(n) \cos\left[\frac{(2n+1)k\Pi}{2N}\right], \tag{3.7}$$

em que, $\beta(k) = \sqrt{\frac{1}{N}}$ para $k = 1$, $\beta(k) = \sqrt{\frac{2}{N}}$ para $k \neq 1$, e $n, k \in 0, 1, ..., N-1$, com n representando amostras no tempo do sinal $x(n)$, e k, *bins* de frequência de $C(k)$.

Por sua vez, inversa da transformada discreta em cossenos (IDCT) é calculada como

$$x(n) = \sum_{k=0}^{N-1} \beta(k) C(k) cos[\frac{(2n+1)k\Pi}{2N}]. \tag{3.8}$$

A grosso modo, um impulso no início de um quadro de análise resulta em uma transformada DCT cossenoidal de frequência relativamente baixa. Em contrapartida,

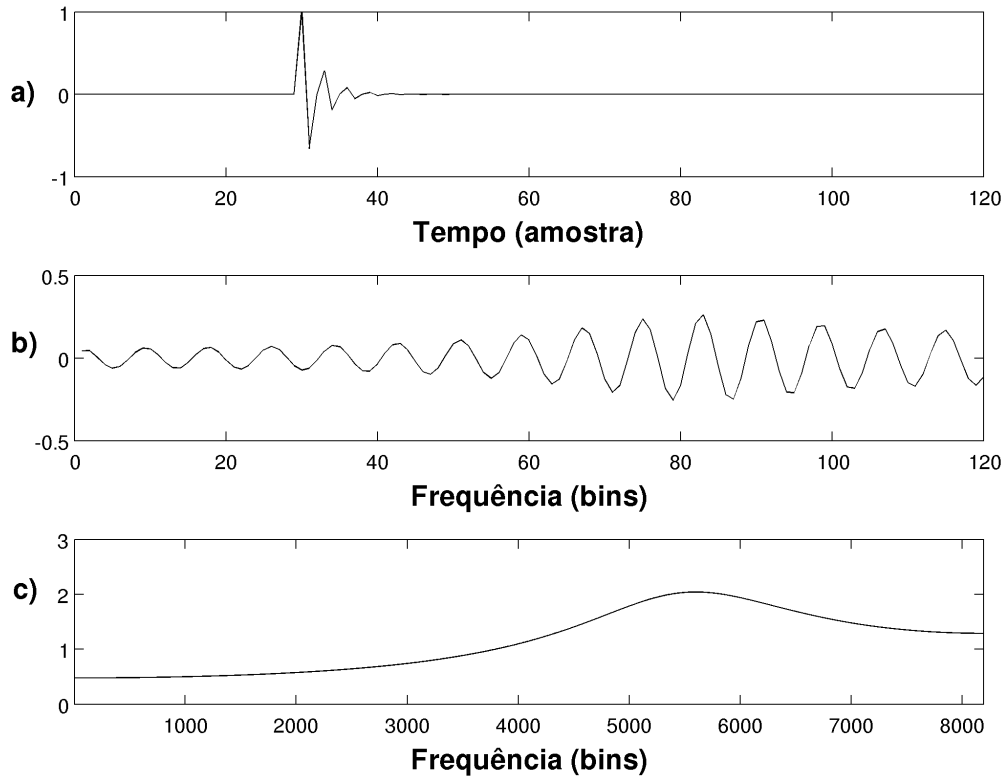

Figura 3.3. Forma de onda (a), DCT (b), e parte positiva da Transformada de Fourier (c) de uma senoide modulada por uma exponencial. Note que, em (c), a Transformada de Fourier foi calculada com alta resolução (*zero padding* de tamanho 16000).

um impulso na região final de um quadro apresenta uma transformada DCT cossenoidal de frequência relativamente alta. Na Figura 3.3 **(a)**, pode ser visto o sinal de uma senoide modulada por um decaimento exponencial, representando um transiente mais realístico do que um impulso. Esse sinal não seria bem representado por modelagem senoidal, visto que apresenta um espectro de frequência da Transformada de Fourier espalhado, como pode ser visto na Figura 3.3 **(c)**. A detecção de pico nesta curva iria retornar apenas um valor. A recomposição do espectro com apenas um valor, acarretaria em uma transformada inversa senoidal sem modulação, falhando assim, na modelagem. Entretanto, na Figura 3.3 **(b)**, onde pode ser visto o espectro da DCT, a curva apresenta um comportamento bem mais apropriado para modelagem senoidal.

No que diz respeito às propriedades matemáticas da DCT, as mais importantes são descritas em Yip (2001). São mostradas duas delas aqui: linearidade e dualidade

multiplicação/convolução. A DCT é linear. Considerando dois sinais, $x(n)$ e $y(n)$, uma transformada T é linear quando satisfaz

$$T(\alpha x(n) + \beta y(n)) = \alpha T(x(n)) + \beta T(y(n)). \tag{3.9}$$

A DCT apresenta também a propriedade de multiplicação/convolução. Essa propriedade é satisfeita quando uma transformada T satisfaz

$$x(n) * y(n) = T^{-1}(T(x(n)) \times T(y(n))). \tag{3.10}$$

Na Equação 3.10, o símbolo $*$ representa a operação de convolução e T^{-1} é a inversa da transformada T.

Neste capítulo, foram apresentados os princípios de funcionamento da técnica de modelagem de transientes TMS. As etapas de processamento do método foram demostradas, e detalhes para implementação e significado dos parâmetros foram apresentados. O capítulo seguinte é dedicado à discussão sobre detecção e caracterização de regiões em sinais de música e fala com presença de componente transiente.

Capítulo 4

Detecção e caracterização de transientes

4.1 Detecção de transientes

Do ponto de vista de modelagem, a detecção de transientes pode determinar o chaveamento da utilização de modelos adequados para sinais contendo apenas componente determinística + ruído, ou a utilização de modelos adequados para a modelagem de transientes. Esta decisão acarreta uma diminuição dos parâmetros, e também evita processamentos desnecessários. Basicamente, os métodos utilizam um dos quatro princípios: variações na energia do sinal, variações na magnitude do espectro de frequência, variações na fase do espectro e detecção por modelagem.

Vários métodos foram propostos. Alguns são pensados e testados especificamente para sinais de música ou de fala. Outros são apresentados como aplicáveis a qualquer sinal de áudio. Em Daudet (2006), foi apresentada uma revisão de alguns métodos de extração de transientes em sinais musicais, extensível para fala e outros sinais de áudio. Friedlander & Porat (1989) propõem a representação de Gabor para detecção de transientes, utilizando janela exponencial de um lado, alegando serem estas mais adequadas a este tipo de sinal. Foi proposto em Röbel (2003) uma abordagem para detecção de quadros transientes através da comparação do centro de gravidade do espectro com um limiar previamente ajustado. À medida em que o quadro se desloca de regiões de transientes para regiões estacionárias, o valor do centro de gravidade se altera.

Na música, Masri & Bateman (1996) usaram características de informações de altas frequências do espectro para detectar transientes, e Duxbury et al. (2001) propuseram a separação de transientes de sinais musicais através de técnicas de análise em multirresolução.

Na fala, Szwoch et al. (2006) propuseram um algoritmo de detecção de transientes em sinais, alegando que sistemas de codificação de fala podem ser melhorados com a detecção e modelagem adequada dos transientes. O algoritmo é baseado na análise em multi-bandas de frequência. Rasetshwane et al. (2006) combinam a abordagem de análise em sub-bandas usando a transformada *wavelet* com taxa de amostragem variável para identificar e modificar transientes em sinais de fala.

São detalhados a seguir dois métodos de detecção de transientes mais utilizados para sinais musicais: um por energia e um por variações na magnitude do espectro de frequência.

4.1.1 Detecção por energia

A presença de transientes em sinais musicais está intimamente relacionada à variação de energia (Grey, 1977). Em muitos casos, variações bruscas de energia levam à produção de transientes. Por este motivo, uma das formas de detecção dos instantes delimitadores da região de transição das notas é a partir da estimação dos máximos da taxa de variação de energia RMS (*Root Mean Square*) dentro da nota. A envoltória de energia RMS é calculada aplicando a Equação 4.1 para cada quadro de análise do sinal. Os quadros normalmente são da ordem de 20 milissegundos para sinais musicais.

$$RMS(q) = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x(n)^2} \tag{4.1}$$

onde N é o número de amostras de um quadro, $x(n)$ é a n-ésima amostra do quadro de análise, e q é o índice do quadro. O método é baseado na análise do contorno da envoltória de energia (Maestre & Gómez, 2005). Considerando o envelope de energia como uma função diferenciável contínua no tempo, os pontos de máxima curvatura são detectados pela derivada segunda. Os mínimos locais da derivada segunda determinam os pontos candidatos. Os dois candidatos escolhidos serão os que apresentarem maior inclinação positiva, medida entre ele e o início da nota, e maior inclinação negativa com o final da nota. Estes dois pontos definem o final do ataque e o início do decaimento. O ataque e o decaimento são as regiões de transição das notas. A detecção por este

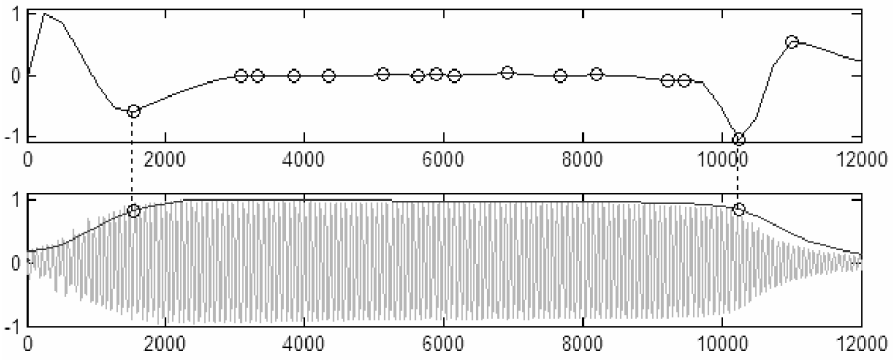

Figura 4.1. Delimitadores de região de transientes detectados pela energia RMS. Gráfico superior contém a derivada segunda de RMS com picos locais marcados com círculos. Gráfico inferior contém o sinal (linha clara), a envoltória RMS (linha escura) e os instantes detectados representados por círculos. Eixo horizontal está em amostras e eixo vertical em energia. (Figura retirada de Loureiro et al. (2008).)

método está ilustrada na Figura 4.1.

4.1.2 Detecção por Fluxo Espectral

O Fluxo Espectral, um dos métodos de detecção de transientes por variações na magnitude do espectro de frequência, é descrito aqui. Esse método é mais utilizado em sinais musicais e, por isto, foi usado neste estudo para detecção de regiões de transição neste tipo de sinal.

O Fluxo Espectral é definido como a correlação da magnitude do espectro de frequência entre quadros consecutivos, e é calculado como

$$F(q) = \frac{1}{M} \sum_{p=1}^{M} |r(X(p)_q, X(p)_{q-1})|, \tag{4.2}$$

onde M é o número de bins do espectro, X é o espectro de frequência de um sinal, r é uma medida de correlação, e $F(q)$ é o fluxo espectral para o quadro q. A curva do Fluxo Espectral é normalmente usada em seu complemento de um $(1 - F(q))$ para facilitar sua visualização.

O valor do Fluxo Espectral tende a aumentar em regiões com pouca variação na evolução temporal do espectro. Portanto, sua variação está geralmente associada à mudança de nota. Porém, como constatado em Loureiro et al. (2008), a estabilização do Fluxo Espectral acontece somente em regiões de sustentação das notas.

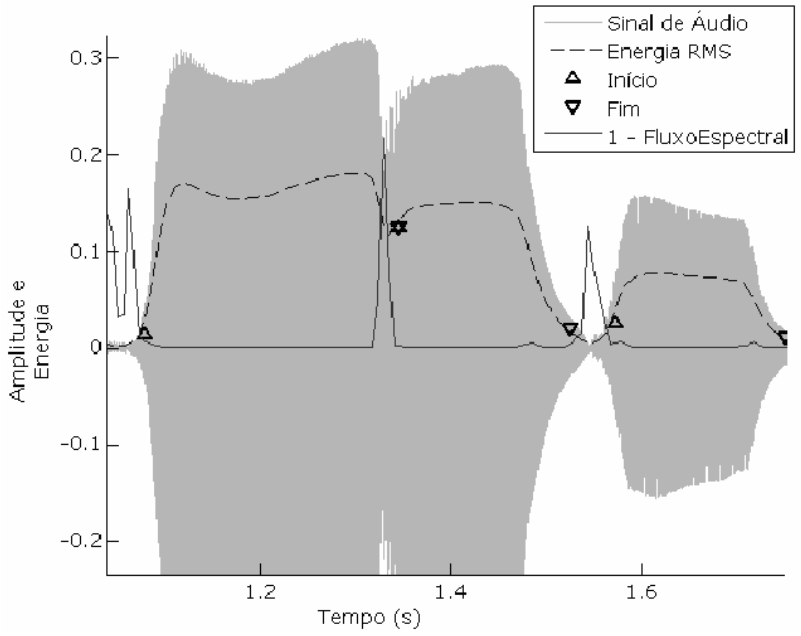

Figura 4.2. Sinal de áudio, complemento de um do Fluxo Espectral, energia RMS, e instantes de início e final de notas de clarineta. (Figura retirada de Loureiro et al. (2008).)

Em regiões de transição, seu valor é menor e instável, como pode ser visto na Figura 4.2.

Isto permite a utilização do Fluxo Espectral para detecção de regiões com presença de transiente. Em Campolina et al. (2009) foi feita uma comparação entre os métodos de detecção por energia e por Fluxo Espectral. Os dois métodos apresentaram detecções próximas na maioria dos casos, com média de 6,4 milissegundos de diferença. Porém, para notas mais longas, houve grandes diferenças (em torno de 500 milissegundos), ocasionadas pelo crescimento mais lento de energia no início das notas. Isto ocasiona uma detecção tardia do método de detecção por energia para notas longas, bem depois da estabilização do espectro. Assim, a detecção de regiões de transição das notas por Fluxo Espectral se mostrou mais adequada.

4.2 Caracterização dos transientes

Na fala, estudos utilizam a caracterização de transientes para diferentes propósitos. Estudos na área de fonoaudiologia (Melo et al., 2012; Bonatto, 2007) diferenciam as oclusivas do português brasileiro por características acústicas dos *bursts*. Duas das caracteristicas importantes são: o intervalo de tempo entre os *bursts* e o início

do vozeamento, e a amplitude dos *bursts*. Em Repp & Lin (1989) foi verificado que características espectrais dos *bursts*, como a frequência central do espectro, ajudam a diferenciar as consoantes.

Na música, vários estudos sobre caracterização das transições entre notas musicais focam na variação da envoltória de energia na região (Maestre & Gómez, 2005). De fato, o formato da envoltória de energia influencia a percepção de duração e ritmo para tons puros (Fastl & Zwicker, 2007). Porém, dada a importância perceptiva dos transientes, a quantidade destes eventos pode influenciar a percepção das transições. Para avaliação quantitativa desta influência, é preciso uma forma de medida. A seguir são descritos índices, propostos neste trabalho, para serem utilizados nestas avaliações.

4.2.1 Índices de Transiência

A separação da componente transiente do sinal possibilita uma medida de comparação de energias. Por esta razão são propostos, neste estudo, três índices: Índice de Transiência Regional (ITR), Índice de Transiência Comparativo (ITC) e Índice de Transiência Global (ITG). O cálculo dos índices depende da determinação das regiões de transição e da separação da componente transiente. Os métodos utilizados para isso, neste estudo, são o Fluxo Espectral e o TMS, respectivamente.

O objetivo do ITR é possibilitar a comparação das energias dos transientes e do sinal apenas na região de transição. Dados dois sinais $s(n)$ e $t(n)$, com $t(n)$ sendo a componente transiente de $s(n)$, o ITR é definido como a razão entre as energias RMS de $t(n)$ e $s(n)$ em uma região de interesse:

$$ITR = \sqrt{\sum_{n=1}^{R} \frac{t(n)^2}{s(n)^2}}, \tag{4.3}$$

em que R é o tamanho da região de interesse onde o índice é calculado.

O ITC é definido visando à comparação entre a energia dos transientes na região de transição e a energia dos transientes em toda a nota. Dados os mesmos $s(n)$ e $t(n)$ da definição de ITR, o ITC é definido como a razão entre a energia RMS de $t(n)$ em uma região de interesse, e a energia RMS de $t(n)$ na região de interesse somada à região complementar a todo o sinal:

$$ITC = \frac{\sqrt{\frac{1}{R}\sum_{n=1}^{R} t(n)^2}}{\sqrt{\frac{1}{R}\sum_{n=1}^{R} t(n)^2} + \sqrt{\frac{1}{N-R}\sum_{n=R+1}^{N-R} t(n)^2}}, \tag{4.4}$$

em que R é o tamanho da região onde o índice é calculado, e N é o tamanho do sinal.

Por último, a definição do ITG objetiva medir a proporção de transientes em toda a extensão do sinal. Tomando mais uma vez os sinais $s(n)$ e $t(n)$, o ITG é definido como a razão entre as energias RMS de $t(n)$ e $s(n)$, em todo o sinal:

$$ITG = \sqrt{\sum_{n=1}^{N} \frac{t(n)^2}{s(n)^2}}. \tag{4.5}$$

Pensando em regiões de interesse como região transição de notas, a combinação dos valores de ITR, ITC, ITG, fornece uma caracterização da quantidade e distribuição de transientes em uma nota musical. Um baixo valor de ITG significa pouca energia de transientes em toda a nota. Quanto mais alto o valor de ITC, maior a concentração dos transientes na região de transição. Em contrapartida, quanto menor o valor de ITC, mais os transientes estão distribuídos dentro da nota. ITR indica a quantidade de energia dos transientes na região transição. A aplicação destes índices faz sentido para qualquer sinal de áudio. Porém, neste trabalho eles são avaliados para notas musicais.

Este capítulo foi dedicado à discussão sobre detecção de regiões com presença de componente transiente em sinais de fala e de música. Dois métodos mais utilizados em sinais musicais foram detalhados. Outros métodos, tanto para sinais de fala quanto para sinais de áudio em geral, foram mencionados. Formas de caracterização dos transientes foram discutidas. Uma delas, proposta neste trabalho, foi a definição de índices para medir a proporção da componente transiente e sua distribuição no sinal. A seguir, são apresentados os resultados dos experimentos e testes realizados neste estudo.

Capítulo 5

Resultados

Neste capítulo, são apresentados os resultados dos testes e experimentos para avaliar os seguintes tópicos: o desempenho do TMS ao modelar sinais com características transientes, comparando-o com a modelagem SMS; a importância de *bursts* para o reconhecimento de consoantes oclusivas; o desempenho do TMS para modelar *bursts*; o desempenho do TMS para separar transientes; e a aplicação dos Índices de Transiência. O capítulo é dividido em duas seções. Na primeira seção, são descritas as abordagens do TMS utilizadas. Na segunda seção, os testes e experimentos são explicados e seus resultados demonstrados.

5.1 Descrição da modelagem TMS

O TMS é utilizado em duas abordagens diferentes nos experimentos e testes. A primeira modela diretamente os sinais de natureza predominantemente transiente, sem a presença significativa de componente determinística ou estocástica. Esta abordagem está representada no diagrama da Figura 5.1. A segunda é utilizada em sinais com presença de outras componentes. Neste caso, a parte determinística do sinal deve ser separada usando alguma modelagem por síntese adequada. O próprio SMS é utilizado nos testes para a separação da parte determinística. O TMS é então aplicado ao resíduo da modelagem SMS. No diagrama da Figura 5.2 está representada a segunda abordagem.

O SMS já está bem desenvolvido na literatura, não sendo o foco deste trabalho. Para o SMS, este estudo utiliza a implementação feita por J. Bonada, X. Serra, X. Amatriain e A. Loscos, disponível em Udo et al. (2011). O código do SMS é modificado para permitir a especificação de um número fixo de senoides com maior intensidade,

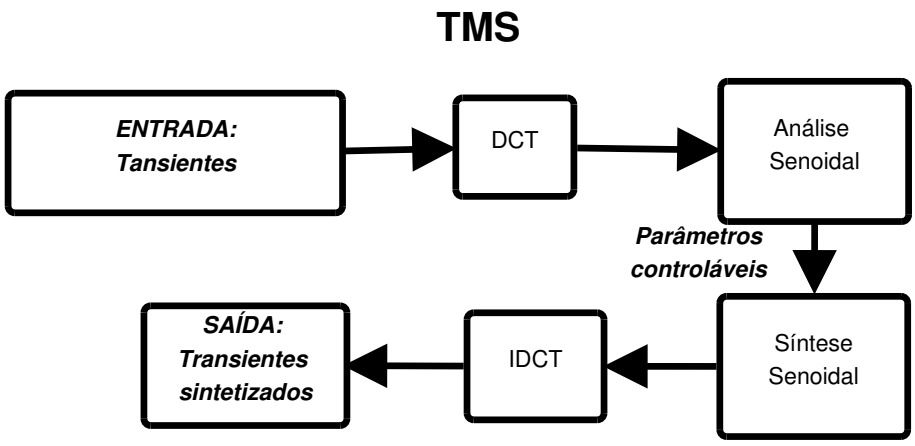

Figura 5.1. Diagrama de blocos do TMS para o caso de modelagem de transientes isolados.

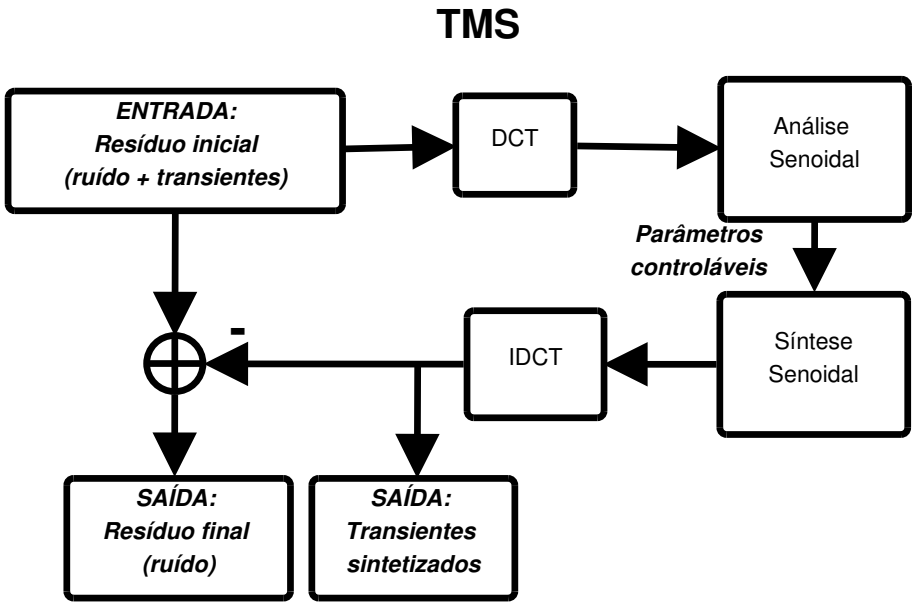

Figura 5.2. Diagrama de blocos do TMS para o caso de modelagem de transientes somados a componentes estocásticas (ruídos).

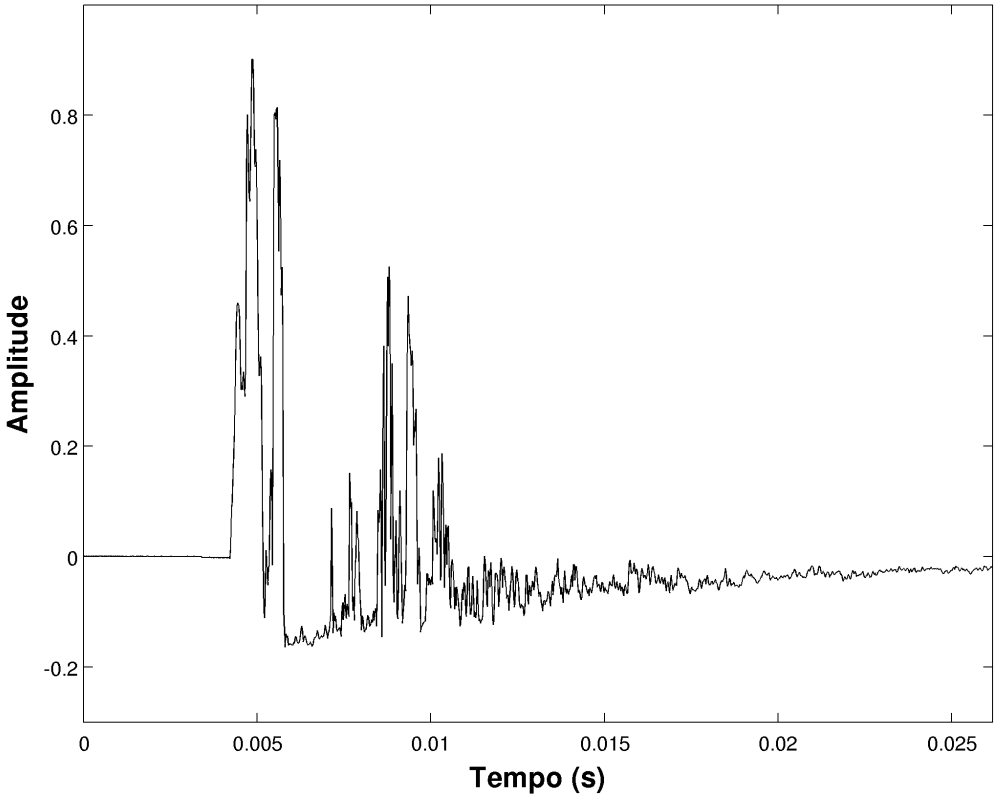

Figura 5.3. Forma de onda da gravação de um estouro de balão em câmara anecoica.

por quadro da modelagem. No código original é passado um limiar de intensidade em decibels, havendo a possibilidade de variação do número de senoides para cada quadro da modelagem. É utilizada a *Fast Fourier Transform* (FFT) para transformação em frequência. As implementações de procedimentos e processamentos de dados foram feitos em MATLAB.

5.2 Avaliação do TMS

Para a avaliação inicial do TMS, é utilizada uma gravação de estouro de balão em câmara anecoica do *Electronic Music Studios*, da universidade de Iowa. A taxa de amostragem é de 44100 Hz. Este caso se enquadra na primeira abordagem do TMS. A forma de onda da gravação está mostrada na Figura 5.3 e, como pode ser observado, apresenta características impulsivas. A Figura 5.4 contém a DCT deste sinal.

A avaliação do desempenho do TMS é feita através da comparação da forma

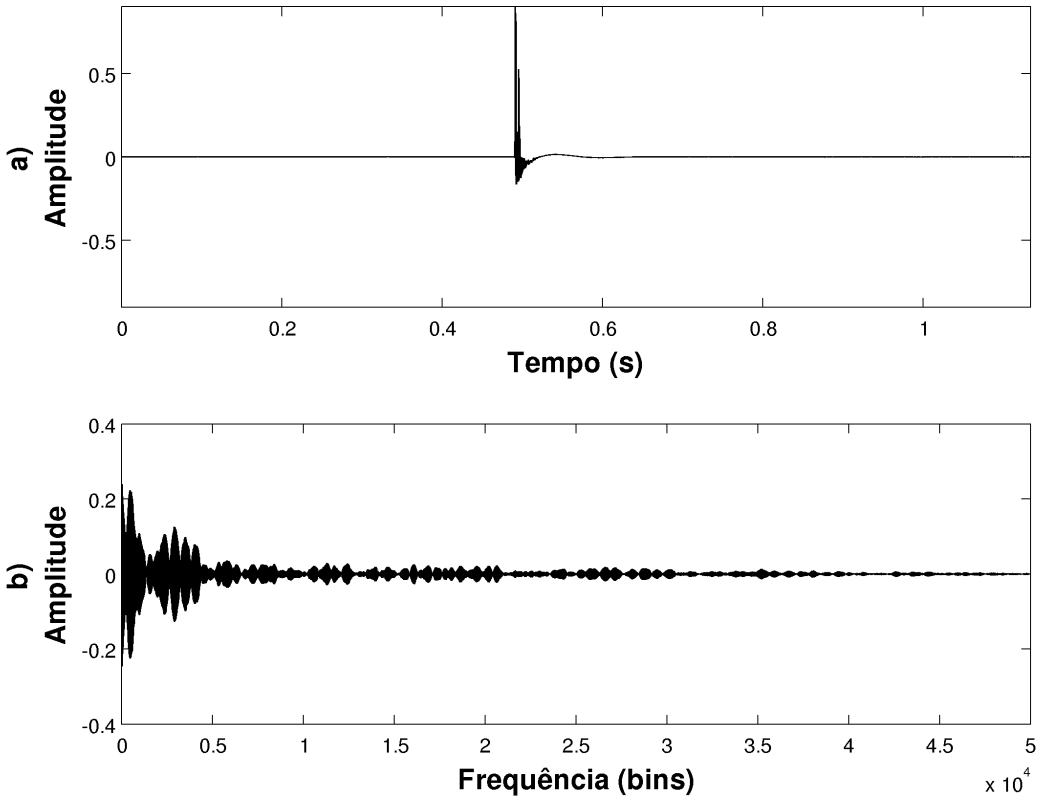

Figura 5.4. (a) Forma de onda da gravação de um estouro de balão e (b) sua DCT.

de onda do sinal original com o sinal sintetizado, variando-se o número de senoides utilizadas na modelagem. A medida de comparação utilizada foi o coeficiente de correlação de Pearson. O mesmo é feito com o SMS. Nos dois casos, o número de senoides utilizadas varia de 1 a 20. As curvas resultantes estão na Figura 5.5.

Com uma senoide por quadro, o TMS (curva contínua da Figura 5.5) apresenta um valor de coeficiente de correlação igual a 0.87, enquanto o SMS (curva tracejada da Figura 5.5) apresenta o valor de 0.51. Com 4 senoides por quadro, a curva do TMS se estabiliza em um valor de coeficiente de correlação igual a 0,89, enquanto que, neste número de senoides para o SMS, o valor do coeficiente de correlação foi de 0.70. Para 20 senoides por quadro, o SMS apresenta um valor igual a 0.82, contra 0.89 para o TMS.

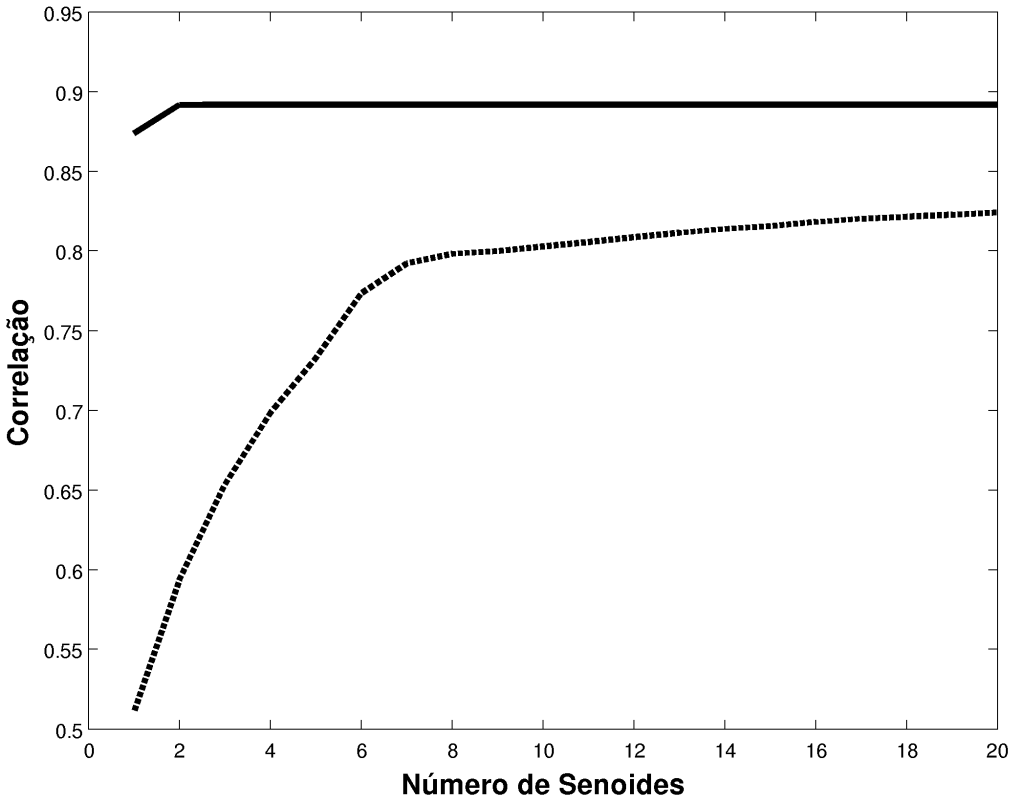

Figura 5.5. Coeficiente de correlação entre sinal original e sinal modelado, avaliado de uma até 20 senoides por quadro, para SMS (linha tracejada) e TMS (linha contínua).

5.3 Experimentos com fala

5.3.1 Experimento com consoantes oclusivas

Um experimento é proposto para a avaliação da importância dos *bursts* para percepção. O experimento mede a influência da ausência e da modelagem adequada dos *bursts* no reconhecimento. É também feita uma avaliação de qualidade das modelagens por SMS e TMS.

A escolha das palavras é feita de forma que as seis consoantes oclusivas [p], [t], [k] (não vozeadas), e [b], [d] e [g] (vozeadas), apareçam em quantidade e forma semelhante. Com este intuito, as palavras utilizadas são: **pago**, **tado**, **cabo**, **baco**, **dato**, e **gapo**. Cada oclusiva ocupa o primeiro fonema da primeira e segunda sílabas, de pelo menos uma das palavras. É usada a frase portadora **Escute ... agora.** tanto para a gravação das palavras quanto para a apresentação aos ouvintes no teste de escuta. A base de

dados do experimento é composta de gravações de quatro locutores adultos diferentes com idade entre 20 e 30 anos (dois homens e duas mulheres) e de 11 avaliações de adultos de idades entre 20 e 63 anos, de ambos os sexos.

5.3.2 Gravação e preparação das amostras

As gravações foram realizadas utilizando um microfone de condensação Brüel & Kjær de 1/2 polegada, posicionado de frente para o locutor, a 50 centímetros de distância dos lábios. Foi utilizado o conversor A/D Creative da Sound Blaster, 24 bits, taxa de amostragem de 44100 Hz. O ambiente utilizado foi a sala de gravação do CEFALA (Centro de Estudos da Fala, Acústica, Linguagem e Música) da Escola de Engenharia da Universidade Federal de Minas Gerais, que apresenta um isolamento de aproximadamente 30 decibels e tratamento acústico.

A escuta dos testes foi feita utilizando um fone de ouvido AKG 414p, com as amostras normalizadas para 0,9 de amplitude. São seis frases: **Escute pago agora**, **Escute tado agora**, **Escute cabo agora**, **Escute baco agora**, **Escute dato agora**, **Escute gapo agora**, de cada locutor.

As palavras e *bursts* foram segmentadas manualmente utilizando o aplicativo Praat versão 5.1.25-1, com auxílio de algumas de suas ferramentas como detecção de região periódica, e detecção de pontos mais próximos a cruzamento por zero. A Figura 5.6 apresenta a marcação da região de *burst* da oclusiva [t] de **tado**.

Depois de segmentados, os *bursts* são modelados por SMS e TMS, ambos utilizando 10 senoides para cada quadro. A escolha do número de senoides utilizadas é baseada no teste da seção anterior. Na Figura 5.5, a diferença entre as duas curvas começa a se estabilizar em torno de 10 senoides.

As modelagens para cada locutor são feitas da seguinte forma: As palavras contendo as oclusivas são colocadas em sequência. Com exceção dos *bursts*, todo o sinal é zerado, sobrando apenas uma sequência de *bursts*. Em seguida, este sinal é submetido aos métodos de modelagem SMS e TMS. A abordagem do TMS utilizada neste caso é a primeira, ilustrada no diagrama da Figura 5.1. Como o sinal a ser modelado é curto (em torno de três segundos), a DCT no TMS é aplicada a todo o sinal, não havendo necessidade de dividi-lo em blocos menores. Em ambos, SMS e TMS, são utilizadas: janelas de 1024 amostras (23,2 ms); deslocamento de 256

Tabela 5.1. mínimos e máximos de duração dos *bursts* para os quatro locutores.

Locutor	*Mínimo (ms)*	*Máximo (ms)*
1	8,7	55,2
2	9,8	51,2
3	8,4	72,9
4	10,7	70,9

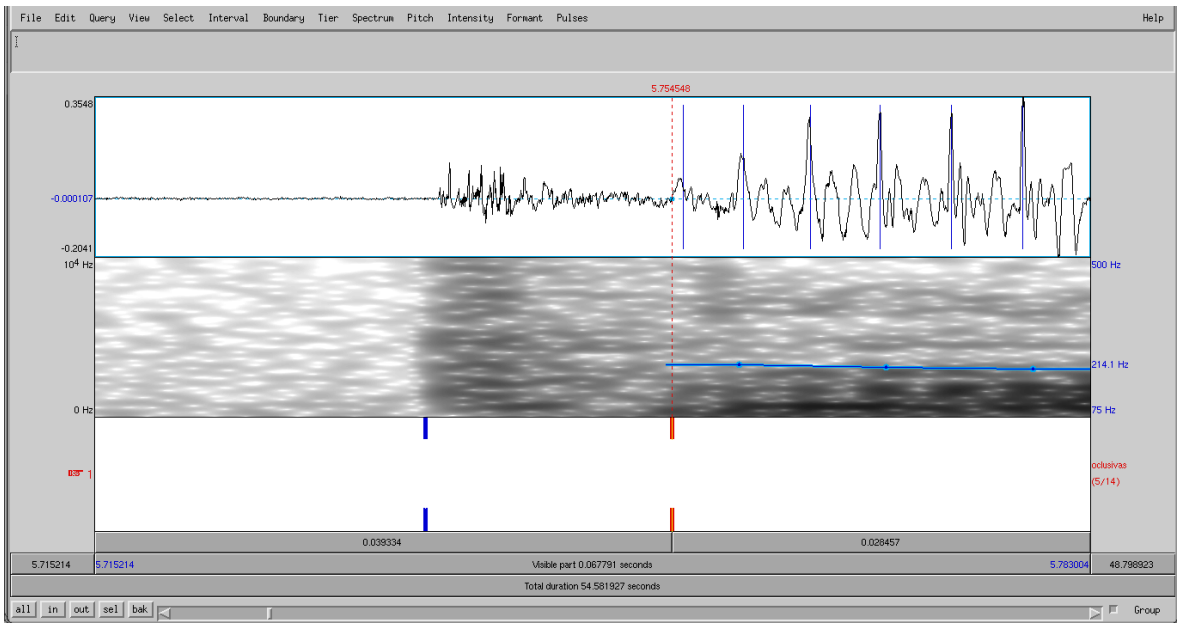

Figura 5.6. Exemplo de segmentação manual dos *bursts* utilizando o Praat: *bursts* da oclusiva [t] de **tado**.

amostras (5,8 ms); janelas de *Blackman-Harris*, recomendadas pelos autores do código do SMS; resolução da FFT de 4096 pontos. Feita a modelagem, os *bursts* são reinseridos de acordo com as situações do experimento. Para a situação S5, o mesmo procedimento é feito, porém apenas os *bursts* são zerados.

Os mínimos e máximos de duração dos *bursts* para os quatro locutores estão na Tabela 5.1. A média entre todos os *bursts* é 23,6 ms.

5.3.3 Reconhecimento de oclusivas

Nos testes de reconhecimento de consoantes oclusivas, seis situações diferentes são apresentadas aos sujeitos. Em cada situação são executadas seis frases portadoras como descritas no item **Experimento com consoantes oclusivas**. Cada frase contém duas consoantes oclusivas a serem identificadas pelos ouvintes. Ao todo, são

Tabela 5.2. Situações de reconhecimento de oclusivas: modificações feitas nas palavras sujeitas a reconhecimento em cada situação. A frase portadora se mantém idêntica nas seis situações.

Situação	*Descrição*
S1	*Bursts* retirados manualmente.
S2	*Bursts* trocados.
S3	*Bursts* modelados por SMS.
S4	*Bursts* modelados por TMS.
S5	*Bursts* reais e restante modelado por SMS.
S6	Sinal original sem alterações.

12 consoantes oclusivas reconhecidas para cada uma das seis situações. O objetivo é comparar as médias de reconhecimento nas diferentes situações descritas, validados com testes estatísticos de comparação entre médias de amostras.

A frase portadora, para cada locutor, é idêntica em todas as situações. As palavras contendo as oclusivas a serem reconhecidas foram permutadas aleatoriamente, com a intenção de eliminar fatores indesejados de memorização. A Tabela 5.2 contém as modificações feitas nas palavras sujeitas a reconhecimento.

Apenas na situação S5 houve modificação em regiões das palavras que não são *bursts*. Nesta situação, toda a palavra foi modelada por SMS com 10 senoides por quadro, com exceção dos *bursts* que foram mantidos inalterados.

Em cada situação, são 11 sujeitos reconhecendo 12 consoantes oclusivas para cada um dos 4 locutores. Nos testes estatísticos, são utilizadas médias de reconhecimento por ouvinte, resultando em 11 médias por situação. São, portanto, seis amostras (S1 a S6) de 11 valores cada.

Para a realização de testes comparativos de diferença de médias de acertos entre as amostras das situações é necessário validar as premissas de normalidade e independência das amostras. A independência entre as amostras é reforçada com a blocagem por locutores e por ouvintes, e fatores espúrios são reduzidos com a aleatorização de apresentação das oclusivas e dos locutores aos ouvintes. Para avaliar a normalidade, é utilizado o teste de Lillieford a 5% de significância, que testa a Hipótese nula (H0) de os dados serem originados de distribuição normal, contra a hipótese alternativa (H1) de não serem originados de distribuição normal. Os

resultados do teste estão mostrados na Tabela 5.3.

Apenas para S6 H0 foi rejeitada com evidências fortes de não normalidade, apresentando valor p 50 vezes menor que a significância. Porém, como pode ser visto na Figura 5.7, as amostras de S6 são altamente concentradas em torno da média de 11,77. A alta concentração de acertos perto do máximo 12 é esperada para esta situação, visto que em S6, o sinal original é apresentado. Por esta razão, nas comparações feitas com S6, é utilizado o valor da média ao invés da amostra, configurando um teste de diferença de média simples.

A amostra S4 também obteve H0 rejeitada, porém com evidências fracas de não normalidade, apresentando alto valor p (aproximadamente metade da significância). Por apresentar alto valor p, possuir tamanho pequeno, e por inspeção visual no gráfico de probabilidade normal da Figura 5.7, a amostra S4 foi considerada oriunda de distribuição normal. Os gráficos de probabilidade normal para as outras amostras podem ser vistos também na Figura 5.7. Gráficos de probabilidade normal permitem uma comparação entre os dados e a distribuição normal que melhor se ajusta ao caso. A distribuição normal ajustada é representada por uma reta no gráfico. Quanto mais próximos da reta, mais os dados podem ser considerados oriundos de distribuição normal.

Validadas as premissas, as médias podem então ser comparadas. Na Tabela 5.4, estão mostrados resultados da aplicação do teste t para comparações das médias entre as amostras das seis situações de reconhecimento. E, na Tabela 5.5, são apresentadas as médias e percentuais de cada amostra.

5.3.4 Teste MOS

O *Mean Opinion Score* (MOS) é um método subjetivo de teste de qualidade. A qualidade do sinal é avaliada por pessoas utilizando uma pontuação que varia de 1 a 5, sendo 1 inaceitável e 5 excelente. A média de pontuação é calculada para a avaliação final. Para pontuações acima de 4 o sinal avaliado é considerado de alta qualidade. Para o caso deste experimento, é feita a comparação da diferença de qualidade entre duas sequências de palavras. O significado das pontuações estão descritos na Tabela 5.7. São três situações diferentes avaliadas, descritas na Tabela 5.6. Cada sequência é formada por seis palavras, as mesmas utilizadas para o reconhecimento descrito na seção *Reconhecimento de oclusivas*. As palavras das sequências são dispostas de maneira

Tabela 5.3. Teste de normalidade de Lillieford para os dados de reconhecimento do experimento com oclusivas: Hipótese nula (H0) dos dados serem originados de distribuição normal, contra a hipótese alternativa (H1) de não serem originados de distribuição normal, ao nível de significância de 5%.

Situação	*Situação da H0*	*Valor p*	*Significado*
S1	Não rejeitada	$0,24$	Sem evidência de não normalidade
S2	Não rejeitada	$0,15$	Sem evidência de não normalidade
S3	Não rejeitada	$0,50$	Sem evidência de não normalidade
S4	Rejeitada	$0,025$	Evidência fraca de não normalidade (alto valor p)
S5	Não rejeitada	$0,10$	Sem evidência de não normalidade
S6	Rejeitada	$0,001$	Evidência forte de não normalidade (baixo valor p)

aleatória e a escuta realizada nas mesmas condições do reconhecimento.

Na situação S7, o sinal original é comparado com ele mesmo para fins de normalização da pontuação, de maneira a eliminar o efeito de fatores espúrios no procedimento de escuta dos testes.

A Figura 5.8 contém um histograma com as pontuações da escala MOS obtidas no teste subjetivo. Foram as mesmas 11 pessoas do teste de reconhecimento. Desta forma, cada situação, S7, S8, e S9, apresenta 44 avaliações: 11 sujeitos avaliando 4 locutores. Na Tabela 5.8, estão mostradas as médias de pontuação para as situações.

5.4 Experimentos com música

5.4.1 Testes do Índices de Transiência

O teste dos Índices de Transiência é realizado através da avaliação de seus valores obtidos para diferentes instrumentos musicais. As amostras utilizadas são gravações em câmara anecoica do *Electronic Music Studios* da Universidade de Iowa. Uma nota, o Dó5 (523 Hz), de cada instrumento é utilizada. Os instrumentos são: cello, clarineta,

Tabela 5.4. Testes comparativos entre médias de acertos de oclusivas em diferentes situações: Hipótese nula (H0) de que as médias são iguais, contra Hipótese alternativa (H1) de que as médias são diferentes.

Situações	*Situação da H0*	*Valor p*	*Significado*
S1 e S6	Rejeitada	$< 0,001$	Evidência forte de diferença entre médias (baixo valor p)
S2 e S6	Rejeitada	$< 0,001$	Evidência forte de diferença entre médias (baixo valor p)
S3 e S6	Rejeitada	$0,002$	Evidência forte de diferença entre médias (baixo valor p)
S4 e S6	Não rejeitada	$0,062$	Sem evidência de diferença entre médias
S5 e S6	Não rejeitada	$0,075$	Sem evidência de diferença entre médias
S3 e S4	Rejeitada	$0,002$	Evidência forte de diferença entre médias (baixo valor p)
S4 e S5	Não rejeitada	$0,56$	Sem evidência de diferença entre médias
S1 e S2	Rejeitada	$< 0,001$	Evidência forte de diferença entre médias (baixo valor p)
S1 e S4	Rejeitada	$< 0,001$	Evidência forte de diferença entre médias (baixo valor p)

Tabela 5.5. Médias de reconhecimento das oclusivas.

Situação	*Média de acertos (Máximo 12)*	*Diferença percentual (Em relação a S6)*
S1	9,43 (78,60%)	19,88%
S2	7,30 (60,80%)	38,03%
S3	10,41 (86,74%)	11,58%
S4	11,41 (95,08%)	3,09%
S5	11,48 (95,64%)	2,51%
S6	11,77 (98,11%)	0%

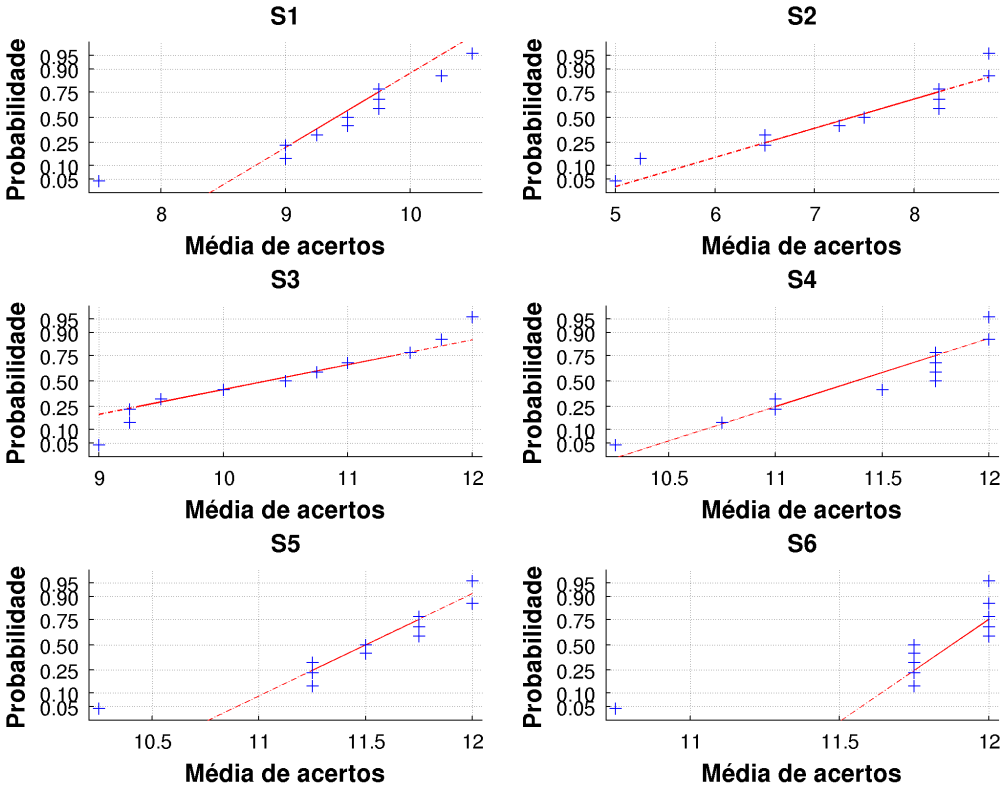

Figura 5.7. Curvas de probabilidade normal para as seis situações de reconhecimento S1 a S6, descritas na Tabela 5.2.

Tabela 5.6. Situações de teste MOS avaliadas de acordo com a escala da Tabela 5.7.

Situação	*Primeira sequência*	*Segunda sequência*
S7	Sinal original	Sinal original
S8	Sinal original	*Bursts* modelados por SMS
S9	Sinal original	*Bursts* modelados por TMS

Tabela 5.7. Escala MOS utilizada do experimento.

Pontuação	*Significado*
1	**Inaceitável:** Não é possível reconhecer a informação contida na segunda sequência.
2	**Ruim:** Há grande distorção na segunda sequência, mas é mantida a inteligibilidade.
3	**Razoável:** Foi percebida diferença entre as sequências, e a primeira é melhor.
4	**Boa:** Há diferença entre as sequências, mas não é possível julgar qual é melhor.
5	**Excelente:** Não foi percebida diferença entre as sequências.

Tabela 5.8. Pontuação MOS para oclusivas. As porcentagens estão em parêntesis.

Situação	*Pontuação média*
S7	4,55 (90,9%)
S8	3,32 (66,4%)
S9	4,16 (83,2%)

oboé, trompete, violino (pizzicato) e flauta.

São utilizados instrumentos de naturezas diferentes para possibilitar a associação dos valores dos índices com características do instrumento, maneiras de execução e forma de excitação. A taxa de amostragem é de 44100 Hz. Os inícios e finais das notas foram detectados manualmente. As gravações são de notas isoladas, e por isso, não estão presentes no sinal perturbações geradas por mudança de digitação no instrumento ou por influência de notas vizinhas.

Neste caso, a modelagem dos transientes é feita depois da separação da componente estacionária. É usada a segunda abordagem do TMS descrita na Figura 5.2. A separação da componente estacionária é feita utilizando o próprio SMS com 20 senoides por quadro na modelagem. Feito isso, o TMS recebe o resíduo da modelagem SMS para modelar os transientes. No TMS são usadas 10 senoides. O resultado é um sinal do mesmo tamanho do sinal original, contendo apenas a parte transiente modelada. Esse processo é realizado em cada nota.

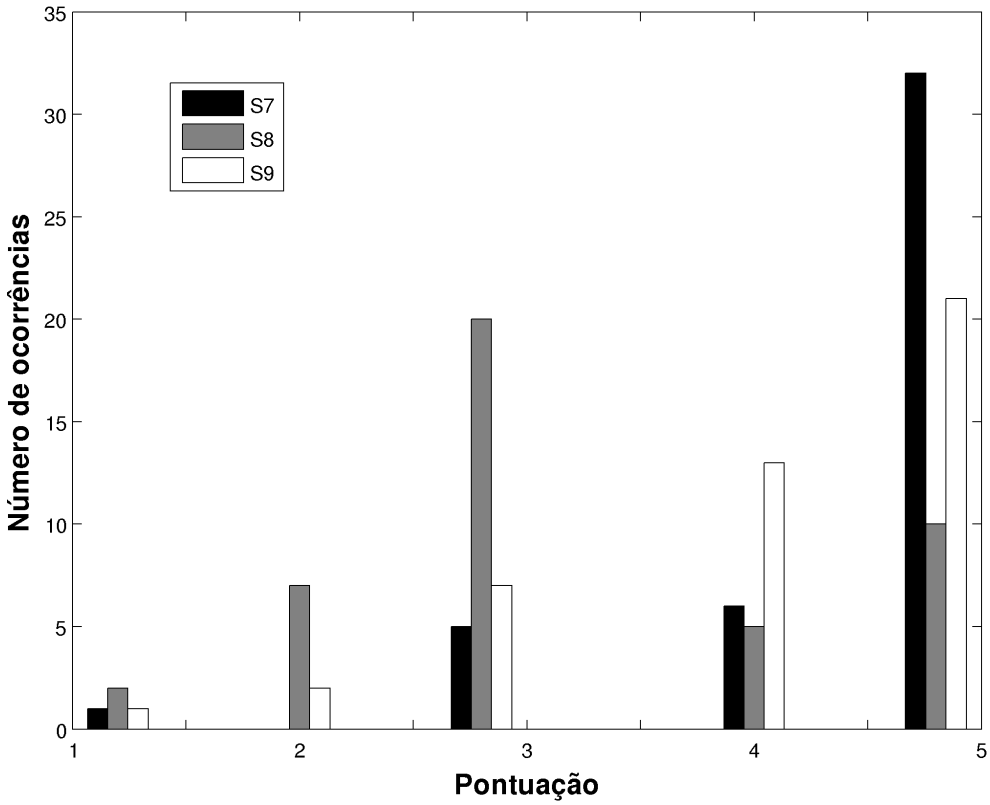

Figura 5.8. Histograma de pontuações MOS: barras pretas, cinzas e brancas representam as situações S7, S8, e S9, respectivamente.

5.4.2 Detecção de regiões de transição

A detecção de regiões de transição é feita utilizando a curva de Fluxo Espectral. É usada a implementação desenvolvida pelo Centro de Estudos do Gesto Musical e Expressão (CEGeME) da Escola de Música da UFMG, descrita em Campolina et al. (2009). O tamanho da janela é de 1024 amostras (23,2 ms), com deslocamento de 256 amostras (5,8 ms). A medida de correlação utilizada no Fluxo Espectral é o coeficiente de correlação de Pearson.

A curva do Fluxo Espectral é comparada com um limiar. O limiar é a média da curva em toda a nota. Regiões maiores que a média são considerados regiões de transição. Na Figura 5.9 estão mostradas as regiões detectadas (curva mais fina) e as curvas do complemento de um do Fluxo Espectral (curva mais espessa), para todos os instrumentos.

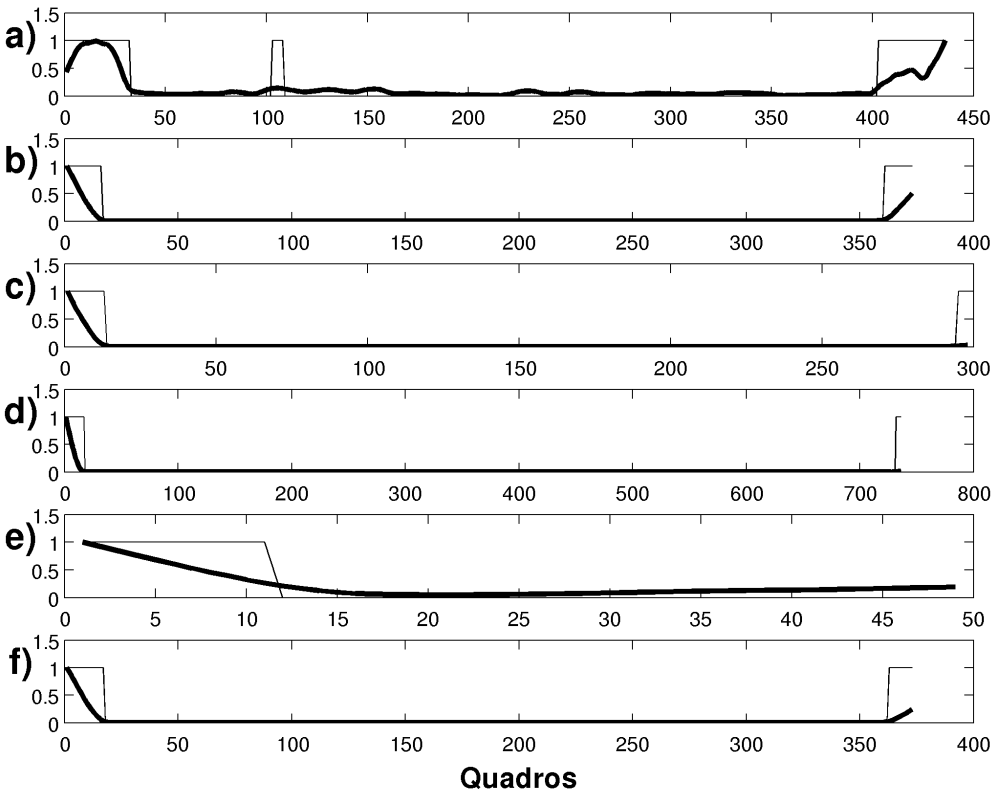

Figura 5.9. Regiões de transição detectadas por fluxo espectral: Complemento de um do Fluxo Espectral mostrado nas linhas espessas e, regiões de transição detectadas representadas por nível alto das linhas finas. As letras dos gráficos identificam os instrumentos. (a) cello, (b) clarineta, (c) oboé, (d) trompete, (e) pizzicato de violino, (f) flauta.

5.4.3 Valores dos índices

Nas figuras 5.10, 5.11 e 5.12 estão mostrados os valores dos Índices de Transiência ITR, ITC e ITG para os instrumentos. Os valores percentuais dos índices estão na Tabela 5.9.

A Figura 5.13 contém as formas de onda do pizzicato de violino e da clarineta, com suas respectivas componentes transientes modeladas por TMS. O pizzicato obteve o maior valor para ITR, enquanto a clarineta obteve o menor.

Na curva superior da Figura 5.14 são mostradas as formas de onda de todos os instrumentos concatenadas. Os resíduos da separação da componente determinística estão mostrados na curva inferior.

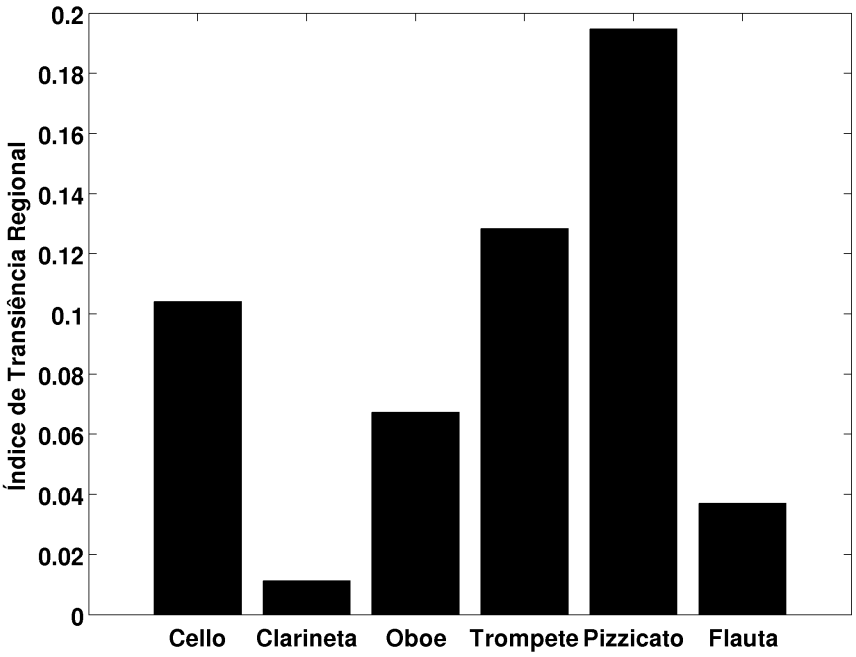

Figura 5.10. Índice de Transiência Regional (ITR) para os instrumentos.

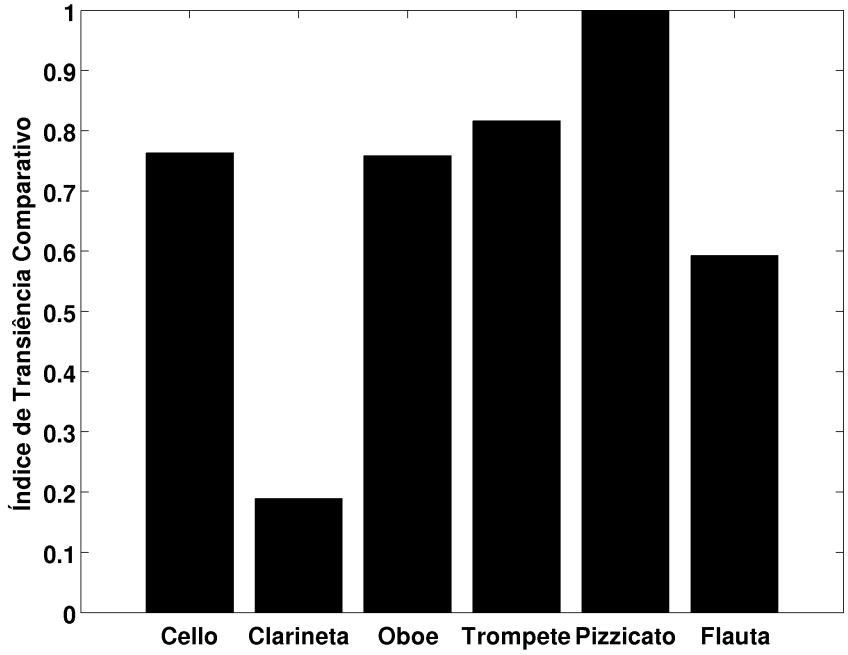

Figura 5.11. Índice de Transiência Comparativo (ITC) para os instrumentos.

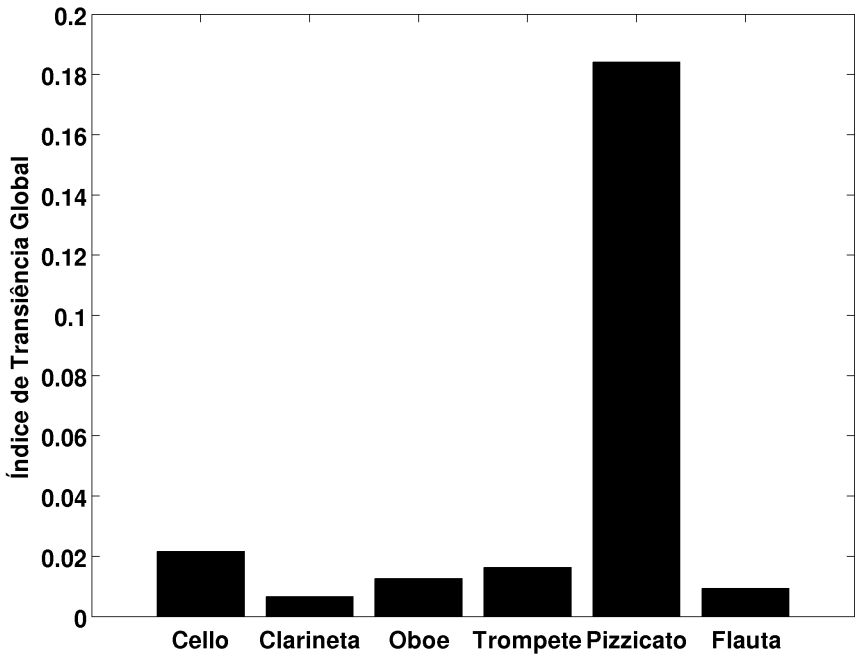

Figura 5.12. Índice de Transiência Global (ITG) para os instrumentos.

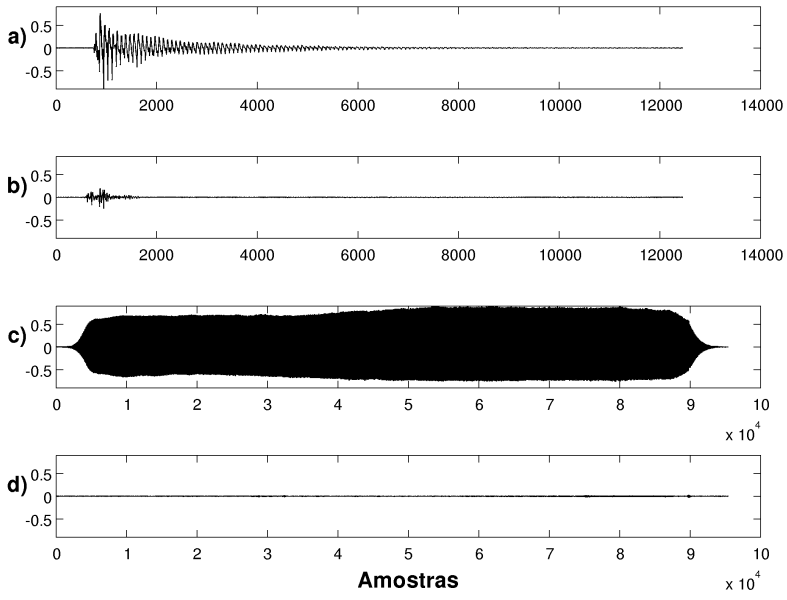

Figura 5.13. Sinais dos instrumentos com maior e o menor ITR: (a) forma de onda do pizzicato de violino, (c) clarineta, (b) e (d) suas componentes transientes modeladas por TMS. O eixo vertical dos gráficos representa a intensidade dos sinais.

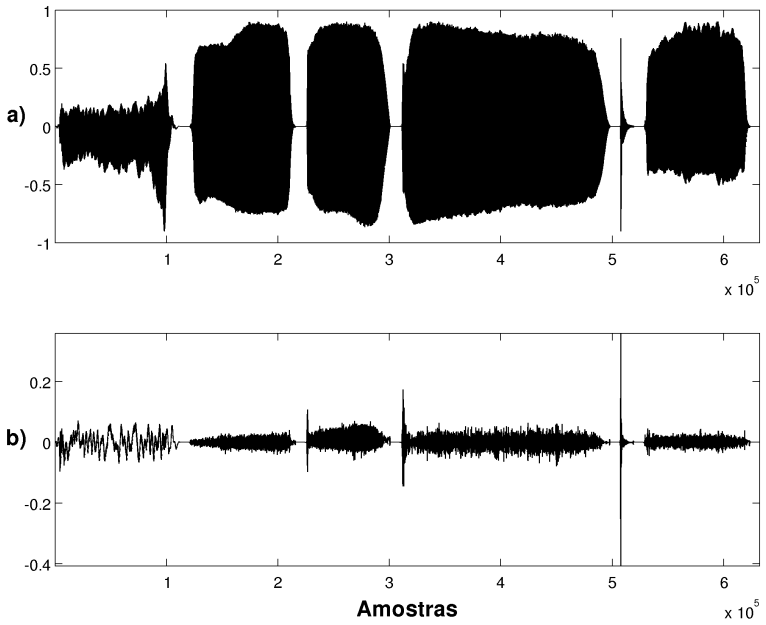

Figura 5.14. Sinais e resíduos dos instrumentos. (a) forma de onda de todos os instrumentos concatenados na sequência: cello, clarineta, oboé, trompete, pizzicato de violino, e flauta. (b) resíduos da separação da componente determinística. O eixo vertical dos gráficos representa a intensidade dos sinais.

Tabela 5.9. Valores percentuais dos Índices de Transiência: ITR, ITC, ITG

Instrumento	*ITR (%)*	*ITC (%)*	*ITG (%)*
Cello	10,41	76,30	2,16
Clarineta	1,13	18,89	0,66
Oboé	6,73	75,82	1,26
Trompete	12,84	81,64	1,64
Pizzicato de violino	19,47	99,92	18,41
Flauta	3,70	59,29	0,93

Na curva superior da Figura 5.15 podem ser vistos os resíduos da separação da componente determinística dos instrumentos concatenados. A componente transiente está mostrada na curva do meio da mesma Figura e os ruídos finais na curva inferior.

Os resultados dos experimentos e testes foram apresentados neste capítulo. A seguir, os resultados são discutidos e, posteriormente, conclusões finais e considerações sobre trabalhos futuros são feitas.

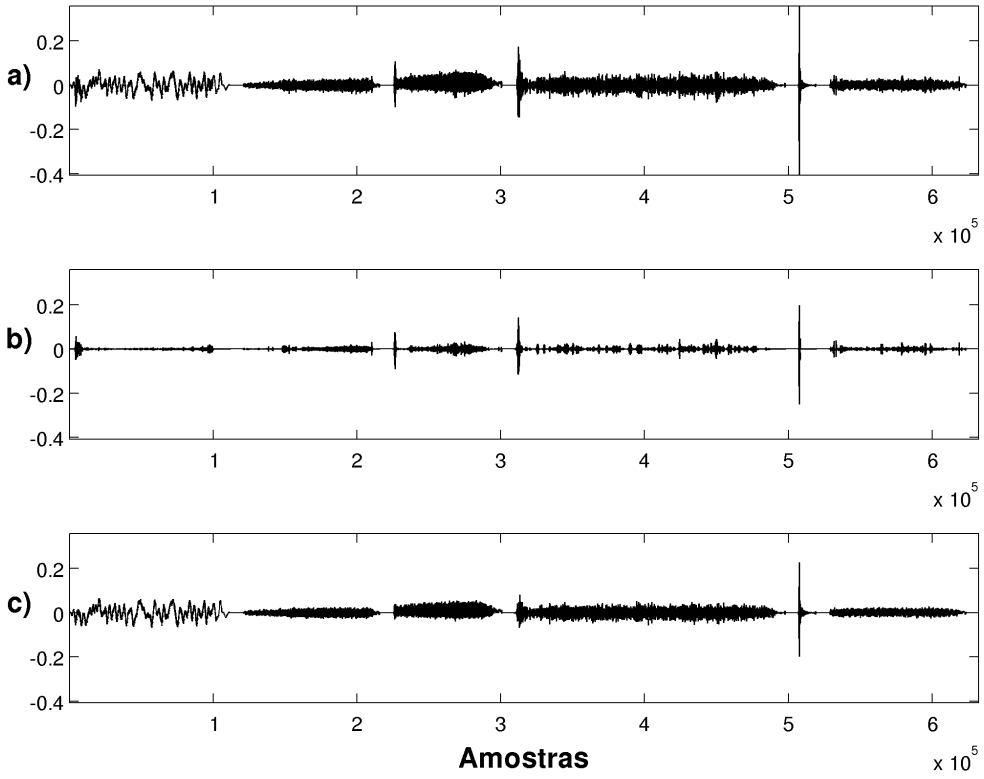

Figura 5.15. Resíduo, transientes e ruído final dos instrumentos. (a) resíduos da separação da componente determinística de todos os instrumentos concatenados na sequência: cello, clarineta, oboé, trompete, pizzicato de violino, e flauta. (b) componentes transientes. (c) ruídos finais. O eixo vertical dos gráficos representa a intensidade dos sinais.

Capítulo 6

Discussão dos resultados

No primeiro teste, a DCT do sinal de estouro de balão apresenta variações mais lentas do que o sinal original, e por isso é mais adequada para a modelagem senoidal (figuras 5.3 e 5.4). Na Figura 5.5 pode ser visto que a modelagem TMS alcançou um valor alto de correlação (0,87) com apenas uma senoide por quadro. O SMS para uma senoide por quadro obteve uma correlação de apenas 0,51, aproximadamente 36% a menos. Com 4 senoides por quadro, a curva do TMS se estabiliza em um valor de coeficiente de correlação igual a 0,89, enquanto que com este número de senoides para o SMS, o valor do coeficiente de correlação foi de 0,70, aproximadamente 20% a menos. A curva de coeficiente de correlação para o SMS não se estabiliza antes de 20 senoides e apresenta, neste caso, um valor igual a 0,82, contra 0,89 do TMS, 6,8% a menos. A curva de correlação para o SMS parece crescer assintoticamente, o que faz sentido, visto que qualquer sinal pode ser representado por uma soma de infinitas senoides. Porém, a assintota para ambos os modelos não deve chegar a unidade, devido a erros na modelagem e na síntese.

Focando agora no teste de reconhecimento de oclusivas, a modelagem dos *bursts* por TMS (S4) não obteve média de acerto significativamente diferente do sinal original (S6). Comparado à média de acerto de 98,1% obtida das elocuções originais, o TMS (S4) atingiu 95,1%, sendo significativamente maior que os 86,7% obtidos com o SMS (S3). Isto mostra a importância de uma modelagem adequada para os *bursts*. A média de acertos cai para 78,6% com a remoção dos *bursts* (S1), mostrando a importância da presença destes eventos para o reconhecimento de oclusivas, mesmo possuindo duração média de apenas 23,6 ms. Com os *bursts* trocados (S2), a média de acertos cai para apenas 60,8%, o que corrobora a existência de informação nos *bursts* que os diferencia e influi diretamente na inteligibilidade.

As situações S3 e S5 são situações opostas. Ora apenas os *bursts* (S3), ora apenas o restante das palavras (S5) são modelados utilizando SMS com 10 senoides por quadro. S3 obteve 9,1% a menos de diferença percentual de acertos em relação a S6. S5 não obteve mádia de acerto significativamente diferentes do sinal original (S6). Isto corrobora a importância de se modelar os transientes de forma diferenciada.

No teste MOS, comparado à pontuação de 4,55 obtida pelo sinal original (S7), a modelagem dos *bursts* por TMS (S9) obteve 4,16, pontuação maior que os 3,32. Pontuações acima de 4 são consideradas de qualidade. Portanto, a modelagem dos *bursts* por TMS com 10 senoides por quadro resultou em uma modelagem de qualidade, enquanto que por SMS resultou em uma modelagem com distorções significativas.

Passando agora à análise dos resultados obtidos para notas musicais, nos testes dos Índices de Transiência, a maioria das regiões detectadas pelo Fluxo Espectral coincide com os inícios e finais das notas, como pode ser visto na Figura 5.9. Sendo assim consideradas, neste trabalho, como regiões de transição.

O pizzicato de violino obteve a maior quantidade de energia dos transientes proporcional ao sinal original, tanto dentro da região de transição (ITR igual a 19,5%), quanto em toda a nota (ITG 18,4%). Valores bem mais altos comparados com a clarineta, que obteve ITR igual a 1,1% e ITG igual a 0,7%. Isto pode ser relacionado com o fato do pizzicato ser a única nota pinçada. É também detectada uma maior concentração de transientes na região de transição para esta nota (ITC igual a 99,9%). O alto valor de ITC pode ser associado ao fato de a excitação no pizzicato ocorrer apenas no início, gerando uma alta concentração de transientes nesta região. Para as outras notas, a excitação acontece durante toda a nota: no cello, com a fricção do arco nas cordas; na clarineta e oboé, pela vibração da coluna de ar através da palheta; no trompete e flauta, pela vibração da coluna de ar através dos lábios e da pressão do sopro no bocal (Fletcher & Rossing, 1998).

O cello, o oboé e o trompete, assim como o pizzicato de violino, apresentam concentração elevada de transientes na região de transição, refletido nos altos valores de ITC (72,3%, 75,8%, e 81,6%), fato que pode ser visualizado na Figura 5.15.

A clarineta obteve um valor muito baixo de ITR e ITG (1,13% e 0,66%), assim como a flauta (3,70 % e 0,93%). Os dois instrumentos são de sopro, o que implica na

presença de um ruído não-desprezível, como pode ser visto na Figura 5.14. Porém, como são notas isoladas, não houve mudança de digitação nem influência de notas vizinhas. Isto leva a associar à situação em que um ataque suave foi executado. Para casos em que ITR e ITG são muito pequenos, não faz sentido a análise de ITC. Nestes casos, a energia da componente transiente é muito menor do que a energia do resíduo, podendo ocorrer erros maiores na modelagem. Uma análise visual da Figura 5.14 e da Figura 5.15 permite uma avaliação de coerência na separação da componente transiente das gravações dos instrumentos testados.

Tanto na fala quanto na música, a importância dos transientes aparece de forma clara, sendo sua modelagem útil para uma análise embasada.

Capítulo 7

Conclusão e trabalhos futuros

Esse estudo é dedicado à análise, modelagem e percepção auditiva de transientes em sinais musicais e de fala. É feita uma revisão da literatura, assim como a apresentação de fundamentos conceituais para o entendimento do tema. Ataques de instrumentos musicais e *bursts* em início de consoantes oclusivas são exemplos de regiões com presença de transientes. É mostrada sua importância para a percepção. Isto justifica a detecção e modelagem adequada da componente transiente de sinais, permitindo maior flexibilidade para sistemas de análise, síntese e transformação.

No estudo, é avaliado o método de modelagem de transientes *Transient Modeling Synthesis* (TMS) ao modelar a componente transiente de sinais musicais e de fala. O TMS é comparado à modelagem senoidal *Spectral Modeling Synthesis* (SMS). Experimentos de reconhecimento e qualidade MOS (*Mean Opinion Score*) são realizados para medir a importância da modelagem adequada dos *bursts* de consoantes oclusivas. Medidas de quantidade relativa e distribuição de transientes no sinal são propostas e avaliadas em um teste com gravações de instrumentos musicais diferentes.

Na análise de sinais de fala, os resultados mostram que a presença dos *bursts* é importante no reconhecimento de consoantes oclusivas. A ausência dos *bursts* reduziu o acerto no reconhecimento de 98%, obtido com sinais originais, para apenas 79%. Além disso, o TMS obteve bom desempenho para modelar os *bursts*. Atingiu 95%, sendo significativamente maior que os 87% obtidos com o SMS. A permutação dos *bursts* reduziu ainda mais o acerto para 61%, corroborando a existência de informação nos *bursts* que os diferencia. Em relação à qualidade perceptiva de modelagem dos *bursts*, o TMS apresenta uma pontuação acima de 4, o que significa um sinal de qualidade. A pontuação MOS do TMS é 4,16, contra 3,32 do SMS.

Três medidas de quantidade relativa de transientes em sinais de áudio são propostas neste trabalho e chamadas de Índice de Transiência Regional (ITR), Índices de Transiência Comparativo (ITC) e Índices de Transiência Global (ITG). Os três índices são testados para sinais musicais. Os índices indicaram maior quantidade de transientes em uma nota pizzicato de violino que em uma nota de cello executada com arco, refletido nos valores de ITR e ITG iguais a 20% e 18% para o pizzicato, e 10% e 2% para o cello, respectivamente. Houve também uma maior concentração dos transientes na região de transição para o pizzicato, refletida no valor de quase 100% para o pizzicato e 76% para o cello.

Em relação a trabalhos futuros, algumas possibilidades merecem ser investigadas. A primeira é o desenvolvimento de sistemas para acoplar o TMS a outras abordagens de modelagem de fala e música. Uma possibilidade de extensão do estudo é a avaliação do TMS em outras situações da fala, sem restrição às consoantes oclusivas. Um detector automático de *bursts* é também uma possibilidade interessante. Uma tarefa mais avançada seria o mapeamento dos parâmetros da modelagem TMS para transformações nos sinais. Outra possibilidade é a utilização de apenas transientes para o reconhecimento de locutores ou instrumentistas. Por último, vale a pena realizar uma avaliação dos Índices de Transiência em contextos musicais para análises sistemáticas.

A modelagem paramétrica de transientes na fala e na música ainda é um tema pouco explorado. Este trabalho apresenta apenas algumas das muitas possibilidades de modelagem. Espera-se que os resultados obtidos despertem o interesse para investigações mais aprofundadas.

Referências Bibliográficas

Ahmed, N.; Natarajan, T. & Rao, K. (1974). Discrete cosine transform. *Computers, IEEE Transactions on*, 100(1):90--93.

Bonada, J. & Serra, X. (2007). Synthesis of the singing voice by performance sampling and spectral models. *Signal Processing Magazine, IEEE*, 24(2):67--79.

Bonatto, M. (2007). A produção de plosivas por crianças de três anos falantes do português brasileiro. *Rev CEFAC*, 9(2):199--206.

Campolina, T.; Loureiro, M. & Mota, D. (2009). Expan: a tool for musical expressiveness analysis. Em *Proceedings of the 2nd International Conference of Students of Systematic Musicology*, pp. 24--27.

Daudet, L. (2006). A review on techniques for the extraction of transients in musical signals. *Computer Music Modeling and Retrieval*, pp. 219--232.

Duxbury, C.; Davies, M. & Sandler, M. (2001). Separation of transient information in musical audio using multiresolution analysis techniques. Em *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland.*

Fastl, H. & Zwicker, E. (2007). *Psychoacoustics: facts and models*, volume 22. Springer-Verlag New York Inc.

Flanagan, J. (1972). *Speech analysis: Synthesis and perception.* Springer-Verlag.

Fletcher, N. & Rossing, T. (1998). *The physics of musical instruments.* Springer Verlag.

Friedlander, B. & Porat, B. (1989). Detection of transient signals by the gabor representation. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 37(2):169--180.

Goodwin, M. (1996). Residual modeling in music analysis-synthesis. Em *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 2, pp. 1005--1008. IEEE.

Grey, J. (1977). Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am*, 61(5):1270--1277.

Hant, J.; Strope, B. & Alwan, A. (1997). A psychoacoustic model for the noise masking of plosive bursts. *The Journal of the Acoustical Society of America*, 101:2789.

Jain, A. (1979). A sinusoidal family of unitary transforms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (4):356--365.

Kent, R. & Read, C. (2002). The acoustic characteristics of consonants. *The acoustic analysis of speech. 2nd ed., Canada: Singular Thomson Learning*, pp. 139--88.

Ladefoged, P. & Maddieson, I. (1996). *The sounds of the world's languages*. Massachusetts: Wiley-Blackwell. p. 47-101.

Levine, S. & Smith, J. (1998). A sines+ transients+ noise audio representation for data compression and time/pitch scale modifications. *Preprints-Audio Engineering Society*.

Liberman, A.; Delattre, P.; Cooper, F. & Gerstman, L. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, 68(8):1.

Loureiro, M.; Borges, R.; Campolina, T.; Magalhães, T.; Mota, D. & de Paula, H. (2008). Extração de conteúdo musical em sinais de áudio para a análise de expressividade. Em *Anais do XXII Encontro da Bociedade Brasileira de Acústica*. SOBRAC.

Loureiro, M. A.; Yehia, H. C.; Paula, H. B.; Campolina, T. A. M. & Mota, D. A. (2009). Content analysis of note transitions in music performance. Em *Proceedings of the 6th Sound and Music Computing Conference (SMC 2009), Porto, Portugal*, pp. 355--359. INESC Porto.

Luce, D. (1963). *Physical correlates of nonpercussive musical instrument tones*. Tese de doutorado, MIT.

Maestre, E. & Gómez, E. (2005). Automatic characterization of dynamics and articulation of expressive monophonic recordings. Em *Proceedings of the 118th Audio Engineering Society Convention*. Citeseer.

Maia, E. (1985). *No reino da fala, a linguagem e seus sons*. São Paulo: Ática.

Masri, P. & Bateman, A. (1996). Improved modelling of attack transients in music analysis-resynthesis. Em *Proceedings of the International Computer Music Conference*, pp. 100--103. Citeseer.

McAulay, R. & Quatieri, T. (1986). Speech analysis/synthesis based on a sinusoidal representation. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(4):744--754.

Melo, R.; Mota, H.; Mezzomo, C.; de Castro Brasil, B.; Lovatto, L. & Arzeno, L. (2012). Desvio fonológico ea dificuldade com a distinção do traço [voz] dos fonemas plosivos: dados de produção e percepção do contraste de sonoridade. *Rev. CEFAC*, 14(1):18--29.

Molla, S. & Torrésani, B. (2004). Determining local transientness of audio signals. *Signal Processing Letters, IEEE*, 11(7):625--628.

Neto, M. U.; Silva, J. E. C.; Gomes, L. C. T.; Silva, D. A.; Campolina, T. A. M.; Sansão, J. P. H.; Yehia, H. C. & Vieira, M. N. (2012). Análise paramétrica de sinais de voz baseada em estimação conjunta do modelo fonte-filtro. Em *Anais do XXX Simpósio Brasileiro de Telecomunicações (SBrT), Brasília DF*. SBrT.

Palombini, C. (2006). O objeto sonoro de pierre schaeffer: duas abordagens. Em *Anais do XVI Congresso da Associação Nacional de Pesquisa e Pós-graduação em Música*, pp. 817--820. ANPPOM.

Rao, K. & Hwang, J. (1996). *Techniques and standards for image, video, and audio coding*. Prentice-Hall, Inc.

Rasetshwane, D.; Boston, J. & Li, C. (2006). Identification of speech transients using variable frame rate analysis and wavelet packets. Em *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pp. 1727--1730. IEEE.

Repp, B. & Lin, H. (1989). Acoustic properties and perception of stop consonant release transients. *Journal of the Acoustical Society of America*, 85(1):379--396.

Risset, J. (1965). Computer study of trumpet tones. *The Journal of the Acoustical Society of America*, 38:912.

Röbel, A. (2003). A new approach to transient processing in the phase vocoder. Em *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*, pp. 344--349.

Schaeffer, P. (1967). *Solfege de l'objet Sonore. INA GRM*. Seuil e GRM.

Serra, X. & Smith, J. (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12--24.

Stevens, K. (2000). *Acoustic phonetics*, volume 30. The MIT press.

Stevens, K.; Massey, N. et al. (1994). *Transients at stop-consonant releases*. Tese de doutorado, Massachusetts Institute of Technology.

Szwoch, G.; Kulesza, M. & Czyzewski, A. (2006). Transient detection for speech coding applications. *International Journal of Computer Science and Network Security*, 6(12):320--325.

Udo, Z. et al. (2011). *DAFX - Digital Audio Effects*. John Wiley & Sons.

Verma, T. & Meng, T. (1998). An analysis/synthesis tool for transient signals that allows a flexible sines+ transients+ noise model for audio. Em *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 6, pp. 3573--3576. IEEE.

Verma, T. & Meng, T. (2000). Extending spectral modeling synthesis with transient modeling synthesis. *Computer Music Journal*, 24(2):47--59.

Yip, P. (2001). *The transform and data compression handbook*. CRC.

Anexo A

Formulários do experimento de fala

Formulário de experimento - Gravação

Nome Completo: _____

Profissão: _____ Escolaridade: _____

Idade: _____ Sexo: _____

Obs.: Os dados pessoais desse formulário não serão usados nominalmente para a divulgação de resultados.

Assinatura: _____

Diga, por favor, a frase: "Escute _____ agora", com as seguintes palavras inseridas no espaço em branco, uma de cada vez, na seguinte ordem:

Pago
Tado
Cabo
Baco
Dato
Gapo
Gapo

Repita, por favor, a etapa anterior, agora com as seguintes palavras:

Gapo
Dato
Baco
Cabo
Tado
Pago
Pago

Obrigado!

Figura A.1.

Formulário de experimento - Escuta

Nome Completo: _____

Idade: _____ Sexo: _____

Obs.: Os dados pessoais não serão usados nominalmente para a divulgação de resultados.

Assinatura: _____

Para as sequências S1 a S6, ouça, por favor, a frase: "Escute _____ agora", e escreva nos espaços abaixo quais consoantes estão sendo ditas. Se não conseguir identificar, escreva o símbolo "Ñ". As palavras possuem quatro letras, duas vogais ("a" e "o"), e duas consoantes.

Para as sequências S7 a S9, marque com um "X", por favor, no número que melhor classifica a relação entre as duas sequências executadas, de acordo com o seguinte padrão:

Pontuação	Significado
1	**Inaceitável**: **Não** é possível **reconhecer a informação** contida na segunda sequência.
2	**Ruim**: **Há** grande **distorção na segunda** sequência, mas é mantida a inteligibilidade.
3	**Razoável: Foi** percebida diferença entre as sequências, e a **primeira é melhor.**
4	**Boa: Há** diferença entre as sequências, mas **não** é possível julgar **qual é melhor**.
5	**Excelente: Não** foi percebida **diferença** entre as sequências.

(Obs.: As sequências possuem ordem aleatória de palavras, devendo ser avaliada apenas a diferença de qualidade e inteligibilidade entre as sequências.)

Locutor _	S1	S2	S3	S4	S5	S6	▒	S7	S8	S9
1	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**1**	___	___	___
2	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**2**	___	___	___
3	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**3**	___	___	___
4	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**4**	___	___	___
5	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**5**	___	___	___
6	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	▒	▒	▒	▒

Locutor _	S1	S2	S3	S4	S5	S6	▒	S7	S8	S9
1	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**1**	___	___	___
2	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**2**	___	___	___
3	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**3**	___	___	___
4	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**4**	___	___	___
5	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**5**	___	___	___
6	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	▒	▒	▒	▒

Figura A.2.

Locutor _	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	▒	**S7**	**S8**	**S9**
1	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**1**	___	___	___
2	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**2**	___	___	___
3	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**3**	___	___	___
4	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**4**	___	___	___
5	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**5**	___	___	___
6	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	▒	▒	▒	▒

Locutor _	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	▒	**S7**	**S8**	**S9**
1	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**1**	___	___	___
2	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**2**	___	___	___
3	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**3**	___	___	___
4	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**4**	___	___	___
5	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	**5**	___	___	___
6	___a___o	___a___o	___a___o	___a___o	___a___o	___a___o	▒	▒	▒	▒

Obrigado!

Figura A.3.