

Anais do 6º Congresso de Engenharia de Áudio

12ª Convenção Nacional da AES Brasil

Proceedings of the 6th AES Brazil Conference

12th AES Brazil National Convention



A Convergência
Surround

05 a 07 de Maio de 2008

Centro de Convenções Rebouças - São Paulo/SP

Coordenador do Congresso | *Conference Chair:* Regis Rossi Alves Faria

Coordenador da Convenção | *Convention Chair:* Joel Brito

Editado por | *Edited by:* José Augusto Mannis e Regis Rossi A. Faria



Audio Engineering Society - Seção Brasil

Comitê Organizador

Coordenador Geral da Convenção:

Joel Brito (AES Brasil)

Coordenador Geral do Congresso:

Regis Rossi A.Faria (AES Brasil / EP-USP)

Comitê executivo

Aldo Soares (AES Brasil)

Coordenação de Programa Técnico:

Adolfo Maia Jr. (IMECC-UNICAMP)

Sidnei Noceti Filho (CTC-UFSC)

Coordenação de Artigos e Pôsteres:

Rubem Dutra R.Fagundes (FENG-PUCRS)

Christian Herrera (CEFET-MG)

Coordenação Editorial:

José Augusto Mannis (IA-UNICAMP)

Regis Rossi A.Faria (AES Brasil / EP-USP)

Programação LaTeX:

Sávio Marcelo Soares

Apoio logístico:

Renan Vital

Rosilene Louro

Editoração e arte:

Lídia Brito

Totum Marketing e Comunicação

Comitê de Programa Técnico:

Adolfo Maia Jr. (IMECC-UNICAMP)

Aníbal J. de S. Ferreira (Univ. do Porto, Portugal)

Christian Herrera (CEFET-MG)

Eduardo R. Miranda (Univ. Plymouth, UK)

Fabrício de Oliveira Ourique (FENG-PUCRS)

Francisco J. Fraga da Silva (UFABC)

José Augusto Mannis (IA-UNICAMP)

José Manuel N. Vieira (Univ. de Aveiro, Portugal)

Julio Cesar Boscher Torres (POLI-UFRJ)

Luiz W. P. Biscainho (POLI & COPPE-UFRJ)

Marcelo Queiroz (IME-USP)

Márcio Gomes (CEFET-SC)

Miguel A. Ramírez (EP-USP)

Phillip M. S. Burt (EP-USP)

Pierre Dumouchel (ETS/UQAM, Canadá)

Regis R.A. Faria (EP-USP)

Rodrigo Cicchelli Velloso (EM-UFRJ)

Rubem Dutra R. Fagundes (FENG-PUCRS)

Sergio Lima Netto (POLI & COPPE-UFRJ)

Sidnei Noceti Filho (CTC-UFSC)

Vinicio Licks (FENG-PUCRS)

Agradecimentos:

AES Board of Governors

Silvia Regina S.Della Torre (Biblioteca EP-USP)

Copyright © 2008
Audio Engineering Society – Brazil Section

Congresso de Engenharia de Áudio 6.: São Paulo: 2008); Convenção Nacional AES Brasil (12.: São Paulo: 2008)
Anais 6. Congresso de Engenharia de Áudio; 12. Convenção Nacional AES Brasil / ed. J.A. Mannis, R.R.A. Faria. – São Paulo: AES Brasil, 2008.
1 CD-Rom.

ISBN 978-85-99997-03-1

1.Engenharia de áudio (Congressos) 2. Computação musical (Congressos)
3. Processamento de sinais (Congressos)
I.Convenção Nacional AES Brasil (12.: São Paulo, 2008) II.Audio Engineering Society. Seção Brasil III.Mannis, José Augusto IV.Faria, Regis Rossi Alves V.t.

CDD621.3828

Os artigos publicados nestes anais foram reproduzidos dos originais finais entregues pelos autores, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo.

Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA,
www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org.

Todos os direitos são reservados. Não é permitida a reprodução total ou parcial dos artigos sem autorização expressa da AES Brasil.

Impresso no Brasil.
Printed in Brazil.

Apoio:



Organização:



PUCRS

Realização:



Sociedade de Engenharia de Áudio

AES – Audio Engineering Society – Brazil Section

Rua Carlos Machado 164, sala 305, Pólo Rio de Cine e Vídeo – Barra da Tijuca
Rio de Janeiro, Brasil – Cep. 22775-042 | e-mail: aesbrasil@aes.org | www.aesbrasil.org
telefone: +55 (21) 2421-0112 | fax: +55 (21) 2421-0112

Administração

Presidente/Chairman: Joel Brito

Vice-Presidente/Vice-Chairman: Regis Rossi Alves Faria

Secretário/Secretary: Aldo Ricardo Soares

Tesoureiro/Treasurer: Guilherme Figueira

Conselheiros/Council members: Carlos Ronconi

Framklim Garrido

Germano Kannenberg

José Anselmo Pereira

Maurício Gargel

Yves Zimelman

Audio Engineering Society, Inc.

International headquarters

60 East 42nd St., Room 2520, New York, NY, 10165-2520, USA

e-mail: hq@aes.org | www.aes.org

telephone: +1 (212) 661-8528 | fax: +1 (212) 661-7829

Sumário

Contents

Prefácio dos organizadores / Organization greetings	7
--	-------	---

Revisores / Reviewers	9
------------------------------	-------	---

Sessões de artigos / Papers sessions

Sessão 1 - Modelamento e processamento de sinais de áudio

(Audio signal processing and modeling)

1.	Alteração das características da voz para uso em mundos virtuais (Second life) <i>Felipe Breve Siola e Francisco José Fraga da Silva</i> 11
2.	Estudo da sonoridade da clarineta através de modelo experimental <i>L.C. Oliveira, R. Goldemberg e J. Manzolli</i> 15
3.	Estudo de guias de ondas com geometria de dupla reflexão parabólica <i>Felipe Menezes e Iuri Pepe</i> 20
4.	Técnicas de processamento de áudio em sinais de voz, para auxílio diagnóstico de doenças laríngeas <i>R.D.R. Fagundes, I.C. Zwetsch e D. Scolari</i> 24

Sessão 2 - Acústica e áudio espacial

(Acoustics and spatial sound)

5.	Análise modal bidimensional de salas com superfícies seriais difusoras <i>José Augusto Mannis</i> 30
6.	Cabeças artificiais e manequins: um resgate histórico <i>Stephan Paul</i> 38

7.	Método analítico para o cálculo da diferença interaural de tempo (ITD) no plano horizontal <i>Camila Tigussa Sato, Marcelo Lapa Espiga, Stephan Paul, Samir Gerges e Pascal Dietrich</i>	46
8.	Análise quantitativa do simulador acústico: RAIOS <i>Julio Cesar B. Torres, Flavia Correia Tovo, Mariane R. Petraglia e Roberto A. Tenenbaum</i>	52

Sessão 3 - Codificação, processamento e edição de som

(Sound edition, processing and coding)

9.	Marcação automática de eventos usando sinal de áudio em transmissões esportivas de TV <i>Luiz G. L. B. M. de Vasconcelos, Sergio L. Netto, Luiz W. P. Biscainho e Charles B. do Prado</i>	58
10.	Equalização e identificação adaptativas para áudio utilizando marca d'água como sinal de supervisão <i>Leandro de Campos Teixeira Gomes, Mário Ulian Neto e João Marcos Travassos Romano</i>	65
11.	FlawQ: um plug-in VST para equalização gráfica digital <i>Felipe C. V. Martins, Leonardo de O. Nunes, Alan F. Tygel e Luiz W. P. Biscainho</i>	72
12.	Avaliação subjetiva de qualidade de áudio:fala vs. música <i>Daniel S. Gerscovich e Luiz W. P. Biscainho</i>	80

Sessão 4 - Análise, classificação e percepção do som

(Sound analysis, classification and perception)

13.	Separação de instrumentos musicais com uma única mistura <i>Diego Barreto Haddad, Mariane Rembold Petraglia e Paulo Bulkool Batalheiro</i>	88
14.	Towards the evaluation of automatic transcription of music <i>Tiago Fernandes Tavares, Jayme Garcia Arnal Barbedo e Amauri Lopes</i>	96

15.	Short-term classification of musical instruments: a critical view <i>Jayme Garcia Arnal Barbedo e Amauri Lopes</i>	100
16.	Automatic estimation of harmonic complexity in audio <i>José Fornari e Tuomas Eerola</i>	108
17.	Classificação automática de sons de instrumentos musicais usando discriminantes lineares <i>Jorge Costa Pires Filho, Paulo Antonio Andrade Esquef e Luiz Wagner Pereira Biscaíno</i>	112
18.	Um modelo para a transcrição automática de melodias para partitura <i>Gabriel Simões Gonçalves da Silva e Antonio Cesar de Castro Lima</i>	119

Sessão 5 - Análise, síntese e sistemas para computação musical

(Analysis, synthesis and computer music systems)

19.	AURAL: ambiente interativo aplicado à sonificação de trajetórias robóticas <i>Artemis Moroni, Josué Ramos, Sidney Cunha e Jônatas Manzolli</i>	128
20.	Applications of group theory on sequencing and spatialization of granular sounds <i>Renato Fabbri e Adolfo Maia Jr.</i>	134
21.	Mecanismo de argumentação e controle de versão para um ambiente cooperativo de prototipação musical <i>Aurélio Faustino Hoppe, Evandro Manara Miletto e Marcelo Soares Pimenta</i>	139

Índice de autores Author index	143
---	-----

Prefácio dos Organizadores

Nesta 12^a edição da Convenção Nacional e 6^a do Congresso de Engenharia de Áudio da AES Brasil abordamos o tema "A Convergência Surround".

Presenciamos um cenário em que novas tecnologias e formatos de som envolvente ou "surround" se fundem nas novas perspectivas de aplicações em serviços de áudio personalizados, nas redes de distribuição de som/música, e no *broadcasting*, como a TV e o rádio digital. A convergência dos sistemas de áudio com os sistemas de informação e telecomunicações abre novas fronteiras para plataformas de autoria e de reprodução de áudio espacial, bem como novos serviços especializados. Com este tema, buscamos fomentar uma ampla discussão e exposição de novidades no setor, e uma análise desta visão futura pelos segmentos envolvidos da cadeia de áudio.

Além de uma extensa programação técnica com diversos seminários e palestras, organizamos também um painel sobre som *surround*, em três sessões. Especialistas no tema traçam um panorama dos mais variados aspectos da produção ao consumo de áudio *surround*, e abordam as referências tecnológicas do momento. A intenção é estimular a disseminação da informação neste tema para os profissionais da indústria nacional, que ora encontra-se em expansão.

Publicamos nesta edição vinte e um artigos científicos, que consolidam a qualidade e a diversidade da produção acadêmica em diversas linhas de pesquisa e desenvolvimento voltadas para o áudio, a música e a voz. Os artigos abordam tópicos que vão do processamento digital de áudio, análise, percepção, classificação e codificação sonora, até aplicações para ambientes virtuais e de computação musical.

Desejamos a todos os congressistas um excelente AES Brasil 2008 e que aproveitem a oportunidade para elaborar novos contatos, absorver novos conhecimentos, e participar ativamente deste evento.

**Comissão Organizadora
AES Brasil 2008**

Organization Committee Greetings

In this 12th edition of the National Convention and 6th of the Audio Engineering Conference of the AES Brazil we approach the subject "The Surround Convergence".

We witness a scenario where new surround technologies and formats fuse in new perspectives of applications towards personalized audio services, as in the music/sound distribution networks, and in the broadcasting, with TV and the digital radio. The convergence of audio systems with the information and telecommunications systems opens new frontiers for authorship platforms and for spatial audio reproduction, as well as new special services. Exploring this theme, we bring an ample discussion and exhibition of new features in the sector, and an analysis of this future vision for the segments of the audio chain.

Beyond an extensive technical program with diverse seminaries and lectures, we also organize a panel on surround sound, in three sessions. Specialists in the subject trace a rich panorama of many aspects from the production to the consumption of surround, and approach the technological references of the moment. The intention is to stimulate the dissemination of the information in this subject for the professionals of the national industry, which is in expansion.

We publish in this edition twenty one scientific papers, that consolidate the quality and the diversity of the academic production in many lines of research and development directed toward the audio, music and voice. The articles approach topics that go from digital audio processing, analysis, perception, classification and sound coding, till applications for virtual and computer music environments.

We desire to all the congressmen an excellent AES Brazil 2008 and that they take the chance to elaborate new contacts, to absorb new knowledge, and to participate actively of this event.

**Organization Committee
AES Brazil 2008**

Revisores

Reviewers

Adolfo Maia Jr. (UNICAMP)
Amaro Azevedo de Lima (IFRJ)
Anibal Ferreira (Univ. do Porto, Portugal)
Christian Herrera (CEFET-MG)
Cristiano Ferreira (UFSC)
Eduardo Miranda (University of Plymouth, UK)
Fabrício Ourique (PUCRS)
Fernando Pacheco (UFSC)
Francisco Fraga da Silva (UFABC)
José Mannis (UNICAMP)
José Vieira (Univ. de Aveiro, Portugal)
Julio Torres (UFRJ)
Luiz Biscainho (UFRJ)
Marcelo Queiroz (USP)
Márcio Gomes (CEFET-SC)
Miguel Ramírez (USP)
Phillip Burt (USP)
Regis Faria (USP)
Rodrigo Velloso (UFRJ)
Rubem Fagundes (PUCRS)
Sergio Netto (UFRJ)
Vinícius Licks (PUCRS)
Walter Gontijo (UNIVALI)

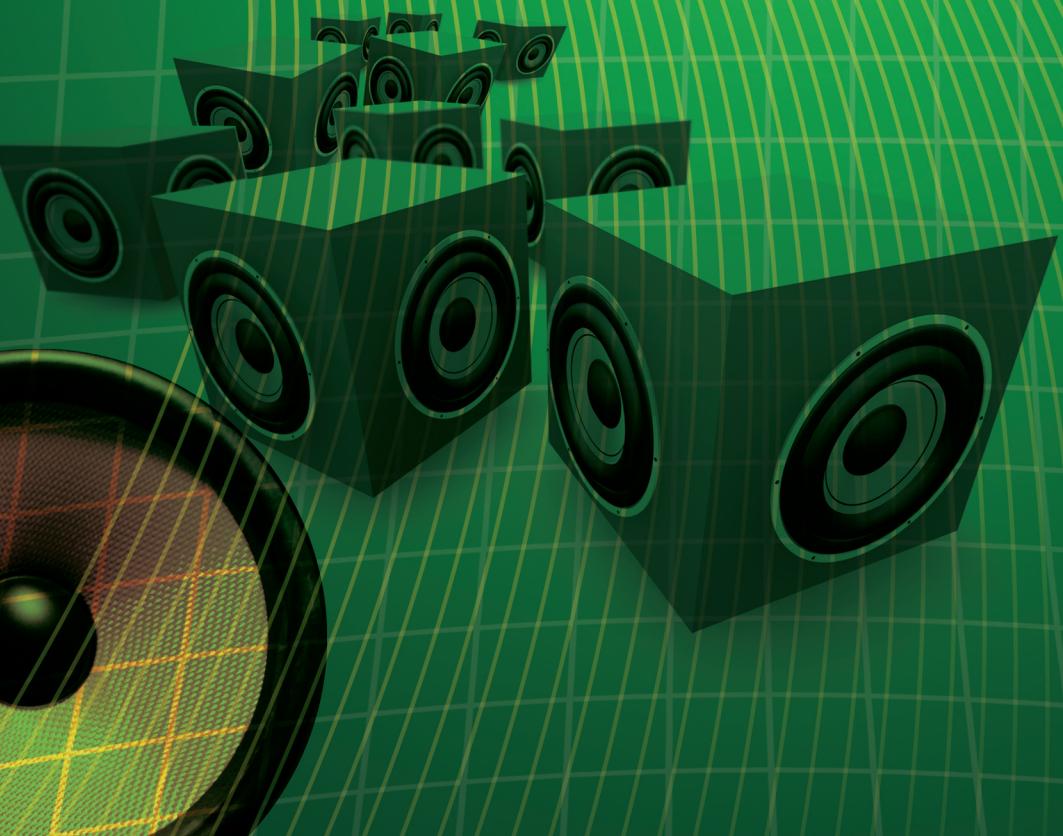
Sessões de Artigos

Papers Sessions

Sessão 1

Modelamento e processamento de sinais de áudio

(Audio signal processing and modeling)





Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Alteração das características da Voz para uso em Mundos Virtuais (Second Life)

Felipe Breve Siola¹, Francisco José Fraga da Silva¹

¹Universidade Federal do ABC (UFABC)
Santo André, São Paulo, CEP 09210-170, Brasil
{felipe.siola, francisco.fraga}@ufabc.edu.br

RESUMO

Um dos objetivos deste projeto é o de poder contribuir, como outros, para melhorar o novo recurso de voz do mundo virtual *Second Life* (SL), alterando as características da voz de um usuário do SL. Foi criado um aplicativo no ambiente MATLAB para implementar os algoritmos de processamento utilizados. Nesta fase inicial da pesquisa foram escolhidas para serem alteradas as características de freqüência de *pitch* e “timbre”. Para o “timbre”, foi utilizado um modelo matemático do pulso glotal no domínio do tempo e da freqüência. Para alteração de *pitch* utilizou-se uma técnica de compressão ou expansão do espectro de freqüências.

0 INTRODUÇÃO

O número de usuários do mundo virtual *Second Life* (SL) na internet, chamados de “Residentes”, tem crescido de maneira impressionante, atingindo mais de doze milhões [1]. Segundo estudo sobre os usuários do mundo virtual realizado em dezembro de 2007, o Brasil agrupa cerca de 5% da população ativa no SL, ocupando a sexta posição do ranking de países com maior número de “Residentes” ativos [2].

Segundo a empresa *Linden Lab*, criadora e controladora do SL, a utilização de voz como meio de comunicação entre os usuários do SL sempre foi um objetivo. Em meados de junho de 2007 esta ferramenta passou a fazer parte do mundo principal do SL, porém ainda em caráter experimental.

No entanto, no *blog* oficial do SL [3], muitos “Residentes” se mostraram receosos de utilizar este recurso porque são de opinião que o uso da voz, sendo algo muito pessoal, os remete diretamente à sua “*First Life*”. Argumentam que, ao buscarem a experiência de ser “Residentes” em um mundo virtual, no qual eles podem esquecer temporariamente as diversas limitações impostas

pela vida real, não querem ser “traídos” e talvez até mesmo identificados pelo uso da voz.

A opção de alterar as características da própria voz, de modo que ela possa soar de maneira inteiramente nova no mundo virtual do SL, surge então como um expediente extremamente desejável e até mesmo como um fator condicionante para o grau de popularidade deste recurso no SL. De fato, a possibilidade de modificar/modular a própria voz foi aventada pelo *Linden Lab* e aparece como uma das perguntas do banco de FAQs relacionadas a este assunto. Em algumas páginas da internet, já é possível encontrar tutoriais “passo a passo” sobre como modificar a própria voz com softwares auxiliares, a fim de utilizar tranquilamente o recurso de voz do SL sem o risco de ser reconhecido [4].

Neste trabalho, ainda em desenvolvimento, foi criado um aplicativo no ambiente MATLAB para implementar os diversos algoritmos de processamento do sinal de fala, com o objetivo de alterar algumas de suas características visando um possível futuro uso no SL. Nesta primeira fase da pesquisa foram escolhidas apenas as características de freqüência de *pitch* e “timbre”.

1 FUNDAMENTAÇÃO TEÓRICA

O modo como é realizada a alteração de “timbre” apresentada neste trabalho está fundamentado no próprio processo de produção da fala. Em particular, na produção dos sons com excitação fonada ou sonora (que é o caso da grande maioria dos fonemas), ou seja, com vibração das cordas vocais, em contraposição aos sons cuja excitação é chamada de surda, isto é, ocorre sem vibração das cordas vocais [5].

A Figura 1 ilustra a forma de onda típica do fluxo de ar através da glote durante um ciclo completo de fonação. Nesta figura destaca-se a relação direta entre a duração do pulso glotal e o “timbre” da fala. As aspas são necessárias, pois o termo timbre refere-se originalmente ao envelope temporal e espectral do som produzido ao se tocar uma nota de um instrumento musical específico. No contexto desta pesquisa, o termo timbre foi utilizado com referência ao envelope espectral dos sons da fala.

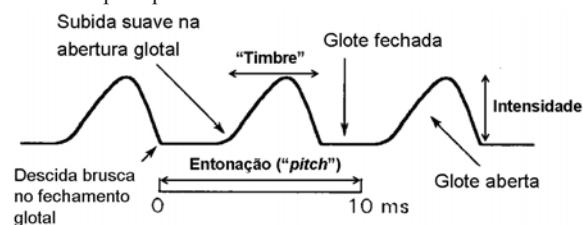


Figura 1 Forma de onda do fluxo de ar através da glote durante a fonação (3 ciclos completos), adaptado de [5].

Existem basicamente três etapas da produção da fala que modificam o envelope espectral dos sons produzidos com excitação fonada: a forma do pulso glotal (em especial sua duração), as ressonâncias produzidas pelo trato vocal (formantes) e o efeito da radiação labial e/ou nasal [6]. Aqui estamos interessados somente na contribuição do pulso glotal: quanto mais largo, maior será a atenuação sofrida pelas harmônicas de alta freqüência, quanto mais estreito, menor será a atenuação. A forma e duração do pulso glotal está diretamente relacionada com o tamanho das cordas vocais e, portanto, também da glote, que nada mais é do que o espaço entre as mesmas. Tipicamente, mulheres e crianças possuem cordas vocais menores. Isso resulta em pulsos glotais estreitos e consequentemente em harmônicas de alta freqüência com intensidade consideravelmente maior que no caso de homens adultos, que possuem pulsos glotais mais largos devido à maior inércia das suas cordas vocais.

Neste trabalho, foi desenvolvido um modelo matemático para a forma de onda do pulso glotal, que será utilizado diretamente para a alteração das características de timbre. A magnitude da Transformada de Fourier do modelo do pulso glotal será aplicada para “diminuir” o timbre, enquanto que o inverso desta magnitude será usado para “aumentar” o timbre, ou seja, para aumentar a proporção entre a intensidade das harmônicas de alta freqüência com relação às de baixa freqüência.

2 PRÉ-PROCESSAMENTO

O sinal de som a ser processado deve ser digitalizado a uma taxa de amostragem de 32 kHz, portanto cobrindo todo o espectro de fala na faixa de 0 a 16 kHz. Todo o processamento é realizado no domínio da freqüência, utilizando a técnica amplamente conhecida na literatura como *phase vocoder* [7]. O sinal de voz no domínio do

tempo é dividido em quadros com duração de 50 ms (1600 amostras), com sobreposição de 75% entre quadros sucessivos (incremento de 12,5 ms a cada quadro). Antes de se aplicar uma Transformada Rápida de Fourier (FFT) de 2048 pontos, cada quadro foi multiplicado por uma janela Hamming [8].

Prevendo um fator de dois como máximo aumento possível no *pitch*, as amostras da FFT de cada quadro de fala correspondentes às freqüências entre 8 e 16 kHz são zeradas (filtragem passa-baixas) a fim de possibilitar uma expansão correspondente do espectro de 0 a 8 kHz.

3 ALTERAÇÃO DE TIMBRE

A aproximação que modela o formato do pulso glotal $g(n)$ no domínio do tempo discreto foi gerada como

$$g(n) = x(P - n) \text{ onde } x(n) = na^n, 0 \leq n \leq P - 1 \quad (1)$$

P é o período de *pitch* (inteiro, em amostras) e a é um número real entre 0 e 1. A Figura 2 contém os gráficos das funções $x(n)$ e $g(n)$, nas partes (a) e (b), respectivamente.

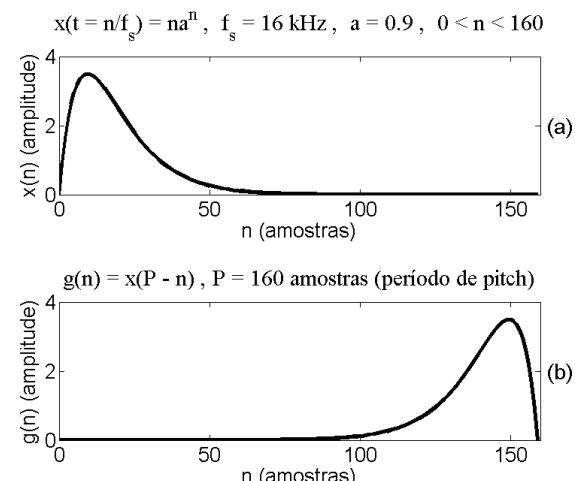


Figura 2 (a) Função geradora do modelo do pulso glotal
(b) Modelo matemático do pulso glotal (masculino)

O modelo matemático $g(n)$ resultante possui todas as características de um pulso glotal real, apresentadas na Figura 1: subida suave na abertura da glote e descida brusca no fechamento glotal [5]. Além disso, este modelo matemático possui a vantagem de ser facilmente tratável, como se verá a seguir. Considerando uma taxa de amostragem de 16 kHz (pois a metade superior do espectro foi zerada conforme explicado na seção anterior), o período de pitch tipicamente feminino de 5 ms (freqüência de pitch de 200 Hz) corresponderá a um valor de $P = 80$ amostras, enquanto que o masculino (em torno de 10 ms) corresponderá a $P = 160$ amostras.

Com um valor da constante $a \leq 0.9$, podemos desconsiderar o truncamento da função $x(n)$, pois esta terá valor praticamente nulo para índices n próximos de P . Ou seja, $x(n) \equiv 0$ para valores $n > 80$, conforme pode ser observado na Figura 2 (a). Assim, a função geradora deste modelo de pulso glotal poderá ser escrita como $x(n) \equiv na^n u(n)$, onde $u(n)$ é o degrau unitário. Logo a Transformada Z do pulso glotal pode ser calculada como

$$g(n) = x(-n + P) \leftrightarrow G(z) = z^P Z[x(-n)] = z^P X(z^{-1}) \quad (2)$$

pelas propriedades da Transformada Z [8]. Então se

$$x(n) \equiv na^n u(n) \leftrightarrow X(z) \equiv \frac{az^{-1}}{(1 - az^{-1})^2} \quad (3)$$

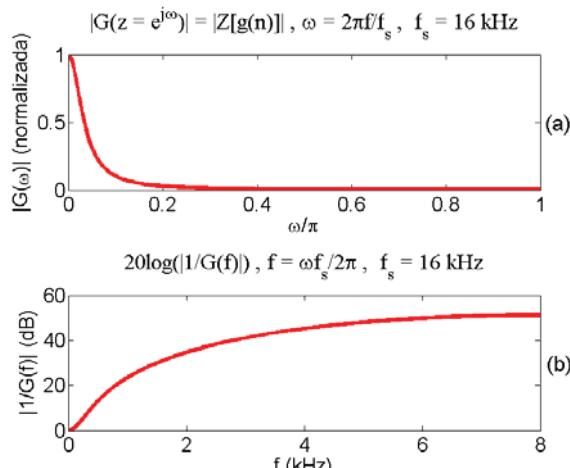
Substituindo (3) em (2) vem

$$G(z) \equiv z^P \frac{az}{(1 - az)^2} = z^{P+1} \frac{a}{(1 - az)^2} \quad (4)$$

Portanto, para obter a expressão do pulso glotal no domínio da freqüência digital ω , basta fazer

$$z = e^{j\omega} \Rightarrow G(\omega) = e^{j\omega(P+1)} \frac{a}{(1 - ae^{j\omega})^2} \quad (5)$$

A Figura 3 (a) apresenta o gráfico da magnitude normalizada do pulso glotal $G(\omega)$ no domínio da freqüência (para $a = 0.9$), conforme a equação (5). Esta curva será multiplicada pelo espectro de tempo curto (quadro) do sinal de fala a fim de “diminuir” o timbre, ou seja, reduzir a amplitude das componentes espectrais de altas freqüências com relações às componentes de baixas freqüências. Na Figura 3 (b) pode-se ver o gráfico do inverso da magnitude $G(\omega)$, porém com as abscissas em f (Hz) e as ordenadas em dB.



Conforme será explicado na seção 5, o usuário do aplicativo poderá escolher em quantos dB ele deseja “diminuir” ou “aumentar” o timbre. Esta medida corresponde à atenuação ou amplificação máxima das componentes espectrais, que ocorre na freqüência de 8 kHz (ou $\omega = \pi$), nos gráficos da Figura 3 (a) e (b), respectivamente. Resta então converter este valor de atenuação ou amplificação em dB para um determinado valor da constante a . A correspondência entre estes valores é mostrada a seguir. Substituindo os valores $\omega = 0$ e $\omega = \pi$ na equação (5) obtém-se

$$|G(0)| = \frac{a}{(1 - a)^2} \quad \text{e} \quad |G(\pi)| = \frac{a}{(1 + a)^2} \quad (6)$$

Para se obter o valor da constante a que corresponde a uma atenuação de A dB na freqüência de 8 kHz (ou então, $\omega = \pi$), basta resolver a equação (7):

$$20 \log_{10}(|G(0)|) - 20 \log_{10}(|G(\pi)|) = A \text{ (dB)} \quad (7)$$

Substituindo (6) em (7) chega-se facilmente à expressão

$$a = \frac{10^{A/40} - 1}{10^{A/40} + 1} \quad (8)$$

4 ALTERAÇÃO DE PITCH

Para a alteração do *pitch* foi utilizada uma técnica de compressão ou expansão do espectro do sinal de fala. Esta consiste na alteração da localização das amostras da FFT, conforme a curva de compressão ou expansão de freqüências desejada. O fator K de compressão/expansão pode variar de 2.0 (reduz a faixa de freqüências à metade) a 0.5 (dobra a faixa de freqüências do espectro).

A partir do fator K , a posição das amostras da FFT do sinal original é alterada, mapeando-as linearmente para outras freqüências, conforme ilustrado na figura 4. Ou seja, valores de K maiores do que 1.0 provocam compressão do espectro de freqüências e, portanto, diminuem a freqüência de *pitch*, ao passo que valores de K inferiores à unidade expandem o espectro e consequentemente aumentam o *pitch*. A Figura 4 mostra uma curva de compressão com fator $K = 1.6$.

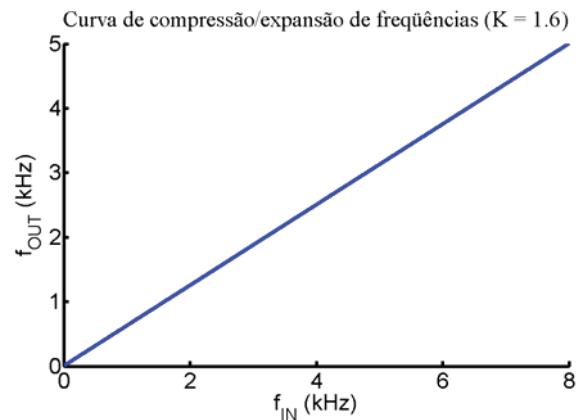


Figura 4 Curva de compressão/expansão de freqüências.

O uso desta técnica, por modificar todo o espectro do sinal, além de causar alteração no *pitch*, causa também alteração na posição das freqüências formantes. No próximo passo da pesquisa, esta técnica será substituída pela técnica de análise por predição linear (LPC).

5 INTERFACE COM O USUÁRIO

Foi desenvolvido um aplicativo com interface gráfica, utilizando o software MATLAB, onde é permitido ao usuário definir as variáveis necessárias para o processamento digital de sinais, isto é, o valor para alteração do timbre, o valor para alteração da freqüência de *pitch* e o nome do arquivo de voz a ser processado.

Conforme mostrado na Figura 5, a interface apresenta um visual amigável, permitindo que qualquer usuário do software, mesmo sem conhecer nada sobre as técnicas de processamento digital de sinais, facilmente entenda o funcionamento da aplicação e possa utilizá-la. A primeira ação a ser realizada pelo usuário é a escolha do arquivo no formato WAV a ser processado. Para isto, basta que o usuário clique sobre o botão “Procurar”, e selecione em sua máquina o arquivo desejado.

A escolha do valor “Alteração de pitch” (entre 0.5 e 2.0 para a variável K) pode ser feita por meio de um cursor ou então digitando diretamente. O último passo é “Alteração de timbre”, onde o usuário poderá escolher um valor de

atenuação A na faixa de -40dB a 40dB, usando o mesmo processo, conforme ilustrado na Figura 5. Valores positivos de A são diretamente utilizados na equação (8) para obtenção de a e geração do pulso glotal. Valores negativos de A são interpretados como desejo de aumentar o timbre. Nesse caso será usado o valor positivo correspondente na mesma equação (8), porém agora o inverso do espectro do pulso glotal é que será utilizado, conforme explicado anteriormente na seção 3.

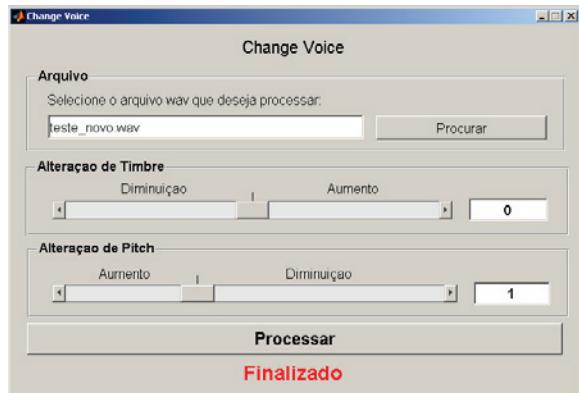


Figura 5 Interface gráfica do software

Depois de selecionados estes valores, ao clicar no botão "Processar" o aplicativo realiza todas as operações desejadas e cria automaticamente um novo arquivo de voz. Este arquivo terá o sufixo "`_chgd_K<#K><#A>dB`" acrescentando ao nome do arquivo original, e será informando ao usuário que o processo foi finalizado com sucesso.

É importante destacar que a compressão ou expansão do espectro correspondente à alteração de *pitch* é realizada antes da modificação de timbre. Ou seja, a envoltória que será alterada pelo aumento ou diminuição do timbre corresponde ao espectro já comprimido ou expandido. No entanto, caso tenha havido expansão do espectro, as amostras do sinal de fala acima de 8000 Hz são novamente zeradas antes de se realizar a alteração de timbre.

6 RESULTADOS

Embora não exista uma maneira objetiva de se avaliar os resultados deste tipo de processamento, podem-se verificar graficamente as alterações provocadas pelo mesmo, conforme ilustrado na Figura 6, onde vemos um espectro de tempo curto correspondente à vogal "i" de um locutor masculino, antes e depois de processado pelo software, nas partes (a) e (b), respectivamente.

Nesse caso, o timbre foi aumentado em 30 dB e foi aplicado o máximo fator de expansão permitido pelo software ($K = 0.5$), duplicando o valor do *pitch* original e também das freqüências formantes dessa vogal.

7 CONCLUSÃO

Tendo em conta a aplicação visada, que é a de alteração das características da fala para um possível uso no SL, pode-se dizer que a finalidade principal da pesquisa foi parcialmente alcançada, uma vez que o aplicativo em MATLAB é capaz de alterar as características da voz. Porém, falta traduzir o código para C++ e integrá-lo ao SL.

Mesmo para uma ligeira alteração no timbre e no *pitch* ($K = 1.3$ e $A = 20$, por exemplo) torna-se praticamente impossível identificar o locutor da frase original. Se for

feita apenas a alteração de timbre, a voz modificada soa perfeitamente natural, porém mais facilmente identificável. No caso da alteração de *pitch* por meio da técnica de compressão ou expansão de freqüências, a fala processada soa de maneira um pouco distorcida, devido à alteração provocada na localização das freqüências formantes. Porém, pretende-se corrigir esta deficiência em uma próxima fase da pesquisa, por meio do uso da técnica de análise por predição linear (LPC), a fim de separar o sinal de excitação das ressonâncias do trato vocal [9][10], possibilitando assim a alteração do *pitch* e das formantes de forma independente.

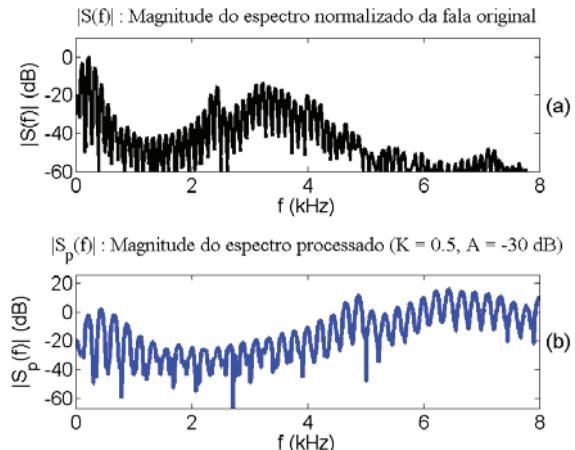


Figura 6 (a) Espectro normalizado da vogal "i" original
(b) Espectro processado ($K = 0.5, A = -30 \text{ dB}$)

8 REFERÊNCIAS

- [1] *Economic Statistics of Second Life*. URL: http://secondlife.com/whatis/economy_stats.php, acessado em março de 2008.
- [2] *Second Life Key Metrics through December 2007*. URL: <http://spreadsheets.google.com/pub?key=pxbDc4B2FH97QWH8L28FHOw&gid=7>, acessado em março de 2008.
- [3] *Second Life Blog, Bringing Voice to Second Life*. URL: <http://blog.secondlife.com/2007/02/27/bringing-voice-to-second-life/#comment-156987>, acessado em março de 2008.
- [4] *My Configuration for Voice Effects in Second Life*. URL: <http://www.erzsabet.com/mySLSound.html>, acessado em março de 2008.
- [5] L. R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ: Prentice-Hall, 1978.
- [6] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, Wiley-IEEE, 1999.
- [7] J.L. Flanagan and R.M. Golden, *Phase vocoder*, Bell Syst. Tech. J., Vol. 45, pp. 1493-1509, 1966.
- [8] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, 2nd edition, 1999.
- [9] E. Moulines and J. Laroche, *Non-parametric techniques for pitch-scale and time-scale modification of speech*, Speech Communication, vol. 16, no. 2, pp. 175–205, February 1995.
- [10] R. C. D. Paiva, L. W. P. Biscainho and S. L. Netto, *A Sequential System for Voice Pitch Modification*, 5.º Congresso de Engenharia de Áudio (AES Brasil 2007), São Paulo, Brasil, 2007.



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Estudo da Sonoridade da Clarineta através de Modelo Experimental

L.C. Oliveira¹, R. Goldemberg², J. Manzolli²

¹FEEC/NICS-UNICAMP

²IA/NICS-UNICAMP

Campinas, SP, CEP: 13083-970, Brasil

{luis,rgoldem,jonatas}@nics.unicamp.br

RESUMO

A sonoridade da clarineta foi investigada empregando-se o Projeto Fatorial de Experimentos. A abertura da boquilha, dureza da palheta, posição na palheta, entre outras, foram as variáveis independentes analisadas. Como resposta, freqüência e amplitude da nota fundamental e dos componentes espectrais foram anotadas. A tessitura do instrumento foi dividida em três regiões: chalumeau, clarino e agudíssima. Em cada uma delas elegeu-se uma nota para observação. Desta análise obtém-se um modelo experimental linear nas variáveis independentes de modo a descrever o efeito delas na sonoridade da clarineta.

0 INTRODUÇÃO

A investigação da sonoridade da clarineta sob o foco experimental já faz parte de nossas atividades desde 2005, como pode ser verificado em OLIVEIRA et al. [1,2,3].

Esta pesquisa vem gradativamente ampliando seu grau de complexidade. Partimos de um ponto de vista exploratório, onde se esperava uma análise com dados de menor precisão, porém que indicasse os principais aspectos do problema.

Tivemos uma noção dos fatores que maiores influências têm sobre o fenômeno e a partir do modelo experimental obtido, linear, esboçamos a síntese do instrumento em tempo real utilizando o software livre Pure Data.

Os resultados iniciais foram motivadores para a seqüência do trabalho. Um refinamento das condições anteriores com relação às variáveis independentes possibilitou chegar a este trabalho, que corresponde a uma parte de todos os dados experimentais desta segunda etapa.

Inicialmente apresentaremos as modificações pontuais para a condução deste novo conjunto de experimentos. Justificam-se as alterações com base nos resultados anteriores, OLIVEIRA et al. [1,2,3].

Os dados obtidos assim como os resultados analisados são apresentados em tabelas possibilitando maior compreensão para a análise.

Uma provável seqüência para a continuidade deste trabalho é apresentada após as conclusões.

1 APARATO E PROCEDIMENTO EXPERIMENTAIS

Este trabalho utiliza o mesmo aparato experimental dos trabalhos anteriores OLIVEIRA et al. [1,2,3]. Algumas modificações localizadas foram empregadas e serão descritas a seguir, porém, da mesma forma que anteriormente eliminou-se um clarinetista profissional para obtenção dos dados, pois o processo de adaptação do instrumentista pode interferir nos resultados.

As modificações envolvidas na coleta de dados estão localizadas nos fatores (variáveis independentes) descritos abaixo.

O mecanismo de alteração para a variação do volume vazio foi alterado. Utilizou-se uma placa de madeira com área pouco menor que da base acrílica. A placa está presa a dois parafusos que possuem molas abaixo dela de modo a possibilitar seu movimento na direção vertical alterando a altura da posição da placa. Com relação ao topo da caixa acrílica estas alturas foram $h_{-1}=10$ cm, $h_0=11,3$ cm e $h_{+1}=12,5$ cm, correspondendo à fração de volume vazio de 58,8, 66,5 e 73,5, respectivamente.

Estes valores proporcionaram maior variação relativa em torno do ponto central, pois conforme foi analisado no primeiro conjunto de experimentos, faz-se necessário aumentar a variação do volume vazio.

Nos trabalhos anteriores foram utilizadas boquinas disponíveis no estúdio de marcas e aberturas distintas. Neste novo conjunto de experimentos foram adquiridas boquinas de uma mesma marca e distintas aberturas: #3, #5 e #7, correspondendo aos níveis -1, 0 e +1, respectivamente.

Com relação à área de contato com a palhetas esta teve variação relativa em torno do ponto central muito grande frente aos demais fatores: 110% contra 40% para a variação da dureza das palhetas e 15% para a variação da fração vazia do volume. Modificou-se então a borracha de contato (na mordedura) com a palhetas de modo a proporcionar áreas de contato, $A_{-1} = 8,4 \cdot 10^{-5} \text{ m}^2$, $A_0 = 9,6 \cdot 10^{-5} \text{ m}^2$, $A_{+1} = 10,8 \cdot 10^{-5} \text{ m}^2$. Nestas condições tem-se uma variação relativa em torno do ponto central de 25%, bem mais próxima da variação do volume que agora atinge 22% nas condições descritas acima.

Para a gravação foi utilizado um microfone com condensador e resposta de frequência entre 20 e 20000 Hz. O software foi o Sound Forge 9.0 com taxa de amostragem de 44kHz e configuração de 16 bits.

As análises espectrais foram feitas utilizando o mesmo software nas mesmas condições dos experimentos anteriores, FFT com 2048 pontos e janela Blackman-Harris.

Com relação ao Projeto Fatorial de Experimentos foi eliminada uma variável independente por não apresentar efeito significativo na sonoridade. Trata-se da quantidade de material absorvente sonoro (estopa) no interior da caixa acrílica.

Desta forma, foi elaborado um projeto praticamente similar ao anterior: fracionado e com resolução III. Porém designado como $2^{5-2} = 2^3$. Como anteriormente teremos um total de oito (8) experimentos para analisar cinco (5) fatores. Ainda, as condições do ponto central são repetidas três (3) vezes para estimar o erro experimental. Neste projeto em particular, o padrão de fusão (*confounding patterns*) é **4=12, 5=13**.

Outra diferença está na escolha das notas em cada região. Neste conjunto foram utilizadas as notas E₅ e E₄, respectivamente para as regiões agudíssima e clarino. Manteve-se a nota E₂ na região grave (chalumeau) como anteriormente. Lembrar que a nota central é C₃.

A título de revisão, a Figura 1 mostra o conjunto do aparato experimental e a Figura 2 o detalhe da “mordedura”.

A Tabela 1 mostra os novos níveis utilizados para os fatores. A Tabela 2 mostra o Projeto Fatorial Fracionado de

experimentos elaborado para as três notas mencionadas e tendo como respostas a amplitude e freqüência da fundamental e dos componentes espectrais.

FATORES	-1	0	+1
1)Volume Vazio do Tanque Pulmão(%)	58,8	66,5	73,5
2)Dureza da Palheta (Nº)	2	2,5	3
3)Posição da Mordedura na Palheta	Interna	Centro	Externa
4)Boquilha	#3	#5	#7
5)Área de Contato com palheta (10^{-5}m^2)	8,4	9,6	10,8

Tabela 1: Níveis dos fatores utilizados nos experimentos

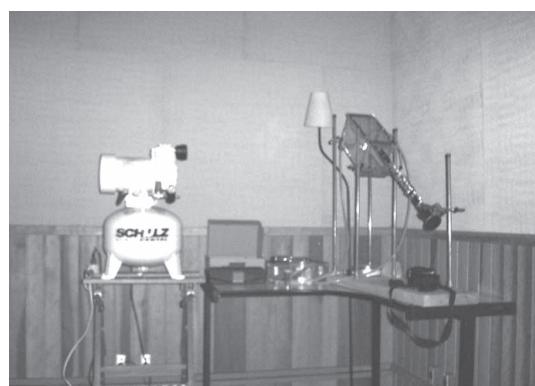


Figura 1: Visão Geral do Aparato Experimental

Experimento Nº	FATORES				
	1	2	3	4=12	5=13
1	-1	-1	-1	+1	+1
2	+1	-1	-1	-1	-1
3	-1	+1	-1	-1	+1
4	+1	+1	-1	+1	-1
5	-1	-1	+1	+1	-1
6	+1	-1	+1	-1	+1
7	-1	+1	+1	-1	-1
8	+1	+1	+1	+1	+1
9	0	0	0	0	0
10	0	0	0	0	0
11	0	0	0	0	0

Tabela 2: Tabela Geral do Projeto de Experimentos

2 RESULTADOS

As Tabelas 3 a 5 mostram os valores da amplitude das notas fundamentais e de seus respectivos componentes espectrais em cada região. As Tabelas 6 a 8 o resultado da

da nota (**E₂**). Os fatores atuam a partir do segundo componente espectral. No entanto, nota-se a significativa influência da abertura da boquilha nesta região. Enquanto que uma boquilha mais aberta tende a abaixar a freqüência dos componentes espetrais ela tende a aumentar a amplitude destes mesmos componentes espetrais.

Nesta mesma região observa-se que a interação (23) entre a dureza da palheta e a posição na mesma é significativa. A análise desta interação deve ser feita em conjunto e a título de exemplo, vamos considerar o efeito desta interação na freqüência do 5º componente espectral: quando a mordedura está na posição mais interna (menor comprimento do tubo) e passa-se de uma palheta #2 para uma #3 nota-se uma queda (6,5 Hz) na freqüência deste componente espectral. E quando se está na posição mais externa a dureza da palheta praticamente não altera a freqüência. Analogamente, quando se utiliza uma palheta #2 e afasta-se a boquilha do instrumentista não se observa alteração na freqüência do 5º componentes espectral. No entanto, nota-se uma elevação da freqüência (7,5 Hz) quando a palheta é #3 ao executar o mesmo movimento.

Para a nota **E₄** da região **clarino** nota-se que a freqüência emitida sofre maior influência da dureza da palheta inclusive sobre a fundamental. Apesar de valores distintos nos diferentes componentes espetrais, a tendência é de diminuir a freqüência dos componentes espetrais nesta região quando se passa de uma palheta #2 para a #3.

Com relação à amplitude da nota emitida observa-se a influência da dureza da palheta na fundamental e na maioria dos componentes espetrais, porém de modo contrário. Percebe-se um aumento da intensidade produzida quando se alterna de uma palheta #2 para uma #3. Não se observa de modo conclusivo as interações entre os fatores na amplitude da nota.

Para a nota **E₅** da região agudíssima nota-se que o volume vazio tem nítida influência sobre a freqüência da fundamental e dos componentes espetrais. A tendência é aumentar os valores das freqüências quando se aumenta a fração de volume vazio. Nesta região notou-se a presença das interações (23), posição na e dureza da palheta bem como destas em conjunto com a fração de volume vazio. Como envolvem interações de ordem superior (123), um projeto de experimentos completo (não fracionado) deve ser elaborado para proporcionar uma análise mais conclusiva.

Com relação às amplitudes, tanto a fração de volume vazio como a alteração de uma posição mais interna para mais externa tendem a reduzi-las, exceção para a fundamental que tende a ser elevada com as mesmas variações.

A seqüência do trabalho deve nortear dois caminhos. Um mais algébrico e outro mais qualitativo.

Para o primeiro caso deve-se mencionar que neste trabalho foram apresentados os resultados de apenas três notas. Porém, outras sete devem ser analisadas, de modo a perfazer um total de dez notas distribuídas entre as três regiões do instrumento. Desta forma, pode-se elaborar um modelo experimental para síntese da sonoridade da clarineta, incluindo extrapolações. As notas e intensidades da fundamental e dos componentes espetrais são determinadas dentro de um desvio seguindo aproximadamente a série de Fourier.

No segundo caso pode-se partir para uma análise subjetiva através de uma análise de JND (Just Noticeable

Difference) multidimensional onde o valor dos parâmetros podem ser alterados e relacionados com a sonoridade percebida. GHOUTI [5] é uma das referências que apresenta a direção a seguir.

4 REFERÊNCIAS

- [1] OLIVEIRA, L.C., GOLDEMBERG, R., MANZOLLI, J. (2005a). Estudo Experimental da Sonoridade Chalumeau da Clarineta através de Projeto Fatorial (I), *Anais da IX Convenção Nacional da AES, SP*.
- [2] OLIVEIRA, L.C., GOLDEMBERG, R., MANZOLLI, J. (2005b). Estudo Experimental da Sonoridade Chalumeau da Clarineta através de Projeto Fatorial (II), *Anais do XV Congresso da ANPPOM, RJ*
- [3] OLIVEIRA, L.C., GOLDEMBERG, R. e MANZOLLI, J. (2006). Estudo Experimental da Sonoridade Chalumeau da Clarineta através de Projeto Fatorial (I), *Anais da IX Convenção Nacional da AES, SP*.
- [4] BOX, G.E.P.; HUNTER, W.G. e HUNTER, J.S. (1978). *Statistics for Experimenters – An Introduction to Design, Data Analysis and Model Building*. John Wiley & Sons, NY.
- [5] GHOUTI, L. e BOURIDANE, A.; HUNTER, J.S. (1978). Towards a Universal Multiresolution-Based Perceptual Model. *Image Processing, 2006 IEEE – International Conference on Issue 8-11 Oct. 2006 – Page(s): 449-452.*



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Estudo de Guias de Ondas com Geometria de Dupla Reflexão Parabólica

Felipe Menezes¹ e Iuri Pepe¹

¹Universidade Federal da Bahia - UFBA
Salvador, Bahia, 40210-340, Brasil

menezes_felipe@yahoo.com.br, mpepe@ufba.br

RESUMO

Este trabalho tem por objetivo avaliar o comportamento de guias de ondas, com geometria de dupla reflexão parabólica, aplicados na resolução do problema de acoplamento das componentes de mais altas freqüências do espectro audível. Esta solução será aplicada em sistemas comerciais do tipo *line array*. Neste sentido é imperativo que seja determinando o padrão de diretividade, em função da variação do comprimento desses guias. Esse estudo foi feito em duas etapas: na primeira buscou-se uma solução para o acoplamento pelo desenvolvimento e otimização da geometria de dupla reflexão parabólica. Na segunda, foi determinada a diretividade dos guias, em função da variação de sua geometria, em particular no que diz respeito ao comprimento (L) dos guias.

0 INTRODUÇÃO

Um importante critério a ser avaliado no desenvolvimento de sistemas sonoros para ambientes fechados é o controle da proporção entre o som direto e o som reverberante. Essa proporção pode ser estrategicamente controlada através do desenvolvimento de sistemas de sonorização capazes de melhor controlar a diretividade de propagação. O uso de sistemas de fontes sonoras em linha (*line arrays*) para distribuição sonora em uma sala de audiência vem crescendo nesses últimos vinte e cinco anos. Contudo a teoria básica a respeito das linhas de fonte sonoras foi desenvolvida no inicio deste século e apresentada na literatura sobre antenas de rádio e óptica [1]. A dificuldade de transpor estes estudos, para a prática da arrumação de fontes em linha, encontrava-se na falta de tecnologia capaz de produzir uma extensão da zona de campo próximo (zona de Fresnel) nas regiões de freqüência mais altas do espectro audível [2]. Esse desenvolvimento tecnológico é recente e acompanha o

próprio crescimento do uso dos sistemas do tipo *line arrays*.

A idéia do uso de guias de ondas para controle da diretividade e o consequente aumento da extensão da zona de campo próximo é um dos frutos desse desenvolvimento, que vem sendo produzido e comercializado, nesses últimos anos, por quase todas as empresas de construção de caixas acústicas. O uso de guias de onda permitiu solucionar o problema físico da limitação da faixa de freqüência, no que diz respeito a sua altura, executadas em campo próximo para os ouvintes da sala. Problema este, estreitamente ligado à impossibilidade de se possuir fontes, que dispostas lado a lado, mantenham pequenas distâncias entre os seus centros acústicos. Este trabalho busca entender e aplicar a solução proposta por Guido Noselli [3], na qual, um guia com geometria de dupla reflexão parabólica, acoplado a um drive de reprodução de altas freqüências é usado para tornar ondas sonoras coerentes, e assim simular ondas produzidas por fontes pontuais. Estendendo portanto a região do espectro audível de modo a aproveitar o

fenômeno da zona de campo próximo para uma melhor reprodução de programas sonoros em salas de audiência.

1 DESENVOLVENDO O GUIA DE ONDAS

Para aplicarmos as propriedades de uma única linha de fonte sonora, em uma distribuição em linha de fontes pontuais, precisamos adotar um dos seguintes critérios [4 e 5]: A área de radiação individual das fontes precisa cobrir mais que 80% da área total da linha das fontes. A distância entre os centros acústicos das fontes sonoras individuais deve ser menor que $1/6 F$, sendo F a freqüência (em kHz) e a distância entre os centros acústicos (em metros). Na solução proposta por Guido Noselli a frente de emissão de onda de um drive de altas freqüências é estendida através de duas reflexões em superfícies parabólicas, fazendo com que a área de cobertura da radiação das fontes individuais se torne superior a 80% da área total das fontes em linha. Mantendo, ainda assim, toda a frente de onda isofásica, componente a componente.

1.1 Dupla Reflexão Parabólica (DRP)

A geometria parabólica possui duas propriedades de extrema importância para o desenvolvimento deste guia de ondas. A primeira reside no fato de que toda onda partindo do foco da parábola é refletida pela superfície côncava, paralelamente ao eixo formado pelo foco e pelo vértice da curva cônica. A segunda, diz que toda onda a ser refletida pela superfície convexa da parábola, sendo ela paralela ao eixo da mesma, é refletida com a mesma direção de uma reta auxiliar que perpassa o ponto da superfície, em que a onda foi refletida, passando também pelo foco.

Utilizando-se destas duas propriedades, o guia possui uma geometria que inicialmente transforma a fonte real (drive de altas freqüências) em uma fonte pontual virtual, graças a uma primeira reflexão com a superfície convexa da parábola central (figura 1), fazendo com que a fonte plana real simule uma fonte pontual centrada no foco da parábola central. Em seguida a onda sonora é novamente refletida, agora pela parábola externa; que tem o mesmo ponto focal que a parábola central; assim, a componente inicialmente refletida, parte para a segunda reflexão com a trajetória de uma onda que, ‘virtualmente’, parte de uma fonte pontual posicionada no foco (figura 2). Desta forma a fonte real plana torna-se outra fonte real plana de maior dimensão.

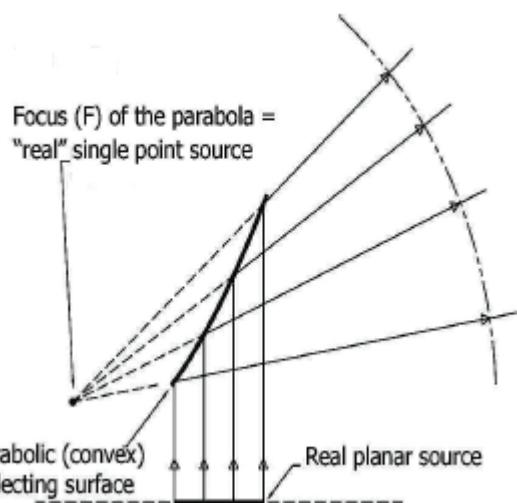


Fig 1: Transformação de uma fonte planar, em uma fonte pontual virtual, através de uma reflexão sobre superfície parabólica (ref. [3]).

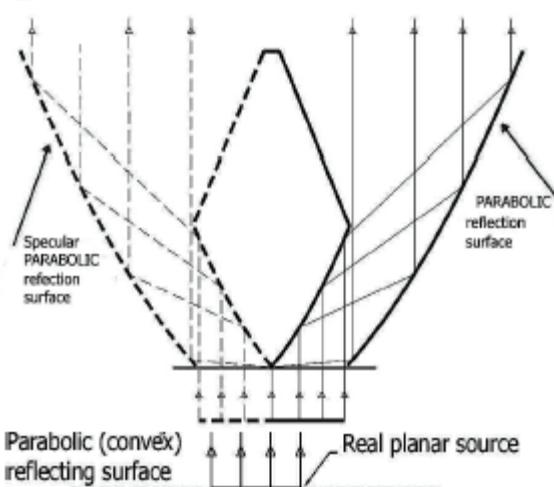


Fig. 2: Abertura a frente de onda de um guia através de uma dupla reflexão parabólica (ref. [3]).

1.2 Execução dos Protótipos

No projeto dos protótipos foi usada matemática simples, cartesiana, para a determinação do traçado das curvas dos guias. Com o interesse de conservar os demais parâmetros fixos, com exceção do comprimento L (figura 3), foram montados quatro protótipos, variando o comprimento de 5 em 5cm. O material usado para a confecção dos guias foi o papelão, com espessura de parede de 5mm.

Dimensões:

D * (diâmetro do guia) = 2'
 H * (altura do guia) = 8'
 L1 (comprimento do guia 1) = 30 cm
 L2 (comprimento do guia 2) = 35 cm
 L3 (comprimento do guia 3) = 40 cm
 L4 (comprimento do guia 4) = 45 cm
 *dimensões fixas para todos os guias

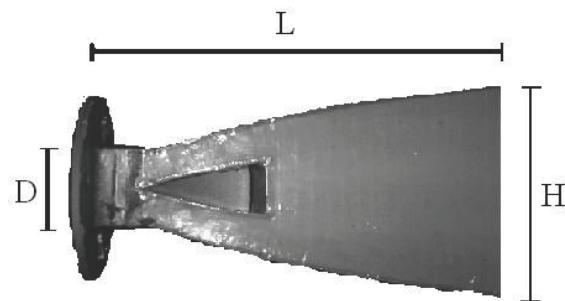


Fig. 3: Guia de ondas e suas dimensões

2 APARATO E PROCEDIMENTOS EXPERIMENTAIS

O levantamento da diretividade dos protótipos foi realizado usando o software ARTA, com o auxílio de uma placa de conexão USB de dois canais e um microfone de resposta plana, omnidirecional. As medidas foram feitas em uma sala com temperatura estabilizada em $24 \pm 1^\circ\text{C}$, de baixa reverberação, sobre um semi-círculo com raio de 1m e com passos angulares de 10°. A fonte de emissão foi colocada no centro geométrico do semi-círculo. Assim, foi possível levantar o padrão da pressão sonora em função da freqüência e do deslocamento no eixo polar. A faixa de freqüência coletada foi de 1kHz à 16kHz, salientando que em série com o transdutor (driver) foi colocado um filtro passa alta de primeira ordem, com freqüência de corte centrada em 1k6 Hz. O nível de sinal trabalhado foi de 115 dB, mensurado na frente do driver, sobre a execução de ruído rosa.

3 DADOS OBTIDOS

A seguir pode-se observar o sonograma dos quatro guias de ondas estudados:

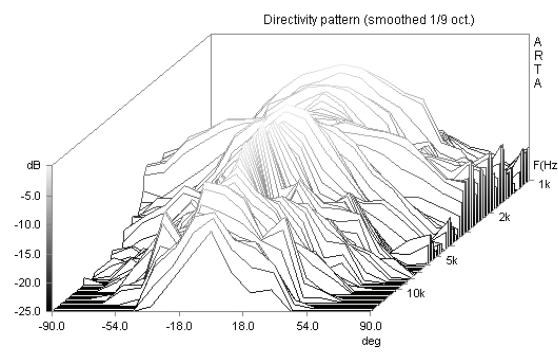


Fig 4: Waterfall do guia de ondas 1.

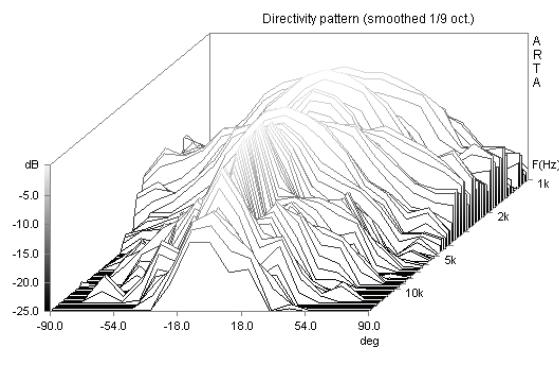


Fig 5: Waterfall do guia de ondas 2.

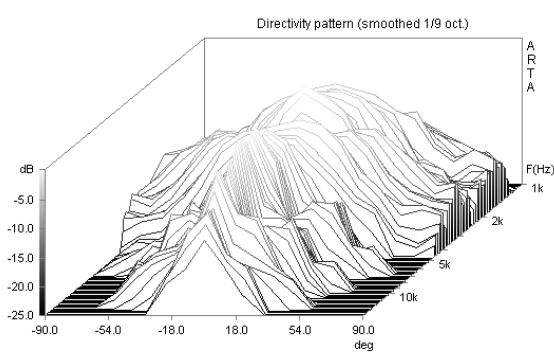


Fig 6: Waterfall do guia de ondas 3.

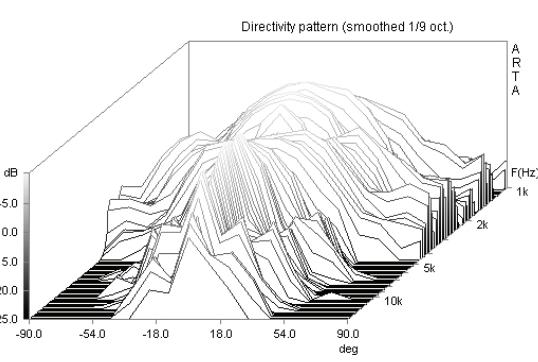


Fig 7: Waterfall do guia de ondas 4.

4 ANÁLISE DOS RESULTADOS

Analisado os waterfalls de diretividade dos guias estudados, observa-se que todos os guias apresentam uma zona de dispersão na região entre 3kHz e 4 kHz. Após esta região o espectro tende a ter uma diretividade bem centrada, para qualquer um dos guias, mostrando que todos os protótipos podem ser aplicados acima de 4kHz.

Pode-se inferir também um aumento de diretividade nas regiões de mais altas freqüências (região acima de 10kHz) em função do aumento do comprimento do guia de ondas e o aumento da dispersão nas regiões de mais baixa freqüência do espectro, em relação a este mesmo parâmetro. Resultado compreensível, já que, com uma maior omnidirecionalidade as freqüências mais baixas se

tornam mais facilmente incoerentes entre si em um guia de onda com maior superfície refletora.

Ao que parece, tem-se uma região de sintonia para a diretividade de cada guia de onda, esta região está deslocada no espectro em função da variação do comprimento dos guias.

5 CONCLUSÃO

Nos dias atuais a utilização de guias de ondas para o controle de diretividade das regiões de alta freqüência do espectro audível e aplicação nos sistemas de linhas de fontes sonoras é uma realidade. O estudo de soluções de controle das propriedades dessas linhas de propagação de fontes sonoras torna-se necessário, tendo em vista a popularidade adquirida por essa tecnologia em todo mundo e em particular no Brasil.

O desenvolvimento deste guia desvendou uma forma de controle da diretividade de um radiador de altas freqüências (a partir de 4 kHz) através da dupla reflexão da onda em superfícies parabólicas.

A engenharia da geometria do guia permite projetá-lo com parâmetros que melhor o adéquam ao projeto do sonofletor no qual o guia irá atuar, levando em conta as dimensões desejadas, além, é claro, da região de freqüências a serem irradiadas pelo guia. Para o ajuste desses parâmetros, deve-se assumir um dado compromisso, levando em conta a aplicação a qual se destina o guia em questão.

A dispersão de diretividade na faixa de freqüência entre 3 kHz e 4 kHz, muito bem marcada nos waterfalls de todos os guias, é um fato que desperta interesse, já que não existe razão aparente para tal comportamento. Podendo essa dispersão ser decorrente do material usado na confecção dos protótipos, ou simplesmente uma influencia de reverberação da sala. Assim, estes questionamentos abrem espaço para a continuação deste estudo.

6 REFERÊNCIAS

- [1] WOLF, I and MALTER, L. *Directional Radiation of Sound*, J. Acoustical Society of America. Vol 2, October, 1930, p.201.
- [2] LEO L. BERANEK. *Acoustics*, Mc Graw- Hill Book Company, Inc. 1954C.
- [3] G. NOSELLI, *Reflective Wave Guides for the reproduction of high frequencies*
- [4] M. URBAN, C. HEIL, AND P. BAUMAN. *Wavefront Sculpture Technology*, Presented at the 111th Aes Convention, New York, September 21-24, 2001.
- [5] HEIL, and M. URBAN. *Sound Fields Radiated by Multiple Sound Sources Arrays*, presented at the 92nd AES Convention, Vienna, March 24-27, 1992



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Técnicas de processamento de Áudio em sinais de voz, para Auxílio Diagnóstico de Doenças Laríngeas

R. D. R. Fagundes¹, I. C. Zwetsch¹ e D. Scolari¹

¹IDEIA-PUCRS, Prédio 30, sala 301-03, Av. Ipiranga 6681, Porto Alegre, RS, 90619-900, Brasil

rubemdrf@sapienstek.com, iuberi@bol.com.br,

RESUMO: O presente trabalho apresenta técnicas de DSP (Digital Signal Processing) para análise de vozes com doenças laríngeas, para obter-se um método computacional eficiente de identificação dos distúrbios da laringe através das alterações na análise Cepstral e do reconhecimento via DD-HMM (discrete density Hidden Markov Model). Esta metodologia de análise e diagnóstico laríngeo obteve resultados com precisão superior a 80% na avaliação de casos reais.

1 INTRODUÇÃO

Este trabalho apresenta um modelo de análise da voz, com uso da técnica Cepstral, como método de caracterização e diagnóstico de distúrbios laríngeos. O diagnóstico destas alterações é atualmente realizado principalmente pelo exame de videolaringoscopia. Certas doenças, mesmo para médicos especialistas experientes, podem trazer dificuldade diagnóstica, pois às vezes são muito parecidas no aspecto, apesar de apresentarem origens e alterações fisiopatológicas diferentes.

Estas dificuldades também são encontradas na realização de técnicas computacionais de processamento de sinais que, em determinados casos, não são eficientes o suficiente para a diferenciação das alterações.

No presente trabalho, aplicou-se a análise Cepstral com DD-HMM em casos normais e nas seguintes doenças: nódulo vocal, cisto vocal, pólipos vocais, edema de Reinke e sulco vocal.

Estas representam a grande maioria dos atendimentos de pacientes com alteração da voz, que não sejam as alterações transitórias por infecções das vias respiratórias, onde temos em alguns dias a melhora do quadro geral.

Na maioria dos casos, as doenças citadas produzem uma rouquidão com características típicas de cada uma, principalmente quando analisadas por ouvintes

mais experientes, tais como médicos otorrinolaringologistas ou profissionais da área da fonoaudiologia.

Várias técnicas de análise de sinais da voz são estudadas para a identificação de alterações da laringe [1][2][3][4][5][6][7][8].

A proposta deste trabalho é utilizar a análise Cepstral com DD-HMM como método de análise das alterações acústicas da voz nas doenças da laringe.

A análise Cepstral do sinal de voz para o estudo de distúrbios da laringe é muito útil, permitindo trabalhar com o sinal da glote (excitação) separadamente das repercuções resonantes do trato vocal, facilitando o entendimento das alterações que as doenças causam nas pregas vocais.

Esta técnica de DSP no estudo do sinal acústico da emissão da voz permitirá detectar modificações nas ondas que se relacionem com as doenças e, consequentemente, a criação de modelos para uma classificação, permitindo a obtenção de ferramenta de diagnóstico não-invasiva.

É utilizado o sistema DD-HMM de reconhecimento de padrões para distinguir as diferenças Cepstrais de cada doença, e com isto, têm-se um sistema totalmente automático, não-invasivo de diagnóstico de alterações laríngeas.

FUNDAMENTAÇÃO TEÓRICA

2.1 ANÁLISE CEPSTRAL

A análise Cepstral do sinal de voz permite trabalhar com o sinal da glote (excitação) e do trato vocal (ressonância) separadamente, pelas suas propriedades homomórficas, facilitando o estudo das alterações que as doenças causam nas pregas vocais. Temos a separação das características do filtro do trato vocal da seqüência de excitação.

O modelo mostrado na figura 1, é o mais freqüentemente usado, onde se assume que o sinal de voz, $s(t)$ é composto por um sinal de excitação $e(t)$ aplicado ao filtro do trato vocal, com uma resposta impulsional $v(t)$.

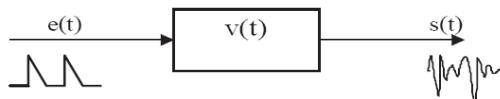


Fig.1: Modelo simplificado da produção da voz no domínio do tempo.

Onde $s(t)$ é a convolução de $e(t)$ com $v(t)$, que é definida por:

$$s(t) = e(t) \otimes v(t) \quad (1)$$

No domínio da freqüência a equação é definida como:

$$S(\omega) = E(\omega) \times V(\omega) \quad (2)$$

Em que $S(\omega)$, $E(\omega)$ e $V(\omega)$ são as transformadas de Fourier das funções contínuas no tempo $s(t)$, $e(t)$ e $v(t)$ ou as transformadas discretas de Fourier das seqüências de amostras temporais $s(n)$, $e(n)$ e $v(n)$. Assim, para executarmos uma análise Cepstral do sinal de voz, o sinal $s(t)$ será processado via Transformada de Fourier (usualmente uma FFT) resultando o espectro de freqüência $S(\omega)$

Contudo, o sinal de excitação $E(\omega)$ e o filtro do trato vocal $V(\omega)$ não podem ser diretamente identificados no sinal de voz $S(\omega)$ resultante, visto que o sinal de voz é a resposta em freqüência do trato vocal (modelado algebricamente como um filtro) pela função excitação $E(\omega)$. Como se deseja determinar as alterações laríngeas a partir dos efeitos analisados no sinal de voz, será necessário dissociar os efeitos da excitação e os feitos do trato vocal diretamente do sinal $S(\omega)$. Neste sentido, a análise Cepstral, descrita a seguir, propiciará esta dissociação:

Lembrando a propriedade matemática dos logaritmos:

$$\log(a \times b) = \log(a) + \log(b) \quad (3)$$

Assim, tomindo o logaritmo de $S(\omega)$ e aplicando (2):

$$\log(S(\omega)) = \log(E(\omega) \times V(\omega)) \quad (4)$$

e, a seguir (3):

$$\log(E(\omega) \cdot V(\omega)) = \log(E(\omega)) + \log(V(\omega)) \quad (5)$$

Na expressão (5) o sinal de voz $S(\omega)$ está sendo apresentado em sua forma logarítmica. No entanto, as componentes da excitação e do trato ainda são indistinguíveis.

Então, lembrando que:

$$\mathfrak{J}^{-1}(f + g) = \mathfrak{J}^{-1}(f) + \mathfrak{J}(g)^{-1} \quad (6)$$

A propriedade aditiva do espectro logarítmico continua a se verificar quando lhe for efetuada a transformada inversa de Fourier, pela aplicação de (5) em (6), sendo o resultado dessa operação chamada de função Cepstral ou Cepstro.

O processo resumido de estimativa do Cepstro pode ser visto no diagrama de blocos apresentado na figura 2:

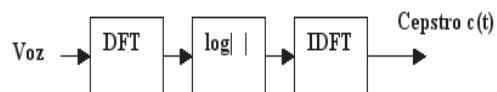


Fig.2: Diagrama de blocos da estimativa do Cepstro[18].

A concentração da componente periódica no espectro logarítmico de um sinal de voz num intervalo de freqüências equivalente ao inverso do período fundamental T, aparece no Cepstro como um pico.

O eixo horizontal, da função Cepstral tem dimensões temporais e o nome de quefrâncias. Com isto, na voz se obtém uma clara distinção entre a componente de excitação e a contribuição do trato vocal, que aparece como um aglomerado de componentes aos baixos valores de quefrâncias afastado da componente do período fundamental que aparece em valores mais altos de quefrâncias.

Na figura 3 vê-se o Cepstro de um segmento de voz onde o pico correspondente ao período fundamental está próximo da quefrância de 10 ms, separado das componentes do trato vocal às de baixas quefrâncias. Nesta figura são apresentadas apenas as componentes do Cepstro superiores a sensivelmente 1 ms, pois as componentes de mais baixas quefrâncias têm valores comparativamente muito superiores aos restantes e a sua apresentação não deixaria claro o pico correspondente à freqüência fundamental [9][10] [11] [12] [13] [14].

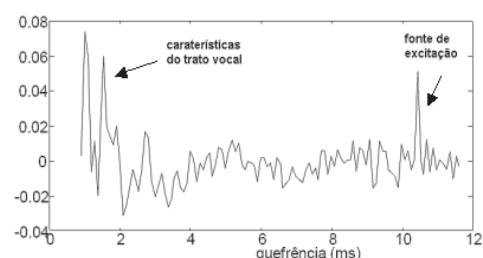


Fig.3: Cepstro de um segmento de fala.

2.2 LIFTTERING

A função de transferência do trato vocal e a função de excitação da voz aparecem em partes separadas da escala de quefrâncias, e podem ser separadas em duas funções, pelo processo de “lifteragem”, facilitando o estudo individualizado das alterações na excitação e da parte ressonantal, como apresentado na figura 3.

2.3 QUANTIZAÇÃO VETORIAL

Na quantização vetorial por amplitude, a idéia básica é indexar um vetor em um código numérico, como mostrado na figura (4), onde os eixos são as amplitudes dos coeficientes [1].

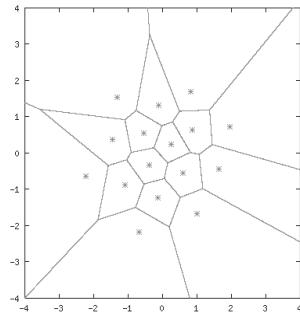


Fig. 4. Exemplo de Quantização Vetorial de 4 bits e 2 dimensões.

Vetores de dimensão N de um determinado espaço vetorial podem ser agrupados em células. Cada célula tem um vetor representante de dimensão N chamado de *code word*, e cada *code word* é representado por um índice numérico de 1 a M, onde M, é o número máximo de células. Para treinar os *code words*, será aplicado o algoritmo de LBG (Linde-Buzo-Gray) utilizando-se distâncias euclidianas. O algoritmo provê todas os *codewords* treinados em uma tabela chamada *codebook* [1,15].

A fig. (4) mostra o sistema de reconhecimento de um DD-HMM utilizando a quantização vetorial.

$$d(x, y) = \sum_{i=1}^p (x_i - y_i)^2 \quad (7)$$

A expressão (7) mostra a fórmula da distância euclidiana, onde x e y são vetores dimensionais N.

Vetores de 10 coeficientes, provenientes do Cepstro, são quantizados pelo Quantizador Vetorial e utilizados no DD-HMM.

2.4 DD-HMM

Um dos problemas do DD-HMM é achar a seqüência de estados $S = S_1, S_2, \dots, S_N$, onde N é o número de estados individuais, que melhor modela a seqüência de observações de entrada

$O = O_1, O_2, \dots, O_M$, onde M é o número de símbolos de observação distintas por estado. É denotado o estado no tempo t como q_t e os símbolos individuais como $V = \{V_1, V_2, \dots, V_M\}$. A seqüência de observação de entrada é composta por *code words*, quantizados pelo

algoritmo de LBG. Um DD-HMM é caracterizado por $\lambda = (A, B, \pi)$, onde $A = \{a_{ij}\}$ é a matriz de probabilidades de transição de estados e $a_{ij} = P(q_{t+1} = j | q_t = i)$, que é a probabilidade de transição do estado i no tempo t para o estado j no tempo t+1, para $1 \leq i, j \leq N$; $B = \{b_j(k)\}$ é a matriz de probabilidade de observação e $b_j(k) = P(O_t = V_k | q_t = j)$, que é a probabilidade de observar o símbolo V_k estando no estado j no tempo t, para $1 \leq k \leq M$ e para $j = 1, 2, \dots, N$. A matriz B define a função de probabilidade de distribuição para cada estado (fig. 5); $\pi = \{\pi_i\}$ é a matriz de probabilidades iniciais e $\pi_i = P(q_1 = i)$, para $1 \leq i \leq N$ [1].

Para treinar as matrizes A e B, é usado o algoritmo de Baum-Welch e para achar a melhor seqüência de estados para uma dada seqüência de observações é utilizado o algoritmo de Viterbi (fig. 6) [1,15].

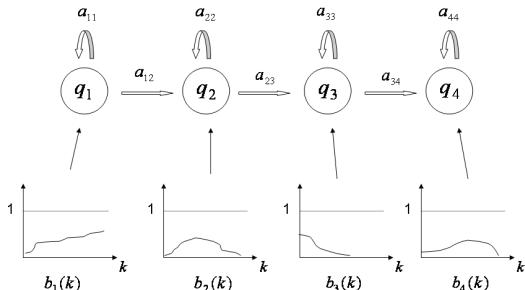


Fig. 5. Exemplo de estrutura DD-HMM de 4 estados.

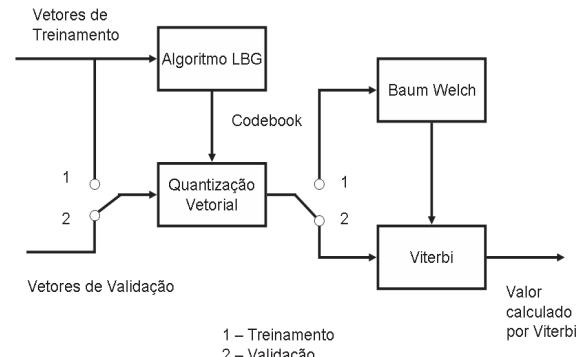


Fig. 6. Estrutura de treino e validação do DD-HMM.

2.5 FISIOLOGIA DAS DOENÇAS DAS PREGAS VOCAIS

A qualidade da voz depende do modo de fechamento e abertura da glote e da vibração das pregas vocais. Certas doenças laríngeas impedem que pregas vocais tenham uma vibração glotal harmônica, resultando em uma área vocal onde o trato vocal é excitado em duas freqüências fundamentais diferentes[10,16].

3 MATERIAIS E MÉTODOS

Foram utilizados sinais acústicos gravados de pacientes com alterações vocais e de normais, atendidos em consultório, que realizaram videolaringoscopia. A distribuição das doenças usadas como modelos das alterações acústicas foi 10 cistos femininos, 5 edemas de Reinke femininos, 7 nódulos femininos, 5 sulcos (3 masculinos e 2 femininos) e 7 pólipos masculinos.

Também foram obtidos 8 sinais femininos e 5 masculinos de vozes sem lesões nas pregas vocais e sem alteração acústica perceptível, que compõem o grupo normal. Todos os sinais de voz foram adquiridos em três vogais, "A", "E" e "I"[17].

O sinal acústico foi digitalizado em modo mono, com frequência de amostragem de 22 KHz(decorrente de gravações das vozes em consultório em 44KHz e que foram submetidas a downsampling para 22KHz) e 16 bits de resolução[1,9, 15,17].

Os dados obtidos foram submetidos à análise de processamento digital de sinais em programa rotina de análise Cepstral no Matlab, que consta dos seguintes passos (Figura 7):

- 1)obtenção de fragmento da vocalização e janelamento
- 2)análise Cepstral
- 3)análise por DD-HMM com as três vogais de cada paciente.

A pré-ênfase que permite a filtragem de sons labiais não foi usada, porque altera também o sinal de excitação (da glote), fenômeno descrito na literatura e constatado também neste trabalho[17].

O janelamento foi ajustado para a análise do frame da vocalização



Fig.7: Esquema em blocos da análise do sinal acústico.

4 RESULTADOS

Os achados Cepstrais da análise das vocalizações de cada alteração diagnosticada foram analisados automaticamente utilizando a técnica de DD-HMM. Com a ajuda das características fisiopatológicas das alterações laringeas já conhecidas foram descritos os achados que podem ser particularmente atribuídas às diferentes doenças.

Na comparação entre as cinco alterações laringeas em estudo é possível observar que existem diferenças no perfil Cepstrográfico, quando comparadas com o grupo normal e também entre si. Tais diferenças são mais evidentes no edema de Reinke e no sulco vocal.

Para a análise automática do DD-HMM, 70% dos dados foram utilizados para treinamento e o restante foi utilizado para validação do sistema. O *codebook* foi treinado com os dados de treino com 32 vetores código (*code vectors*), cada vetor com 10 elementos, apresentado a distribuição na figura 8, com eixos representando o centróide e sua ocupação de vetores, pois uma distribuição

equilibrada é fundamental para o treinamento de DD-HMM:

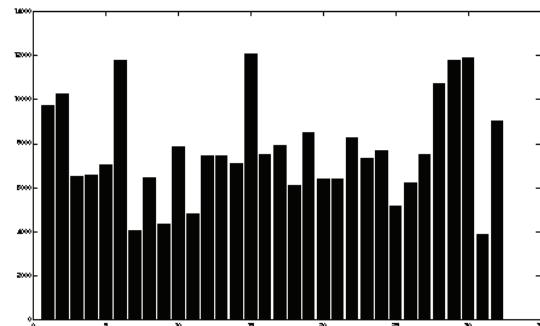


Fig.8: Gráfico de Ocupação do codebook apóis treino.

O sistema de DD-HMM formado por 9 estados, analisou as vogais "A", "E" e "I" e treinou as matrizes de transição de estados e a matriz de probabilidade de observação pelo algoritmo de Baum-Welch. Em seguida o sistema foi validado pelo algoritmo de Viterbi, o qual compara a probabilidade da seqüência analisada ser representada por uma doença específica.

O resultado obtido de validação foi 81,9% de acerto no máximo. Esse resultado valida significativamente o método proposto, contribuindo de forma efetiva para a elaboração de um protocolo de diagnóstico laringeo, baseado na técnica de análise Cepstral.

5 CONCLUSÃO

Algumas doenças geram alterações significativas nos achados do Cepstro outras nem tanto.

Os achados Cepstrográficos do sulco vocal e do edema de Reinke, são os que mais se diferenciam de todas as outras doenças. Destaca-se também que algumas doenças apresentam alterações características e constantes que servem como método diagnóstico. A tabela 1, apresenta de maneira condensada, os dados das alterações e os achados do método proposto, comparando as cinco doenças e um caso normal com determinados fragmentos da vocalização

Com os resultados da proposta de análise Cepstral com DD-HMM do sinal da voz iniciados neste trabalho, permite-se prever que este método será uma ferramenta diagnóstica muito útil e promissora, pois é um método não-invasivo, de custo mais baixo e fácil execução. Para uma maior qualificação deste método, deve-se realizar trabalhos com a inclusão de mais amostras, de outras doenças mais raras e de realização de estudo científico estatisticamente adequado para a validação, como um estudo prospectivo duplo cego.

Tabela 1. Doenças das Pregas Vocais, suas características principais e aspectos diagnósticos pelos métodos convencionais e o proposto na pesquisa.

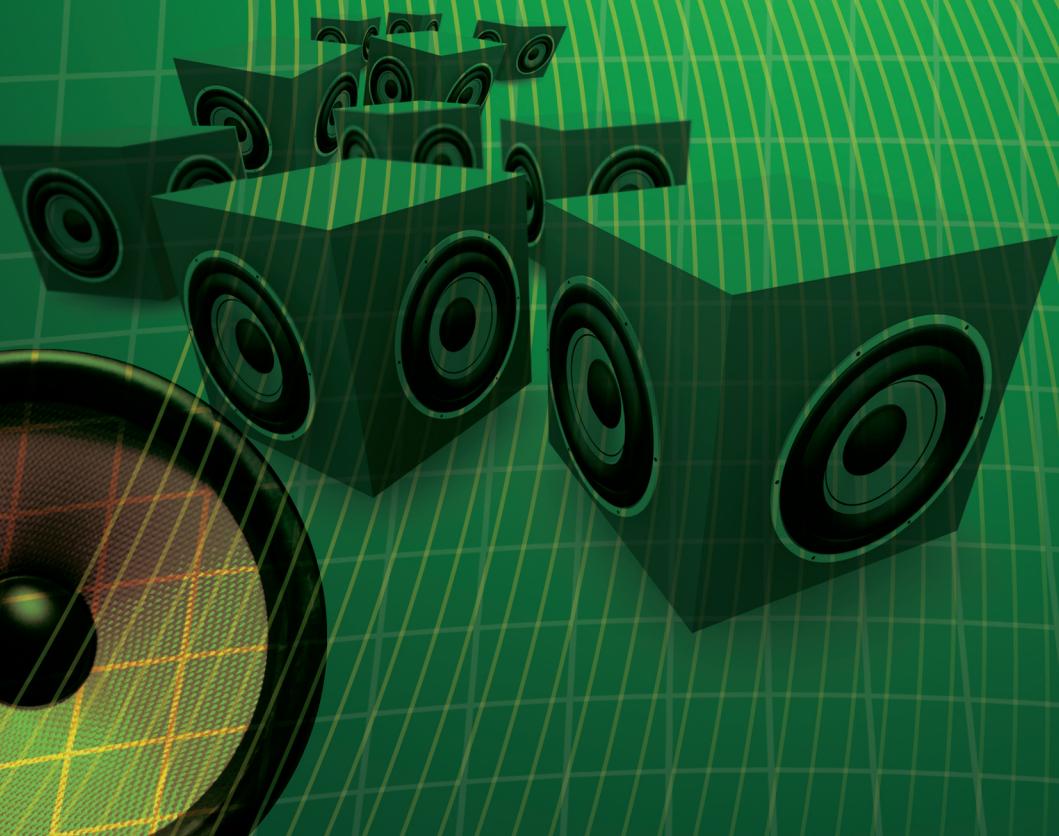
Definição ^(2,3,5)	Características acústicas ^(2,3,5)	Imagen da prega vocal	Sistema Cepstral de análise da voz (proposto)
NORMAL	Pregas vocais sem lesões	Som perceptivelmente adequado	
Cisto	aumento da massa Lesão: cística, vibratória, rigidez e com fluido em geral unilateral	Lesão: cística, vibratória, rigidez e com fluido em geral unilateral	
EDEMA	aumento da massa, onda aperiódica, rigidez diminuída Lesão: edema da mucosa das pregas	Lesão: edema da mucosa das pregas	
Sulco	diminuição da massa Lesão: falha em forma de sulco, uni ou bilateral	Lesão: falha em forma de sulco, uni ou bilateral	
Nódulo	Interfere na vibração dependendo de sua firmeza, aumento de massa Lesão: quase sempre bilateral e simétrica	Lesão: quase sempre bilateral e simétrica	
PÓLIPO	aumento da massa vibratória, aumento da rigidez em geral, vibração assimétrica e aperiódica Lesão: em geral unilateral	Lesão: em geral unilateral	

6 REFERÊNCIAS

- [1]. Rabiner RL, Schafer R. *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
- [2]. Rabiner R.L.,Gold B. *Theory and Application of Digital Processing*, Prentice Hall, 1975.
- [3]. Gomez P, Godino JI, Rodriguez F, et al. *Evidence of Vocal Cord Pathology from the mucosal wave Cepstral contents*, IEEE, Madrid, 2004;45-52.
- [4]. Wilpon JG, Rabiner LR, Lee CH et al. "Automatic recognition of keywords in unconstrained speech using Markov Models", *IEEE Transactions on Acoustic, Speech and Signal Processing* , Vol.38, N° 11, Nov.1990;1870-1878.
- [5]. Hadjittodorov S, Mitev, P. "A Computer system for acoustic of pathological voices and laryngeal diseases screening, Technical note", *Medical Engineering , & Physics* , Sofia, n. 24, 2002;419-429.
- [6]. Manfredi C, D`aniello M, Bruscaglioni P, et al. "A comparative analysis of fundamental frequency estimation methods with application to pathological voices", *Medical Engineering & Physics*, Firenze, Italy,n. 22, 2000, 135-147.
- [7]. Wszołek W, Tadeusiewicz R, Izworski A, et al. "Automated understanding of selected voice tract pathologies based on the speech signal analysis", *Proceedings of the 23rd Annual EMBS International Conference, EMBS, Istanbul, October*, 2001;25-28.
- [8]. Mitev P, Hadjittodorov S. "Fundamental frequency estimation of voice of patients with laryngeal disorders", *Information Sciences* , Sofia, n. 156, 2003;3-19.
- [9]. Rosa MO, Pereira JC, GrellerM and CarvalhoA."Signal processing and statistical procedures to identify laringeal pathologies", *IEEE Transactions on Biomedical Engineering IEEE/EESC-USP*, 1999; 423-426.
- [10]. Dedivitis RA.Barros, APB.Métodos de Avaliação e Diagnóstico de Laringe e Voz, 2^{ed} Lovise, São Paulo, 2002.
- [11]. Minoru H. Diane MB. Exame videoestroboscópico da laringe; Porto Alegre, Artes Médicas,1997.
- [12]. Erich CM, LupercioLB, Osíris CB, et al. Incidência de lesões laríngeas não neoplásicas em pacientes com queixas vocais; Revista Brasileira de Otorrinolaringologia, vol.67, n.6, nov/dez 2001;788-94.
- [13]. Khul I. Manual prático de Laringologia;Editora da universidade,Porto Alegre1982.
- [14]. Martinez CE, Rufiner HL. "Acoustic analysis of speech for detection of laryngeal pathologies", *Proceedings of the 22nd Annual EMBS International Conference*, EMBS, Chicago, July 2000,23-28.
- [15]. Fagundes R.D.R. Reconhecimento de Voz, Linguagem Contínua , usando Modelos de Markov, Dissertação de Mestrado, Universidade de São Paulo, São Paulo, 1993.
- [16]. Hansen JH, Gavdida-Ceballos L, Kaiser RJF. "A Nonlinear operator-Based Speech Feature Analysis Method with Application to Vocal Fold Pathology Assessment". *IEEE Transactions on Biomedical Engineering*, Vol. 45, N°. 3, March 1998;937-940.
- [17]. Furui S. *Digital Speech Processing, Synthesis and Recognition*, Marcel Dekker,Inc.,2001.
- [18]. Deller, J, Hansen J, Proakis J, Discrete-Time Processing of Speech Signals, IEEE Press, 1999.

Sessão 2

Acústica e áudio espacial
(Acoustics and spatial sound)





Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Análise modal bidimensional de salas com superfícies seriais difusoras

José Augusto Mannis

Instituto de Artes - Universidade Estadual de Campinas (Unicamp)

Campinas, SP, 13083-854, Brasil

jamannis@uol.com.br

RESUMO

O escopo deste artigo restringe-se à influência das superfícies seriais difusoras sobre o comportamento dos modos normais de uma sala a partir de dados obtidos por comparação entre resultados de análise modal de uma *sala de referência* quadrangular (QUAD) com esta mesma, porém modificada em uma de suas partições: uma vez com superfície serial linear (LNSS), e outra com uma superfície serial com elementos semicilíndricos (SCSS). Os resultados finais indicam a banda de freqüência a partir da qual se constata a influência das superfícies seriais sobre os modos normais, bem como a atuação das irregularidades semicilíndricas de SCSS. Observa-se ainda uma distribuição variada de ventres de máximos e mínimos de pressão, além da tendência destes a se localizarem junto às superfícies seriais.

0 INTRODUÇÃO

As superfícies seriais difusoras são produto de pesquisa de doutorado deste autor [1] tendo por objetivo melhorias para o conforto e a estética acústica de locais para audições e performances musicais através do espalhamento sonoro sem perda de energia, tendo sido objeto de depósito de pedido de patente pela Unicamp. Estas superfícies são concebidas através de processo serial inspirado nos procedimentos de composição musical com 12 sons idealizado por Arnold Schoenberg [2], na primeira metade do Séc. XX, e nas seqüências numéricas empregadas nos difusores baseados na reflexão com interferência de fase [4][5][6][7], idealizada por Manfred R. Schroeder, na década de 1970. As superfícies seriais difusoras possuem séries numéricas aplicadas em diversos parâmetros geométricos de *design*, atuando na origem e no desenvolvimento de sua estrutura. Além do espalhamento sonoro buscando potencializar a reverberação e dar máxima abrangência às fontes sonoras em todas as localizações em que estas se situarem, as superfícies seriais difusoras, por alterar o perfil dos limites do volume interno, terão impacto sobre os modos normais de uma

sala. Para que a ação dos modos normais não seja prejudicial à resposta acústica de uma sala é necessário que a quantidade de modos (axiais, tangenciais e oblíquos) [8] acumulada em cada banda de freqüência seja uniforme e regularmente crescente, dos graves aos agudos, de acordo com o critério de Bonello [9][10]. Quanto à distribuição e à dinâmica dos ventres de pressão, quanto menor for o acúmulo sistemático de ventres de máximos e mínimos de pressão num mesmo local, melhor a resposta acústica e mais homogênea será a escuta em variados pontos da sala, o que é desejável para a performance e audição musical. Salas quadrangulares podem soar bem quando têm distribuição e dinâmica dos ventres de pressão favoráveis em função das medidas do local, estas podendo ser analisadas tanto pelo critério de Bonello quanto pelo diagrama de Bolt, Beranek e Newman [11]. Porém, Nieuwland e Weber [12] realizaram pesquisas sobre câmaras reverberantes no *Philips Research Laboratories of Eindhoven* (Países Baixos) e concluíram que em salas não retangulares, a estrutura espacial do nível de pressão sonora dos modos é irregular. Nesse sentido, a presença de superfícies seriais difusoras pode justamente ter impacto positivo sobre o comportamento dos modos normais em

uma sala quadrangular, o que pode ser avaliado através de análise modal comparativa. As superfícies seriais envolvidas neste experimento são do tipo: LNSS – *superfície serial difusora linear* composta de painéis de larguras variáveis, se articulando por eixos paralelos em uma dimensão, perpendiculares ao sentido do alinhamento dos painéis; SCSS – *superfície serial difusora com elementos semicilíndricos*, consistindo no alinhamento inclinado de uma série de tubos em meia cana fixados contra uma superfície de fundo.

1 DESIGN DO CONTORNO DAS SALAS E PREPARAÇÃO PARA ANÁLISE MODAL

Através de aplicativo de *design* e análise do Laboratório de Mecânica Computacional da Faculdade de Engenharia Mecânica da Unicamp, permitindo predição de comportamento de sistemas mecânicos, foi efetuada simulação para observar os modos atuando em uma *sala de referência* quadrangular (QUAD) (Figura 1), sem tratamento, comparativamente a duas soluções de superfícies seriais difusoras: LNSS (Figura 3) e SCSS (Figura 4), onde somente uma das paredes recebeu tratamento para proporcionar difusão. Para avaliar a atuação da serialização dos detalhes em SCSS, foi ainda simulada uma sala contendo, ao invés da superfície de elementos semicilíndricos, uma parede sobre os eixos dos círculos que seccionam os cilindros RLN_SCSS (Figura 5).

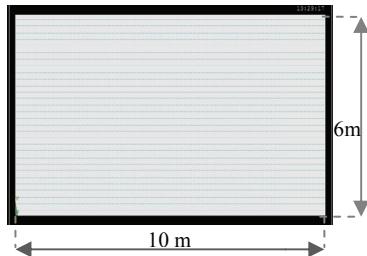


Figura 1 - Sala de referência quadrangular (QUAD).

Sala de Referência (QUAD)	Regiões
Hz	
0 a 28	X
28 a 150	A
150 a 599	B
a partir de 599	C

Tabela 1 – As quatro regiões (X, A, B e C) resultantes das medidas selecionadas para a *sala de referência* em análise: 10x6x4m

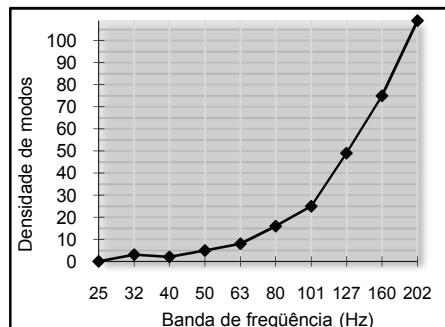


Figura 2 – Sala de referência (QUAD): Densidade de modos por banda de freqüência – Critério de Bonello, Cálculo considerando os 7 primeiros parciais das ondas estacionárias em cada dimensão.

As medidas da *sala de referência* quadrangular QUAD foram definidas não buscando uma sala defeituosa, mas, ao contrário, considerando ao mesmo tempo o Critério de Bonello (Figura 2) e o Diagrama de Bolt, Beranek e

Newman. O resultado selecionado foi 10x6x4m (L x P x H). De acordo com Everest [7] cada sala possui uma região A dominada pelos modos normais, caracterizada por um comportamento do som como onda; uma região C caracterizada por trajetórias com reflexões especulares, ou seja, comportamento do som como raio; uma região B dominada pela difração e difusão, na qual o som se comporta de forma transitória entre onda e raio; e uma região X onde não se sabe bem o que acontece. As medidas aqui selecionadas implicam, conforme Everest [8], nas regiões detalhadas na Tabela 1.

Para efeito da análise modal bidimensional o parâmetro altura foi desconsiderado após os cálculos.

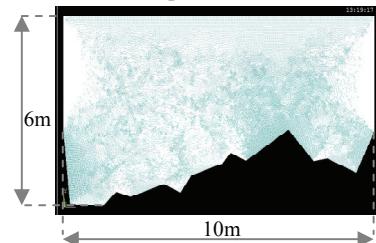


Figura 3 – LNSS: Sala com uma parede modificada – Superfície serial linear.

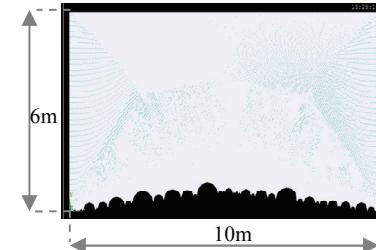


Figura 4 – SCSS: Sala com uma parede modificada – Superfície serial com elementos semicilíndricos.

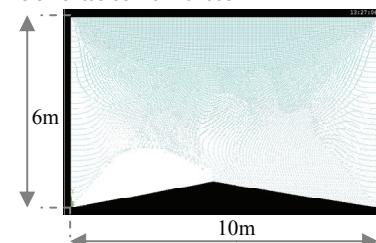


Figura 5 – RLN_SCSS: Transformação da superfície SCSS (Figura 4) reduzida a modelo linear, contendo somente os eixos unindo os centros dos círculos que seccionam os cilindros.

15,06°C	temperatura	
c	340 m/s	velocidade de propagação do som no ar
p ₀	2.10 ⁻⁵ Pa	pressão de referência
ρ	1,2 Kg/m ³	densidade do ar no nível do mar
	0	admitância dos limites do contorno
α	0 [Sabines métricos]	coeficiente de absorção do material de revestimento das superfícies

Tabela 2 – Parâmetros considerados nos cálculos preliminares e na análise modal.

2 RESULTADOS DA ANÁLISE MODAL

Os máximos e mínimos de pressão aparecem em tom escuro. Tons claros indicam menor variação de pressão. Contudo, o que importa na análise é como se comportam as regiões mais escuras (os ventres com máximos e mínimos de pressão) destacadas e, portanto, facilmente visualizadas em todas as figuras.

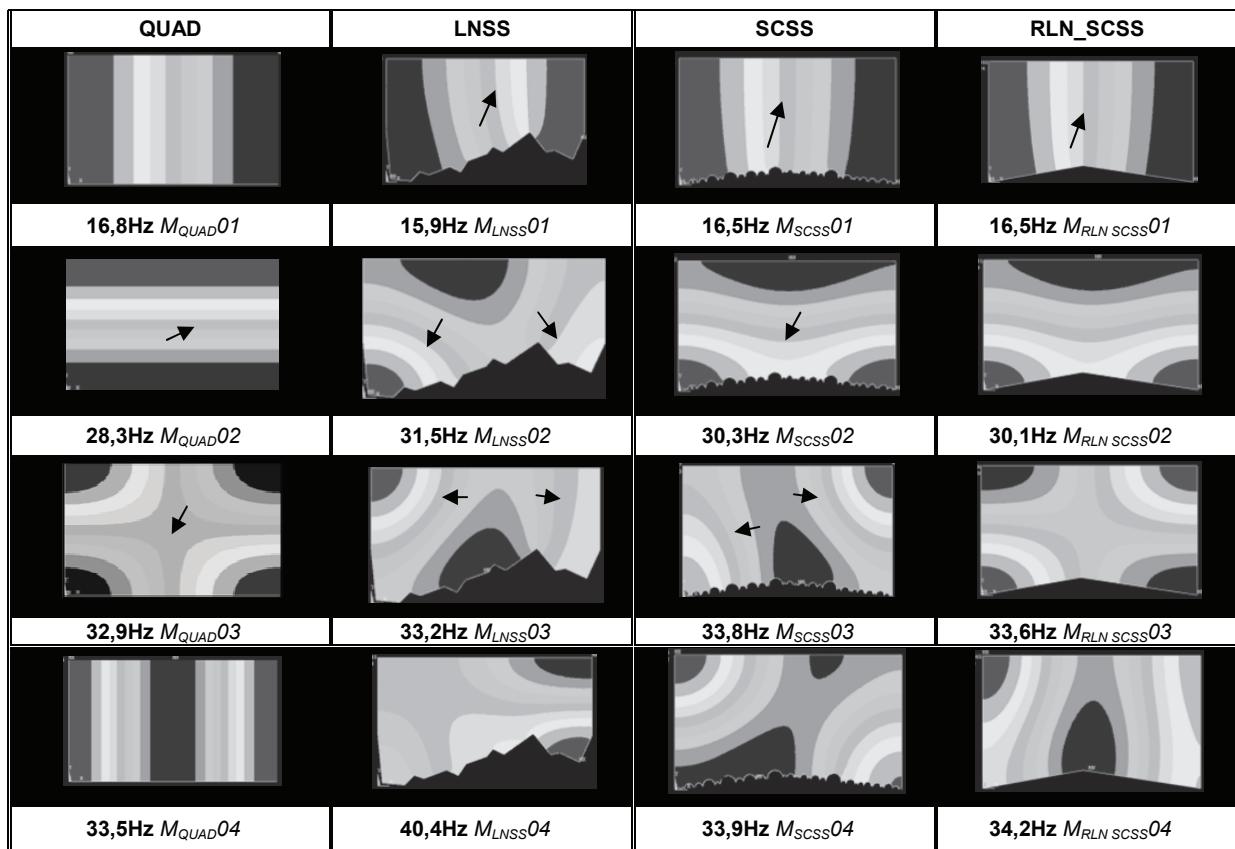


Figura 6 – Análise modal: primeiros quatro modos das quatro salas – QUAD (quadrangular), LNSS (superfície serial difusora linear), SCSS (superfície serial difusora com elementos semicilíndricos), RLN_SCSS (SCSS sem os elementos semicilíndricos). Os máximos e mínimos de pressão aparecem em tom escuro. Cada seta indica faixa onde ocorre menor variação de pressão. N.B.: Não há correspondência direta entre a variação de pressão e a variação de tom claro-escuro de modo que o tom mais claro seja o que represente menor variação de pressão.

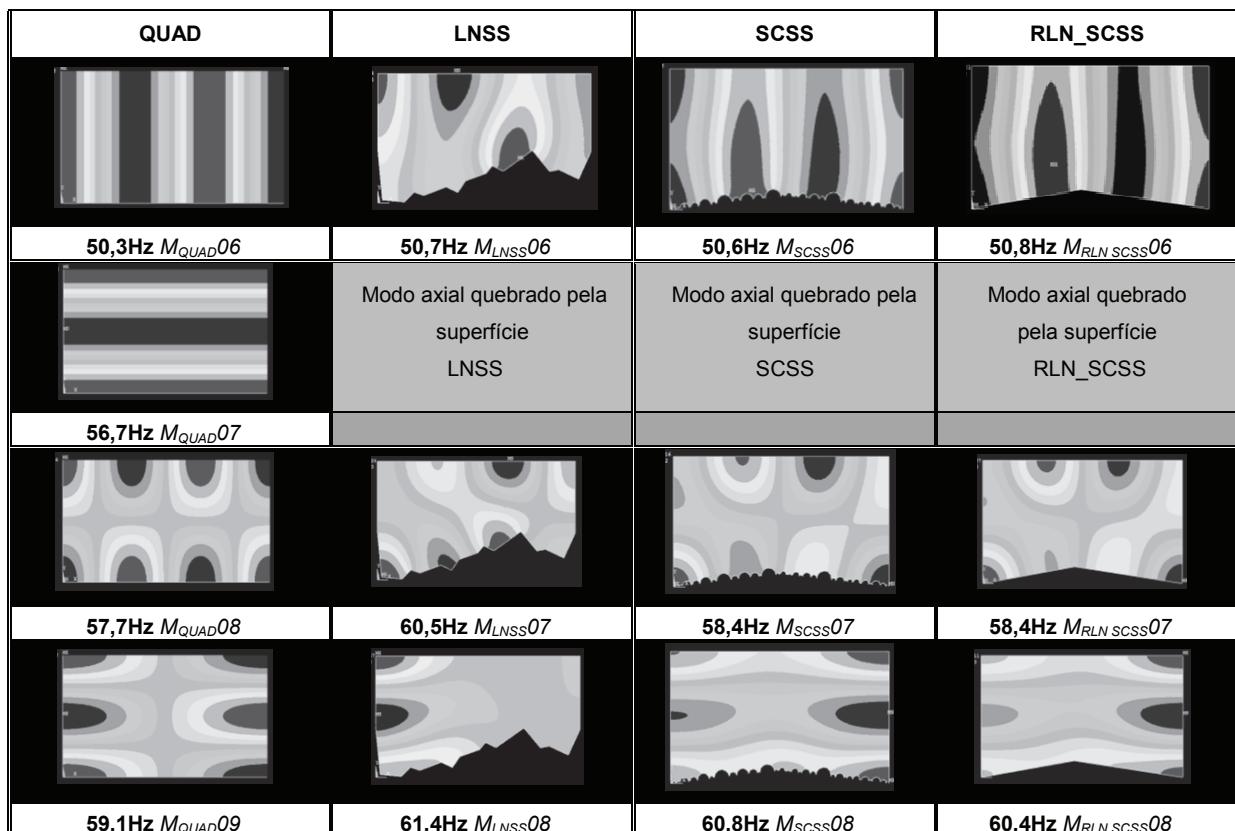


Figura 7 – Análise modal: resultados entre 50 e 61Hz, com destaque para o modo axial $M_{QUAD}07$ (56,7Hz) quebrado pelas superfícies serials difusoras.

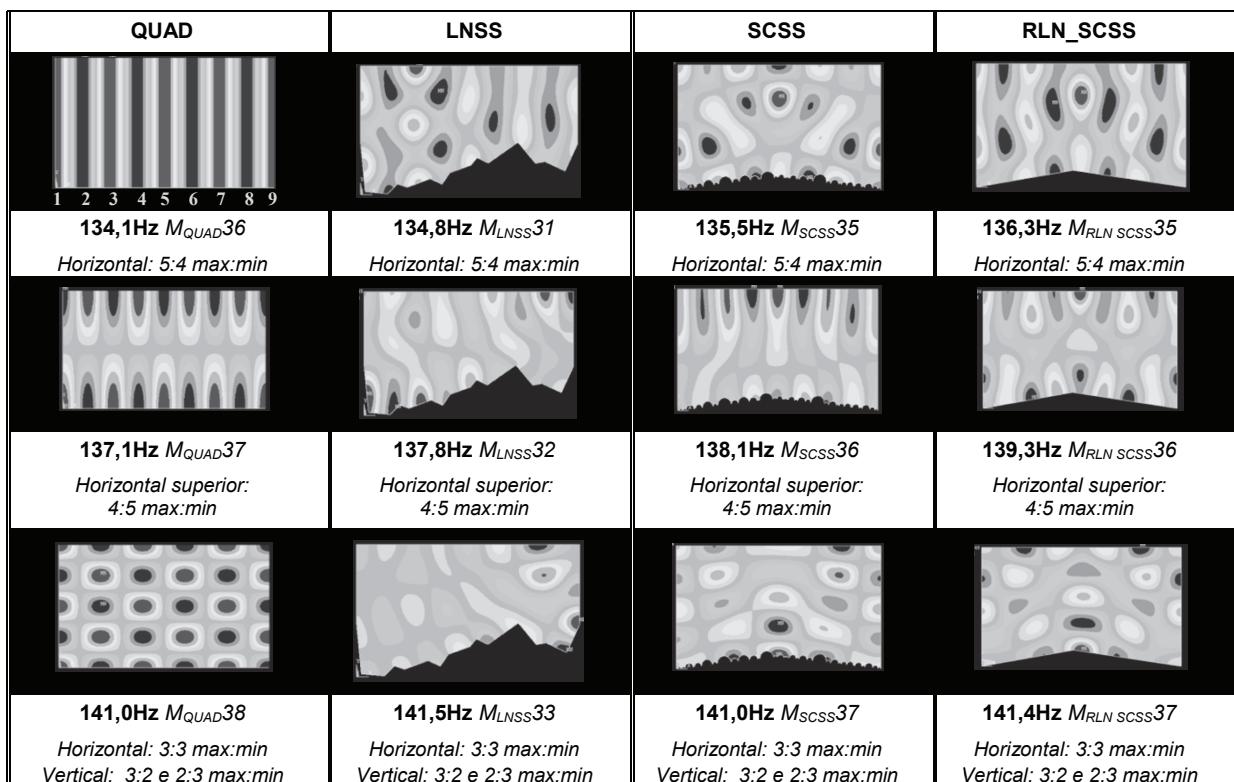


Figura 8 – Modos minimizados pelas superfícies seriais sendo reconhecidos e associados ao modo da sala sem tratamento por terem comportamento semelhante, identificado pela freqüência do modo e pela distribuição de máximos e mínimos em linhas e colunas da figura resultante da análise modal. A relação entre máximos (max) e mínimos (min) observada no resultado colorido está aqui indicado em forma de proporção, por exemplo: na primeira ilustração ao alto à esquerda – M_{QUAD36} (134,1Hz) – temos 5:4 que são 5 máximos para 4 mínimos. A configuração seqüencial das colunas escuras é, portanto (1) Max – (2) min – (3) Max – (4) min – (5) Max – (6) min – (7) Max – (8) min – (9) Max.

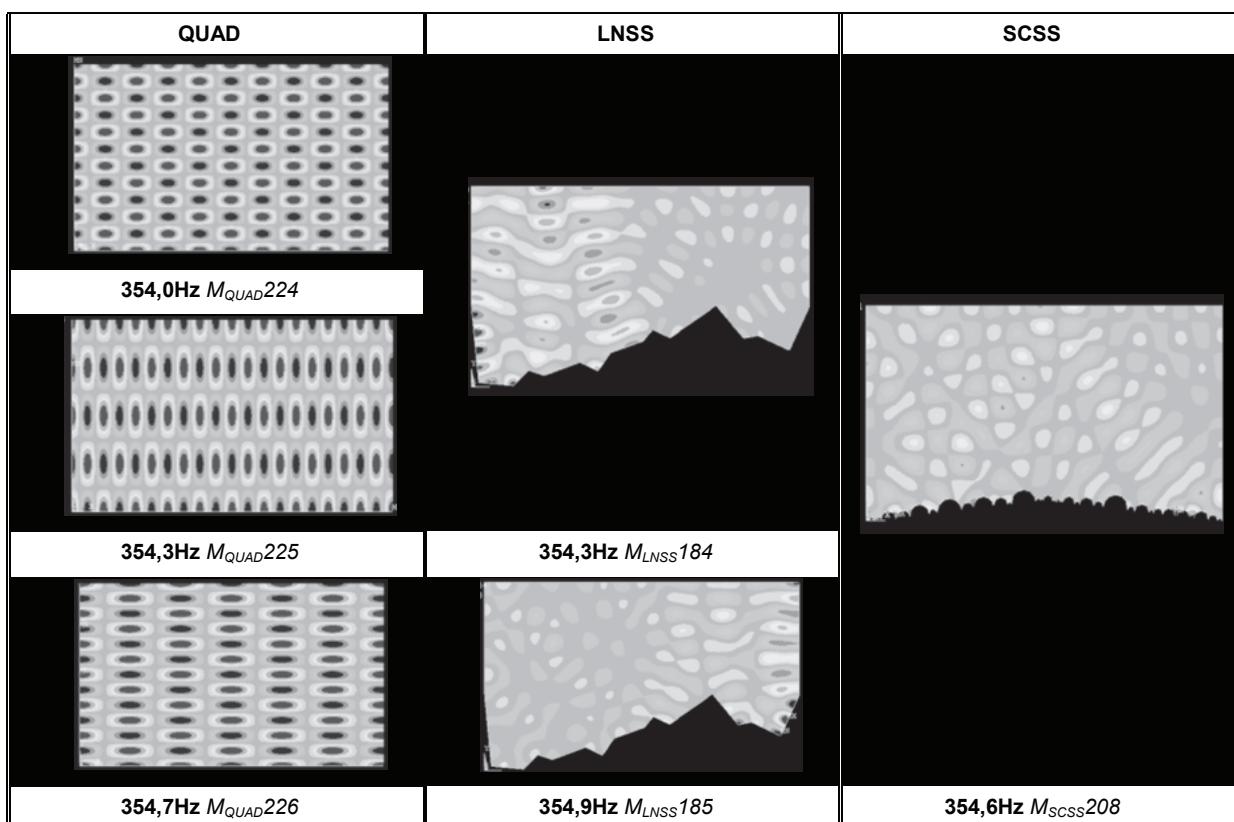


Figura 9 – Resultados da análise modal em 354Hz. Os três modos $M_{QUAD224}$, $M_{QUAD225}$ e $M_{QUAD226}$ foram totalmente minimizados após a instalação de uma superfície serial difusora.

3 AVALIAÇÃO DOS RESULTADOS DA ANÁLISE MODAL

A avaliação a seguir foi realizada pensando com as formas observadas nos resultados da análise modal. Após a familiarização com as figuras e seus comportamentos, as formas e suas dinâmicas adquiriram sentido. Se a coerência dos processos é reflexo de comportamentos sistemáticos, dela parece emergir, através da morfologia das imagens, uma linguagem assimilada durante a observação, graças à qual é possível compreender a dinâmica do processo, como que “lendo” nas próprias formas observadas. Os recursos verbais de raciocínio lógico e a tipologia de simetrias serviram para traduzir essas impressões, não representando-as mas procurando posicionar o leitor diante dos objetos, sugerindo como e para onde olhar, de modo a enxergar o que vê, perceber o que enxerga e a compreender o que percebe. As formas dizem o que estão acontecendo com elas e o olhar comparativo entre figuras permite-nos assimilar essa “coerência”, aprender essa “sintaxe” e a “ler” sobre as dinâmicas.

3.1 Tipologia da simetria

Simetria translacional: Trata-se da repetição de um elemento mantido de forma idêntica à sua figuração original, sendo somente transladado.

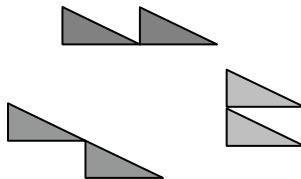


Figura 10 – Três exemplos de translação de um elemento geométrico.

Simetria axial: também conhecida por simetria *bilateral* ou por *reflexão*, ocorre quando há um eixo de simetria.

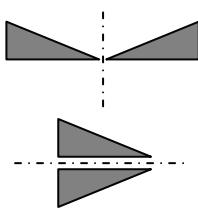


Figura 11 – Simetria Axial: espelho.

Simetria rotacional: Quando o elemento é rodado em torno de um eixo central ortogonal.

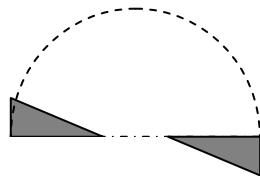


Figura 12 – Simetria Rotacional: elemento rodado.

3.2 Distribuição variada dos máximos e mínimos de pressão

De maneira geral, observa-se como nas salas com superfícies LNSS, SCSS e RLN_SCSS os ventres com máximos e mínimos de pressão variam constantemente suas posições. A localização variada, espalhada, regular e

uniforme dos ventres de máximos e mínimos em função da freqüência contribui para que a resposta acústica seja homogênea, desejável para uma sala destinada a performance e audição musical.

Contrariamente, na *sala de referência* (QUAD) os ventres estão preponderantemente localizados em determinados pontos, linhas e regiões, mais do que em outros. Sobre qualquer eixo de simetria da figura geométrica há acúmulo de ventres de pressão, e o número de eixos de simetria na *sala de referência* é superior às demais. Esses locais devem ser evitados. Por exemplo, em medições acústicas, os pontos de posicionamento do microfone devem estar a uma distância mínima das superfícies das partições, pois nelas há acúmulo sistemático de ventres de pressão e levam a resultados que só se verificam localmente e não podem ser generalizados para todo recinto em medição.

3.3 Identificação de modos com simetrias

Os modos 07 comparados resultam em dois grupos, um formado pelo modo $M_{QUAD}07$ (linhas horizontais) e o outro formado pelos modos $M_{LNSS}07$, $M_{SCSS}07$ e $M_{RLN_SCSS}07$ com alguns ventres circulares e com simetrias rotacionais tanto em relação ao eixo vertical quanto horizontal. Porém, este tipo de simetria é justamente o do modo $M_{QUAD}08$.

Os modos 08 comparados resultam igualmente em dois grupos, um formado pelo modo $M_{QUAD}08$, com simetria rotacional em eixos vertical e horizontal com total de 8 ventres de pressão, e o outro formado pelos modos $M_{LNSS}08$, $M_{SCSS}08$ e $M_{RLN_SCSS}08$ com elementos de simetria axial em relação ao eixo vertical e tendendo a 6 ventres de pressão. Porém, essa é a exata simetria do modo $M_{QUAD}09$.

3.4 Alteração dos modos normais

Ao comparar modos correlacionados na Figura 8 é possível perceber como os modos $M_{QUAD}36$ (134,1Hz), $M_{QUAD}37$ (137,1Hz) e $M_{QUAD}38$ são ‘dissolvidos’ tornando-se quase que irreconhecíveis uma vez instalada no local uma superfície serial difusora. O mesmo se constata na Figura 9.

3.5 Quebra de modos pelas superfícies seriais difusoras

Identificada a semelhança entre os resultados da análise modal dos modos $M_{LNSS}07$, $M_{SCSS}07$ e $M_{RLN_SCSS}07$ (Figura 7) e o modo $M_{QUAD}08$, bem como dos modos $M_{LNSS}08$, $M_{SCSS}08$ e $M_{RLN_SCSS}08$ e o modo $M_{QUAD}09$, pode-se atribuir a essas identidades comportamentos acústicos igualmente similares. Como os modos estão encadeados continuamente em suas progressões individuais, constata-se, então, que o comportamento acústico observado no modo $M_{QUAD}07$ está ausente das famílias de modos M_{LNSS} , M_{SCSS} , M_{RLN_SCSS} . Pode-se concluir que esses comportamentos foram inibidos ao instalar os difusores LNSS, SCSS, RLN_SCSS. De forma figurativa, adotou-se o termo “quebrar” um modo quando um difusor instalado em um local elimina determinado comportamento modal presente no recinto original.

Observa-se, então, a quebra do segundo modo axial no sentido da profundidade da sala, modo $M_{QUAD}07$ (56,7Hz).

3.6 Ventres de pressão: diminuição quantitativa em unidades e na área específica ocupada

No grupo dos primeiros modos $M_{QUAD}01$ a 05 (16 a 47Hz) (Figura 6) observa-se uma melhora em relação ao modo $M_{QUAD}02$ (axial), que pela ação das superfícies seriais difusoras perde a regularidade das faixas horizontais

de pressão. O mesmo ocorre, de forma mais pronunciada, em relação aos modos $M_{LNSS}03$ e 04 , $M_{SCSS}03$ e 04 , onde as superfícies seriais diminuíram significativamente a área ocupada pelos ventres de pressão, bem como o número total de ventres, de 4 para 3 em $M_{SCSS}03$ e de 4 para 2 em $M_{LNSS}03$.

O modo $M_{LNSS}08$ (61,4Hz) apresenta uma diminuição de 6 para 2 ventres em relação ao modo $M_{QUAD}09$ (59,1Hz). Ambos os modos $M_{SCSS}09$ (62,4Hz) e $M_{RLN_SCSS}09$ (61,9Hz) diminuíram os ventres de pressão de 9 para 3 em relação ao modo $M_{QUAD}10$ (65,8Hz).

Os modos $M_{QUAD}37$ (137,1Hz) – $M_{LNSS}32$ (137,8Hz) – $M_{SCSS}36$ (138,1Hz) e $M_{RLN_SCSS}36$ (139,3Hz), são todos caracterizados e unidos por possuírem estrutura derivada de um original ($M_{QUAD}37$) com 9 ventres acima (4 máximos e 5 mínimos) e abaixo (5 máximos e 4 mínimos). Percebe-se claramente que as superfícies difusoras praticamente “desmancharam” os ventres de pressão da linha inferior da configuração inicial da sala (QUAD), mas guardam o embrião de 9 ventres com simetria rotacional que pode ser visualizado apesar dos ventres de pressão da linha inferior estarem bastante debilitados.

A atuação das superfícies seriais difusoras é cada vez mais forte na medida em que os modos vão se elevando. Na Figura 8 $M_{LNSS}33$ (141,5Hz) quase que perde as características visuais de observação que o unem a $M_{QUAD}38$. O que permitiu a identificação neste caso foi a freqüência ao redor de 141Hz, o número de máximos e mínimos em alinhamento vertical ao lado direito de $M_{LNSS}33$ e a semelhança da região direita deste com $M_{SCSS}37$ e $M_{RLN_SCSS}37$. Comparando M_{QUAD} , M_{LNSS} e M_{SCSS} , na banda de 352 a 354Hz, além de modos quebrados, o número de ventres máximos e mínimos diminuiu em torno de 90%, sendo reduzidos a alguns pontos esparsos, com área igualmente reduzida.

3.7 Atuação das irregularidades semicilíndricas em SCSS

A análise modal da superfície RLN_SCSS foi efetuada para poder comparar os resultados com SCSS e identificar a partir de que freqüência as irregularidades semicilíndricas serializadas passam a agir significativamente. Nos primeiros modos, apesar de ter diferentes formas resultantes do comportamento modal, a atuação principal é a da inclinação dos eixos alinhando os centros dos círculos que seccionam os cilindros, dispostos em “v” invertido, que se observa na superfície inferior de RLN_SCSS.

Há semelhanças entre os *patterns* das figuras formadas nos modos iniciais de M_{SCSS} e M_{RLN_SCSS} , porém a partir de $M_{SCSS}53$ (171,1Hz) e $M_{RLN_SCSS}52$ (170,1Hz) (Figura 13) foi observado o enfraquecimento da correlação entre ambas, sendo exatamente nesse ponto que as irregularidades semicilíndricas começam claramente a atuar: os resultados da análise modal possuem a mesma base morfológica, mas as transformações das variações de pressão próximas à superfície SCSS são significativas, dissolvendo praticamente as colunas de máximos e mínimos de pressão. O mesmo se verifica entre os modos $M_{SCSS}62$ (188,0Hz) e $M_{RLN_SCSS}61$ (187,0Hz) (Figura 13).

Comparando-se as seqüências de modos $M_{SCSS}66$ a 68 e $M_{RLN_SCSS}66$ a 68, constata-se como a superfície serial semicilíndrica minimizou sensivelmente a formação dos dois ventres de pressão ao centro de $M_{RLN_SCSS}67$.

Verifica-se, assim, que a partir da banda de freqüência de 170-190Hz ($\lambda = 2,00$ a 1,79m) ($\lambda/4 = 50\text{cm}$ a 45cm) há

influência das irregularidades circulares (diâmetros entre 10 e 50cm) serializadas no comportamento modal do sistema todo.

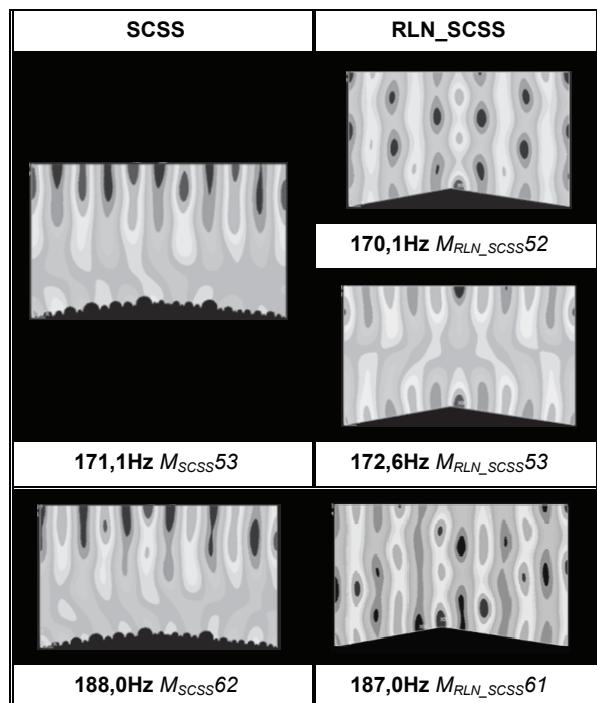


Figura 13 – Atuação das irregularidades semicilíndricas de SCSS constatada a partir de 170Hz comparando os resultados de SCSS aos de RLN_SCSS.

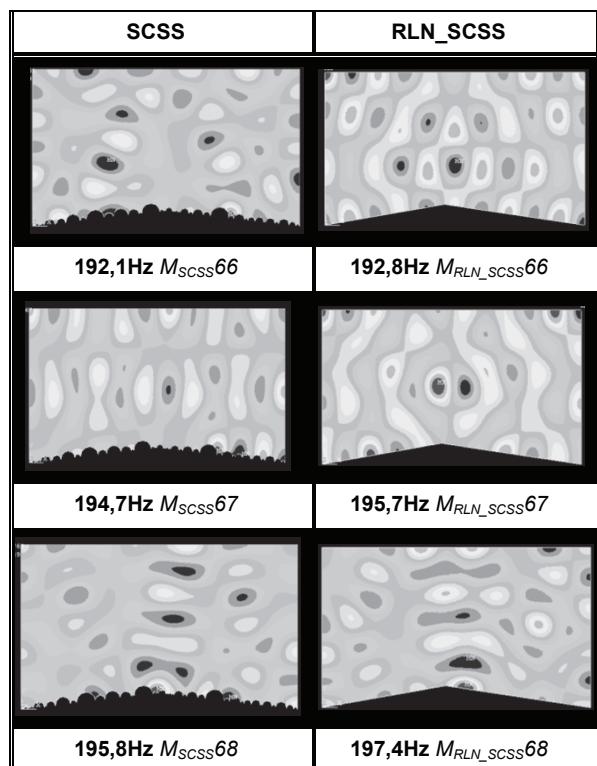


Figura 14 – Seqüência contínua dos modos $M_{SCSS}66$ a 68 e $M_{RLN_SCSS}66$ a 68 mostram como pela ação das irregularidades semicilíndricas os ventres de máximo e mínimo de pressão ao centro de $M_{RLN_SCSS}67$ foram minimizados.

3.8 Tendência dos ventres de máximos e mínimos de pressão a estarem próximos às superfícies seriais difusoras

Quanto aos ventres de pressão, além da redução em número destes nos contornos das plantas das salas alteradas em relação à sala de referência, nota-se que os remanescentes se situam próximos às superfícies difusoras, muitas vezes junto a elas, cabendo a recomendação de evitar esses arredores, prevendo um recuo desta região onde condições acústicas não são adequadas. Medições aí efetuadas estarão sujeitas a valores discrepantes. Pontos mais distantes se caracterizam por melhor uniformidade, homogeneidade e equilíbrio na distribuição e variação das posições dos máximos e mínimos de pressão.

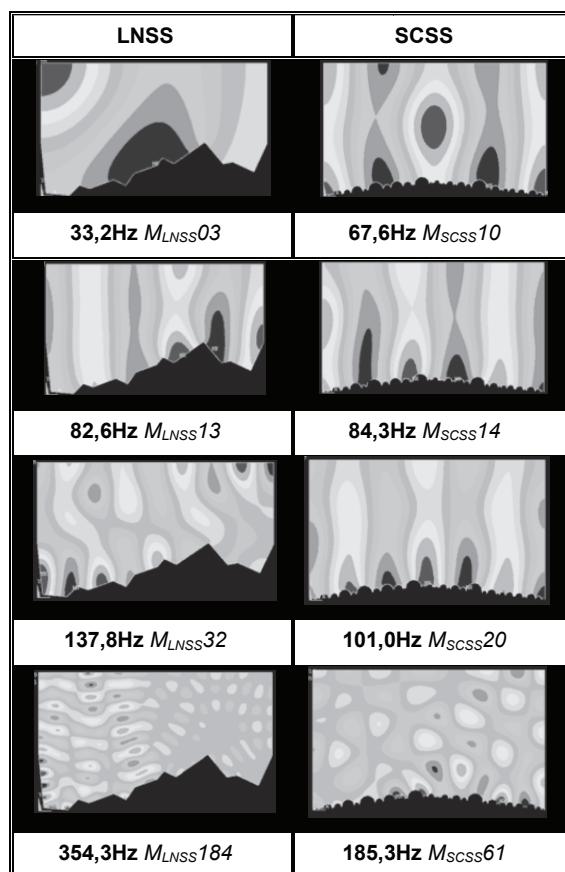


Figura 15 – Tendência dos Máximos e Mínimos de pressão a se localizarem nas proximidades ou diretamente sobre as superfícies seriais difusoras.

Essa tendência das superfícies seriais difusoras em “atraírem” os ventres de pressão para si representa, portanto, uma vantagem e ao mesmo tempo um alerta: a vantagem é que, além de diminuir a quantidade de ventres, os poucos que restam estarão próximos a ela, bem como ao nos afastarmos da superfície difusora estarão significativamente diminuídos os máximos e mínimos de pressão em número e em área ocupada. O alerta é para evitar qualquer ponto de escuta e captação muito próximo a uma superfície serial difusora, estando sujeito a resultados inadequados. Para conhecer com precisão a distância que deve ser mantida, será necessário um trabalho complementar envolvendo medições e análises. Contudo, observando as figuras é possível estimar que a aproximadamente 1,0 a 1,5m estejamos livres do acúmulo de ventres de pressão.

Conhecendo a tendência dos ventres de pressão se situarem próximos às superfícies seriais, não só sabemos onde eles provavelmente estão como também, pela manipulação das superfícies, podemos deslocá-los solidariamente a estas.

4 CONCLUSÃO

Constata-se na análise modal bidimensional de salas contendo uma superfície serial difusora (LNSS ou SCSS) a atuação destas a partir de 30Hz aumentando proporcionalmente à frequência; a significativa diminuição do número de ventres de pressão, chegando a 90%, bem como da área total por eles formada; a influência das irregularidades semicilíndricas de SCSS a partir de 170Hz; a tendência dos ventres de pressão a se situarem próximos à superfícies seriais. Dessa forma, conclui-se que além do espalhamento sonoro buscado as superfícies seriais difusoras desta análise têm um impacto positivo sobre os modos normais de uma sala.

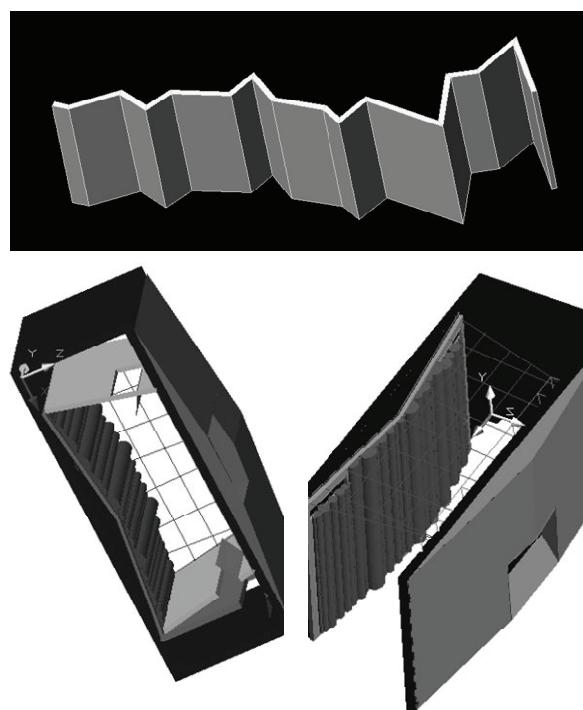


Figura 16 – Exemplos de superfícies seriais difusoras: acima LNSS; abaixo SCSS.

5 AGRADECIMENTOS

Prof. Dr. José Roberto de França Arruda e Alexander Mattioli Pasqual do Laboratório de Mecânica Computacional da Faculdade de Engenharia Mecânica da Unicamp.

6 REFERÊNCIAS

- [1] MANNIS, José A. *Difusores sonoros projetados a partir de processo serial: adequação acústica de pequenas salas à performance e audição musical*. 2008. 400p. Tese (Doutorado em Música) – Universidade Estadual de Campinas, Campinas, 2008.
- [2] SCHOENBERG, Arnold. *Style and idea*. New York : Philosophical Library, 1950. vii, 224 p.
- [3] SCHROEDER, Manfred R. *Binaural dissimilarity and optimum ceilings for concert halls: more lateral*

- diffusion. *J. Acoust. Soc. Am.*, v. 65, n. 4, p. 958-963, Apr., 1979.
- [4] SCHROEDER, Manfred R. *Number theory in science and communication: with applications in cryptography, physics, digital information, computing, and self-similarity*. 3. ed. New York: Springer-Verlag, 1997. (Spring series in Information Sciences, v. 7) 363p. (1. ed. em alemão, Berlin : Springer, 1984)
 - [5] D'ANTONIO, Peter; KONNERT, John H. The Schroeder quadratic-residue diffusor: design theory and application. In: AES CONVENTION, 74., 1983, New York. *Proceedings...* 26 p. [1999 (C-4)]
 - [6] D'ANTONIO, Peter; KONNERT, John H. The reflection phase grating diffusor: design theory and application. *J. Audio Eng. Soc.*, v. 32, n. 4, p.228-238, Apr., 1984
 - [7] COX, T. J.; D'ANTONIO, Peter. *Acoustic absorbers and diffusers: theory, design and application*. London: Spon, 2004. 405 p.
 - [8] EVEREST, F. Alton; SHEA, Mike. Acoustics of small rooms. In: BALLOU, Glen (Ed.) *Handbook for sound engineers : the new audio cyclopedia*. Indianapolis (EUA): Howard W. Sams & Co., 1988. cap. 3, p. 41-60.
 - [9] BONELLO, Oscar J. Acoustical evaluation and control of normal room modes. *J. Acoust. Soc. Am.*, suppl.1, v. 66, n. 2, 1979. In: ACOUSTICAL SOCIETY OF AMERICA, 98th meeting, 1979, Salt Lake City, Utah, EUA.
 - [10] BONELLO, Oscar J. A new computer aided method for the complete acoustical design of broadcasting and recording studios, In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 1979. *Anais...* v. 4, April, p. 326-329, 1979.
 - [11] DAVIS, Don; DAVIS, Carolyn. *Sound system engineering*. 2. ed. Indianapolis (EUA): Howard W. Sams & Co., 1987. 665 p.
 - [12] NIEUWLAND, J. M. van; WEBER, C. Eingenmodes in nonrectangular reverberation rooms. *Noise Control Engineering*, v. 13, n. 3, p. 112-121, November-December, 1979.



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Cabeças artificiais e manequins: um resgate histórico

Stephan Paul¹

¹ Universidade Federal de Santa Catarina, Lab. de Vibrações e Acústica, Dep. de Eng. Mecânica
Campus Trindade, Florianópolis, Santa Catarina, 88040-900, Brasil

stephan.paul.acoustic@gmail.com

RESUMO

Recentemente ouve-se e lê-se muito sobre tecnologia *binaural*, biaural ou biauricular de gravação ou medição, sendo muitas vezes também chamada de *dummy head technique* ou *artificial head recording technique*. Por falta de informação, muitas pessoas a consideram uma tecnologia recente e restritamente vinculada a uma subárea da engenharia acústica, a qualidade sonora. Apesar de ser de interesse dessa área, a tecnologia biauricular tem origens em outros campos da engenharia e é igualmente importante em outras áreas e a busca por informações sobre essa tecnologia vem crescendo junto ao interesse na criação de ambientes virtuais acústicos. Ao longo de muitos anos de existência, ela passou por vários desafios, sendo que alguns deles não foram resolvidos até hoje. Em um resgate histórico, o artigo mostra as origens da tecnologia biauricular, desde os primeiros ensaios nos Laboratórios Bell nos E.U.A. até os anos 80, relatando também um pouco sobre os desafios e as relações e diferenças entre esta tecnologia e a estereofonia.

0 INTRODUÇÃO

Ultimamente observou-se um interesse novo frente à tecnologia de obtenção de sinais biauriculares. No Brasil ela é chamada de tecnologia de gravação *binaural*, biaural, biauricular ou até biáurea¹. O inte-

resse nessa tecnologia de gravação renasceu provavelmente por várias razões, entre elas a popularização de *blogs* na internet, nos quais os usuários discutem assuntos de seu interesse e disponibilizam arquivos para os demais internautas, como por exemplo amostras de sons biauriculares. Apesar de que nem sempre o que está sendo chamado de gravação biauricular realmente o é, a informação encontrada na internet muitas vezes passa a impressão de que se trata de uma tecnologia recente e muitas empresas que vendem equipamentos de

¹Gravação biaural ou biauricular seriam as denominações certas considerando o Português brasileiro. Gravação binaural é uma adaptação direta e não muito correta do nome em inglês, *binaural recording*, considerando que o prefixo bin- não existe na língua portuguesa.

gravação biauricular dão uma impressão ultra-moderna aos seus produtos.

O presente artigo resgata um pouco da história sobre uma parte importante da gravação biauricular, a das cabeças artificiais ou manequins. Estes utilizados para obtenção de sinais biauriculares, considerando especialmente os anos anteriores a 1990.

1 PASSOS INICIAIS

A partir dos primeiros estudos documentados sobre o fenômeno da audição biauricular, no século XIX (por exemplo Dove [7], Seebeck[20], Steinhauser[22], Thompson [23, 24, 25], J.W Strutt (Lord Rayleigh) [18]), ficou claro que o simples fato de seres humanos possuírem dois ouvidos é o maior responsável pela sofisticação da audição humana. A fim de proporcionar transmissões mais fidedignas de peças musicais, começou-se a experimentar o uso de microfones espaçados. Desta forma nasceu um antecessor à estereofonia, tendo o inventor comparado o efeito audível ao efeito visível, conhecido naquela época como esteroscopia.

A primeira transmissão “estereofônica” de uma peça musical da ópera de Paris, em 1881, foi realizada por Adler [1]. Os sinais foram obtidos com vários pares de microfones de carbono usados nos telefones da época. Transmitidos via duas linhas telefônicas, uma para cada sinal, eles podiam ser escutados com dois fones de ouvido monaurais. A técnica que faz uso de dois microfones simples está sendo utilizada até hoje, mas ela não é mais considerada como uma técnica biauricular de gravação, como ocorreu muitas vezes no passado e erroneamente ocorre até hoje, sobretudo por falta de informação correta.

Há relatos (por ex. [8]) que a estação de rádio WPAJ na cidade de New Haven, estado de Connecticut, E.U.A., por volta de 1925, microfonou utilizando dois microfones e transmitiu ambos os sinais em freqüências diferentes². Assim, as pessoas podiam ouvir os dois sinais conectando um fone de ouvido monaural a um rádio sintonizado na primeira freqüência de transmissão e o segundo fone monaural ao outro aparelho sintonizado na segunda freqüência de transmissão. Naquela época utilizava-se quase exclusivamente fones de ouvido monaurais, já que os alto-falantes ainda eram extremamente primitivos. Como não foram achadas fontes originais da época, não se pode afirmar que estes sinais realmente eram o que se considera hoje sinais biauriculares. Provavelmente foram utilizados simplesmente dois microfones no lugar dos ouvido, e não um manequim com microfones, e essa técnica de obtenção foi batizada, como também a obtenção dos sinais feita por Adler em 1881, erroneamente como biauricular (*binaural* em [8]).

²Vide também as informações sobre a data de 6 de agosto na página <http://www.wdrccb.com/history.html>

1.1 O patente de W. Bartlett Jones

Com o passar do tempo os conhecimentos sobre a audição biauricular foram aprimorados e surgiu a idéia de copiar mais partes dos fatores anatômicos responsáveis pelo fenômeno da audição biauricular, por exemplo a presença da cabeça e não apenas o número de receptores para o som.

Em 1927, o estadunidense W. Bartlett Jones, de Chicago, submeteu uma patente considerando um sistema de captação, gravação e reprodução de sinais biauriculares. A patente, finalmente concedida em abril de 1932 sob o número 1855149 [15], mostra desenhos simples de uma “cabeça artificial” em forma de cilindro ou esfera, com microfones nas posições que correspondem aos ouvidos. Estes microfones deveriam ser inclinados para frente para melhor reproduzir, conforme o autor da patente, a inclinação dos pavilhões auditivos na cabeça humana. A patente também descreve, em forma muito simplificada, aparelhos de gravação e reprodução de sinais biauriculares em disco e propõe meios de reprodução dos sinais de tal forma a preservar os indicadores biauriculares necessários para localização de fontes sonoras (principalmente a diferença interauricular de tempo e de intensidade). Porém, com os conhecimentos que se têm atualmente sobre audição e tecnologia biauricular, pode-se afirmar que as idéias de Jones não conseguiriam reproduzir de forma muito fiel a localização de uma fonte e os sinais obtidos não seriam considerados verdadeiramente biauriculares.

1.2 O manequim Oscar

As primeiras referências bibliográficas achadas pelo autor, que dizem respeito ao uso de um manequim para copiar alguns dos fatores responsáveis pela audição biauricular, situa esse fato no início dos anos 1930 [9, 21]. Conforme o próprio Harvey Fletcher (em entrevista em 1963)³, um grupo de pesquisadores dos laboratórios Bell trabalhou, no início de 1930, na melhoria da transmissão de voz pelo telefone a fim de conseguir uma impressão sonora com alta fidelidade. Após exaustivas tentativas, eles perceberam que a audição biauricular humana é um dos fatores mais importantes para a alta fidelidade do som ao vivo e que a transmissão de apenas um sinal monaural é a responsável pela falta de fidelidade. Fletcher e seus colegas utilizaram então um manequim de madeira, chamado de “Oscar”. Este tinha dois microfones colocados nas partes laterais de sua cabeça, perto ou dentro dos ouvidos. A posição exata não pode ser definida, já que parte das fontes bibliográficas afirmam que os microfones estavam dentro da orelha [9] e outras fontes apontam que estavam apenas perto das orelhas [8, 16]. Considerando o fato de que Fletcher e seus colegas realmente utilizaram um manequim, seu trabalho pode ser considerado sem dúvida o marco inicial da tecnologia biauricular.

³O vídeo destas entrevistas estava disponível em http://auditorymodels.org/jba/BOOKS_Historical/FletcherVideo/ em fevereiro de 2008.

“Oscar” foi apresentado ao público em 1932, na feira internacional de Chicago (*1932 World's fair*) [9, 21], onde pessoas podiam ouvir, através fones de ouvido, os sinais obtidos com o seu uso. Conforme as mesmas fontes, o público ficou bastante impressionado. Posteriormente, os laboratórios Bell dedicaram-se à tarefa de reproduzir o mesmo efeito de alta fidelidade utilizando alto-falantes. Destes esforços derivou-se um método que emprega três microfones simples em vez da cabeça “Oscar”. Este método, bem como os demais métodos que utilizaram apenas dois microfones, foram chamados de estereofonia.

A partir do manequim “Oscar”, desenvolveu-se seu sucessor: “Oscar II”. Segundo a Figura 1, a cabeça “Oscar II” realmente tinha os microfones no lugar dos ouvidos, mas estes microfones tinham um diâmetro muito grande e o canal auditivo não foi modelado.



Figura 1: A cabeça artificial “Oscar II” sendo utilizada por H. Fletcher em experimentos de localização de fontes sonoras. Nestes ensaios, chamados de *double dome research*, uma pessoa tinha a cabeça artificial fixada à sua e ela escutava os sinais obtidos e alterados pelos microfones da cabeça artificial. A pessoa foi requerida a apontar à direção de onde aparentemente vinha o som captado. Fonte: AT&T Corporate Archives.

Nos anos seguintes, não há mais relatos do uso do manequim “Oscar” mas sim de diferentes arranjos estereofônicos simples de microfones. Isso provavelmente decorre do fato de que a tecnologia biauricular daquela época não era, apesar dos esforços dos pesquisadores dos *Bell Laboratories*, compatível com a reprodução por alto-falantes. Pois, mesmo utilizando fones de ouvido, a posição aparente da fonte sonora movia-se juntamente com a cabeça do ouvinte, problema este só resolvido posteriormente com o uso de um dispositivo que determina a posição e orientação da cabeça do ouvinte⁴ e ajusta os sinais por meio da aplicação da função de transferência relacionada à cabeça (*HRTF*⁵) correspondente.

⁴Este dispositivo é conhecido como *head-tracker*.

⁵Do inglês: *head related transfer function*.

2 ESTEREOFONIA, UMA TECNOLOGIA MAIS SIMPLES

As técnicas de gravação que utilizam vários microfones mas desconsideram a cabeça, já utilizada por Adler em 1881, eram mais simples e mais compatíveis com a reprodução por alto-falantes, pelo menos quando se tratava de sinais gravados em ambientes grandes ou em condições parecidas às de campo livre.

Snyder [21] relata que um dos grandes desafios daquela época era a gravação de vários canais sem que a informação de fase entre os sinais se perdesse⁶. Segundo ele, isso só começou a ser possível com a chegada dos gravadores em fita magnética, no final dos anos quarenta. Porém, mesmo com maior precisão de alinhamento das cabeças que magnetizam a fita, considerando um erro de $\pm 1.25/1000$ polegadas no posicionamento, a localização da imagem sonora sofria erros na ordem de 20° .

Nos anos 50, impulsionados pela disponibilidade de meios de gravação que permitem o armazenamento de sinais com a preservação da fase e com ampla faixa dinâmica em fita magnética, até em condições difíceis como no interior de um automóvel em andamento [2], surgiram novos usos para a tecnologia estereofônica⁷ e biauricular e novas cabeças artificiais.

No seu artigo de 1953, Snyder [21] concluiu que uma tecnologia que preservasse todos os aspectos importantes do som teria aplicações não apenas em gravações e reproduções de peças musicais, mas também na análise de sons, promessa que será comprovada posteriormente.

3 ESTEREOFONIA COM CORPO DE SEPARAÇÃO

André Charlin, pioneiro francês de áudio, apresentou a sua “cabeça artificial” em 1954, a *tête Charlin* (Figura 2). Essa “cabeça”, um balão revestido com material absorvente, possuía dois microfones de campo de pressão nos lugares que correspondem aproximadamente à posição dos ouvidos humanos.

Em 1955, a empresa alemã Schoeps apresentou um microfone parecido à *tête Charlin*: uma esfera de alumínio de 20 cm de diâmetro (Figura 3) com dois microfones de campo de pressão, estes que possuíam cápsulas com características omnidirecionais. A Schoeps produziu apenas alguns exemplares deste microfone.

⁶A informação de fase entre os sinais esquerdo e direito é um indicador importante para a localização da fonte sonora.

⁷David C. Apps e colegas [2] da General Motors nos E.U.A. podem ser considerados os primeiros a utilizar gravações estereofônicas na indústria automobilística, no início dos anos 50. O título do artigo “The Use of Binaural Tape Recording in Automotive Noise Problems” sugere, como aconteceu muitas vezes naquela época e ainda acontece, erroneamente que se tratava de gravações biauriculares. Porém, ao ler o artigo, que também relata o uso de comparações subjetivas dos sinais obtidos, fica claro que se tratava de gravações estereofônicas com microfones simples e sem corpo de separação ou cabeça.

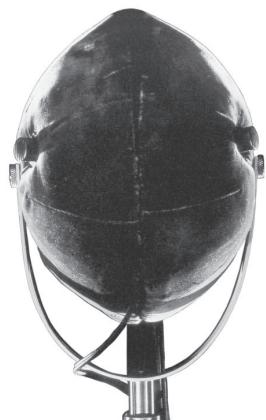


Figura 2: A *tête Charlin*. Fonte: [6].



Figura 3: O microfone da Schoeps. Fonte: [28].

A abordagem de Jones, de André Charlin, de Schoeps e de muitos seguidores deve ser considerada como microfonia com corpo de separação e não exatamente como gravação biauricular. Isso porque nem todos os elementos importantes para a audição humana foram modelados, como por exemplo o pavilhão e canal auditivo, parte do tronco, a impedância da pele, do canal auditivo e da membrana timpânica.

Nos anos seguintes surgiram mais alguns conjuntos de microfones com corpo de separação, mas nenhum fabricante começou a fabricação em série, já que outras técnicas mais simples eram consideradas suficientes, em função do tipo de gravação, e nenhuma empresa queria investir em pesquisa.

Dessa forma, tem-se duas tecnologias parecidas à verdadeira tecnologia de gravação biauricular, a estereofonia e os microfones estereofônicos com corpo de separação.

4 NOVOS AVANÇOS DA TECNOLOGIA BIAURICULAR

Alguns laboratórios de acústica apresentaram novos trabalhos relacionados ao assunto de audição, medição e gravação biauricular no final dos anos 60.

Em 1966, Bauer, Torick e colegas do laboratório CBS [3, 26] apresentaram um manequim (Figura 4), baseado em dados antropométricos de astronautas⁸. Ele foi idealizado para medições da transmissão sonora de capacetes de astronautas, testes de equipamentos de comunicação, tais como fones de ouvido, e também para calibração de microfones em campo acústico da voz humana. Para isso, o manequim apresentado possuía dois ouvidos artificiais e voz artificial, esta implementada utilizando um alto-falante.



Figura 4: O manequim desenvolvido por Bauer, Torick e colegas. Fonte: [26].

Os ouvidos foram construídos para responder a critérios da impedância acústica de ouvidos reais, levantados por pesquisadores como Zwischenberger. A princípio foi obtida a pressão sonora correspondente à pressão nas membranas timpânicas de uma pessoa real. Posteriormente, foi integrado um circuito elétrico para calcular a pressão sonora na entrada do canal auditivo, importante na comparação com dados experimentais. Outro circuito opcional foi responsável pela ponderação das pressões sonoras obtidas com as curvas isofônicas levantadas por Fletcher& Munson, um procedimento muito útil para uma época em que o processamento de sinais ainda era difícil.

Em 1969, a empresa alemã Sennheiser apresentou um manequim, também chamado de "Oskar", (Figura 5). Ele foi destinado à gravação de rádio-novelas e peças musicais em estúdios de rádio. Ouvindo os sinais com fones de ouvido, foi possível localizar tanto a posição lateral como vertical da fonte apesar dos erros cometidos, como a localização desta dentro da cabeça, confusão com relação à parte anterior e posterior, entre outros.

Em 1972, a Industrial Research Inc., uma subsidiária da Knowles Inc., apresentou o *Knowles Acoustics Manikin for Acoustic Research - KEMAR*⁹ (Figura 6), destinado ao desenvolvimento e à avaliação

⁸A pesquisa de Torick, Bauer e colegas foi financiada pela agência aeroespacial dos E.U.A.

⁹O KEMAR está sendo comercializado hoje pela empresa dinamarquesa G.R.A.S. b.v.



Figura 5: O manequim “Oskar” da Sennheiser. Fonte: www.sennheiser.com

de aparelhos auditivos *in-situ* [17, 5, 27]. Seu tronco, cabeça e orelha, com o pavilhão auditivo e o canal auditivo, foram dimensionados conforme dados antropométricos. Além disso modelou-se a impedância do conjunto canal auditivo e membrana timpânica e a impedância da pele.

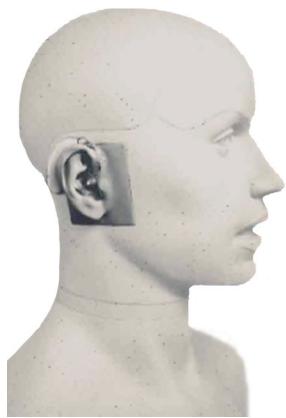


Figura 6: Kemar - o pimeiro manequim representativo. Fonte: [17].

Com geometria baseada em dados antropométricos, o manequim KEMAR tornou-se o primeiro manequim representativo e foi adotado como manequim de referência para medições *in situ* de aparelhos auditivos conforme a norma IEC 959 - *Technical Report - Provisional head and torso simulator for acoustic measurements on air conduction hearing aids* [14].

Até hoje as funções de transferência relativas à cabeça (*HRTF*) obtidas através de medições que o utilizam [10], servem como referência e são disponibilizadas na internet.

A impedância do canal auditivo e da membrana timpânica foram modeladas de tal forma que correspondessem a orelhas reais, tanto abertas quanto parcialmente ou totalmente obstruídas, condição importante para a avaliação de aparelhos auditivos. Para permitir

estudos do efeito de tamanhos diferentes de pavilhões auditivos e do canal auditivo, este conjunto podia ser facilmente trocado.

Com a feira internacional de radiocomunicação¹⁰ de 1973 em Berlim, Alemanha, deu-se início o uso da tecnologia biauricular para gravações de rádio-novelas na Alemanha. As novelas gravadas desta maneira foram um grande sucesso. Porém, um dos grandes problemas daquela época, era a baixa razão sinal-ruído do sistema de gravação, este que ficou muito evidente com a utilização de fones de ouvido para escutar os sinais. Com o uso de alto-falantes o ruído era bem menos perceptível, mas os sinais gravados não eram muito compatíveis a este tipo de reprodução. Essa não compatibilidade refere-se a problemas devido à falta de equalização ou à aplicação de equalização errada, resultando na mudança considerável de timbre do sinal reproduzido pelos alto-falantes com relação ao original. Além disso ocorreu o que se conhece como efeito da fala cruzada, fazendo com que o sinal emitido pelo alto-falante esquerdo também chegassem ao ouvido direito e *vice versa*, comprometendo a localização da fonte.

Com a utilização de fones de ouvido, que a princípio impede o efeito da fala cruzada, ocorreram também problemas com a localização da fonte sonora. A impressão que se teve foi que ela estava muito perto ou até dentro da cabeça do ouvinte (*in-head localization*). Além disso, houve a inversão da localização no plano mediano (fontes que estavam a frente são localizadas atrás quando escutadas) e não foi possível definir a posição da fonte em relação ao plano horizontal [13].

Apesar desses problemas, era grande o interesse de profissionais de áudio na tecnologia biauricular de gravação e reprodução naquela época e, para atender a demanda a um preço razoável, a Sennheiser lançou, em 1974, o primeiro *headset* biauricular, o MKE 2002. Este, parecido com um estetoscópio, possibilitou gravações biauriculares quando colocado na cabeça de uma pessoa ou de um boneco, como na Figura 7. Mais tarde outras empresas do ramo de áudio também lançaram sistemas de gravação biauricular, por exemplo as japonesas Sony e JVC. Todos eles não estão sendo mais produzidos.

Nos anos 70, muitos estudos foram desenvolvidos a fim de melhorar a fidelidade das reproduções biauriculares, tanto com alto-falantes quanto com fones de ouvido. Estes estudos mostraram que a presença de elementos como ombros e tronco e também a modelagem suficientemente precisa dos pavilhões auditivos, são importantes para uma reprodução do som tal qual como foi gravado *true-to-original*. Ficou também evidente a necessidade da equalização correta de toda a cadeia de gravação e reprodução e de uma razão sinal-ruído bem melhor.

Em 1975, a empresa alemã Neumann apresentou a cabeça KU 80, destinada sobretudo a gravações

¹⁰De alemão: *Internationale Rundfunkausstellung*.



Figura 7: O primeiro *headset* biauricular, o MKE 2002 da Sennheiser, em uma cabeça de boneco.

Fonte: www.literaturcafe.de/bilder

de peças musicais. Microfones muito pequenos e confiáveis, disponíveis no final dos anos 70, possibilitaram a obtenção de sinais no interior dos ouvidos de pessoas reais, abrindo espaço para aquisição de sinais com pessoas por técnica intra- ou extrauricular. Baseando-se em resultados de pesquisas com minimicrofones posicionados nos canais auditivos de pessoas, melhorou-se a cabeça da Neumann e, em 1981, a cabeça KU 80, com equalização de campo-livre, foi substituída pela KU 81. Esta tinha modelos mais fidedignos dos pavilhões auditivos e possuía equalização de campo-difuso [13] em vez de equalização de campo livre, objetivando melhor compatibilidade dos sinais com reprodução via alto-falantes (com cancelamento da fala-cruzada) e a redução dos erros de localização.



(a) KU 80



(b) KU 81

Figura 8: As cabeças artificiais Neumann KU 80 e KU 81. Fonte: www.neumann.com

A equalização de campo difuso ou campo livre dos sinais biauriculares foi necessária para garantir que o sinal não fosse modificado duas vezes durante o caminho *fonte - cabeça artificial - alto-falante - ouvido* da pessoa, por uma função de transferência relacionada à cabeça. A equalização retira totalmente ou parcialmente a influência do manequim sobre o sinal para que,

no caso da reprodução via alto-falantes, apenas a função de transferência relacionada à cabeça da pessoa que escuta os sinais fosse aplicada. Isso não é necessário caso a pessoa escute os sinais com fones de ouvido, o que fez com que, no princípio, o sinal transmitido só pudesse ser compatível com a apresentação via fones ou alto-falantes.

5 OS ANOS 80, NOVAS APLICAÇÕES

Nos anos 80 houve um interesse crescente em gravações e medições de sons técnicos, diz-se ruídos, de uma forma fidedigna, ou seja, comparável à audição humana. Ao mesmo tempo exigiu-se que os sinais pudessem ser analisados por meios comuns de análise de sinais e comparados com sinais obtidos através de microfones simples. Essa exigência paradoxal foi cumprida parcialmente pela aplicação de curvas de equalização que retiram, totalmente ou parcialmente, a influência do manequim sobre o sinal, possibilitando sua análise. Naturalmente, para reproduções via fones de ouvido os sinais precisam ser reequalizados.

Entre 1980 e 1982, Genuit e colaboradores, no *Institut für elektrische Nachrichtentechnik* (instituto de telecomunicações) da universidade técnica de Aachen (Alemanha), desenvolveram, em cooperação com uma montadora alemã de automóveis, um sistema de gravação biauricular que ficou conhecido mais tarde como AachenHead. Este sistema apresentava uma razão sinal-ruído muito aceitável e foi o primeiro a possuir duas interfaces de equalização de gravação, uma de equalização de campo livre e outra de campo difuso. Assim, pela primeira vez havia a possibilidade de escolha da equalização de gravação mais adequada, em função das condições acústicas encontradas no ambiente de gravação ou medição.



(a) AachenHead



(b) HMS II

Figura 9: O manequim AachenHead e o sucessor HMS II. Fontes: www.kettering.edu e foto do autor

O AachenHead deu lugar a outro manequim, conhecido como HMS II, por volta de 1986, com uma geometria simplificada porém representativa. Novamente utilizou-se dados antropométricos, desta vez levantados pela Universidade de Essen na Alemanha, mas optou-

se pela simplificação da geometria do manequim, desprezando todos os elementos que não modifcassem os sinais no seu caminho aos receptores. Desta forma não foram modelados nariz, boca, etc e chegou-se a uma geometria extremamente estética. O ouvido externo, por sua vez, foi modelado por elementos de geometria simples e matematicamente descriptíveis, sem que houvesse diferenças significativas na função de transferência relacionada à cabeça *HRTF*, considerando fontes distantes [11, 12].

Um caminho similar de simplificação por superfícies geométricas foi escolhido para o manequim 4128 [4] e mais tarde o 4100, da fabricante dinamarquesa Brüel&Kjær.



Figura 10: O primeiro manequim 4128 da Brüel&Kjær.
Fonte: [4].

Comercializado a partir de 1985, o manequim 4128 inclui todo o torso humano e possui modelos dos pavilhões auditivos bastante fidedignos, similares às do KEMAR. Tem como opção o uso de simuladores da impedância acústica do ouvido humano que podem ser conectados aos pavilhões auditivos. Além disso, apresenta voz artificial que, junto à simulação da impedância, qualificou-o para testes de fontes próximas, tais como aparelhos auditivos e telefones.

Naquela época, a Brüel&Kjær era bastante cautelosa quanto à finalidade de uso deste produto, pois também comercializava um sistema completo de medição de telefones, o 3356 e 3357 [4]. Neste mesmo período, a fabricante informava apenas a função de transferência relativa à cabeça do manequim 4128, em campo livre, sem que essa curva fosse utilizada para equalização dos sinais. Isso a diferenciava de outros manequins, como por exemplo o HMS II, que possuía equalização variável, e o KU 81, com equalização fixa para campo difuso.

6 OS ANOS 90

Nos anos 90, empresas que já ofereciam manequins ou cabeças artificiais na década passada, como a Brüel&Kjær, a HEAD-acoustics, a Knowles e a Neumann, continuaram melhorando os seus produtos, como

por exemplo com a implementação de processamento digital para equalização.

Outras empresas como a Cortex¹¹, da Alemanha, começaram a produzir manequins próprios e muitos laboratórios de Acústica desenvolveram cabeças ou manequins. Um destes é o manequim desenvolvido por Schmitz e colegas [19], no Instituto de Acústica Técnica da Universidade de Aachen, Alemanha. Este manequim tinha seus pavilhões auditivos escolhidos conforme um critério de melhor performance para localização, diferente dos pavilhões auditivos do KEMAR, por exemplo, que são escolhidos por representar uma média de pessoas.

Já a partir dos anos 80, as empresas apresentaram conceitos diferentes para os seus produtos. A Neumann, por ser uma fabricante de microfones de estúdio, dedicou-se ao ramo musical, sendo a cabeça KU 100, sem dúvida, a mais conhecida neste ramo. A Knowles e a Brüel&Kjaer fabricam manequins sobretudo para medição de fontes próximas, tais como aparelhos auditivos, de telefone e fones de ouvido. E a HEAD-acoustics deu prioridade para aplicações na área automobilística.

7 CONCLUSÕES

O presente artigo resgatou um pouco da história de uma parte importante da tecnologia biauricular, a história das cabeças artificiais ou manequins, que por muitas pessoas são considerados invenções muito recentes. Assim foi dada atenção a contribuições importantes, porém muitas vezes esquecidas, de pessoas como Fletcher e colegas, entre muitos outros. Mesmo existindo há bastante tempo, essa tecnologia atravessou e ainda está passando por muitos desafios para conseguir lidar com problemas como a individualidade das funções de transferência relacionadas à cabeça, entre outros.

8 AGRADECIMENTOS

Agradeço muito ao Prof. Michael Vorländer, do Instituto de Acústica Técnica da Universidade de Aachen, pelas dicas e orientações. Sou também muito grato pelas publicações que consegui por meio dos colegas Pascal Dietrich e Gottfried Behler, ambos do Instituto de Acústica Técnica da Universidade de Aachen. Agradeço também ao Prof. Arcanjo Lenzi, do Lab. de Vibrações e Acústica da Universidade Federal de Santa Catarina, por sempre me deixar consultar as obras de sua biblioteca pessoal. Finalmente sou muito grato pelo esforço da minha colega Camila Sato nas correções de português.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Scientific American. The telephone at the Paris Opera. *Scientific American*, 31:422–423, December 1881. <http://earlyradiohistory.us/1881opr.htm>.

¹¹Hoje parte da 01dB, do grupo francês Areva.

- [2] David C. Apps. The use of binaural tape recording in automotive noise problems. *J. Acoust. Soc. Am.*, 24(6):660–662, 1952.
- [3] B. B. Bauer, A. J. Rosenheck, and L. A. Abbagnaro. External-ear replica for acoustical testing. *J. Acoust. Soc. Am.*, 42(1):204–207, Jul 1967.
- [4] Brüel&Kjær. Electronic instruments master catalogue. Catalogue, 1986.
- [5] Mahlon Burkhard. A manekin useful for hearing aid tests - revisited. In *Proc. International Congress of Acoustics ICA*, 2004.
- [6] A. Charlin. Techniques phonographiques – la compatibilité. *Toute l'Electronique*, pages 468 – 471, novembre 1965.
- [7] H.W. Dove. *Repertorium der Physik*, volume 3. Veit & Comp., 1839.
- [8] M. Ericson, W. D'Angelo, E. Scarborough, S. Rogers, P. Amburn, and D. Ruck. Applications of virtual audio. In *Proceedings of the IEEE 1993 National Aerospace and Electronics Conference (NAECON)*, volume 2, pages 604–611, 1993.
- [9] Stephen H. Fletcher. Harvey Fletcher 1884-1981 A biobliographical memoir by Stephen H. Fletcher. Memoir, 1992.
- [10] Bill Gardner and Keith Martin. HRTF measurements of a KEMAR dummy. Technical report, MIT Media Lab, May 1994.
- [11] Klaus Genuit. *Ein Modell zur Beschreibung von Außenohrübertragungseigenschaften*. PhD thesis, RWTH Aachen, 1984.
- [12] Klaus Genuit. Eine systemtheoretische Beschreibung des Aussenohres. In *Fortschritte der Akustik - DAGA'85*, pages 459–462, 1985.
- [13] Herbert Hudde and Jürgen Schröter. Verbesserungen am Neumann Kunstkopfsystem. *Rundfunktechnische Mitteilungen*, 25(1):1–6, 1981.
- [14] IEC. IEC 959 - technical report - provisional head and torso simulator for acoustic measurements on air conduction hearing aids. International standard.
- [15] W. Bartlett Jones. Method and means for the ventriloquial production of sound. U.S. Patent 1855149, April 1932. filed April 13, 1927.
- [16] John Klepko. *5-Channel Microphone Array with Binaural-Head for Multichannel Reproduction*. PhD thesis, Faculty of Music, McGill University, Montreal, 1999.
- [17] Hugh S. Knowles. Manikin measurements. In *Manikin Measurements*, 1978.
- [18] J.W.S. Rayleigh. *The Theory of Sound*, volume 2 of *Dover Classics of Science and Mathematics*. Dover, unabridged second revised edition edition, 1945.
- [19] A. Schmitz. Ein neues digitales Kunstkopfmesssystem. *Acustica*, 81:416–420, 1995.
- [20] A. Seebeck. *Repertorium der Physik*, volume 8. Veit & Comp., 1844.
- [21] Ross H. Snyder. History and development of stereophonic sound recording. *J. Audio Eng. Soc.*, 1(2):176–179, 1953.
- [22] A. Steinhauser. The theory of binaural audition. *Phil. Mag.*, 7:181–197, 261 – 274, 1877.
- [23] S.P. Thompson. On binaural audition. *Phil. Mag.*, 4(5):274–276, July-December 1877.
- [24] S.P. Thompson. *Phil. Mag.*, VI:385, 1878.
- [25] S.P. Thompson. On the function of the two ears in the perception of space. *Phil. Mag.*, 13:406–416, 1882.
- [26] E.L. Torick, A. Di Mattia, A.J. Rosenheck, L. A. Abbagnaro, and B.B. Bauer. An electronic dummy for acoustical testing.pdf. *J. Audio Eng. Soc.*, 16:397–493, 1968.
- [27] Michael Vorländer. Past, present and future of dummy heads. In *Proc. Acústica*, Guimarães, Portugal, 2004.
- [28] Jörg Wuttke. *Mikrofonaufsätze*. Schoeps, 2000.



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil

12ª Convenção Nacional da AES Brasil

5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Método analítico para o cálculo da Diferença Interaural de Tempo (ITD) no plano horizontal

Camila Tigussa Sato,¹ Marcelo Lapa Espiga,¹ Stephan Paul,¹ Samir N.Y. Gerges¹
e Pascal Dietrich²

¹ Universidade Federal de Santa Catarina, Laboratório de Vibrações e Acústica
Florianópolis, Santa Catarina, Brasil

² RWTH Aachen University, Institute of Technical Acoustics
Aachen, Alemanha

milatsato@gmail.com, espigamarcelo@gmail.com, stephan.paul.acoustic@gmail.com,
samir@emc.ufsc.br, pdi@akustik.rwth-aachen.de

RESUMO

Sistemas surround de reprodução de som objetivam envolver o ouvinte fazendo este localizar-se em relação ao evento sonoro. Um fator que contribui para a localização de fontes sonoras é a diferença do tempo de chegada de uma mesma frente de onda sonora a cada ouvido, esta chamada de *ITD (Interaural Time Difference)*. Desenvolveu-se um modelo analítico para calcular esta diferença no plano horizontal, que simplifica a cabeça como um cilindro, e fez-se ensaios práticos para a validação deste. A similaridade entre os resultados obtidos valida o modelo analítico e a aproximação da cabeça como um cilindro para o caso do fator *ITD*.

0 INTRODUÇÃO

A evolução de sistemas de reprodução de áudio é crescente, tornando necessário o desenvolvimento de tecnologias que possibilitem reproduções cada vez mais fiéis. A percepção dos ouvintes frente aos estímulos sonoros reproduzidos deve ser a mesma da obtida com o evento sonoro real. Para isso, é necessário se preocupar com diversos fatores que influenciam nessa percepção, como a intensidade sonora, o timbre e a capacidade de

localização da fonte sonora.

A percepção das fontes sonoras é um recurso que ajuda os seres humanos na percepção do ambiente em que se encontram ou na percepção de sons desejáveis dentro de um cenário com muito ruído de fundo (*cocktail-party-effect*). O fato do homem possuir dois ouvidos, atrelado a elementos como o tronco, a cabeça, os pavilhões auditivos, entre outros, induz a vários in-

dicadores biaurais e monaurais¹ nos sinais que chegam às membranas timpânicas. A diferença no tempo de chegada de uma mesma frente de onda sonora em cada um dos ouvidos é um dos fatores que proporciona a percepção da localização da fonte. A esse fenômeno dá-se o nome de Diferença Interaural de Tempo, ou *ITD* (*Interaural Time Difference*).

Vários modelos analíticos foram propostos ao longo de mais de 100 anos de pesquisa sobre audição biauricular para o cálculo da *ITD*. Os mais simples, desconsideravam a cabeça, deixando os receptores livres no espaço, e desta forma não consideravam a difração do som na cabeça. Em seu livro, de 1962, Woodworth & Schlosberg [5] apresentam um modelo analítico para o cálculo da diferença interaural de tempo, no plano horizontal, de uma esfera rígida e com os receptores no eixo de simetria. Este modelo, que calcula a *ITD* de forma independente da freqüência do som incidente mas considera a difração do som na esfera, é dado por [4]:

$$ITD = \frac{a}{c}(\sin \phi + \phi) \quad (1)$$

sendo a o raio da esfera, c a velocidade do som e ϕ o ângulo de incidência em radianos.

Segundo Minaar e colegas [?] este modelo foi expandido por Lercher & Jot para incluir elevações da fonte e também por Savioja *et al.* para melhor adequar aos dados experimentais.

O modelo apresentado por Woodworth & Schlosberg, assim como as extensões propostas por Lercher & Jot e Savioja *et al.* consideram que a distância da fonte r é muito maior do que o raio R da cabeça ou esfera.

1 MODELO ANALÍTICO PARA CÁLCULO DA DIFERENÇA INTERAURAL DE TEMPO (ITD)

O modelo desenvolvido neste trabalho considera a cabeça como um cilindro perfeito (de raio igual a 19cm), com os ouvidos localizados em posições exatamente opostas no plano coronal², e a fonte sonora posicionada a uma distância fixa qualquer do centro do cilindro, movendo-se em torno deste no mesmo plano horizontal em que estão contidos os ouvidos.

O cálculo da *ITD* é baseado na Equação (2), sendo $\Delta d = |d_1 - d_2|$, considerando d_1 e d_2 as distâncias percorridas pela onda sonora da fonte (P) a cada ouvido, e c a velocidade do som no ar. Já que a velocidade do som no ar pode ser considerada uma constante ($c = 343[\text{m/s}]$), nesta situação, é possível determinar a

ITD calculando Δd .

$$ITD = \frac{\Delta d}{c} \quad (2)$$

O cálculo de Δd depende da posição da fonte no plano horizontal, descrito pela ângulo (ϕ). Assim, o método é dividido em três situações diferentes de ϕ .

1.1 Cálculo de Δd para $0 \leq \phi < \alpha$

O primeiro caso analisado refere-se à valores de ϕ em que ocorre difração do som am ambos os lados da cabeça.

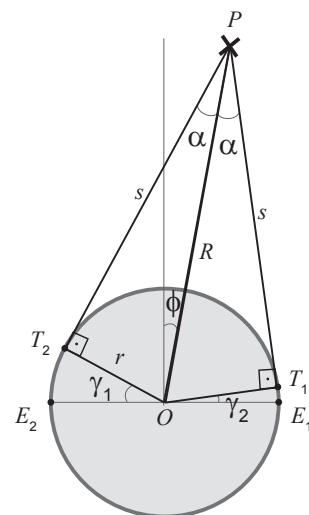


Figura 1: Representação do sistema com $0 \leq \phi < \alpha$

Observando a Figura 1, tem-se que P representa a posição da fonte sonora, O é o ponto central do cilindro, R é a distância entre P e O , ϕ é o ângulo formado entre P e o plano mediano, E_1 e E_2 são as posições dos ouvidos, r é o raio do cilindro, s é a reta que parte da posição da fonte e tangencia o cilindro, T_1 e T_2 são os pontos onde s toca o cilindro e α é o ângulo entre as retas \overline{OT} e \overline{OE} . Assim, as distâncias percorridas pelo som até cada ouvido serão:

$$d_i = s_i + r\gamma_i. \quad (3)$$

Para $\phi = 0$, tem-se que $\gamma_1 = \gamma_2 = \alpha$ e que $\alpha = \text{arcsen}(r/R)$ é sempre constante (o triângulo retângulo formado por R , r e s é mantido para qualquer ϕ). A reta s será calculada através da aplicação do teorema de Pitágoras, pois se conhece o valor de R e r , e γ será o mesmo para ambos os lados resultando na igualdade entre as distâncias d_1 e d_2 .

Com a variação de ϕ , s mantém seu valor porém os valores de γ serão diferentes para cada lado, com $\gamma_1 = \alpha - \phi$ e $\gamma_2 = \alpha + \phi$. Então Δd é calculado através da Equação 4.

$$\begin{aligned} \Delta d &= r(\alpha + \phi) - r(\alpha - \phi) \\ &= 2r\phi \end{aligned} \quad (4)$$

¹Também chamados e biauriculares e monauriculares.

²Em anatomia, plano coronal ou frontal é o plano que divide o corpo em duas partes: anterior e posterior. Na acústica, o plano frontal é considerado o plano vertical em que se situam os ouvidos. Neste trabalho, considerar-se-á plano coronal o que divide a cabeça em duas metades exatamente iguais e o plano frontal o plano em que estão os ouvidos (nos seres humanos, este plano em é posterior ao plano coronal).

1.2 Cálculo de Δd para $\phi = \alpha$

O ângulo ϕ será igual a α quando o ponto tangente T_1 coincidir com o ouvido E_1 , como mostra a Figura 2.

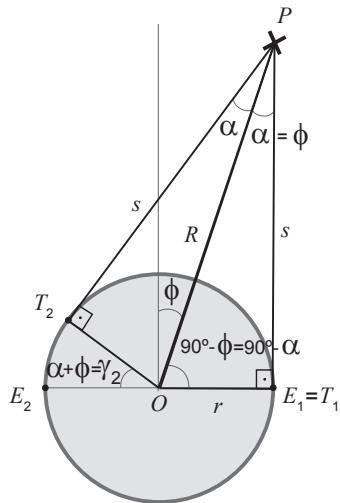


Figura 2: Representação do sistema com $\phi = \alpha$

Neste caso, Δd será obtido através da Equação (5)

$$\Delta d = 2r\phi = 2r\alpha \quad (5)$$

1.3 Cálculo de Δd para $\alpha < \phi \leq 90^\circ$

A Figura 3 esquematiza a situação para $\alpha < \phi \leq 90^\circ$, sendo $\gamma_2 = \alpha + \phi$ e $d_2 = \sqrt{(r - R \sin \phi)^2 + (R \cos \phi)^2}$.

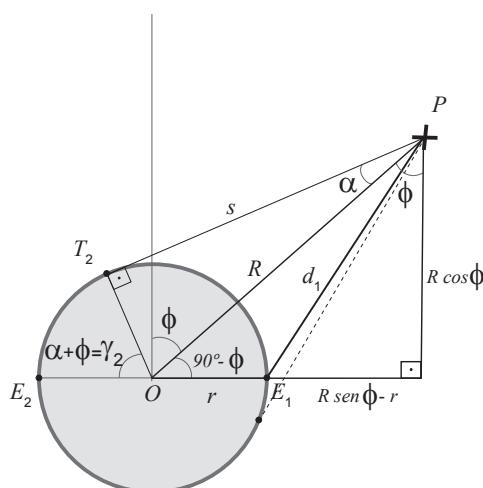


Figura 3: Representação do sistema com $\alpha < \phi \leq 90^\circ$

Assim, para este caso, Δd é calculado através da

Equação (6).

$$\begin{aligned} \Delta d &= d_1 - d_2 = (r(\alpha + \phi) + s) - d_2 \\ &= r(\alpha + \phi) + s - d_2 \\ &= r[\arcsen(\frac{r}{R}) + \phi] + \sqrt{R^2 - r^2} \\ &\quad - \sqrt{(r - R \sin \phi)^2 + (R \cos \phi)^2} \\ &= r[\arcsen(\frac{r}{R}) + \phi] + \sqrt{R^2 - r^2} \\ &\quad - \sqrt{r^2 - 2rR \sin \phi + R^2} \end{aligned} \quad (6)$$

Observa-se que este caso também é válido para $\phi = \alpha$.

1.4 Obtenção da diferença interaural do tempo ITD

Utilizando a Eq. (2) e as Eqs. (4) e (6) obtém-se, para todos os ângulos $0^\circ \leq \phi \leq 90^\circ$, a diferença interaural de tempo ITD:

$$ITD(\phi) = \begin{cases} \frac{1}{c}2\phi r & \text{se } 0 \leq \phi < \alpha \\ \frac{1}{c}(r[\arcsen(\frac{r}{R}) + \phi] \dots \\ + \sqrt{R^2 - r^2} \dots \\ - \sqrt{r^2 - 2rR \sin \phi + R^2}) & \text{se } \alpha \leq \phi < 90^\circ \end{cases} \quad (7)$$

Considerando a simetria do cilindro, pode-se expandir o método para posições da fonte que tem ângulos de incidência maiores que 90° utilizando:

$$ITD(\phi) = \begin{cases} ITD(\phi) & \text{se } 0 \leq \phi < 90^\circ \\ ITD(180^\circ - \phi) & \text{se } 90^\circ \leq \phi < 180^\circ \\ -ITD(\phi - 180^\circ) & \text{se } 180 \leq \phi < 270^\circ \\ -ITD(360^\circ - \phi) & \text{se } 270 \leq \phi < 360^\circ \end{cases} \quad (8)$$

e assim, a Figura 4 é obtida. A função dada pela Eq. (8) é uma função contínua, porém não diferenciável em $\phi = 90, 180, 270, 360^\circ$.

2 VALIDAÇÃO EXPERIMENTAL

Para validar o modelo analítico, as mesmas condições a ele aplicadas foram implementadas em um experimento prático, como esquematizado na Figura 5. As medições foram feitas na câmara semi-anecóica, onde o piso reflexivo foi coberto por espuma, evitando reflexões vindas das paredes e do chão que interferiam nas medições. O posicionamento da fonte foi variado com $\Delta\phi = 5^\circ$, no plano horizontal. Para a validação foram utilizados dois mois modelos diferentes da cabeça humana: o primeiro foi um cilindro, igual ao cilindro considerado no modelo analítico, com dois microfones capacitivos simulando as orelhas e o segundo utilizando um manequim.

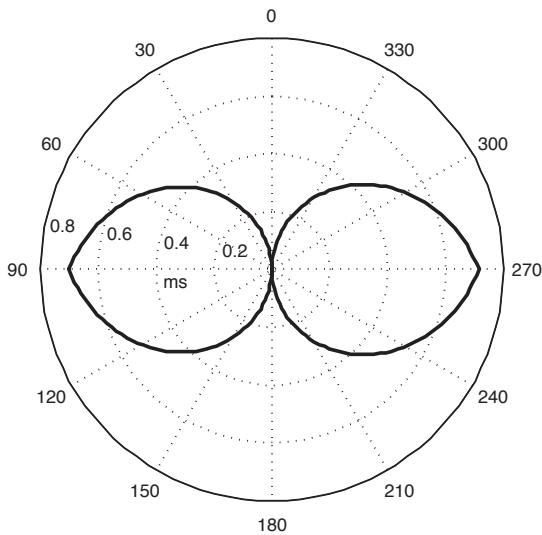


Figura 4: *ITD*, em milisegundos, obtida através do modelo analítico, no plano horizontal

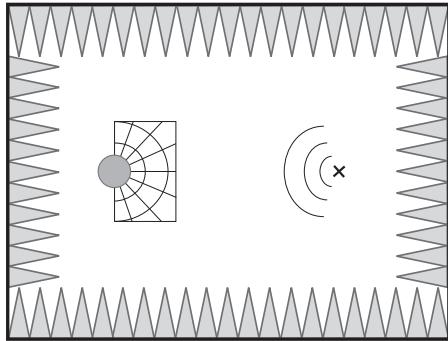


Figura 5: Esquema da montagem experimental

2.1 Comparação com um cilindro

O cilindro utilizado tinha um diâmetro de 19 cm e foi feito de papelão. Dois microfones capacitivos de meia polegada foram posicionados em lados opostos no eixo de simetria, simulando a cabeça humana com ouvidos. A fonte sonora foi posicionada a 4 metros de distância do centro do cilindro, no mesmo plano dos microfones (plano horizontal). Dos diferentes métodos de medição da resposta impulsiva [1] optou-se pelo método mais clássico, utilizando um ruído impulsivo como sinal de excitação. Este foi gerado pela batida de dois pedaços de madeira, uma forma clássica de se obter um sinal impulsivo em medições de acústica de salas. O sinal obtido desta forma possui energia em todas as freqüências audíveis e o procedimento experimental adequa-se ao modelo analítico que é independente da freqüência do sinal emitido pela fonte.

Outras técnicas de medição poderiam utilizar uma varredura senoidal, o que melhoraria muito a razão sinal/ruído além de possibilitar a obtenção de mais informações, tais como diferenças de intensidade,

correlação cruzada e análise em freqüência. Porém, o sinal utilizado é propriamente aplicável neste caso.

A diferença de tempo de chegada da frente de onda sonora entre os sinais dos microfones foi tomada a partir do primeiro vale de sinal no domínio do tempo, pois a primeira onda sonora provém obrigatoriamente da fonte e não de possíveis reflexões. A Figura 6 exemplifica um dos casos analisados. As diferenças de tempo obtidas através desse procedimento foram plotadas com relação ao ângulo da posição da fonte/de incidência do som, como mostrado na Figura 7. Como era esperado, o valor mínimo da *ITD*, 0ms, foi encontrado quando a fonte sonora estava posicionada exatamente à frente do cilindro (0°), e o máximo, 0,69ms, quando o ângulo de incidência da fonte era de 90° em relação ao cilindro.



Figura 6: Sinal no domínio do tempo obtido com $\phi = 25^\circ$

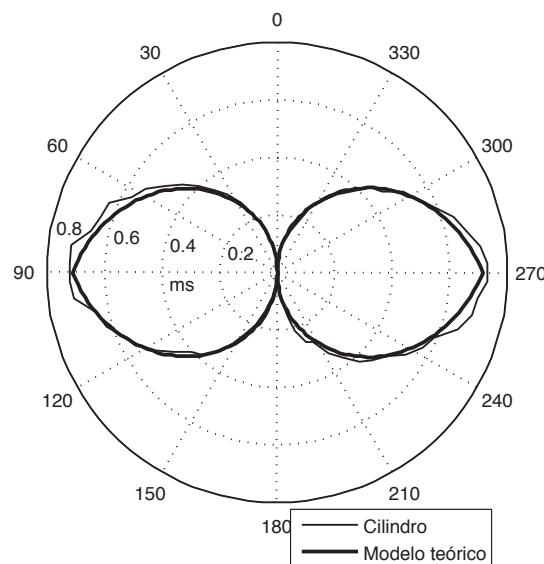


Figura 7: *ITD*, em milisegundos, obtida através do modelo teórico e do experimento prático, no plano horizontal

2.2 Comparação com um manequim

Para comparar os dados obtidos a partir do modelo analítico com dados que representassem melhor a *ITD* de seres humanos, foram feitas medições utilizando um manequim (HMS III.0) da Head-acoustics (Figura 8) nas mesmas condições do experimento anterior.



Figura 8: Vista lateral do manequim HMS III.0

Observa-se que o centro da cabeça foi definido como o ponto médio da reta que divide a cabeça em dois hemisférios (esquerdo e direito), sendo este ponto anterior à linha dos ouvidos.

O resultado obtido com o ensaio utilizando o manequim é mostrado na Figura 9.

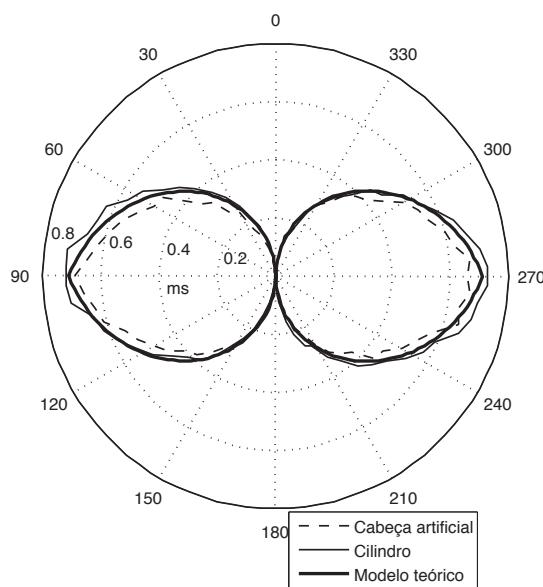


Figura 9: *ITD*, em milisegundos, obtida através do modelo teórico, experimento prático com o uso do cilindro e do experimento prático com o uso da cabeça artificial, no plano horizontal

2.3 Análise dos resultados

Calculando-se a diferença entre as *ITDs* encontradas no modelo analítico e nas validações experimentais ($\Delta ITD = ITD_{an} - ITD_{med}$) obteve-se a Figura 10.

Analizando a Figura 10, nota-se que as diferenças nos valores da *ITD* são pequenas (< 10%). Assim

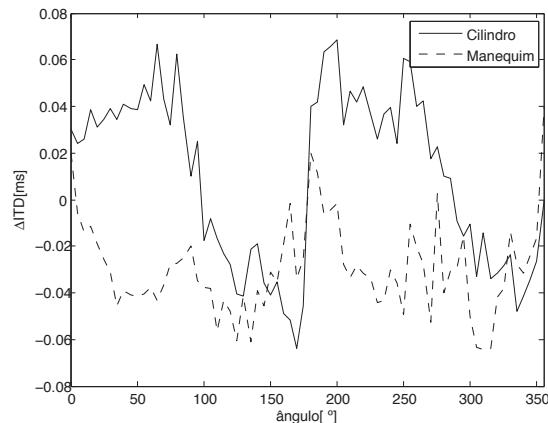


Figura 10: Diferença entre a *ITD* calculada com o modelo analítico e a *ITD* medida com o uso do cilindro e manequim ($ITD_{an} - ITD_{med}$)

como o esperado, o valor médio das diferenças encontrado para o cilindro ($\Delta ITD_m = 0,01 \text{ ms}$) foi menor que o valor encontrado para o manequim ($\Delta ITD_m = 0,023 \text{ ms}$), considerando que o método analítico foi desenvolvido a partir da simplificação da cabeça humana como um cilindro. Vale ressaltar que os três métodos de obtenção da *ITD* independem da freqüência.

As prováveis fontes responsáveis pela discrepância encontrada entre os valores da *ITD* medida e calculada foram erros humanos na variação do posicionamento da fonte, imperfeições do cilindro e o formato do manequim que difere bastante de um cilindro.

3 CONCLUSÕES

Como visto na Figura 10, a diferença entre a *ITD* obtida com o modelo analítico e com o resultado experimental utilizando o cilindro foi bastante pequena (menor que 0,1 ms), mostrando que o modelo desenvolvido foi bastante satisfatório. Além disso foi possível encontrar valores semelhantes aos das literaturas [2] e [4].

Por outro lado, foi surpreendente a *ITD* encontrada com o uso do manequim. Esperava-se que a diferença de resultados seria maior já que o manequim, assim como a cabeça real, tem os ouvidos localizados no plano frontal e não no plano coronal, como foi simplificado no modelo analítico e teórico. Além disso, a forma da cabeça do manequim é muito diferente de um cilindro. Porém, isso não se provou verdadeiro e os três sistemas desenvolvidos para a obtenção de valores de *ITD* foram bastante semelhantes, como mostram as Figuras 9 e 10.

Pode-se dizer então que o modelo analítico apresentado (Eqs. (7) e (8)) é válido em estudos referentes à Diferença Interaural de Tempo para a localização de fontes sonoras no plano horizontal, tendo como principal diferencial o fato de ser também válido para fontes próximas, ao contrário das soluções apresentadas por Woodworth & Schlosberg [5] e Miller *et al.* [3]. Ao considerar $r \gg R$, a solução descrita neste artigo é igual à encontrada por Woodworth & Schlosberg [5].

4 AGRADECIMENTOS

Agradeçemos muito ao Prof. Michael Vorländer, do Instituto de Acústica Técnica da Universidade de Aachen, pelas dicas e orientações.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Akustik – Anwendung neuer Messverfahren in der Bau- und Raumakustik (Iso 18233:2006); Deutsche Fassung EN ISO 18233:2006. Technical standard, March 2003.
- [2] G.F. Kuhn. Model for the interaural time differences in the azimuthal plane. *J Acoust Soc Am*, 82:157–167, 1977.
- [3] J.D. Miller. Modeling Interaural Time Difference Assuming a Spherical Head. Technical report.
- [4] Pauli Minaar, Jan Plogsties, Søren Karup Olesen, Flemming Christensen, and Henrik Møller. The interaural time difference in binaural synthesis. In *Proc. AES Convention*, Paris, February 2000.
- [5] R.S. Woodworth and H. Schlosberg. *Experimental Psychology*. Holt, Rinehart and Winston, 1962.



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso de Engenharia de Áudio

12ª Convenção Nacional da AES Brasil

5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Análise Quantitativa do Simulador Acústico RAIOS

Julio Cesar B. Torres,¹ Flavia Correia Tovo,¹ Mariane R. Petraglia¹ e
Roberto A. Tenenbaum²

¹ Universidade Federal do Rio de Janeiro, Escola Politécnica,
Centro de Tecnologia, Cidade Universitária, RJ, CEP 21945-970, Brasil

² Universidade do Estado do Rio de Janeiro, Instituto Politécnico, Depto. Eng. Mecânica - Friburgo,
RJ, Brasil

julitorres@ufrj.br, flaviatovo@ieee.org, mariane@pads.ufrj.br, tenenbaum@iprj.uerj.br

RESUMO

Neste artigo é apresentada uma análise de desempenho do simulador acústico RAIOS 3, através da comparação dos parâmetros de qualidade acústica obtidos por simulação e por medição *in loco*. A medição foi realizada criteriosamente em uma sala da POLI/UFRJ utilizando o software AcMus, desenvolvido pela USP. Foram realizadas medições monoaurais (microfone) e binaurais (cabeça artificial). Os resultados obtidos mostraram-se consistentes e permitiram avaliar a confiabilidade do simulador acústico quanto à obtenção dos principais parâmetros de qualidade acústica de salas. Os maiores desvios foram encontrados nas baixas freqüências, devido ao método híbrido (traçado de raios e troca de energia) implementado no simulador não modelar perfeitamente a propagação de sons de baixas freqüências.

INTRODUÇÃO

A simulação numérica tornou-se uma das principais ferramentas para o projeto acústico de ambientes. Diversos fatores foram determinantes para o crescente uso de programas capazes de simular a propagação do som em recintos. Dentre eles, pode-se citar a substituição dos modelos em escala reduzida [1, 2], a facilidade de modificação das características acústicas e geométricas da sala, sem a necessidade de construção real e a possi-

bilidade de pré-ouvir o som em um ponto do ambiente (auralização) [3].

Através de simulação é possível prever o comportamento acústico em diversos pontos do ambiente e avaliar a sua qualidade acústica, antes de sua construção. Contudo, é fundamental que o simulador acústico seja capaz de modelar com precisão os fenômenos acústicos inerentes à propagação do som no interior do ambiente real e que as características dos materiais a serem em-

pregados na sala real correspondam aos parâmetros utilizados pelo simulador.

A qualidade acústica do ambiente pode ser determinada através da avaliação de diversos parâmetros acústicos [4] obtidos com base na resposta impulsiva. Os principais parâmetros são o Tempo de Reverberação (T_{30}), a Definição (D_{50}) e o Fator de Clareza (C_{50})[5].

O objetivo deste trabalho é comparar os resultados produzidos pelo simulador acústico RAIOS 3, em desenvolvimento na UFRJ [6], com os valores obtidos através de medição. Foram medidas as respostas impulsivas para três pontos de uma sala da POLI/UFRJ, utilizando um sistema de geração e gravação de sinais específicos para medição de respostas impulsivas de salas. Este trabalho apresenta os primeiros resultados de um conjunto de testes que estão sendo realizados para a validação e desenvolvimento do simulador RAIOS.

CARACTERÍSTICAS DO AMBIENTE

Para este trabalho utilizou-se uma pequena sala retangular, de $10,5 \text{ m}^2$. O piso era constituído de tacos de madeira, três paredes eram feitas de divisórias até o teto, revestidas com fórmica, e a outra parede era de alvenaria com pintura acrílica. O teto possuía um rebaixo em placas de lã de vidro revestidas com material plástico. No teto, devido à ausência de uma das placas de lã de vidro, havia um buraco, que foi também inserido no modelo do simulador. O modelo acústico/geométrico da sala utilizado no simulador RAIOS é apresentado na Fig. 1.

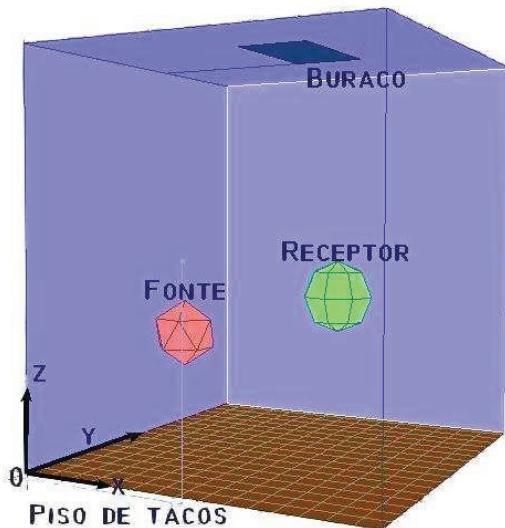


Figura 1: Modelo acústico/geométrico da sala de testes.

As medições e simulações foram realizadas para uma posição de fonte e três posições de receptor, sendo

que para cada posição de receptor os sinais foram gravados usando microfone e posteriormente a cabeça artificial. Dessa forma, para cada par fonte/receptor tem-se uma resposta impulsiva monoaural e uma resposta binaural (uma função para cada ouvido do manequim). Para assegurar que os dados obtidos são consistentes, cada gravação foi executada 10 vezes seguidas, sem que nenhuma parte do sistema fosse modificada, de forma automatizada. As dimensões da sala e as coordenadas das posições da fonte e dos receptores são mostradas na Tab. 1.

Objeto	x	y	z
Sala (dimensões)	3.50	3.00	3.46
Fonte	1.70	0.50	1.00
Receptor 1	1.70	2.00	1.45
Receptor 2	0.50	0.50	1.45
Receptor 3	3.01	2.53	1.45

Tabela 1: Dimensões e posições de fonte e receptores na sala (unid. metro).

PROCEDIMENTO DE MEDIÇÃO

O procedimento de aquisição dos sinais necessários para o cálculo das respostas impulsivas é mostrado na Fig. 2.

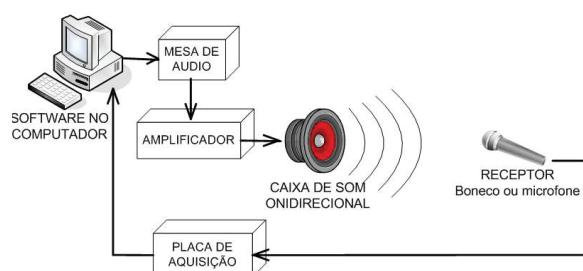


Figura 2: Esquema para medição das respostas impulsivas.

A aquisição dos sinais foi realizada utilizando o programa AcMus [7] desenvolvido pela USP. Neste sistema, um sinal de varredura senoidal é usado para excitar a sala. Esse sinal é gerado no computador e enviado para uma mesa de áudio, onde recebe uma pré-amplificação, e posteriormente é enviado para um amplificador de potência. O sinal amplificado é enviado a um conjunto de alto-falantes dispostos de forma a produzir aproximadamente a mesma pressão sonora ao seu redor (fonte onidirecional). O som que é propagado no interior da sala é captado por microfones, posicionados em locais específicos da sala. Foram utilizados um microfone para medição, modelo ECM 8000, e um manequim desenvolvido no PEE/COPPE/UFRJ que pos-

sui orelhas de silicone e microfones de eletreto posicionados nas entradas dos ouvidos artificiais. Os sinais captados pelos microfones são enviados para a placa de aquisição, que converte o áudio para digital de volta para o computador. O nível de pressão sonora medido em dBC dentro dos recintos foi de aproximadamente 100 dB. Utilizou-se este nível de pressão sonora para minimizar a influência do ruído de fundo, que se encontrava na ordem de 65 dB. As posições de medição foram escolhidas a fim de evidenciar efeitos diferentes. Na posição 1, mostrada na Fig. 3, o manequim encontrava-se de costas para a fonte, de modo que distância entre cada um dos ouvidos e a fonte fosse aproximadamente igual. Na posição 2, o ouvido esquerdo do manequim encontrava-se voltado para a fonte, recebendo o som direto da fonte enquanto o ouvido esquerdo captava predominantemente as reflexões da parede. A terceira posição foi escolhida para se ter o maior atraso no tempo de chegada do som direto e por estar numa área sob influência dos modos naturais (canto da sala). Para todas as posições a face do manequim esteve voltada para a parede 1. As posições dos receptores, as respectivas orientações e a posição da fonte podem ser observados na Fig. 3.

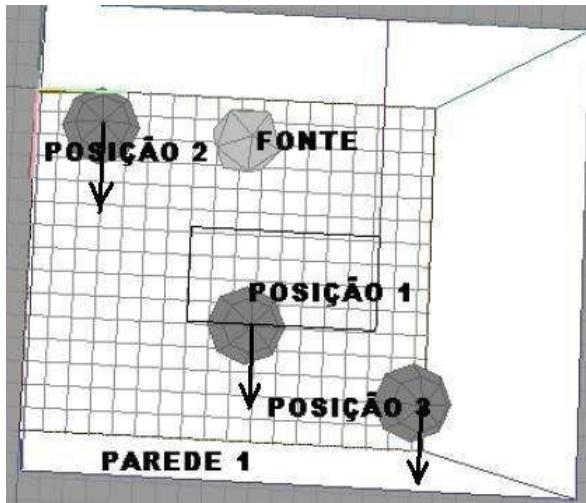


Figura 3: Posições da fonte e dos receptores na sala.

SIMULAÇÃO

A simulação foi realizada utilizando o simulador RAIOS 3, que incorpora um método híbrido de propagação sonora [6, 8], composto pela técnica de traçado de raios [9, 10] e pelo método de transição de energia [11, 12]. Para a simulação foi criado um modelo geométrico da sala, onde cada superfície é representada por um plano delimitado por áreas. Cada superfície, na sala real, possui um coeficiente de absorção e espalhamento que, na prática, não se pode medir *in*

loco. Assim, utilizou-se para a simulação coeficientes de absorção previamente medidos, cujos valores por freqüência para cada material encontram-se na Tab. 2. Para o piso utilizou-se madeira, na parede 1 foi usado alvenaria, nas demais paredes usou-se fórmica, o forro de lá de vidro foi utilizado para o teto e a absorção total foi associada ao buraco existente no teto. Os coeficientes de difusão foram estimados para esta simulação em 0,01, para todas as freqüências e materiais, uma vez que não havia como determinar, sem análise dos materiais, os verdadeiros coeficientes de difusão. Dessa forma, não há, praticamente, espalhamento dos raios ao atingir uma superfície.

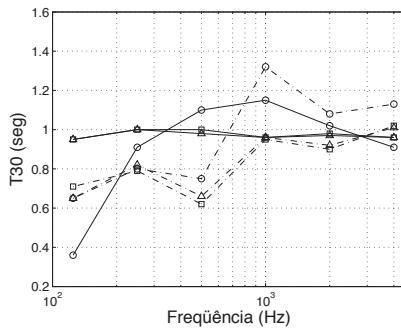
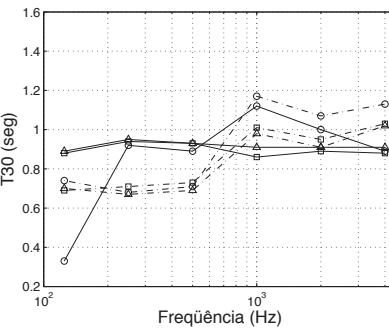
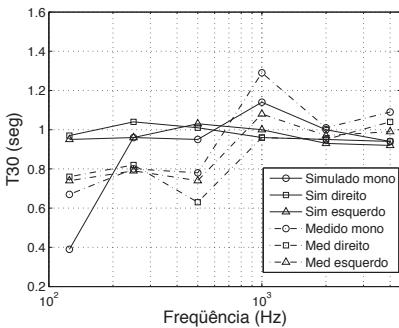
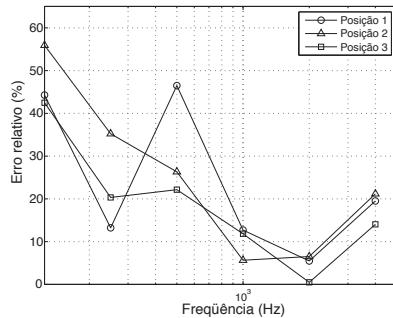
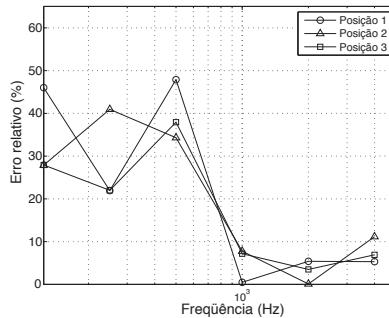
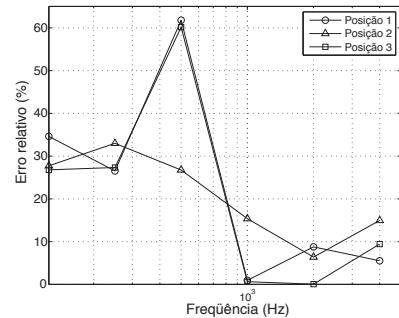
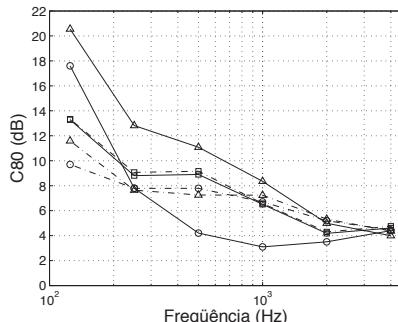
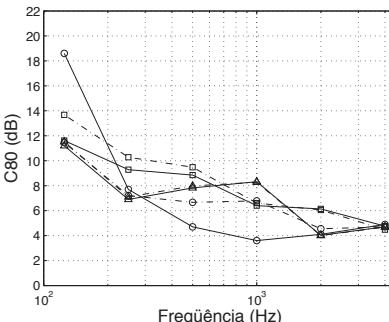
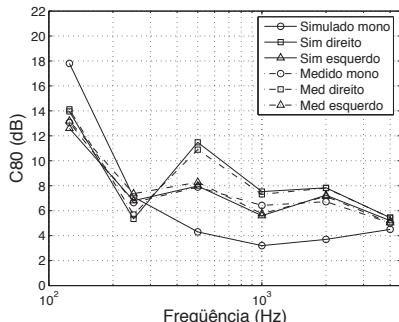
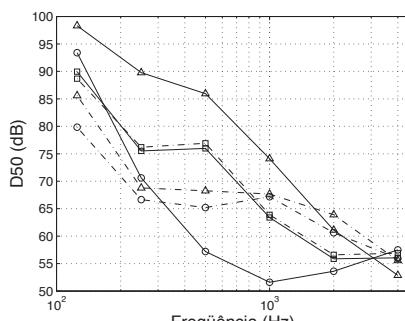
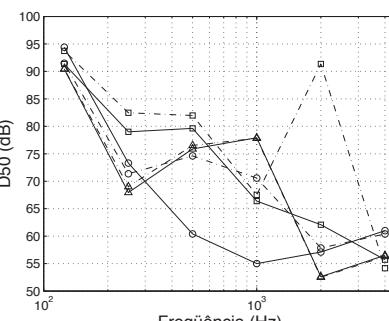
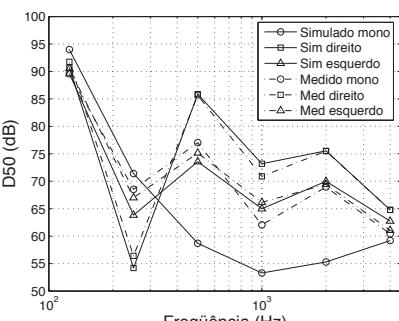
Material	Coeficiente de absorção por banda de oitava					
	125	250	500	1k	2k	4k
madeira	0,20	0,15	0,10	0,10	0,05	0,10
fórmica	0,28	0,08	0,07	0,07	0,09	0,09
alvenaria	0,02	0,02	0,02	0,02	0,02	0,02
forro	0,50	0,35	0,15	0,05	0,05	0,01
buraco	0,99	0,99	0,99	0,99	0,99	0,99

Tabela 2: Coeficientes de absorção utilizados na simulação.

Diversos parâmetros também devem ser considerados na simulação e que influenciam no resultado. Um deles é o número de raios emitidos pela fonte. Quando o número de raios é pequeno, da ordem de milhares de raios, a resposta impulsiva possui uma característica com picos espaçados, devido a poucos raios atingirem o receptor. Fazer o número de raios extremamente grande, da ordem de milhões de raios, eleva consideravelmente o tempo de processamento e não acrescenta informação útil à resposta impulsiva, pois cada raio parte da fonte com uma energia muito pequena. Assim, através de ensaios, verificou-se que o número de raios deve variar de 100 a 500 mil para que a resposta impulsiva simulada esteja mais próxima da real.

ANÁLISE DOS RESULTADOS

Foram analisados os parâmetros acústicos T_{30} , C_{80} e D_{50} para cada posição de receptor, tanto para as respostas impulsivas medidas quanto para as simuladas. Cada medição foi realizada 10 vezes e os valores analisados corresponderam à média desses valores. O desvio padrão obtido para os parâmetros medidos é da ordem de 10^{-3} , o que verifica a precisão e a validade dos dados medidos. Na Fig. 4 são apresentados os gráficos de T_{30} para cada uma das três posições, onde podem ser comparados os valores obtidos na simulação com os da medição. Os erros relativos absolutos entre os valores medidos e simulados de T_{30} são apresentados na Fig. 5. Pode-se observar que os maiores erros ocorrem nas frequências abaixo de 500 Hz. Na Fig. 6 são apresentados os gráficos do parâmetro C_{80} para cada posição de receptor, comparando medição e simulação.

4.1: T_{30} para posição 14.2: T_{30} para posição 24.3: T_{30} para posição 3Figura 4: Comparação entre medição e simulação para T_{30} em cada posição.5.1: T_{30} para microfone5.2: T_{30} para ouvido esquerdo5.3: T_{30} para ouvido direitoFigura 5: Erros entre medição e simulação para T_{30} em cada posição.6.1: C_{80} para posição 16.2: C_{80} para posição 26.3: C_{80} para posição 3Figura 6: Comparação entre medição e simulação para C_{80} em cada posição.7.1: D_{50} para posição 17.2: D_{50} para posição 27.3: D_{50} para posição 3Figura 7: Comparação entre medição e simulação para D_{50} em cada posição.

Como pode ser observado através das Figs. 4 a 5, o erro relativo absoluto possui altos valores para baixas freqüências, principalmente na banda de 500 Hz. A partir de 1 kHz o erro, em média, encontra-se na faixa de 15%, o que é um erro esperado, dada a incerteza dos valores dos coeficientes de absorção. O erro elevado nas baixas freqüências (erro médio de aproximadamente 50%) deve-se ao método de propagação utilizado no simulador. Os métodos de traçado de raios e de transição de energia não consideram os efeitos de interferência e difração do som, que podem ocorrer devido ao comprimento de onda ser suficientemente grande em relação aos objetos e às dimensões da sala. Ou seja, neste caso a acústica geométrica não produz resultados satisfatórios. Além disso, o erro elevado que se observa no tempo de reverberação para a banda de 500 Hz nas posições 1 e 3 não ocorre para a posição 2 de receptor. Dependendo da freqüência, pode estar ocorrendo um efeito de cancelamento ou amplificação do som. Isto ocorre devido aos modos naturais de vibração da sala, cujas dimensões favorecem o reforço de freqüências múltiplas de aproximadamente 50 Hz. A frequência de Schroeder calculada para esta sala é de aproximadamente 345 Hz, o que justifica o erro elevado nas freqüências abaixo de 500 Hz, uma vez que até 345 Hz os modos de vibração da sala são predominantes.

CONCLUSÕES

Neste artigo foi apresentada uma comparação entre os parâmetros de qualidade acústica obtidos através de medição e de simulação. O experimento foi realizado cuidadosamente, para evitar que erros de medição pudessem interferir na comparação. Na simulação foram analisados diversos parâmetros, tais como número de raios, coeficientes de absorção e difusão, discretização de áreas, entre outros, para compreender como suas variações alteram a resposta impulsiva e consequentemente os parâmetros de qualidade acústica. Através da análise dos resultados, observou-se que o simulador acústico RAIOS apresentou um desempenho confiável para a geração de respostas impulsivas monoaurais e binaurais, das quais foram obtidos os parâmetros desejados, tais como T_{30} , D_{50} etc. O simulador contudo apresentou melhores resultados para freqüências acima de 1 kHz do que para baixas freqüências, onde o erro foi mais elevado. Isto se deve ao fato dos métodos de propagação utilizados serem falhos para baixas freqüências, onde a acústica geométrica (traçado de raios) não é mais aplicável devido aos fenômenos de interferência e a atuação dos modos naturais de vibração serem predominantes nesta faixa de freqüência.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Leo L. Beranek, "Acoustical modeling as a tool in problem solving," *Journal of the Audio Engineering Society*, vol. 17, no. 2, pp. 151–155, Apr 1969.
- [2] J.H. Rindel, "Modelling in auditorium acoustics. from ripple tank and scale models to computer simulations," *Revista de Acústica*, vol. XXXIII, no. 34, pp. 31–35, 2002.
- [3] J. C. B. Torres, M. R. Petraglia, and Roberto A. Tenenbaum, "An efficient wavelet-based hrtf for auralization," *Acustica/Acta Acustica*, vol. 90, no. 1, Jan 2004.
- [4] L. Faiget, C. Legros, and R. Ruiz, "Optimization of the impulse response length: Application to noisy and highly reverberant rooms," *Journal of the Audio Engineering Society*, vol. 46, pp. 741–750, 1998.
- [5] L. L. Beranek, *Music, Acoustics and Architecture*, John Wiley, Nova York, 1962.
- [6] R. A. Tenenbaum, T. S. Camilo, J. C. B. Torres, and S. N. Y Gerges, "Hybrid method for numerical simulation of room acoustics with auralization: part 1 - theoretical and numerical aspects," *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 29, pp. 211–221, 2007.
- [7] F. FIGUEIREDO and F. IAZZETTA, "Parâmetros aústicos em salas de música: análise de resultados e novas interpretações.,," *Anais do 4.o Congresso de Engenharia de Áudio. São Paulo: Audio Engineering Society - Seção Brasil*, vol. 1, pp. 66–71, 2006, website : <http://gsd.ime.usp.br/acmus/>.
- [8] R. A. Tenenbaum, T. S. Camilo, J. C. B. Torres, and S. N. Y Gerges, "Hybrid method for numerical simulation of room acoustics with auralization: part 2 - validation of the computational code raios 3," *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 29, pp. 222–231, 2007.
- [9] J. J. Embrechts, "Randomly traced sound ray techniques," *Acustica*, vol. 51, pp. 285–295, 1982.
- [10] A. Kulowski, "Algorithmic representation of the ray tracing technique," *Applied Acoustics*, vol. 18, pp. 449–469, 1984.
- [11] E. Kruzin and F. Fricke, "The prediction of sound fields in non-diffuse spaces by a "random walk" approach," *Journal of Sound Vibration*, vol. 81, pp. 549–564, 1982.
- [12] J. L. B. Coelho, D. Alarcão, A. M. Almeida, T. Abreu, and N. Fonseca, "Room acoustics design by a sound energy transition approach," *Acustica - Acta Acustica*, vol. 86, pp. 903–910, 2000.

Sessão 3

Codificação, processamento e edição de som
(Sound edition, processing and coding)



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Marcação Automática de Eventos Usando Sinal de Áudio em Transmissões Esportivas de TV

Luiz G. L. B. M. de Vasconcelos¹, Sergio L. Netto², Luiz W. P. Biscainho² e Charles B. do Prado¹

¹Departamento de Pesquisa e Desenvolvimento, TV Globo
Rio de Janeiro, RJ, 22460-000, Brasil

²Programa de Engenharia Elétrica/COPPE, DEL/Poli, Universidade Federal do Rio de Janeiro
CP 68504, Rio de Janeiro, RJ, 21941-972, Brasil

luiz.vasconcelos@tvgloblo.com.br, sergioln@lps.ufrj.br, wagner@lps.ufrj.br, charles.prado@tvgloblo.com.br

RESUMO

Este artigo descreve um método para localizar os melhores momentos da transmissão de um jogo de futebol a partir do áudio, com base na energia e na freqüência fundamental da voz do narrador. Para isso, implementou-se um aplicativo com interface gráfica que permite classificar o sinal de forma rápida e prática. O sistema mostrou-se capaz de identificar 100% dos momentos de interesse para o mesmo narrador utilizado no treinamento, ao custo de uma taxa de falsa identificação em torno de 50%. O processo de seleção comprime o sinal de vídeo em cerca de 90% para uma posterior classificação semi-automática.

0 INTRODUÇÃO

Cada vez mais, em nossa sociedade, aumenta a demanda por entretenimento, tal como acesso à Internet, peças teatrais e cinematográficas, *shows* de música, prática de esportes e viagens. Nesse contexto, se inserem as emissoras de TV, que, além de informar, também têm o objetivo de entreter. Uma parcela substancial do entretenimento televisivo é a transmissão de programas esportivos, tais como partidas de futebol, e ainda a posterior exibição de eventos específicos, tais como gols, pênaltis, oportunidades de gol etc. O interesse por transmissões esportivas é tão grande que há canais com programação dedicada a elas, que também exibem eventos específicos ocorridos em outras transmissões e mesmo programas secundários que noticiam apenas esses eventos. Preparar esse tipo de programação requer um enorme consumo de tempo e esforço para cobrir todas as transmissões, já que se requer uma seleção bastante

criteriosa dos eventos de interesse. Atualmente, é necessário um operador acompanhando cada transmissão e marcando os eventos específicos para posterior recuperação, o que torna interessante o desenvolvimento de tecnologias que automatizem ou simplifiquem este processo.

A princípio, pode-se considerar o processamento das imagens para detectar padrões em transmissões televisivas. Porém, retirar informações tão específicas do sinal de vídeo seria bastante complexo, pois cada esporte tem características visuais próprias, além do fato de este tipo de processamento envolver um volume muito grande de dados. Sendo assim, já há trabalhos [1] que analisam o áudio para encontrar trechos desejados de um sinal de vídeo.

Analizando a Figura 1, tem-se que o modelo de produção da voz humana considera um sinal de excitação processado por um filtro que modela o trato vocal. A excitação,

proveniente dos pulmões, caracteriza um aspecto da sonoridade associado à vibração (trecho sonoro) ou não (trecho surdo) das cordas vocais. Para todos os efeitos práticos, em processamento de voz, a frequência de vibração das cordas vocais é denominada de *pitch*. O sistema aqui proposto de classificação de “bons momentos” se baseia nas informações de energia e de *pitch* do sinal de voz. De modo geral, esses dois parâmetros se elevam de forma significativa durante os eventos de interesse.

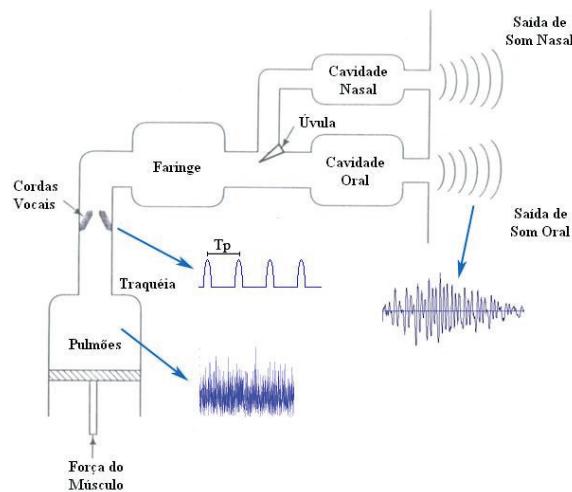


Figura 1 Representação em blocos do processo de geração da voz humana [2].

Nesse contexto, é feita uma análise da evolução destes dois aspectos ao longo dos trechos de interesse para um dado sinal de treinamento. Posteriormente, um módulo de decisão utiliza estes sinais para classificar o trecho em questão como sendo de interesse ou não. Uma etapa final é responsável por agrupar trechos muito próximos, que representariam o mesmo evento, e também verificar a correta duração (início e fim) dos eventos previamente selecionados. O sistema incorpora uma *interface* gráfica para facilitar o acompanhamento do processo por parte do usuário, que pode ainda fazer pequenos ajustes para melhorar o processo de classificação.

Para apresentação completa do sistema, este artigo obedece a seguinte estruturação: A Seção 1 inclui uma descrição do desenvolvimento do sistema e de seu funcionamento geral. Na Seção 2, é descrita a ferramenta gráfica desenvolvida para a aplicação em questão, destacando suas principais funcionalidades no processo de classificação e edição do sinal de vídeo resultante. Na Seção 3, é caracterizado o desempenho do sistema em termos da capacidade de detecção dos “bons momentos”; são considerados sinais do mesmo narrador usado no desenvolvimento do sistema e também de outros narradores. Por fim, na Seção 4, são apresentadas as conclusões do trabalho, ressaltando-se suas principais contribuições.

1 DESENVOLVIMENTO DO SISTEMA

Por se tratar de uma aplicação bastante particular de processamento de voz, esta seção descreve o desenvolvimento do método proposto. Inicialmente, o sistema foi modelado usando-se um único sinal da base de dados, para ao final ser generalizado para outros sinais (do mesmo narrador ou não).

1.1 Base de Dados

Os sinais que compõem a base de dados usada no desenvolvimento e teste do sistema foram cedidos pela TV Globo, e são descritos na Tabela 1. Trata-se de sinais digitais de vídeo com áudio *embedded*, onde o *stream* de áudio foi amostrado à taxa de 48 kHz com 16 bits por amostra em dois canais, sendo o esquerdo referente à narração e o direito, ao ambiente. O sistema proposto é baseado apenas no sinal de narração que possui um mínimo nível de ruído ambiente. O Sinal I foi utilizado para o desenvolvimento do método. Os demais sinais, do mesmo narrador que o Sinal I ou não, foram utilizados na etapa de testes de desempenho. A Tabela 2 descreve a quantidade de “bons momentos” em cada sinal da Tabela 1. Estes valores foram obtidos de forma tradicional, isto é, determinados visualmente por um operador humano.

Tabela 1 Sinais que compõem a base de dados usada no desenvolvimento e teste do sistema.

Sinal	Partida	Narrador	Nome
Sinal I	Vasco x Flamengo 1°T	Narrador I	Eduardo Moreno
Sinal II	Vasco x Flamengo 2°T	Narrador I	Eduardo Moreno
Sinal III	Chivas x San Jose	Narrador I	Eduardo Moreno
Sinal IV	Botafogo x Vasco	Narrador II	Galvão Bueno
Sinal V	Brasil x Chile	Narrador II	Galvão Bueno
Sinal VI	Boca Jrs. x Grêmio	Narrador III	Cléber Machado

Tabela 2 Número de “bons momentos” para cada sinal da base de dados descrita na Tabela 1.

Sinal	Bons Momentos
Sinal I	14
Sinal II	15
Sinal III	20
Sinal IV	28
Sinal V	9
Sinal VI	6

1.2 Energia do Sinal de Voz

A alta energia de um sinal de voz pode indicar se o trecho de vídeo correspondente é de interesse ou não. Para se minimizar a quantidade de dados processados, divide-se o sinal de voz $x(n)$ em blocos de N amostras e determina-se a energia E do bloco por

$$E = \sum_{n=1}^N x^2(n). \quad (1)$$

O valor adequado para N pode ser determinado de forma experimental para a aplicação em questão. Valores pequenos geram um número excessivo de blocos, o que aumenta o custo computacional do método de classificação; por outro lado, valores excessivos para N acarretam a não-detecção de alguns trechos de interesse do sinal de vídeo. A Figura 2 ilustra dois exemplos de fala intensa com durações bastante distintas: 200 ms e 2 s. Com base nisso, foram considerados blocos de durações 250ms, 500ms e 1000ms para se verificar qual destes valores gera um sinal de energia que melhor destaca os momentos de interesse.

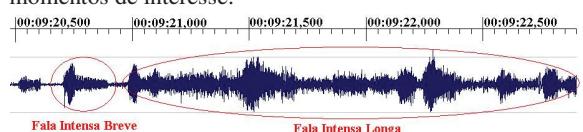


Figura 2 Exemplo de sinal de voz com trechos de interesse de diferentes durações (cerca de 200 ms e 2 s).

A segmentação do sinal de voz pode ser feita ainda usando-se blocos superpostos ou não, como indicado na Figura 3. A não-superposição ocorre quando o deslocamento M do bloco é maior ou igual ao número de amostras N que o compõem. Já a superposição resulta da condição $M < N$. Com o deslocamento superposto, o sistema carrega mais informação a respeito das variações de energia do sinal. Por outro lado, o deslocamento não-superposto é muito mais leve computacionalmente. Porém, para $M = 1$ a questão computacional do cálculo da energia do bloco é facilmente resolvida com a aplicação do algoritmo de *buffer* circular [3,4]. Nesse caso, a energia do bloco atual é determinada pela energia do bloco anterior adicionada à energia da amostra atual $x(n)$ e subtraída da energia da amostra $x(n-M)$. Para se determinar o deslocamento que melhor realça os “bons momentos”, serão considerados os casos $M = 1$ e $M = N$.

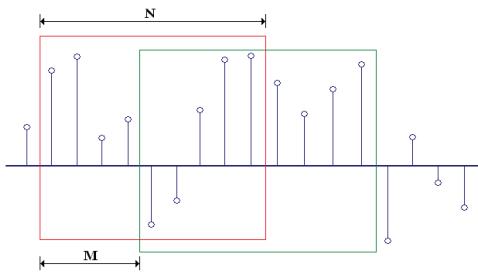


Figura 3 Deslocamento do bloco no domínio do tempo.

Desta forma, seis variações de segmentação foram testadas na classificação do Sinal I: com durações de 250, 500 e 1000 ms, e ainda $M = 1$ e $M = N$ para cada caso anterior. Para cada variação, foi feito um histograma das distribuições da energia dos blocos associados aos eventos de interesse ou não. De modo geral, em todos os casos foi possível observar uma boa separação dos histogramas associados a cada tipo de bloco, como é visto na Figura 4 para a duração de 250 ms e $M = 1$. As Figuras 5 ($M = 1$) e 6 ($M = N$) mostram a taxa de classificação correta de cada tipo de bloco (evento de interesse ou não) em função do limiar de decisão escolhido para a energia do bloco, para as seis variações acima descritas. Destas figuras, conclui-se que todas as variações possuem desempenho semelhante, com uma pequena vantagem em termos de taxa de classificação para o caso $M = 1$ com duração de 1000 ms. Para este caso, privilegiando-se a identificação correta dos “bons momentos” em detrimento de uma identificação incorreta de alguns momentos normais, pode-se estipular o limiar de energia como sendo de 0,04. Naturalmente, este valor é altamente dependente do nível de gravação do sinal de entrada, mas uma normalização apropriada pode ser feita para torná-lo de uso mais geral.

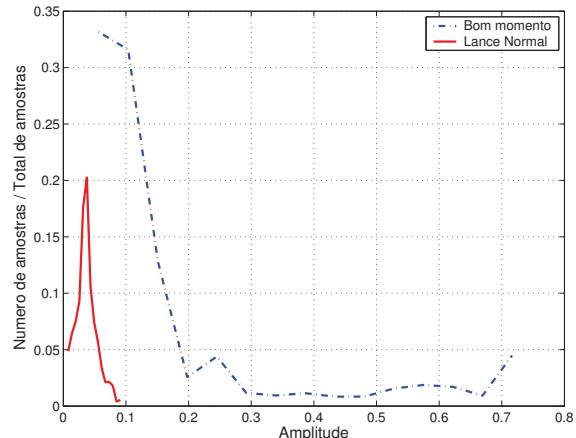


Figura 4 Distribuição estatística do sinal de energia calculada com duração de 250 ms e janela superposta.

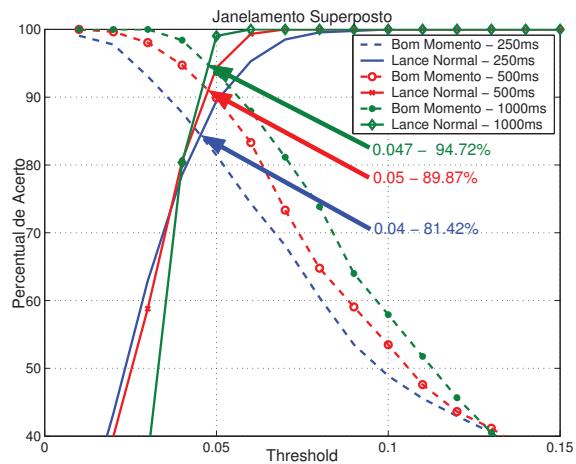


Figura 5 Taxas de acerto de classificação em função do limiar de energia para $M = 1$ e duração de 250, 500 e 1000 ms.

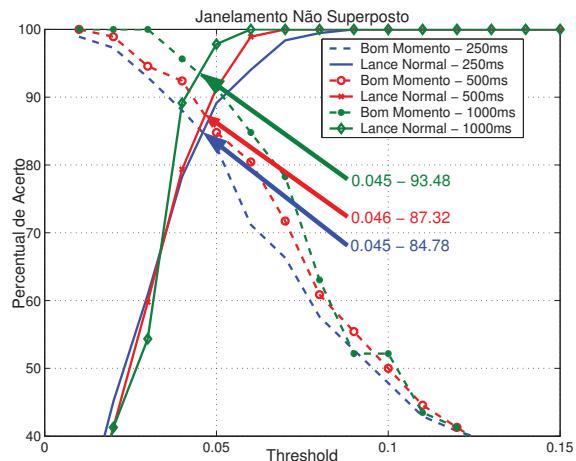


Figura 6 Taxas de acerto de classificação em função do limiar de energia para $M = N$ e duração de 250, 500 e 1000 ms.

1.3 Pitch do Sinal de Voz

O período de *pitch* da voz é determinado pelos movimentos quase periódicos das cordas vocais na faringe, e é o inverso da freqüência fundamental da voz percebida pelo sistema auditivo humano [2].

Uma maneira de se extrair a freqüência fundamental da voz é determinar sua periodicidade a partir da função de autocorrelação [5,6]:

$$R_{xx}(\tau) = \sum_n x_n x_{n-\tau}. \quad (2)$$

A Figura 7 ilustra o aspecto da função de autocorrelação para trechos do sinal de voz associados a momentos de interesse ou não. A partir desta figura, é possível perceber que os picos mais proeminentes da autocorrelação, que determinam o período de *pitch* do trecho de voz correspondente, ficam mais próximos entre si nos caso de um “bom momento”.

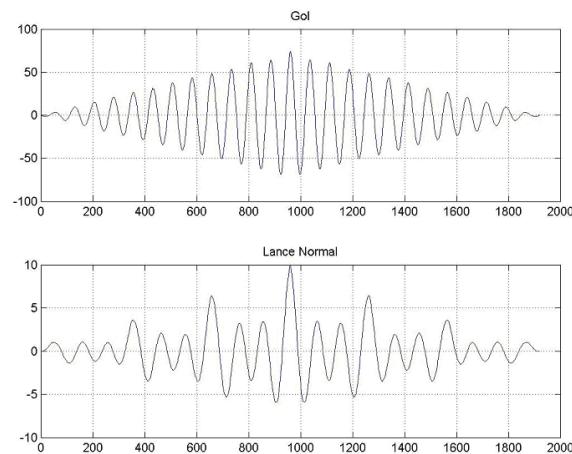


Figura 7 Autocorrelação do sinal de voz para um trecho de gol e outro de lance normal.

O cálculo da autocorrelação pode ser feito no domínio da freqüência a partir da relação [7]:

$$R_{xx}(\tau) = IDFT\{|DFT[x(n)]|^2\}. \quad (3)$$

A freqüência fundamental da voz masculina está em geral em torno de 150 Hz, ou ao menos acima dos 80 Hz. Assim, um período de *pitch* será no máximo de 12,5 ms. A fim de realizar o cálculo do *pitch* pela autocorrelação de forma precisa, é interessante ter pelo menos três ciclos no sinal de voz. Para forçar uma margem de segurança, foram utilizados blocos de 40 ms, correndo-se um pequeno risco de modelar pequenas variações de *pitch* dentro de um único bloco. Para evitar interferências provenientes de outras fontes, antes de qualquer cálculo foi feita uma filtragem passa-baixas limitando a banda do sinal de voz em 1 kHz. Para evitar cálculos desnecessários, foi feita uma detecção de silêncio usando-se um limiar de 0,1 para a energia de cada bloco de 40 ms do sinal de voz em questão. Este valor limite foi determinado a partir de uma análise estatística da energia para os blocos de silêncio ou não em todo o Sinal I, como ilustrado na Figura 8.

A Figura 9 exibe os histogramas do valor de *pitch* dos segmentos marcados como “bons momentos” ou não. É fácil ver que o cruzamento das distribuições se encontra na freqüência de *pitch* igual a 225 Hz. Porém, mais uma vez, por ser mais importante classificar corretamente todos os eventos de interesse, é desejável utilizar um limiar de classificação ligeiramente menor. Usando-se 200 Hz,

87,5% dos blocos de “bom momento” e 3,5% dos lances normais estão sendo marcados. Estes índices podem ser considerados satisfatórios, pois os demais blocos de interesse podem ser identificados pela continuidade do sinal.

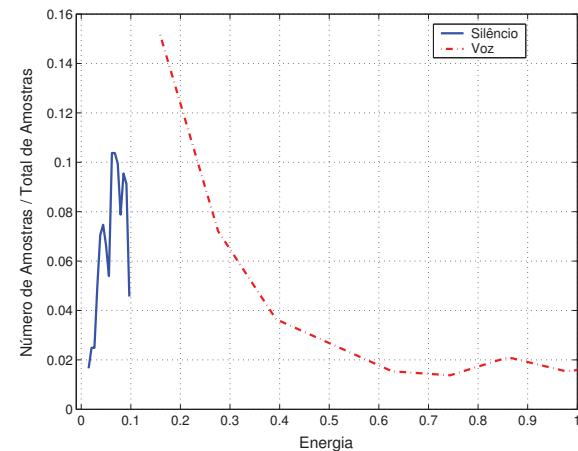


Figura 8 Histogramas de energia para trechos de silêncio e voz do Sinal I.

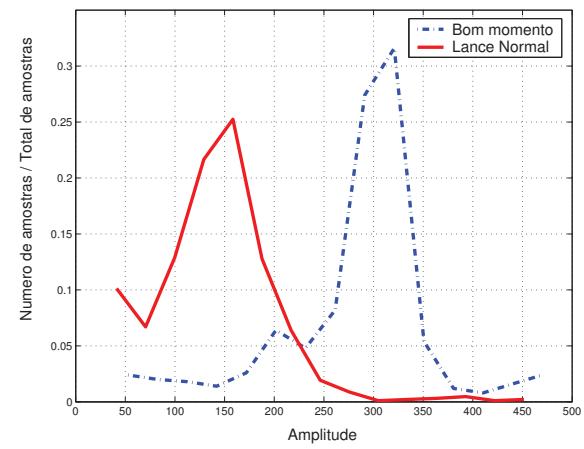


Figura 9 Distribuição do pitch em blocos de eventos de interesse ou não para o Sinal I.

Os limiares aqui encontrados para o valor de *pitch* devem ser válidos para quaisquer sinais do mesmo narrador usado no Sinal I. Para outros narradores, uma análise similar deve ser feita *a priori*, ou ainda de forma automática a partir de um trecho curto do sinal.

1.4 Módulo de Decisão

Os limiares de energia e *pitch* determinados anteriormente servem para um primeiro nível de classificação de um dado bloco como sendo de “bom momento” ou não. A Figura 10 ilustra um exemplo de marcações de um trecho do sinal de voz, onde é possível constatar que a marcação bloco-a-bloco funcionou de forma semelhante para as duas características (energia e *pitch*). De modo geral, para os limiares pré-determinados acima, observa-se que o sinal de energia foi mais conservador no sentido de que suas marcações estavam quase sempre corretas, porém demoravam mais a

identificar um trecho de interesse. Então, foi utilizado um algoritmo que buscasse pelas regiões de bom momento através da energia, para posteriormente confirmar e definir seus limites a partir do sinal de *pitch*.

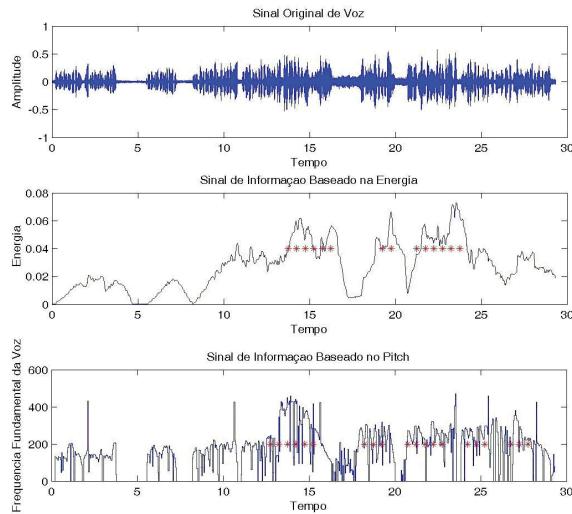


Figura 10 Sinais de informação com marcações instantâneas.

Foi feito ainda um estudo de quantas marcações de energia em seqüência são necessárias para se caracterizar de forma efetiva um “bom momento” no Sinal I. O gráfico da Figura 11 mostra que quanto maior a exigência no número mínimo de marcações em seqüência, menor é o percentual de “bons momentos” identificados. Para garantir a identificação de todos os trechos de interesse e eliminar alguns trechos erroneamente identificados anteriormente, foi então adotada a exigência mínima de três marcações em seqüência para classificar um trecho como “bom momento”.

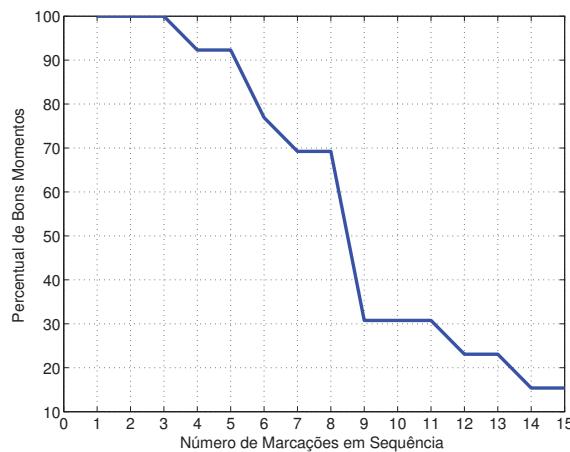


Figura 11 Percentual de “bons momentos” identificados em função do número mínimo de marcações em seqüência.

A Figura 12 é o resultado da aplicação do algoritmo no exemplo da Figura 10. É possível notar que o trecho que foi marcado pela energia com apenas duas marcações em seqüência foi descartado, e que em ambos os trechos marcados pela energia o *pitch* foi útil para determinar o início do bom momento. Porém, apenas no último trecho ele foi utilizado para determinar o fim.

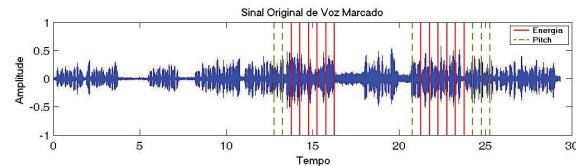


Figura 12 Sinal original de voz com as marcas de energia e de pitch que serão consideradas.

Num último estágio, o algoritmo de decisão une trechos de interesse que estejam muito próximos. Por exemplo, na Figura 12 há dois trechos separados por cerca de cinco segundos, o que pode indicar um único trecho de interesse pelo pequeno intervalo de tempo. Assim, realizando um estudo nos trechos marcados separadamente que fazem parte de um mesmo “bom momento” no Sinal I, descobriu-se que somente 5% desses intervalos foram maiores que oito segundos e que 80% foram menores que cinco segundos. Foi estipulado, então, um intervalo-límite de dez segundos a partir do qual trechos marcados separadamente são mantidos separados. Com este artifício aplicado ao Sinal I, os 53 trechos anteriormente marcados foram agrupados em apenas 24.

2 FUNCIONAMENTO DO SISTEMA

O método de classificação de eventos de interesse descrito na Seção 1 foi desenvolvido numa plataforma denominada MelhoresMomentos. O sistema foi desenvolvido em C++ com base em [8,9], utilizando MFC 8.0, biblioteca do Windows [10], IT++ 4.0.0, e a biblioteca para processamento de sinais [11], que utiliza a biblioteca MKL 9.1.027 da Intel [12].

A interface gráfica do sistema MelhoresMomentos é representada na Figura 13, cujas principais funcionalidades destacadas são:

- (1) Sinal de vídeo sendo analisado;
- (2) Barra de tempo deslizante para rápido avanço ou recuo do sinal de vídeo;
- (3) Botões de “tocar” e “parar” o sinal de vídeo;
- (4) Janela indicativa de marcação ou não do sinal sendo mostrado;
- (5) Indicativo de início de trecho marcado;
- (6) Indicativo de término de trecho marcado;
- (7) Contador do trecho atual em relação ao total de segmentos;
- (8) Lista de “bons momentos” detectados;
- (9) Opção de limpeza da lista de trechos marcados.

Através do aplicativo, o usuário é capaz de abrir um arquivo de vídeo, tocar e parar, selecionar um trecho, exportar tanto áudio como vídeo, e detectar os melhores momentos existentes no trecho selecionado. O usuário ainda pode ajustar os limiares de energia e *pitch* para fazer um ajuste fino no desempenho do sistema.

Em termos de tempo de processamento, o sistema MelhoresMomentos necessitou de cerca de 3 minutos para detectar os melhores momentos de 45 minutos do Sinal I em um processador Intel Pentium Dual Core 3.06 GHz.



Figura 13 Interface gráfica do sistema MelhoresMomentos.

3 TESTES DE DESEMPENHO

Para uma avaliação mais criteriosa do método descrito na Seção 1, foram utilizadas duas medidas de desempenho: O percentual de “bons momentos” marcados corretamente (%BMM) e o percentual de trechos marcados que são efetivamente “bons momentos” (%TMC). A Figura 14 ilustra o que os parâmetros representam dentro dos resultados obtidos. A idéia é que o método marque todos os bons momentos, tendo um %BMM próximo de 100%, mesmo que alguns lances normais sejam também assinalados, gerando um %TMC abaixo de 100%.

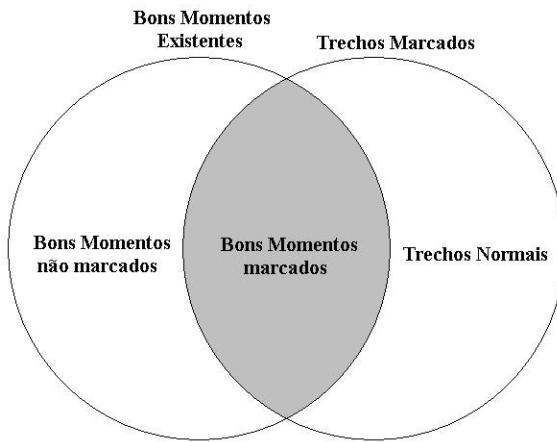


Figura 14 Diagrama de Venn ilustrando os parâmetros que servirão para visualização dos resultados.

A Tabela 3 expressa os resultados iniciais para todos os sinais da base de dados descrita nas Tabelas 1 e 2. O Sinal I, que foi utilizado no desenvolvimento do método, teve todos os trechos desejados devidamente marcados, com alguns trechos indesejados também marcados. Na prática, para este sinal, cerca de 5 minutos de vídeo foram selecionados pelo método como contendo trechos de interesse. Os trechos indevidamente marcados podem ser eliminados de forma semi-automática por um operador humano. Esse processamento adicional, porém, fica extremamente facilitado pela alta seletividade do método, que reduziu o tempo de marcação em cerca de 90%. De modo geral, o mesmo desempenho se repetiu para todos os sinais do mesmo narrador atuando no Sinal I.

Tabela 3 Resultados iniciais do sistema MelhoresMomentos.

Sinal	Narrador	%TMC	%BMM
Sinal I	Narrador I	58,3	100
Sinal II	Narrador I	65,2	100
Sinal III	Narrador I	44,2	100
Sinal IV	Narrador II	87,5	43,8
Sinal V	Narrador II	0	0
Sinal VI	Narrador III	10,5	50,0

Nos sinais IV, V e VI, com locutores e captação diferentes do Sinal I, os resultados foram ruins, principalmente pelo fato de o %BMM ter sido abaixo de 100%. Isto se deve aos limiares utilizados para o Narrador I serem inadequados às características de voz dos demais narradores. Realizando-se um ajuste empírico dos limiares de energia e *pitch*, de modo a se obter %BMM = 100%, obtém-se os resultados indicados na Tabela 4.

Tabela 4 Resultados normalizados para diferentes narradores.

Sinal	Narrador	%TMC	%BMM
Sinal IV	Narrador II	35,2	100
Sinal V	Narrador II	23,4	100
Sinal VI	Narrador III	7,3	100

Mesmo com este ajuste, o funcionamento do sistema se manteve precário, já que os valores de %TMC tornaram-se extremamente baixos. Isto indica que um ajuste criterioso de todos os limiares determinados na Seção 1 deve ser feito para cada diferente narrador.

Procurou-se determinar a causa dos erros de classificação e percebeu-se que estes erros podem ser agrupados em três classes: (i) erros por emoção, onde o narrador aplica emoção à sua voz, porém em trechos descorrelacionados com a partida ou que não se caracterizam como um trecho de interesse, tais como anúncios, *replays* muito após o lance, início e término de partida etc.; (ii) erros devidos a outra pessoa, que são erros ocorridos em trechos de outros narradores, como comentaristas ou repórteres de campo; (iii) outros tipos de erros que não se encaixam nas duas categorias anteriores. A classificação dos erros ocorridos nos diferentes sinais é apresentada na Tabela 5.

Tabela 5 Distribuição dos erros por categorias.

Sinal	Narrador	%Emoção	%Outra Pessoa	%Sem Motivo
Sinal I	Narrador I	60	40	0
Sinal II	Narrador I	62,5	25	12,5
Sinal III	Narrador I	78	17,4	4,6
Sinal IV	Narrador II	53,3	16,7	30
Sinal V	Narrador II	41,7	16,6	41,7
Sinal VI	Narrador III	32	8	60

De modo geral, podemos concluir que o método aqui apresentado funciona muito bem como um detector de emoção do narrador para o qual o método foi treinado. Outros narradores, porém, requerem um ajuste dos limiares de classificação, para minimizar os erros pertencentes aos grupos (ii) e (iii).

Além da marcação correta dos “bons momentos”, foi avaliado se o início e o fim dos bons momentos foram marcados satisfatoriamente (%BMS). Este tipo de análise

possui um caráter subjetivo, contando com a ajuda de um operador experiente. Os resultados indicados por este operador encontram-se na Tabela 6, que mostra que uma boa parcela dos trechos selecionados foi marcada satisfatoriamente. Na prática, percebeu-se que a principal razão de uma marcação inapropriada era a demora do narrador em aplicar emoção à voz. Aqui, mais uma vez, mostrou-se necessária a intervenção do usuário para redefinir os limites dos bons momentos que não foram marcados satisfatoriamente. Esta tarefa, porém, fica facilitada pelas funcionalidades presentes na interface gráfica da plataforma MelhoresMomentos.

Tabela 6 Percentual de Bons Momentos que tiveram seus limites marcados satisfatoriamente pelo método.

<i>Sinal</i>	<i>Narrador</i>	<i>%BMS</i>
Sinal I	Narrador I	64,3
Sinal II	Narrador I	73,3
Sinal III	Narrador I	73,7
Sinal IV	Narrador II	68
Sinal V	Narrador II	76,5
Sinal VI	Narrador III	66,6

4 CONCLUSÕES

Este artigo apresentou um método semi-automático de determinação dos melhores momentos de uma partida de futebol através do áudio do narrador. O método gera dois sinais de informação, um baseado na energia e outro no *pitch*, que realçam a possível ocorrência de “bons momentos”. Um módulo de decisão utiliza ambas as informações para determinar os trechos de interesse, demarcando seus limites, e possivelmente agrupando trechos adjacentes correspondentes a um mesmo evento.

Os resultados foram satisfatórios, apesar de no estágio atual o método se mostrar dependente do locutor utilizado no seu desenvolvimento. A generalização do método exigiria um treinamento para cada narrador, montando-se um banco de narradores, ou ainda fazendo-se um ajuste automático dos limiares de decisão baseado em uma análise preliminar de curta duração.

Na opinião de um profissional de TV, com a generalização do método para outros narradores, será possível que um único operador seja responsável por editar os melhores momentos de diversas partidas que ocorram simultaneamente. De qualquer forma, em seu estado atual de desenvolvimento, o sistema MelhoresMomentos já é capaz de ser utilizado operacionalmente, de forma semi-automática, reduzindo o tempo de análise em cerca de 90% para os sinais com o mesmo narrador usado no seu desenvolvimento.

5 REFERÊNCIAS

- [1] H. Christensen, Y. Gotoh, S. Renals, A Cascaded Broadcast News Highlighter , IEEE Trans. Audio, Speech, and Language Processing, 16(1), 1558-7916, Jan. 2008.
- [2] D. Rocchesso, Introduction to Sound Processing, [http://www.mondo-estremo.com], Mondo Estremo, 20/03/2003.
- [3] P. S. R. Diniz, E. A. B. da Silva, S. L. Netto, Processamento Digital de Sinais – Projeto e Análise de Sistemas, Bookman Editora, 2004.
- [4] Wikipedia, [http://en.wikipedia.org/wiki/Circular_buffer]
- [5] J. H. Deller, J. R. Proakis, J. G. Hansen, Discrete-Time Processing of Speech Signals, Prentice Hall, 1987.
- [6] P. Z. Peebles, Probability, Random Variables, and Random Signal Principles, McGraw-Hill, 2001.
- [7] T. Tolonen, M. Karjalainen, A Computationally Efficient Multipitch Analysis Model, IEEE Trans. Speech Audio Processing, 8(6), 708-716, Nov. 2000.
- [8] P. M. Embree, D. Danieli, Algorithms for Digital Signal Processing.
- [9] N. M. Josuttis, The C++ Standard Library – A Tutorial and Reference, Addison-Wesley, Nov. 2006.
- [10] Microsoft Development Network, [http://www.msdn.com].
- [11] IT++ 4.0.0, [http://itpp.sourceforge.net/], 14/10/2007.
- [12] Intel Math Kernel Library 9.1.027, [http://www.intel.com/cd/software/products/asmo-na/eng/307757.htm].



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Equalização e Identificação Adaptativas para Áudio Utilizando Marca d'Água como Sinal de Supervisão

Leandro de Campos Teixeira Gomes¹, Mário Ulian Neto^{1,2} e
João Marcos Travassos Romano²

¹ Centro de Pesquisa e Desenvolvimento em Telecomunicações (CPqd)
Rod. Campinas–Mogi-Mirim (SP 340), km 118,5, Campinas, SP, CEP 13086-902, Brasil

² Universidade Estadual de Campinas (Unicamp), Faculdade de Engenharia Elétrica e de Computação (FEEC), Lab. de Proc. de Sinais para Comunicações Móveis (DSPCom)
Caixa Postal 6101, Campinas, SP, CEP 13083-970, Brasil

tgomess@cpqd.com.br, uliani@cpqd.com.br, romano@fee.unicamp.br

RESUMO

Apresentamos um método de equalização e identificação adaptativas para áudio baseado no uso de uma marca d'água como sinal de supervisão. A marca d'água é transmitida ininterruptamente através do canal, juntamente com o áudio. No receptor, a marca é processada pelo filtro equalizador/identificador, permitindo o ajuste contínuo dos coeficientes deste último. Um modelo psicoacústico, em conjunto com um algoritmo de conformação espectral, é empregado para evitar distorções audíveis. Resultados de simulação ilustram o método e demonstram sua viabilidade.

0 INTRODUÇÃO

Métodos tradicionais de equalização e identificação adaptativas supervisionadas baseiam-se no uso de seqüências de treinamento como sinal de referência. Nestes métodos, a transmissão de informação é periodicamente interrompida para o envio de uma seqüência de treinamento conhecida no receptor. A comparação entre a seqüência de treinamento original e a rece-

bida permite ajustar os coeficientes do filtro equalizador/identificador.

Uma alternativa para evitar a interrupção periódica da transmissão é o uso de técnicas cegas ou não-supervisionadas, que não contam com sinal de referência, baseando-se geralmente na análise de estatísticas de ordem superior do sinal. Um dos grandes desafios enfrentados por estas técnicas é a possibi-

lidade de convergência para soluções sub-ótimas, além do seu custo computacional muitas vezes superior ao de técnicas supervisionadas.

É proposto aqui um método de equalização e identificação para áudio baseado no uso de uma *marca d'água* como sinal de referência. A marca d'água é um sinal de natureza similar à do sinal de informação (dito *hospedeiro*) e é continuamente inserida neste último ao longo do tempo. Ambos os sinais são simultaneamente transmitidos através do canal que se deseja equalizar ou identificar. Assim como uma seqüência de treinamento tradicional, a marca d'água é conhecida no receptor, sendo utilizada para estimar as características do canal. Para evitar que a inserção da marca d'água no áudio provoque distorções perceptíveis, é empregado um modelo psicoacústico em conjunto com um algoritmo de conformação espectral.

Uma técnica similar à que propomos é conhecida na literatura como *superimposed training* [1, 2]. Esta abordagem baseia-se na sobreposição ao sinal de informação de uma seqüência de dados piloto que é detectada no receptor e usada para identificar o canal de transmissão. Sua principal diferença com relação aos métodos de marca d'água é o conceito de *transparência* que norteia estes últimos: a marca d'água deve ser construída de forma tal que, quando encarada como um ruído adicionado ao sinal hospedeiro, sua influência na detecção e/ou percepção do sinal transmitido seja insignificante. Esta preocupação inexiste no *superimposed training*.

O artigo está estruturado como segue. A seção 1 apresenta uma breve revisão sobre marca d'água em áudio. Nas seções 2 e 3, são revisados os métodos de equalização e identificação adaptativas utilizando marca d'água, apresentados originalmente em [3, 4, 5, 6]. Na seção 4, é mostrado como esses métodos podem ser aplicados a sinais de áudio. A seção 5 apresenta resultados experimentais que demonstram a viabilidade dos métodos propostos e analisam a inaudibilidade da marca d'água com base em avaliações objetivas de qualidade de áudio. Finalmente, a seção 6 traz conclusões e perspectivas de trabalhos futuros.

1 MARCA D'ÁGUA DE ÁUDIO

Um sistema de marca d'água pode ser modelado como a transmissão de informação através de um canal de comunicação: o sinal de marca d'água é responsável por transportar informação útil, enquanto o sinal hospedeiro é encarado como ruído. Em canais de comunicação tradicionais, a potência do sinal de informação é geralmente maior do que a do ruído, e este último é muitas vezes assumido como sendo gaussiano e branco. No caso da marca d'água de áudio, porém, estas premissas não são válidas: para evitar degradações audíveis, o sinal de marca d'água deve ter, via de regra, uma potência muito inferior à do sinal de áudio; além disso, este último é geralmente não-estacionário e fortemente colorido.

Existem diversos métodos para a geração do sinal de marca d'água e sua inserção no áudio. Um dos métodos mais empregados consiste em partir de uma marca d'água branca (com energia espalhada em todo o espectro de áudio) e a submeter a um processo de conformação espectral guiado pelas propriedades de mascaramento freqüencial do áudio [7]. A marca assim produzida é adicionada ao áudio no domínio temporal.

Modelos psicoacústicos podem ser empregados para determinar propriedades de mascaramento freqüencial. Para cada quadro de áudio, estes modelos fornecem uma curva, no domínio da freqüência, denominada *limiar de mascaramento*. Essa curva limita a densidade espectral de potência de um ruído para que este possa ser adicionado ao áudio original sem produzir degradações perceptíveis. Tais modelos são utilizados em codificadores de áudio que aliam alta fidelidade sonora a forte compressão de dados.

Por meio de uma operação de conformação espectral, pode-se fazer com que a densidade espectral de potência da marca d'água esteja sempre abaixo do limiar de mascaramento do áudio, embora próxima deste em todas as freqüências do espectro. Com isso, a potência do sinal de marca d'água é maximizada, aumentando a robustez do sistema, ao mesmo tempo em que as distorções audíveis são minimizadas, garantindo a transparência da marca.

2 EQUALIZAÇÃO ADAPTATIVA UTILIZANDO MARCA D'ÁGUA

O sistema de equalização supervisionada baseado em marca d'água encontrado na figura 1. Por simplicidade, os sinais e sistemas analisados são supostos reais.

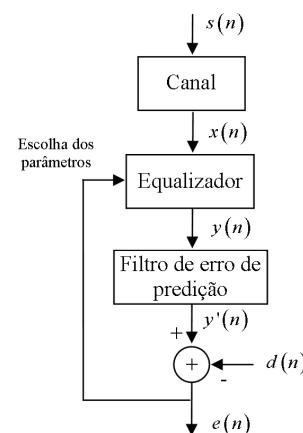


Figura 1: Equalização baseada em marca d'água.

O sinal transmitido $s(n)$ é obtido pela sobreposição da marca d'água $m(n)$ e do sinal de informação $t(n)$:

$$s(n) = t(n) + \alpha_m m(n) \quad (1)$$

onde α_m é um fator de escala. A marca d'água é, por definição, um sinal branco e de potência normalizada,

sendo utilizada como referência no filtro equalizador. Isto leva a uma função de erro quadrático correspondente ao critério de Wiener:

$$J = E\{[y(n) - d(n)]^2\} \quad (2)$$

onde $d(n) = \alpha_m m(n)$ é o sinal de referência e $y(n)$ o sinal na saída do equalizador. Os coeficientes ótimos \mathbf{w}_o do equalizador são obtidos pela minimização desta função de erro.

O canal é modelado como um sistema FIR de ordem N . O equalizador, por sua vez, é também um filtro FIR de ordem M . A matriz de autocorrelação \mathbf{R} na entrada do equalizador, de dimensão $M \times M$, é expressa como:

$$\mathbf{R} = \alpha_m^2 \mathbf{H} \mathbf{E} [\mathbf{m}(n) \mathbf{m}^T(n)] \mathbf{H}^T + \mathbf{H} \mathbf{E} [\mathbf{t}(n) \mathbf{t}^T(n)] \mathbf{H}^T \quad (3)$$

onde α_m^2 é a variância da marca d'água, $\mathbf{m}(n)$ e $\mathbf{t}(n)$ são, respectivamente, vetores contendo M amostras da marca d'água e do sinal de informação a partir do instante n , e $\mathbf{H} \in \mathbb{R}^{M \times (M+N-1)}$ é uma matriz de convolução do canal. O vetor de correlação cruzada \mathbf{p} entre a entrada do equalizador e a sua saída desejada é dado por:

$$\mathbf{p} = \alpha_m^2 \mathbf{H} \mathbf{E} [\mathbf{m}(n) m(n-\delta)] \quad (4)$$

onde δ é o atraso ótimo de equalização. Os coeficientes do equalizador que levam ao ponto de mínimo da função de erro J são:

$$\mathbf{w}_o = \mathbf{R}^{-1} \mathbf{p} = \alpha_m^2 (\alpha_m^2 \mathbf{H} \mathbf{E} [\mathbf{m}(n) \mathbf{m}^T(n)] \mathbf{H}^T + \mathbf{H} \mathbf{E} [\mathbf{t}(n) \mathbf{t}^T(n)] \mathbf{H}^T)^{-1} \mathbf{H} \mathbf{E} [\mathbf{m}(n) m(n-\delta)] \quad (5)$$

A presença de $\mathbf{t}(n)$ na equação (5) implica em uma dependência da solução ótima com relação ao sinal de informação. Por simplicidade, este sinal será inicialmente suposto descorrelacionado, introduzindo-se em seguida um elemento no sistema que permitirá estender os resultados da análise a sinais de informação correlacionados.

Pelo fato de a marca d'água ser descorrelacionada e de potência normalizada, o termo $E[\mathbf{m}(n) \mathbf{m}^T(n)]$ torna-se uma matriz identidade. Assumindo um sinal de informação descorrelacionado, o termo $E[\mathbf{t}(n) \mathbf{t}^T(n)]$ é também uma matriz identidade multiplicada pela variância σ_t^2 de $\mathbf{t}(n)$. Com isso, a matriz \mathbf{R} pode ser reescrita como:

$$\mathbf{R} = \alpha_m^2 \mathbf{H} \mathbf{H}^T + \sigma_t^2 \mathbf{H} \mathbf{H}^T = (\alpha_m^2 + \sigma_t^2) \mathbf{H} \mathbf{H}^T \quad (6)$$

Agora, a matriz de correlação \mathbf{R} não está mais subordinada ao sinal $\mathbf{t}(n)$, exceto pela variância σ_t^2 , que pode ser obtida de forma aproximada a partir da potência do próprio sinal recebido. Desse modo, os coeficientes ótimos do equalizador podem ser expressos como:

$$\mathbf{w}_o = \frac{\alpha_m^2}{\alpha_m^2 + \sigma_t^2} (\mathbf{H} \mathbf{H}^T)^{-1} \mathbf{H} \mathbf{E} [\mathbf{m}(n) m(n-\delta)] \quad (7)$$

Exceto por um fator de escala que depende das variâncias dos sinais transmitidos e que pode ser compensado no receptor, os coeficientes do equalizador tendem à solução que minimiza o erro quadrático de equalização.

Para o caso de sinais de informação correlacionados, tem-se:

$$\mathbf{R} = \alpha_m^2 \mathbf{H} \mathbf{H}^T + \mathbf{H} \mathbf{E} [\mathbf{t}(n) \mathbf{t}^T(n)] \mathbf{H}^T \quad (8)$$

Empregando-se um método de branqueamento, pode-se fazer com que o sistema convirja para a solução que minimiza o erro quadrático de equalização. No esquema da figura 1, o filtro de erro de predição tem por objetivo branquear a saída do equalizador. Por meio de um preditor linear, as componentes periódicas de $y(n)$ são isoladas, sendo em seguida subtraídas deste sinal. Desta forma, isola-se a parcela descorrelacionada $y'(n)$ do sinal de saída do equalizador. Tem-se assim:

$$y'(n) = y(n) - \sum_{k=1}^P u_k y(n-k-\Delta) \quad (9)$$

onde u_k são os coeficientes do preditor com P elementos de atraso. Conforme o erro quadrático do equalizador tende ao seu valor mínimo, $y'(n)$ tende às componentes descorrelacionadas do sinal de informação, mais a marca d'água. A análise feita sob a hipótese de um sinal de informação descorrelacionado torna-se portanto mais precisa à medida que o equalizador se aproxima da solução ótima. O atraso Δ do filtro de predição deve ser grande o bastante para que seja possível remover eficientemente a correlação do sinal $y(n)$.

3 IDENTIFICAÇÃO ADAPTATIVA UTILIZANDO MARCA D'ÁGUA

A figura 2 ilustra o sistema de identificação supervisionada baseado em marca d'água. Como no caso da equalização, a marca d'água é um sinal branco e normalizado. O sinal transmitido é obtido pela sobreposição da marca d'água ao sinal de informação, conforme a equação (1). A marca escalonada é aplicada à entrada do filtro identificador. A função de erro quadrático é dada novamente pela equação (2), mas se trata agora do erro entre a saída do filtro identificador $y(n)$ e a saída do canal $d(n) = x(n)$.

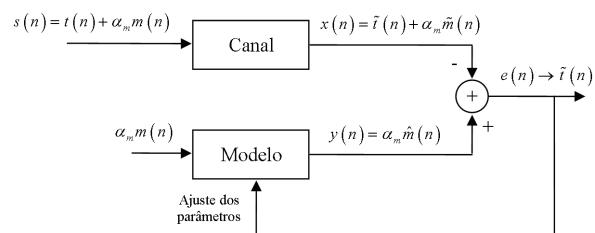


Figura 2: Identificação baseada em marca d'água.

A função de erro quadrático pode ser escrita como:

$$J(\mathbf{w}) = \sigma_x^2 - \mathbf{p}_m^T \mathbf{R}_m^{-1} \mathbf{p}_m + \\ + (\mathbf{w} - \mathbf{R}_m^{-1} \mathbf{p}_m)^T \mathbf{R}_m (\mathbf{w} - \mathbf{R}_m^{-1} \mathbf{p}_m) \quad (10)$$

onde σ_x^2 é a variância do sinal na saída do canal e os termos \mathbf{R}_m e \mathbf{p}_m expressam, respectivamente, a matriz de autocorrelação do sinal de marca d'água e o vetor de correlação cruzada entre o sinal de marca d'água e este mesmo sinal após a ação do canal.

O identificador ótimo é obtido pela minimização do erro quadrático com respeito aos coeficientes \mathbf{w} do modelo de identificação. Da equação (10), infere-se que o erro mínimo ocorre quando o termo $(\mathbf{w} - \mathbf{R}_m^{-1} \mathbf{p}_m)$ se anula:

$$\mathbf{w}_o - \mathbf{R}_m^{-1} \mathbf{p}_m = \mathbf{0} \Rightarrow \mathbf{w}_o = \mathbf{R}_m^{-1} \mathbf{p}_m \quad (11)$$

Observa-se que os coeficientes correspondentes ao ponto de erro quadrático mínimo não guardam relação com o sinal de informação, dependendo apenas das características do sinal de marca d'água; basta portanto o conhecimento prévio deste último para que seja possível ajustar o modelo de identificação.

4 APLICAÇÃO DOS MÉTODOS DE EQUALIZAÇÃO E IDENTIFICAÇÃO A CANAIS DE ÁUDIO

Para que os métodos descritos nas seções anteriores possam ser aplicados à equalização/identificação de canais de áudio, é necessário evitar a degradação da qualidade do áudio pela inserção da marca d'água. Conforme descrito na seção 1, este objetivo pode ser atingido por meio de um processo de conformação espectral. A figura 3 ilustra esta abordagem: a marca d'água branca $m(n)$ é processada por um filtro de conformação espectral $H(f)$ cuja resposta em freqüência é guiada por um modelo psicoacústico, dando origem à marca d'água filtrada $w(n)$; esta é adicionada ao áudio $t(n)$, produzindo o sinal marcado $s(n)$.

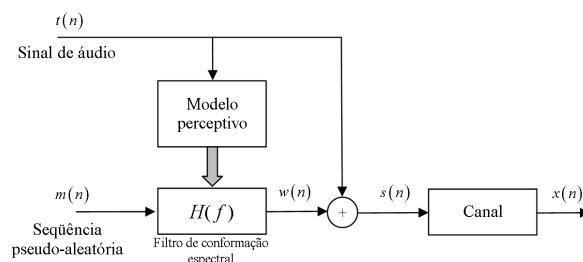


Figura 3: Inserção de marca d'água em áudio.

No receptor, o sinal $x(n)$ (áudio + marca distorcidos pelo canal) sofre primeiramente um novo processo de conformação espectral por meio de um filtro $G(f)$, cujo objetivo é reverter o efeito da conformação espectral efetuada no transmissor. Este processo é ilustrado nas figuras 4 (equalização) e 5 (identificação). O filtro

$G(f)$ é também calculado através de um modelo psicoacústico, tomando-se por base o próprio sinal $x(n)$ para a obtenção do limiar de mascaramento. Supõe-se aqui que a potência da marca d'água é muito inferior à do áudio, não afetando significativamente o limiar de mascaramento, e que a distorção do canal não é forte o bastante para inutilizar o limiar de mascaramento calculado a partir do sinal recebido.

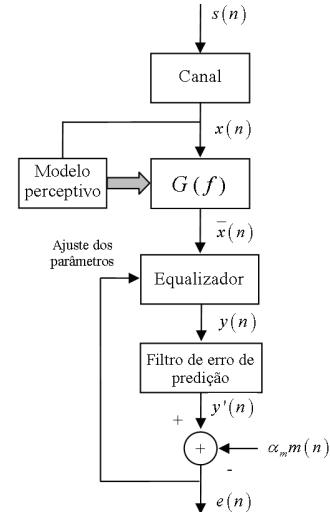


Figura 4: Equalização utilizando marca d'água em áudio.

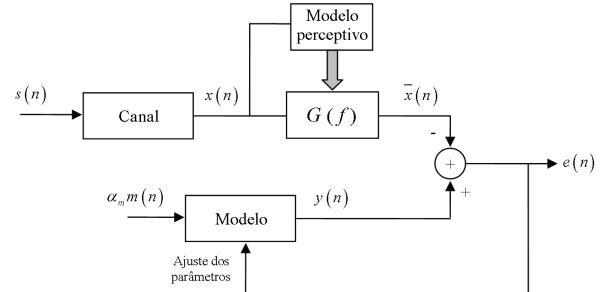


Figura 5: Identificação utilizando marca d'água em áudio.

O sinal $\bar{x}(n)$ na saída do filtro $G(f)$ contém portanto uma estimativa da marca $m(n)$ distorcida pelo canal, mais o áudio distorcido pelo canal e filtrado por $G(f)$. Este sinal é então utilizado nos processos de equalização e identificação, conforme discutido nas seções 2 e 3.

5 RESULTADOS EXPERIMENTAIS

As simulações efetuadas empregaram o modelo psicoacústico número 1 do padrão MPEG-1, disponibilizado pela ISO em implementação de referência. O filtro de conformação espectral, de ordem 50, foi obtido a partir do limiar de mascaramento para quadros de sinal de 512 amostras, utilizando o algoritmo de Levinson [7].

Foram realizadas simulações com materiais de áudio variados e potencialmente sensíveis a degradações, tomando por base avaliações realizadas pelo *Communications Research Centre* do Canadá [8]. O material selecionado, apresentado na tabela 1, inclui voz e instrumentos musicais, tanto isolados como em conjunto.

Sinal	Descrição	Dur.	Fonte
dires	rock	10 s	CD 7599-25264-2 (tr. 6)
svega	canto	10 s	AT&T mix
trump	trompete	10 s	Univ. Miami
symp	orquestra	10 s	EBU SQAM CD (tr. 17)

Tabela 1: Material de áudio utilizado nas avaliações de qualidade.

Para avaliação da qualidade do áudio após a inserção da marca d'água, foi empregada uma implementação comercial do algoritmo de avaliação objetiva PEAQ [9]. Foram utilizados arquivos de áudio no formato PCM linear, 16 bits por amostra, mono, com taxa de amostragem de 44,1 kHz (aceita pelo algoritmo PEAQ graças a uma extensão proprietária da implementação utilizada).

Qualidade do áudio com marca d'água

O indicador de qualidade de áudio do PEAQ, denominado *Objective Difference Grade* (ODG), varia em uma escala contínua desde 0,0 (degradação imperceptível) até -4,0 (degradação muito incômoda). Degradações na faixa de -1,0 a -0,1 podem ser consideradas imperceptíveis para ouvintes comuns (sem treinamento específico para detecção de degradações em áudio), enquanto degradações entre -0,5 e -0,1 são em geral imperceptíveis até mesmo para especialistas em áudio [10].

A figura 6 apresenta a curva de ODG em função da relação de potência sinal-marca d'água (SWR) para todas as amostras de áudio analisadas, nas quais foi inserida uma marca d'água pelo processo de conformação espectral guiado por um limiar de mascaraamento. A figura inclui ainda uma curva representando a média dos resultados para todos os sinais. Observa-se que, para SWR = 20 dB, tem-se um valor ODG médio de -0,8, dentro da faixa de imperceptibilidade para um ouvinte comum. Para SWR = 23 dB, tem-se um valor ODG médio de -0,45, dentro da faixa de imperceptibilidade para especialistas em áudio. A faixa de SWR entre 20 e 23 dB é portanto adequada para sistemas de marca d'água de áudio, com um ajuste fino dependente da aplicação em questão.

A necessidade do modelo psicoacústico e do algoritmo de conformação espectral é justificada pelas curvas da figura 7, onde se vê a medida de dírtorção ODG em função de SWR para as mesmas amostras de áudio da figura anterior, mas utilizando uma marca d'água branca (i.e. totalmente espalhada no espectro de áudio).

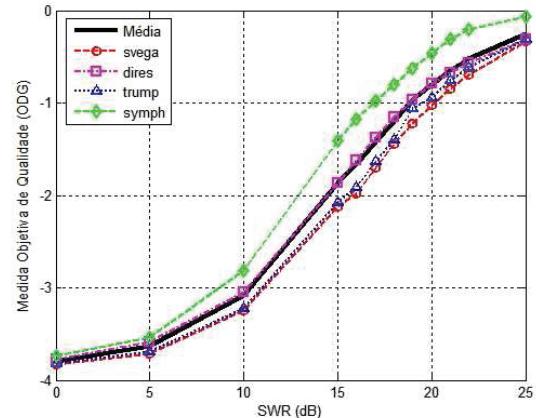


Figura 6: Medida objetiva média de qualidade versus SWR usando modelo psicoacústico.

Esta abordagem se assemelha à empregada nas técnicas de *superimposed training*. Para SWR = 20 dB e 23 dB, respectivamente, o ODG médio está em torno de -3,60 e -3,25, ambos na faixa classificada como degradação “muito incômoda”. Para uma SWR = 25 dB, a medida ODG média aproxima-se de -3,0, limiar entre as faixas de degradação “muito incômoda” e “incômoda”.

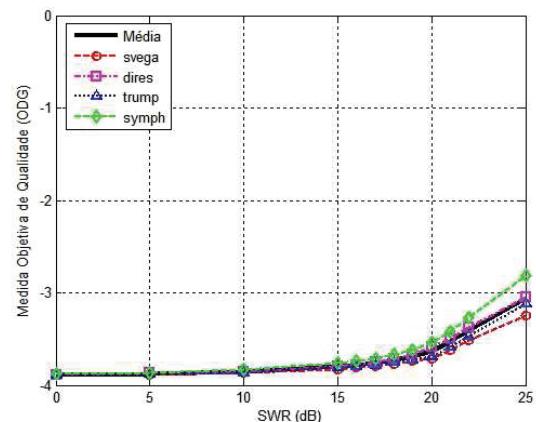


Figura 7: Medida objetiva média de qualidade versus SWR para marca d'água branca.

A comparação entre os resultados das figuras 6 e 7 evidencia a diferença entre a técnica proposta e os métodos de *superimposed training*, não focados na transparência do sinal de referência. Para que o *superimposed training* pudesse ser aplicado a sinais de áudio sem introduzir degradações perceptíveis, seria necessário reduzir a potência do sinal de referência a níveis praticamente indetectáveis.

Sistema de Equalização

Apresentamos a seguir resultados experimentais para o sistema de equalização de canais de áudio utilizando marca d'água. O canal foi simulado por um

filtro FIR de quatro coeficientes e fase mista:

$$H(z) = 1 + 1,2z^{-1} - 0,3z^{-2} + 0,8z^{-3} \quad (12)$$

Foi empregado um equalizador adaptativo com cinco coeficientes inicializados na origem.

Para branqueamento dos sinais de áudio, utilizou-se um filtro de erro de predição linear com 750 coeficientes e um atraso Δ de 2.200 amostras. Este atraso foi determinado supondo uma correlação significativa dos sinais de áudio numa faixa de até 50 ms, a uma taxa de 44,1 kHz. O equalizador foi otimizado através do algoritmo RLS com fator de esquecimento $\lambda = 1$. O filtro de erro de predição foi otimizado através do algoritmo LMS com passo $\mu = 0,0001$. Esta combinação apresentou 100% de convergência global nos casos analisados.

A tabela 2 apresenta o erro quadrático médio (EQM) entre o sinal original e o sinal equalizado para SWR = 20 dB e 23 dB, tendo sido empregados o modelo psicoacústico e o filtro de conformação espectral para minimização da distorção audível introduzida pela marca d'água. A título de comparação, o equalizador ótimo de Wiener de cinco coeficientes, calculados analiticamente, fornece os seguintes valores de EQM para os sinais em questão: 0,0823 para *dires*, 0,0791 para *svega*, 0,0595 para *trump* e 0,0796 para *sympf*. Para SWR = 20 dB, o EQM apresenta valores relativamente baixos e próximos dos valores obtidos com o equalizador ótimo. Para SWR = 23 dB, o EQM apresenta um ligeiro aumento, porém permanece relativamente próximo dos valores ótimos, encontrando-se em um patamar satisfatório para a maioria das aplicações.

A tabela 3 apresenta medidas similares para uma marca d'água branca. Apesar de uma ligeira redução no EQM, os resultados são praticamente equivalentes aos obtidos com o uso do modelo psicoacústico e do filtro de conformação espectral, indicando que a introdução destes elementos não prejudicou a convergência do sistema.

Sistema de Identificação

Apresentamos a seguir resultados experimentais para o sistema de identificação de canais de áudio utilizando marca d'água. O canal foi simulado por meio do filtro FIR de quatro coeficientes especificado na equação (12). Foi utilizado um modelo de identificação também com quatro coeficientes, inicializados na ori-

Sinal	EQM (20 dB)	EQM (23 dB)
<i>dires</i>	0,179	0,211
<i>svega</i>	0,176	0,206
<i>trump</i>	0,161	0,192
<i>sympf</i>	0,169	0,201

Tabela 2: EQM do sinal equalizado com uso de modelo psicoacústico.

Sinal	EQM (20 dB)	EQM (23 dB)
<i>dires</i>	0,173	0,202
<i>svega</i>	0,168	0,197
<i>trump</i>	0,154	0,187
<i>sympf</i>	0,165	0,192

Tabela 3: EQM do sinal equalizado com marca d'água branca.

gem. O modelo de identificação foi otimizado através do algoritmo RLS com fator de esquecimento $\lambda = 1$.

Para avaliação do desempenho do algoritmo de identificação, foi utilizada uma medida de desvio dos coeficientes do filtro identificador em relação aos coeficientes do canal:

$$D = \frac{\|\mathbf{h}_o - \mathbf{h}\|}{\|\mathbf{h}\|} \quad (13)$$

sendo \mathbf{h}_o a estimativa dos coeficientes do canal obtida pelo algoritmo e \mathbf{h} o valor exato desses coeficientes. Quanto mais próximo de zero o desvio D , melhor é o desempenho do algoritmo.

A tabela 4 apresenta o desvio D dos parâmetros do filtro identificador para SWR = 20 dB e 23 dB, tendo sido empregados o modelo psicoacústico e o filtro de conformação espectral para minimização da distorção audível introduzida pela marca d'água. Para uma mesma SWR, os valores obtidos para o desvio D foram muito próximos entre si para todos os sinais analisados. O desvio máximo admissível depende da aplicação em questão; no entanto, os valores obtidos encontram-se dentro de um patamar geralmente considerado satisfatório.

Sinal	Desvio D (20 dB)	Desvio D (23 dB)
<i>dires</i>	0,0586	0,0779
<i>svega</i>	0,0464	0,0656
<i>trump</i>	0,0373	0,0513
<i>sympf</i>	0,0428	0,0639

Tabela 4: Desvio da identificação com uso de modelo psicoacústico.

A tabela 5 apresenta medidas similares para uma marca d'água branca. Apesar de uma ligeira redução no desvio D , os resultados são praticamente equivalentes aos obtidos com o uso do modelo psicoacústico e do filtro de conformação espectral. Isso indica que, como no caso do sistema de equalização, a introdução destes elementos não prejudicou a convergência do sistema.

6 CONCLUSÕES

Neste trabalho, foram propostos métodos adaptativos de equalização e identificação de canais de áudio utilizando uma marca d'água como sinal de supervisão. Para evitar distorções audíveis, a marca d'água é submetida a um processo de conformação espectral guiado por um modelo psicoacústico.

Sinal	Desvio D (20 dB)	Desvio D (23 dB)
dires	0,0521	0,0703
svega	0,0394	0,0589
trump	0,0318	0,0446
symp	0,0382	0,0576

Tabela 5: Desvio da identificação com marca d'água branca.

Avaliações objetivas de qualidade de áudio mostraram que a marca d'água é imperceptível para uma relação de potência áudio/marca acima de um limiar entre 20 e 23 dB. Além disso, resultados experimentais indicaram que a conformação espectral não prejudicou significativamente o desempenho dos sistemas adaptativos em relação ao que se teria com uma marca d'água branca de mesma potência.

Como perspectiva de trabalhos futuros, pretende-se avaliar a aplicabilidade da abordagem proposta a problemas correlatos de processamento de áudio, como por exemplo o cancelamento de eco e reverberação. Os métodos apresentados podem ainda ser adaptados a outras classes de sinais com interpretação sensorial, tais como imagens e vídeo.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] A. R. Varma, L. L. H. Andrew, C. R. N. Athaudage e J. H. Manton, “Iterative algorithms for channel identification using superimposed pilots”, *Australian Communications Theory Workshop*, fevereiro de 2004.
- [2] J. K. Tugnait e W. Luo, “On channel estimation using superimposed training and first-order statistics”, *IEEE Communications Letters*, vol. 7, no. 9, setembro de 2003.
- [3] M. Uliani Neto, L. de C. T. Gomes, J. M. T. Romano e M. Bonnet, “Adaptive equalization based on watermarking”, *VI International Telecommunications Symposium (ITS 2006)*, Fortaleza, Brasil, 3-6 de setembro de 2006.
- [4] M. Uliani Neto, L. de C. T. Gomes, J. M. T. Romano e M. Bonnet, “Égalisation et synchronisation utilisant un tatouage comme signal de référence”, *XXI Colloque GRETSI*, Troyes, França, 11-14 de setembro de 2007.
- [5] M. Uliani Neto, “Equalização e identificação adaptativas utilizando marca d'Água como sinal de supervisão”, Dissertação de mestrado, Universidade Estadual de Campinas (Unicamp), Campinas, Brasil, 2008.
- [6] M. Uliani Neto, L. de C. T. Gomes e J. M. T. Romano, “Identificação adaptativa supervisionada utilizando marca d'Água digital”, *XXV Simpósio Brasileiro de Telecomunicações (SBrT 2007)*, Recife, Brasil, 3-6 de setembro de 2007.
- [7] L. de C. T. Gomes, “Tatouage de signaux audio”, Tese de doutorado, Universidade René Descartes (Paris V), Paris, França, 2002.
- [8] G. A. Soulodre, T. Grusec, M. Lavoie e L. Thibault, “Subjective evaluation of state-of-the-art 2-channel audio codecs”, *AES 104th Convention*, 1998.
- [9] ITU-R Recommendation BS.1387-1, “Method for objective measurements of perceived audio quality”, 1998.
- [10] M. Arnold, “Subjective and objective quality evaluation of watermarked audio tracks”, *Proceedings of the Second International Conference on WEB Delivering of Music (WEDELMUSIC'02)*, 2002.



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

FlawQ: Um *Plug-in* VST para Equalização Gráfica Digital

Felipe C. V. Martins,¹ Leonardo de O. Nunes,¹ Alan F. Tygel¹ e
Luiz W. P. Biscainho¹

¹ LPS - DEL/Poli & PEE/COPPE, UFRJ
Caixa Postal 68504 - Rio de Janeiro, RJ, 21941-972, Brazil

felipecvmartins@lps.ufrj.br, lonnes@lps.ufrj.br, alan@lps.ufrj.br, wagner@lps.ufrj.br

RESUMO

Este trabalho consiste na modificação e implementação do projeto de um equalizador gráfico digital destinado à operação até 20kHz. A arquitetura se baseia num projeto industrial originalmente voltado para implementação em tempo real via DSP, com canais espaçados de forma aproximadamente linear por oitava. A solução modificada aqui proposta foi concebida como um *plug-in* VST com interface gráfica amigável, e inclui como facilidade adicional a visualização da resposta de freqüência resultante do equalizador com os ganhos escolhidos pelo usuário.

0 INTRODUÇÃO

A evolução dos processadores digitais no último quarto do século XX permitiu a aproximação entre as aplicações de ciência avançada e o usuário comum. Na área de áudio, o processamento digital permeia desde os equipamentos domésticos de som até os diversos aplicativos para manipulação e reprodução de áudio disponíveis para computadores pessoais. É possível montar um sistema doméstico relativamente sofisticado de processamento de áudio a baixo custo.

Este trabalho tem como objetivo mostrar o uso de uma ferramenta avançada de filtragem numa aplicação típica de áudio que possa ser facilmente utilizada por

um profissional sem a necessidade de conhecimento especializado em processamento de sinais. Em particular, será apresentado o procedimento de projeto de um equalizador gráfico digital de 6 oitavas baseado em uma estrutura multi-taxa. Esta solução foi escolhida por sua baixa complexidade computacional, uma vez que uma das especificações do equalizador era a operação em tempo-real. A fim de permitir a fácil utilização e portabilidade do sistema, utilizou-se o padrão de *plug-in* VST¹, amplamente aceito por fabricantes e usuários de aplicativos de áudio profissional. Como incremento ao

¹A marca VST (Virtual Studio Technology) é propriedade da Steinberg Co.

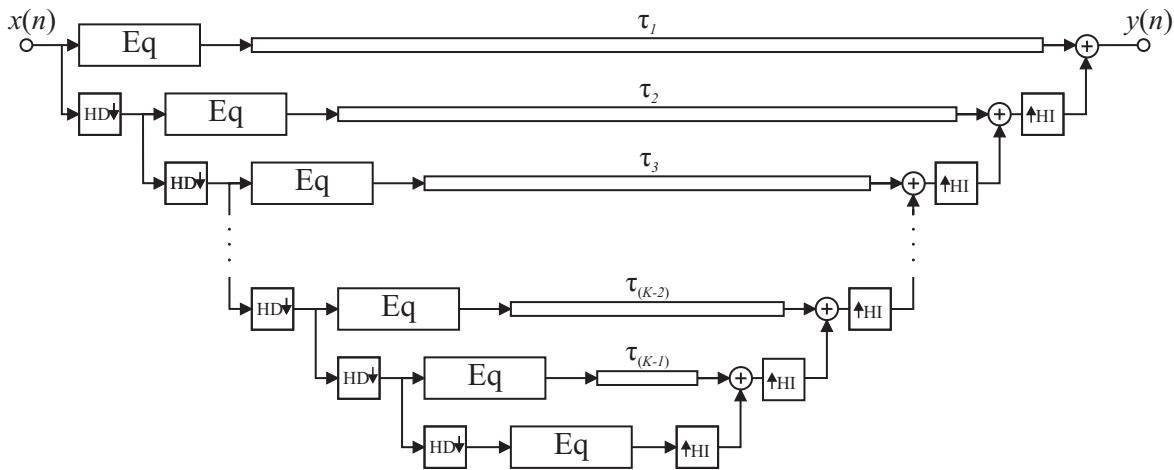


Figura 1: Diagrama de blocos da estrutura em multi-taxa. Adaptado de [1]. Os blocos $\text{HD}\downarrow$ e $\uparrow\text{HI}$ são expandidos na Figura 2.

projeto, a resposta em freqüência verdadeira resultante do equalizador com os ganhos escolhidos pelo usuário é exibida na interface gráfica do *plug-in*.

Após esta Introdução, o artigo é organizado da seguinte forma: É realizada uma descrição da arquitetura adotada e das modificações sobre ela propostas. Em seguida, a implementação em *software* desta estrutura é detalhada. É apresentada a implementação do *plug-in* em VST correspondente. Por fim são mostradas as conclusões.

1 EQUALIZADOR MULTI-TAXA

Esta seção realiza uma breve descrição da estrutura do equalizador implementado neste trabalho, originalmente proposta em [1], e das modificações realizadas sobre aquele projeto inicial.

O equalizador foi concebido na forma de uma estrutura modular de banco de filtros em multi-taxa, mostrada nas Figuras 1 e 2 para um número arbitrário K de sub-bandas. Partindo de um primeiro ramo que opera

cada nível. Cada um dos ramos (exceto o primeiro) é composto por:

filtro anti-aliasing – filtro FIR (do inglês, *finite-length impulse response*) passa-baixas com freqüência de corte igual a $0,5\pi$ rad;

decimador por 2 ($\downarrow 2$) – módulo que remove uma de cada duas amostras do sinal no domínio do tempo;

equalizador linear (Eq) – equalizador gráfico linear com os centro de suas bandas de atuação espaçados linearmente na freqüência;

bloco de atraso (τ_k) – módulo que atrasa o sinal no tempo de um número de amostras pré-especificado para o ramo k ;

interpolador por 2 ($\uparrow 2$) – módulo que insere um zero entre cada duas amostras do sinal no tempo;

filtro anti-imagem – filtro FIR passa-baixas com freqüência de corte igual a $0,5\pi$ rad.

O equalizador linear (Eq) presente em cada ramo possui 9 canais, o primeiro e o último com largura de banda igual à metade da largura de banda dos demais. A faixa total de operação desses equalizadores é função do nível ao qual pertence, que também define sua taxa de processamento. Para um equalizador no k -ésimo ramo, a faixa total de operação vai de 0 a $2^{-(k-1)}\pi$ rad.

Essa estrutura permite um número reduzido de operações, pois cada ramo opera apenas na metade da taxa do ramo imediatamente superior. Além disso, sua modularidade permite que o número de subdivisões em ‘meias-bandas’ possa ser facilmente aumentado, com o consequente aumento na resolução de freqüência fornecida ao usuário para montar as curvas de equalização.

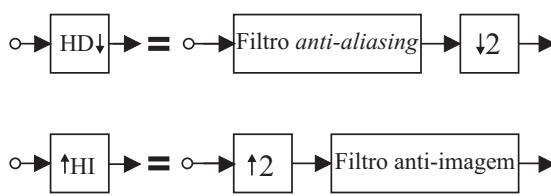


Figura 2: Expansão dos blocos $\text{HD}\downarrow$ e $\uparrow\text{HI}$ da Figura 1.

sobre todo o espectro do sinal $x(n)$, o sistema é desdobrado em ramos aninhados que separam seqüencialmente e processam a metade inferior do espectro. O sinal de saída $y(n)$ é a soma dos sinais processados em

Todos os filtros da estrutura foram projetados como FIR de fase linear [2].

Da mesma forma que no artigo original [1], os filtros *anti-alias* e anti-imagem são passa-baixas idênticos. As especificações aqui adotadas para eles foram:

- faixa de passagem até 0,4535 rad;
- faixa de rejeição a partir de 0,5 rad;
- ordem 140.

O projeto foi feito por otimização *least-squares*, e atingiu a resposta mostrada na Figura 3, com uma atenuação na banda de rejeição maior que 96 dB.

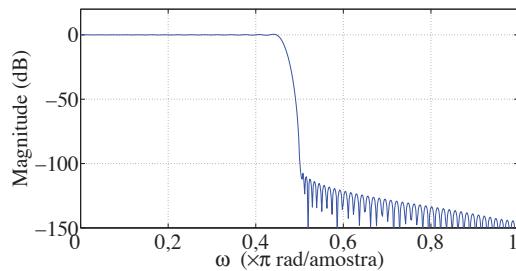


Figura 3: Resposta em freqüência em magnitude do filtro utilizado como *anti-alias* e anti-imagem.

A operação em diferentes taxas gera atrasos diferentes em cada nível da estrutura multi-taxa. Para que a soma dos sinais oriundos de dois ramos diferentes seja coerente, é necessário sincronizar a saída do ramo com a saída do ramo imediatamente abaixo na estrutura. Por isso, é necessário aplicar um atraso adequado na saída dos equalizadores em cada ramo, dependente do atraso resultante do ramo imediatamente abaixo e do atraso já implícito nos demais blocos do ramo em questão. Como todos os filtros envolvidos são filtros FIR de fase linear, esses atrasos são facilmente calculados. A expressão abaixo mostra o cálculo do atraso τ_k do k -ésimo ramo em função dos atrasos τ_{k+1} do ramo seguinte, τ_{eq} do equalizador, τ_{aa} do filtro *anti-alias* e τ_{ai} do filtro anti-imagem:

$$\tau_k = 2\tau_{k+1} + \tau_{eq} + \tau_{aa} + \tau_{ai}. \quad (1)$$

Deve-se observar que o atraso do último ramo pode ser feito nulo, isto é, $\tau_K = 0$. Neste projeto, $\tau_{aa} = \tau_{ai} = 70$.

O número de operações aritméticas por amostra do sinal de entrada por ramo pode ser obtido a partir do comprimento dos filtros *anti-alias* e anti-imagem e do número de operações realizadas pelo equalizador. O número de adições e multiplicações reais por amostra do sinal de entrada no ramo k é

$$\nu_k = \frac{\nu_{eq} + \nu_{aa} + \nu_{ai} + 1}{2^{k-1}}, \quad (2)$$

onde ν_{eq} é o número de operações aritméticas realizadas pelo equalizador, que será calculado na próxima

subseção; e ν_{aa} e ν_{ai} são os números de operações aritméticas realizadas pelos filtros *anti-alias* e anti-imagem, respectivamente, calculados pela expressão

$$\nu_x = 2N_x - 1, \text{ com } x = aa \text{ ou } ai, \quad (3)$$

sendo N_x o comprimento do filtro x em amostras. Neste projeto, $\nu_{ai} = \nu_{aa} = 281$ operações por amostra. O número de operações aritméticas da estrutura completa é, então:

$$\nu = \sum_{k=1}^K \nu_k. \quad (4)$$

Percebe-se que a complexidade computacional cresce de forma linear com o comprimento dos filtros e a complexidade do equalizador. Este projeto utilizou 6 ramos na estrutura multi-taxa, fazendo com que a faixa espectral de operação do equalizador seja subdividida sucessivamente em cinco meias-bandas. Isso implica um pequeno acréscimo na complexidade do sistema em relação ao projeto original, que se compunha de apenas 4 ramos.

1.1 Equalizador Linear

Nesta subseção, o chamado ‘equalizador linear’ utilizado na estrutura multi-taxa e descrito em [1] é brevemente revisto. Seu projeto foi inteiramente preservado, a menos da especificação do filtro-protótipo, como se verá adiante.

O equalizador linear divide o espectro num total de 9 faixas de freqüência de igual largura, à exceção da primeira e da última, com metade da largura das demais. Além disso, os canais resultantes possuem fase linear e podem ser calculados utilizando um número reduzido de operações aritméticas. A magnitude da resposta em freqüência dos 9 canais deste equalizador pode ser vista na Figura 4.

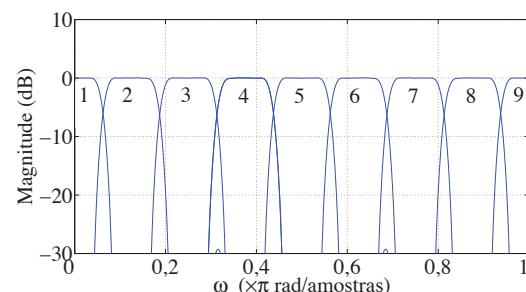


Figura 4: Canais do equalizador linear.

É utilizada a estrutura em árvore da Figura 5, em que os ganhos de saída de A_1 a A_9 são controlados pelo usuário. Em seu projeto, que se baseia na técnica de FRM [2] (do inglês *Frequency Response Masking*), um filtro-protótipo $F_B(z)$ é interpolado de modo que ele e o seu filtro complementar gerem a divisão desejada do espectro. As réplicas indesejadas na resposta do filtro que gera determinada banda, decorrentes da interpolação,

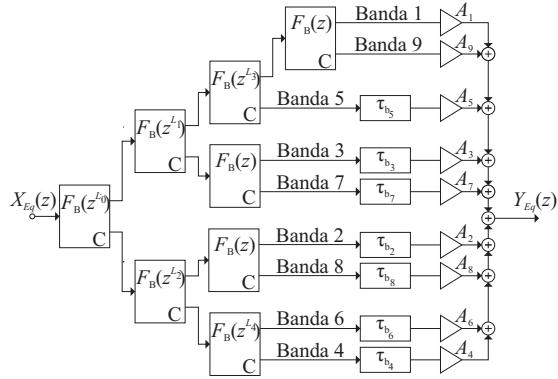


Figura 5: Diagrama de blocos do equalizador linear. “C” indica a saída do filtro complementar. Adaptado de [1].

são estruturalmente eliminadas nos níveis subseqüentes da árvore. A Figura 6 ilustra o processo de geração do primeiro canal do equalizador linear. Deve-se observar

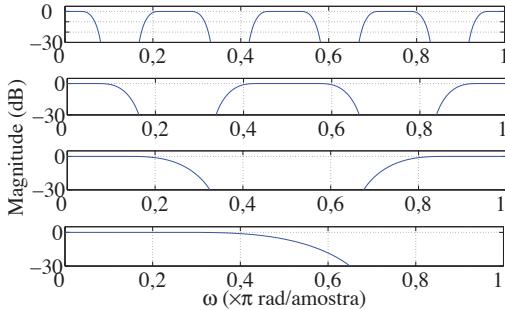


Figura 6: Construção do Canal 1 do equalizador linear a partir das versões modificadas do filtro protótipo. Os gráficos representam, de cima para baixo, as respostas de magnitude na freqüência dos filtros $F_B(z^{L_0})$, $F_B(z^{L_1})$, $F_B(z^{L_3})$ e $F_B(z)$. O filtro resultante é o Canal 1 da Figura 4.

que esta estrutura possui ganho unitário se os ganhos na saída de todos os canais forem escolhidos iguais a 1.

Para reduzir a complexidade computacional, o equalizador utiliza filtros de meia-banda simétricos com ordem par. Apenas metade dos coeficientes desses filtros são não-nulos, o que permite reduzir o número de multiplicações e adições necessárias a um quarto da ordem do filtro. Além disso, o uso de filtros complementares, relacionados pela expressão

$$F_B(z) + \overline{F_B}(z) = 1,$$

evita operações redundantes. A saída $\bar{y}(n)$ do filtro complementar $\overline{F_B}(z)$ para uma entrada $x(n)$ pode ser obtida através de:

$$\bar{y}(n) = x(n) - y(n),$$

onde

$$y(n) = (f_b * x)(n)$$

é a própria a saída do filtro $F_B(z)$.

Em lugar de realizar o projeto do filtro-protótipo meia-banda por janelamento [2] como em [1], o presente trabalho optou por um projeto *equiripple* de ordem 14 com especificações bem mais exigentes. A resposta de magnitude na freqüência do filtro utilizado pode ser vista na Figura 7.

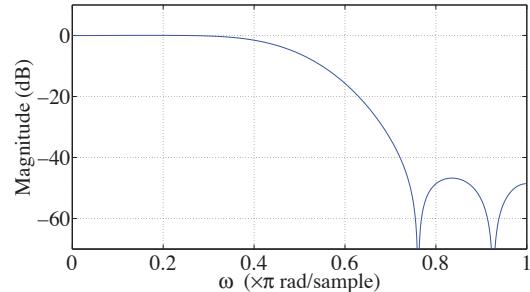


Figura 7: Magnitude da resposta de freqüência do filtro protótipo, $F_B(z)$.

As diversas versões de $F_B(z)$ utilizadas nos diversos ramos da árvore são interpoladas pelos fatores [1] mostrados na Tabela 1.

Tabela 1: Fatores de interpolação do equalizador linear.

L_0	L_1	L_2	L_3	L_4
8	4	2	2	3

É necessário inserir atrasos adequados (τ_{b_2} a τ_{b_8}) nos canais, uma vez que o sinal pode percorrer números desiguais de filtros, e estes podem ter diferentes comprimentos devido às diversas taxas de interpolação. Para a estrutura adotada no projeto, o atraso provocado por um filtro é dado por $l\tau_b$, onde l é o fator de interpolação do filtro e

$$\tau_b = \frac{N_b - 1}{2} \quad (5)$$

é o atraso do filtro-protótipo, suposto de comprimento N_b . Neste projeto, $\tau_b = 70$ amostras, o que implica um atraso total para o equalizador de $\tau_{eq} = 105$ amostras. A Tabela 2 exibe os valores obtidos para os atrasos em cada canal em amostras.

Tabela 2: Comprimento dos atrasos de cada canal em amostras.

τ_{b_2}	τ_{b_3}	τ_{b_4}	τ_{b_5}	τ_{b_6}	τ_{b_7}	τ_{b_8}
28	14	14	7	14	14	28

O número de operações aritméticas (adições e multiplicações reais) realizadas pelo equalizador linear é função do número de operações realizadas pelo filtro protótipo

$$\nu_b = \frac{N_b - 1}{2}. \quad (6)$$

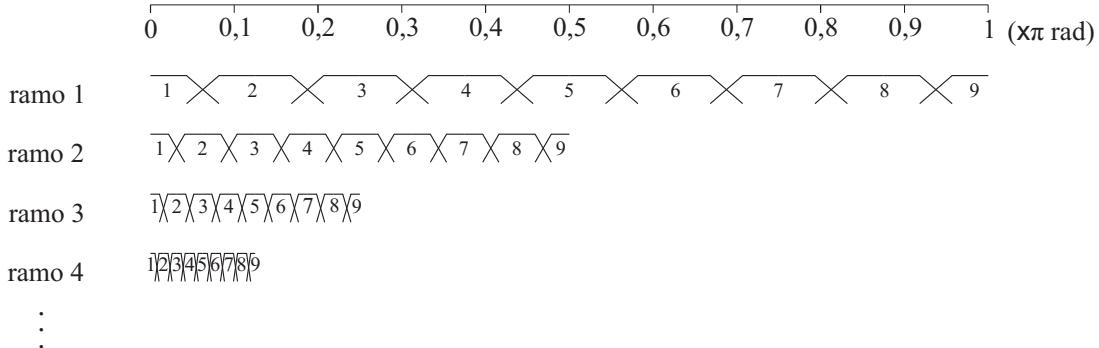


Figura 8: Divisão do eixo da freqüência entre os diferentes canais dos diferentes ramos da estrutura multi-taxa.

O total de operações por amostra do sinal de entrada no equalizador linear é, então, $8\nu_b$. Como $N_b = 15$ para este projeto, o equalizador linear realiza 120 operações aritméticas por amostra do sinal de entrada.

1.2 Estrutura Completa

Combinando-se os resultados obtidos nas duas subseções anteriores, pode-se determinar que:

- os atrasos da estrutura da Figura 1 podem ser vistos na Tabela 3;
- o atraso total do equalizador é de 7700 amostras;
- o número total de operações aritméticas é de 790 por amostra do sinal de entrada.

Tabela 3: Atrasos da estrutura multi-taxa em amostras.

τ_1	τ_2	τ_3	τ_4	τ_5
7595	3675	1715	735	245

A estrutura multi-taxa descrita nas seções anteriores proporciona uma divisão do eixo das freqüências conforme exibe a Figura 8. Como se pode ver, as faixas de operação de equalizadores lineares de ramos diferentes da estrutura se interceptam. Com isso, o ganho de uma determinada faixa pode ser influenciado por mais de um canal, cada um proveniente de um equalizador linear distinto.

A resposta de magnitude na freqüência do equalizador completo quando se escolhe ganho unitário na saída de todos os canais pode ser vista na Figura 9. As respostas dos diferentes canais são somadas, e em particular nas faixas de freqüência onde ocorre superposição. Além de não resultar em ganho unitário, essa superposição dificulta muito a escolha dos ganhos, não só por aumentar excessivamente o número de parâmetros oferecidos para controle do usuário. A modelagem de um *notch* (uma rejeição sintonizada de uma faixa) em baixas freqüências nessas condições pode se tornar um exercício extenuante.

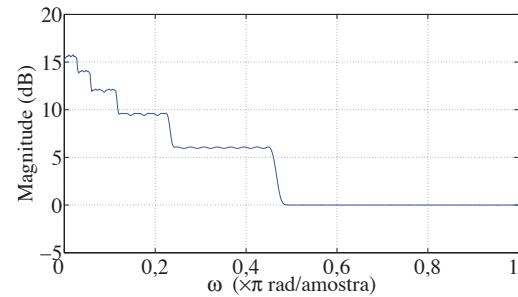


Figura 9: Magnitude da resposta em freqüência da estrutura em multi-taxa utilizando ganho unitário em todos os canais dos equalizadores lineares.

Para contornar este problema, em todos os casos em que ocorria superposição o efeito das bandas superiores foi eliminado pela atribuição de ganho nulo a elas. Assim, são utilizados apenas os canais: de 1 a 6 no primeiro ramo; 4, 5 e 6 nos quatro ramos seguintes; e de 4 a 9 no último ramo. Como efeito colateral, perde-se a resposta plana na situação de ganho unitário, uma vez que a complementaridade entre filtros adjacentes de ramos diferentes não é garantida pelo projeto original. Este é um problema estrutural, que poderia ser resolvido modificando-se a filosofia de projeto dos filtros que compõem a estrutura — o que ultrapassa o objetivo deste trabalho.

Por outro lado, o fato de o *plug-in* implementado neste trabalho mostrar visualmente a resposta real do filtro resultante permite que o usuário tente moderar o problema acima no processo de modelagem da resposta do equalizador.

2 IMPLEMENTAÇÃO

A implementação do equalizador multi-taxa apresentado na Seção anterior foi feita na linguagem C++ [3]. Seu objetivo foi criar um conjunto de classes de C++ que realizassem a filtragem seqüencial do sinal no tempo, aproveitando ao máximo a eficiência computacional da estrutura. Essas classes serão utilizadas para a criação do *plug-in* VST.

A primeira classe criada implementa os filtros de meia-banda interpolados utilizados no equalizador. Esses filtros possuem uma estrutura muito particular, com a maior parte dos coeficientes iguais a zero. O projeto aqui realizado é uma adaptação da classe criada em [4]. Os filtros foram implementados na forma direta, multiplicando-se a saída da memória pelo seu respectivo coeficiente e somando os resultados, apenas para os coeficientes não-nulos e não-unitários. Para tal, foi necessária uma estrutura de dados que levasse em conta o posicionamento dos zeros, de modo a acessar a memória diretamente (sem precisar percorrer toda a estrutura), além de poder ‘deslocar a memória’ alterando apenas um elemento.

Foi criada uma lista encadeada circular modificada, esquematizada na Figura 10, de modo a atender essas especificações. Cada elemento da lista contém um ponteiro para o seu antecessor, e mais quatro ponteiros para os elementos situados a 2^{L-i} amostras e a 2^{L-i-1} amostras, tanto à sua esquerda quanto à sua direita. Essas distâncias correspondem aos elementos não-nulos, sendo que para os coeficientes $h_B(1)$ e $h_B(-1)$ a distância é a metade. Um ponteiro sempre é mantido no elemento da memória correspondente ao coeficiente em z^0 e outro no elemento correspondente à amostra mais recente. Dessa maneira, a lista pode ser deslocada com apenas uma troca de ponteiros, e os elementos não-nulos podem ser acessados diretamente.

A classe `hBandFilter` utiliza essa lista encadeada para implementar a memória do filtro. Os coeficientes não-nulos e não-unitários são armazenados num vetor estático, membro da classe.

Os dois principais métodos da `hBandFilter` são o `set_param`, no qual são passados o fator de interpolação e os coeficientes do filtro; e o `filter`, que recebe um valor real correspondente à entrada e retorna a amostra filtrada por ele e pelo seu complementar.

Uma classe auxiliar chamada `hMyStack` foi criada para implementar os atrasos. Esta classe possui uma pilha do tipo `vector` da biblioteca padrão de C++ como membro. Esta pilha possui comprimento igual ao comprimento do atraso e possui como propriedade o tempo constante de inserção de um elemento no topo da pilha e de leitura de um elemento do final da pilha. A classe possui um método para cada uma dessas duas ações.

A classe `hBandTree` implementa o equalizador linear de cada ramo. Esta classe possui um vetor de 8 objetos da classe `hBandFilter` que implementam os 8 filtros de meia-banda da estrutura. O outro membro desta classe é um vetor de 7 objetos da classe `hMyStack` que representam os 7 atrasos da estrutura. Todos os objetos são inicializados no construtor da classe, sendo os coeficientes do filtro-protótipo lidos de um arquivo externo ao programa. Os atrasos inseridos em cada canal são calculados em função do número de coeficientes do filtro-protótipo durante a execução do construtor da classe. Um método chamado `getDelay` retorna o atraso total do equalizador linear. O método `filter`

recebe como entrada a amostra a ser filtrada e retorna uma amostra filtrada (e convenientemente atrasada).

Uma classe auxiliar para os filtros *anti-alias* e anti-imagem também foi criada, e é chamada `hFilter`. Esta classe implementa filtros FIR de ordem par simétricos na forma direta. Assim como a `hBandFilter`, esta classe possui dois métodos: um chamado `set_param`, que recebe como entrada os coeficientes do filtro e seu comprimento; e outro chamado `filter`, que recebe como entrada a amostra a ser filtrada e retorna uma amostra filtrada.

A estrutura multi-taxa foi implementada na classe `hEqualizerFull`. Esta classe possui um vetor contendo objetos da classe `hBandTree` representando os equalizadores dos ramos, um vetor contendo os atrasos (através da classe `hMyStack` em cada ramo) e dois vetores contendo os filtros anti-imagem e *anti-alias* em cada ramo. Todos os objetos são inicializados no construtor da classe, e os coeficientes dos filtros *anti-alias* e anti-imagem são lidos de um arquivo externo ao programa. Os atrasos de cada ramo são calculados em função do número de coeficientes dos filtros *anti-alias* e anti-imagem e do atraso do equalizador (obtido através do método `getDelay` da classe `hBandTree`). O número de ramos da estrutura é escolhido em tempo de compilação.

O método `filter` da classe `hEqualizerFull` implementa a filtragem, recebendo uma amostra do sinal de entrada e retornando uma amostra filtrada. A implementação dos decimadores e interpoladores foi realizada observando-se a estrutura multi-taxa. O aninhamento dos decimadores provoca uma periodicidade na chamada das funções `filter` dos equalizadores de cada ramo. O k -ésimo ramo, por exemplo, é executado com um período igual a 2^{k-1} . Considerando-se todos os K ramos, a estrutura possui um período total igual a 2^{K-1} . Esse período foi utilizado no método `filter` da classe `hEqualizerFull` para se saber quais ramos estão ativos para uma determinada chamada do método. Um vetor de comprimento 2^{K-1} é criado, do qual cada elemento indica quais ramos estão ativos. Por exemplo, para $K = 3$ seria gerado o vetor $[3, 1, 2, 1]$, indicando que na primeira chamada da função os 3 ramos estão ativos, na segunda chamada apenas o primeiro ramo está ativo e assim por diante. A utilização deste vetor se dá através de uma variável auxiliar que aponta para a posição atual dentro do vetor. Esta variável é incrementada (módulo 2^{K-1}) toda vez que o método `filter` é chamado, fazendo com que o número de ramos ativos seja modificado. Dessa maneira a interpolação e a decimação são feitas de uma maneira eficiente, sem a necessidade de se armazenar o estado de cada ramo. Além disso, o vetor indicando quais são os ramos ativos é calculado durante o construtor da classe; as únicas operações necessárias durante a execução do método `filter` são uma soma e uma leitura da memória.

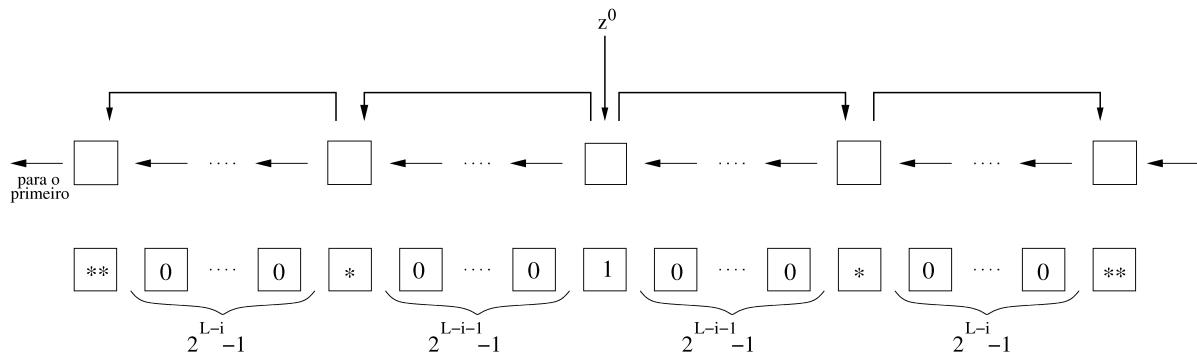


Figura 10: Diagrama da organização da memória de um sub-filtro do nível i , mostrando sua correspondência com os coeficientes do filtro (abaixo). As casas marcadas com asterisco indicam os coeficientes não-nulos. As setas indicam os ponteiros. Figura baseada em [4].

3 PLUG-IN VST

A grande maioria dos efeitos comerciais de áudio digital é vendida hoje na forma de *plug-ins*. Para usá-los, o usuário precisa dispor de um programa *host*, normalmente chamado de DAW (*Digital Audio Workstation*). As DAWs oferecem ferramentas básicas de edição de áudio (multi-trilhas, visualização da forma de onda, alteração de ganhos etc.), enquanto que os *plug-ins* se encarregam dos efeitos mais elaborados (reverberação, equalização, sintetizadores etc.). A comunicação entre *plug-in* e *host* se dá através de algum protocolo previamente estabelecido, que basicamente regula procedimentos como a troca de amostras de áudio e a passagem dos valores dos parâmetros.

Dos padrões existentes hoje no mercado, o VST se mostra mais atraente por dois motivos:

1. Seu SDK (do inglês *Software Development Kit*, conjunto de rotinas necessárias para a implementação do *plug-in*) pode ser obtido gratuitamente no site do fabricante [5];
2. O padrão pode ser usado nas plataformas Mac OSX e Windows.

Este projeto foi implementado na forma de *plug-in* VST, com uma interface gráfica que recebe os ganhos do usuário e apresenta a resposta em freqüência resultante do filtro. O processamento é feito *online*, ou seja, a equalização é feita enquanto a DAW toca as amostras de áudio.

Uma das grandes vantagens da programação em VST é a possibilidade de utilização de uma arquitetura em camadas. Desta maneira, foi possível desacoplar a implementação do *plug-in* da implementação do equalizador.

A função principal de um *plug-in* VST chama-se *processReplacing*: é ela quem recebe as amostras e as devolve ao *host*. No caso deste projeto, ela passa as amostras a um objeto da classe *hEqualizerFull* junto com os ganhos recebidos pela interface gráfica. O objeto, por sua vez, devolve-lhe as amostras processadas, para que possam então ser enviadas de volta ao *host*.

Aí reside o atraso inerente ao sistema, provocado pelo filtro: enquanto o *host* fornece a amostra atual, recebe aquela que estava na saída do filtro, atrasada pelo processamento (de cerca de 7700 amostras, neste projeto).

Um dos desafios na implementação do *plug-in* foi a interface com o usuário.

Na configuração original, mantendo as interseções entre diferentes canais do equalizador, seria necessário apresentar o controle para os 54 ganhos da estrutura (9 ganhos para cada um dos 6 ramos), o que por si só já era uma dificuldade. Quando se decidiu evitar a superposição entre bandas, o número de *faders* a apresentar se reduziu a 24; todos puderam ser dispostos na mesma tela, de maneira bem intuitiva.

A resposta de magnitude do filtro é apresentada de forma realística para o usuário. A cada mudança no valor dos ganhos, o *plug-in* instancia um objeto da classe *hEqualizerFull*, fornecendo um impulso e a configuração de ganhos. Com a resposta ao impulso em mãos, calcula-se a resposta na freqüência por uma FFT (do inglês *Fast Fourier Transform*) de 16384 pontos. O cálculo é feito através da biblioteca FFTW [6]. A interface gerada pode ser vista na Figura 11.

3.1 Características do *Plug-in*

O *plug-in* gerado opera sobre sinais mono com taxa de amostragem igual a 44,1 kHz. Nesta taxa de amostragem, o atraso do equalizador é equivalente a, aproximadamente, 17 ms. A distribuição dos canais é mostrada na Tabela 4. A faixa dinâmica de operação é de 40 dB, o usuário pode aplicar um ganho ou uma atenuação de até 20 dB em cada um dos canais.

O desempenho do *plug-in* atendeu às expectativas, realizando as operações *online*. A sensibilidade dos *faders* não é instantânea, pois a cada mudança de valor nos ganhos é necessário calcular novamente a resposta do filtro. Contudo, a visualização da resposta em freqüência real do equalizador foi apontada como um fator positivo em relação aos equalizadores normalmente disponíveis.



Figura 11: Interface gráfica do FlawQ.

Tabela 4: Localização dos centro dos canais.

Canal	Centro (Hz)	Canal	Centro (Hz)
1	22	13	2067
2	86	14	2756
3	172	15	3453
4	258	16	4134
5	344	17	5512
6	430	18	6809
7	516	19	8270
8	689	20	11025
9	861	21	13781
10	1034	22	16538
11	1378	23	19294
12	1723	24	21361

4 CONCLUSÃO

Este artigo apresentou o projeto de um equalizador gráfico digital. Um equalizador multi-taxa com estrutura em árvore da literatura foi modificado e então implementado em C++ na forma de *plug-in* VST, com uma interface gráfica que permite a visualização da resposta em freqüência real do filtro. Eliminaram-se as superposições entre canais da arquitetura original pelo bem da usabilidade, ao custo da perda da complementaridade entre filtros adjacentes de oitavas distintas. A estrutura dos filtros em árvore permitiu uma implementação rápida que garantiu o funcionamento *on-line* do *plug-in*.

A principal meta deste trabalho foi realizar um estudo preliminar do emprego de uma arquitetura rápida no projeto de equalizadores digitais com interface amigável e alto desempenho. A sua continuação deve passar por uma modificação no projeto da estrutura multi-taxa, de modo a resgatar a complementaridade dos canais adjacentes de oitavas distintas. Pode-se,

ainda, tentar combinar um banco de filtros altamente seletivo como se utilizou em [4] com a estrutura rápida deste artigo.

5 AGRADECIMENTOS

Os autores agradecem o apoio financeiro do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior e da FAPERJ (Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro).

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Ragnar Hergum, “A low complexity, linear phase graphic equalizer,” in *Presented at the 85th AES Convention*, Los Angeles, USA, November 1988, Preprint 4706.
- [2] Paulo S. R. Diniz, Eduardo A. B. da Silva, and Sergio L. Netto, *Digital Signal Processing: System Analysis and Design*, Cambridge, 2002.
- [3] Bjarne Stroustrup, *A Linguagem de Programação C++*, Bookman, 3rd edition, 2000.
- [4] Leonardo O. Nunes, Alan F. Tygel, Rafael A. de Jesus, and Luiz Wagner P. Biscainho, “Equalizador gráfico digital de alta seletividade em VST,” in *Anais do IV Congresso de Engenharia de Áudio*, São Paulo, Brazil, Maio 2006, pp. 47–52.
- [5] “Steinberg media technologies (vst);” http://www.steinberg.net/324_1.html.
- [6] M. Frigo and S. G. Johnson, “The design and implementation of FFTW3,” *Proceedings of the IEEE*, vol. 93, no. 2, pp. 216–231, February 2005.



Sociedade de Engenharia de Áudio Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Avaliação Subjetiva de Qualidade de Áudio: Fala vs. Música

Daniel S. Gerscovich¹ e Luiz W. P. Biscainho¹

¹PEE/COPPE, Universidade Federal do Rio de Janeiro,
Caixa Postal 68504, 21945-970 Rio de Janeiro, RJ, Brasil
danielgerscovich@gmail.com, wagner@lps.ufrj.br

RESUMO

Avaliadores objetivos de qualidade de áudio são aferidos pela avaliação subjetiva correspondente. Este trabalho investiga se um ouvinte quantifica diferentemente fala em seu idioma e áudio de outra natureza. Gravações de fala em português e de música foram codificadas com perdas, formando-se pares fala-música para os quais um avaliador objetivo consagrado desse a mesma nota, varrendo toda a escala. A avaliação auditiva dos mesmos sinais sugere que os indivíduos são menos rigorosos com sinais de fala que com os demais.

0 INTRODUÇÃO

O grande lema da engenharia é buscar a melhor produção com o menor custo possível. Na área das comunicações, um dos grandes desafios é fazer com que a maior quantidade possível de informação seja manipulada com o menor custo. O avanço das técnicas de processamento digital de sinais permitiu que o custo da informação (em termos da quantidade de bits necessários para representá-la) fosse bastante reduzido. Diversos métodos de compressão foram desenvolvidos com tal objetivo; no entanto, se o procedimento não for criterioso, pode reduzir a qualidade do sinal a ponto de inviabilizar sua utilização.

Em processamento digital de fala, especificamente, tem-se o exemplo da rede de telefonia convencional analógica (também conhecida como PSTN, do inglês *Public Switched Telephone Network*). Estabelecida mundialmente desde o início do séc. XX, ela definiu um paradigma de qualidade para a conversação à distância. Após a digitalização da rede na década de 70, suas taxas de transmissão ainda permanecem bastante elevadas (64 kbps). Novas aplicações como telefonia móvel, teleconferência e VoIP (voz sobre IP) precisam operar a taxas menores (2 kbps a 16 kbps), porém com qualidade ao

menos similar à PSTN. Para isto, é necessária uma compressão acentuada do sinal, sem distorcê-lo perceptivelmente. Os codificadores de fala mais eficientes se baseiam em modelos da produção de voz.

Já em processamento digital de áudio, um exemplo atual é o compartilhamento de músicas via internet. Em busca de rapidez na transferência, tornam-se necessários codificadores capazes de comprimir a informação e consequentemente diminuir o tamanho dos arquivos. O sinal de música original, ao ser codificado, sofre modificações em suas características. No entanto, as perdas que sua compressão implica não podem tornar seu conteúdo esteticamente inaproveitável. Os codificadores de música (áudio em geral) mais eficientes são perceptuais, isto é, baseados em Psicoacústica. Por precisarem lidar com sinais de extrema variabilidade e ainda preservar sua fidelidade, tais codificadores tendem a atingir taxas de compressão menores que os de fala.

Evidencia-se, nesse contexto, a necessidade de realizar de forma confiável e sistemática a avaliação de qualidade dos sinais de áudio que sofreram codificação com perdas. No entanto, a principal dificuldade envolvida nisso é o fato de ‘qualidade’ ser uma medida muito abstrata.

Pode-se dizer que áudio se refere a qualquer fenômeno perceptível à audição humana, enquanto que a voz falada é um conceito específico de comunicação e um subconjunto do áudio. Para sinais de áudio, em geral, o principal atributo de qualidade é a reprodução fiel do sinal como um todo; já para sinais de fala, um novo parâmetro se agrega ao primeiro: a inteligibilidade. Esta é função de diversos fatores, como o idioma e a cultura do par falante/ouvinte. E ao final, um sinal de baixa fidelidade (e qualidade, portanto) pode ser considerado altamente inteligível.

Atualmente, a tecnologia já admite que aplicações de fala possam usufruir do mesmo grau de fidelidade que música e áudio em geral, se possível aproveitando todo o espectro audível (no mínimo até 20 kHz). Isso aponta para a necessidade de abordagens genéricas robustas para o problema de avaliação de qualidade de áudio.

Neste trabalho, deseja-se investigar o efeito da expectativa de inteligibilidade na avaliação de qualidade de sinais de fala, quando comparada à avaliação de sinais de áudio¹ com fidelidade comparável.

Na Seção 1 desse artigo, apresentam-se métodos subjetivos e objetivos de qualidade de interesse no trabalho. Na Seção 2, descrevem-se as bases de dados montadas para o experimento descrito na Seção 3. Os resultados são analisados na Seção 4.

1 MÉTODOS DE AVALIAÇÃO DE QUALIDADE

Na última década, foram elaborados diversos métodos para avaliar sistematicamente a qualidade de sinais de fala e áudio. Estes métodos podem ser agrupados em dois subconjuntos: métodos subjetivos e métodos objetivos.

1.1 Métodos Subjetivos

Os chamados métodos subjetivos realizam a avaliação de qualidade da forma mais natural: pela audição dos sinais por pessoas. Para atingirem sua meta, entretanto, requerem a definição cuidadosa do procedimento de teste, e a opinião de um grande número de indivíduos que o realizem sob condições idênticas. Isso os tornam caros e demorados.

A ITU (*International Telecommunication Union*) define nas Recomendações P.800 [1] e P.830 [2] as condições e procedimentos necessários para a realização de testes utilizando métodos subjetivos.

Num teste convencional, cada ouvinte avalia a qualidade percebida e lhe confere uma nota que varia de 1 a 5. Em seguida, a média das notas dadas pelos ouvintes é calculada e tabelada, segundo a escala MOS (*Mean Opinion Score*).

O problema deste método, denominado ACR (*Absolute Category Rating*), é a variação da nota dada para um mesmo sinal por diferentes ouvintes, sob as mesmas condições. Para conseguir maior precisão no resultado da tabela MOS, pode-se inserir um sinal de referência x_{ref} , que representa o sinal a ser avaliado x_{test} sem degradação (idealmente, com MOS igual a 5). Logo, o sinal passa a ser avaliado não mais pela sua qualidade, mas sim pela sua degradação. Este método é denominado DCR (*Degradation Category Rating*). Neste caso, as notas

¹ A partir daqui, por simplicidade, o que se chamará de sinais de áudio excluirá os sinais de fala.

variam entre 0 e -4 na escala SDG (*Subjective Difference Grade*). Desta forma, o caráter pessoal e abstrato da percepção de cada ouvinte é reduzido.

A Tabela 1 relaciona os dois métodos mencionados anteriormente. Um mapeamento simples entre as duas escalas torna possível a comparação de resultados utilizando os diferentes métodos.

Tabela 1 – Escalas MOS e SDG/ODG.

MOS	Qualidade percebida	SDG /ODG	Grau de Degradação
5	Excelente	0	Imperceptível
4	Bom	-1	Perceptível, porém aceitável
3	Regular	-2	Perceptível e desagradável
2	Ruim	-3	Perceptível e irritante
1	Péssimo	-4	Inaceitável

Métodos subjetivos podem ser usados para a avaliação de sinais de áudio e sinais de fala, indistintamente. Hoje são utilizados principalmente para geração de dados comparativos que permitam validar os modernos métodos objetivos.

1.2 Métodos Objetivos

Com o intuito de diminuir custos e aumentar a velocidade na obtenção de dados, foram propostos diversos métodos objetivos, que utilizam diferentes modelos para prever a qualidade subjetiva dos sinais de fala e áudio. Além de apresentarem baixo custo, esses métodos podem permitir monitoração contínua e em tempo real do sinal de interesse. Pode-se dizer que os métodos objetivos fornecem resultados ‘immediatos’, se comparados com os subjetivos.

Os primeiros algoritmos avaliadores de qualidade objetiva de sinais de fala e áudio utilizavam um sinal de referência x_{ref} , além do sinal a ser avaliado x_{test} ([3],[4]). Estes métodos ficaram conhecidos como métodos intrusivos, pois era preciso intervir diretamente na entrada do sistema para extrair as informações de x_{ref} necessárias para gerar a nota de x_{test} . A principal limitação do método intrusivo é a impossibilidade de avaliar problemas no canal de transmissão, pois é necessário extrair informações nas duas extremidades do canal. Outro fator negativo é a duplicação da quantidade de informação, já que é preciso fazer a análise de dois sinais para cada sinal avaliado. Na literatura, métodos intrusivos também são conhecidos como *double-ended* ou métodos com sinal de referência.

Posteriormente, surgiu uma segunda família de métodos de avaliação objetiva de qualidade de fala e áudio: os métodos não-intrusivos ([5],[6]). Neste caso, o avaliador processa apenas o sinal de interesse, x_{test} . Isto permite que se avalie o sinal sem interferir no sistema. Métodos não-intrusivos também são conhecidos como *single-ended* ou métodos sem referência. A Figura 1 ilustra ambos os métodos².

² Rigorosamente, é possível distinguir o conceito de intrusão/não-intrusão do de referência/não-referência. Aqui, seguimos a tendência da literatura, que é confundir as duas classificações.

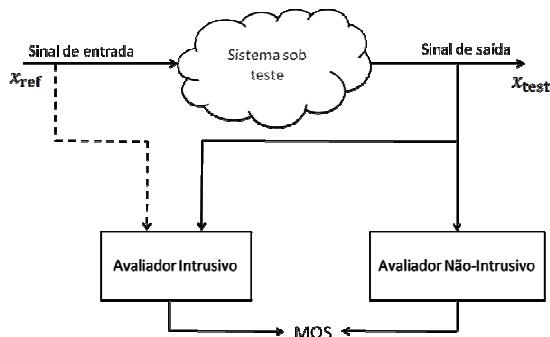


Figura 1 – Modelo intrusivo X Modelo não-intrusivo.

No entanto, os avaliadores objetivos disponíveis no mercado são específicos apenas para determinadas situações. Dentre os que sofreram descrição formal, podem-se citar: PESQ [3], avaliador intrusivo de sinais de fala em banda de telefonia (3,1 kHz – com uma versão estendida até 7 kHz); PEAQ [4], avaliador intrusivo de sinais de áudio em banda completa (24 kHz); P.563 [5], avaliador não-intrusivo de sinais de fala em banda de telefonia.

Considerando a diversidade de aplicações possíveis e a consequente diversidade de defeitos a elas inerentes, ainda há muito que desenvolver nesse campo. Em particular, a implementação de um avaliador genérico, capaz de avaliar sinais de fala e áudio de alta fidelidade ainda é um desafio para pesquisadores. Parte da dificuldade decorre da diferença envolvida na percepção da qualidade dessas duas famílias de sinais.

A fim de fazer uma investigação desse efeito, este trabalho adota o PEAQ como avaliador objetivo.

1.3 Recomendação BS.1387 - PEAQ

A Recomendação BS.1387, conhecida como PEAQ (do inglês *Perceptual Evaluation of Audio Quality*) e aprovada pela ITU-R em 2001, descreve um método para a avaliação objetiva de qualidade percebida de áudio de alta fidelidade. Este método apóia-se em uma série de medidas psicoacústicas e princípios cognitivos para determinar uma medida da diferença de qualidade entre o sinal sob teste (x_{test}) e um sinal de referência (x_{ref}). Um estudo detalhado sobre Psicoacústica pode ser encontrado em [7].

Sua principal aplicação é na avaliação de quanto codificadores com perdas degradam um sinal de referência. Segundo a Recomendação, o algoritmo PEAQ também poderia ser usado na avaliação de sinais contaminados com ruído.

Para permitir a avaliação pelo algoritmo PEAQ, x_{ref} e x_{test} devem estar rigorosamente sincronizados e obedecer as condições da Tabela 2.

Tabela 2 – Requisitos dos sinais de entrada do PEAQ.

Freqüência de amostragem	48000 Hz
Resolução de amplitude	PCM linear a 16 bits
Tamanho do sinal	10,0 a 20,0 s
Nível médio do sinal	92 dB _{SPL}

A descrição detalhada do algoritmo pode ser encontrada em [4] e [8]. Diversos processamentos de fundamentação psicoacústica, tais como a separação em bandas críticas, a adaptação de nível, o mascaramento na freqüência e no tempo e a aplicação da curva de audibilidade, são realizados a fim de preparar as entradas para o cálculo das chamadas MOVs (*Model Output Variables*). Estas são as entradas de uma rede neural treinada para gerar um resultado simples que corresponde à diferença percebida entre x_{ref} e x_{test} , e quantifica a degradação do sinal segundo uma ODG (*Objective Difference Grade*), definida na Tabela 1.

Tendo em vista que em sua concepção original o PEAQ foi treinado com sinais de áudio e fala, ele é potencialmente capaz de avaliar corretamente a perda de fidelidade para as duas famílias de sinais. Por isso, tal avaliador foi adotado neste trabalho para fornecer a avaliação objetiva ‘de referência’ contra a qual se compararão os testes subjetivos específicos.

2 BASE DE DADOS

A base de dados consiste em um ponto de extrema importância para o sucesso ou insucesso de um estudo de avaliação de qualidade. Neste trabalho, a base de dados pode ser dividida em dois conjuntos: o conjunto dos sinais de áudio e o conjunto dos sinais de fala.

2.1 Base de áudio

Uma base de sinais de áudio, em termos gerais, envolve categorias distintas de sinais (por exemplo, ruído, voz, música etc). Definida uma categoria específica conforme a aplicação, no nosso caso sinais de música, ainda assim é preciso garantir generalidade suficiente em termos tanto de eventos temporais e freqüenciais quanto de aspectos perceptivos. Um sinal típico deve ter duração entre 10 e 20 segundos.

A base de áudio utilizada foi organizada pelo Grupo de Processamento de Áudio do Laboratório de Processamento de Sinais da UFRJ em 2008. Possui um total de 25 sinais, com aproximadamente 10 segundos cada, armazenados em formato WAVE mono, com freqüência de amostragem de 44,1 kHz e 16 bits de resolução. A Tabela 3 descreve brevemente os sinais utilizados. Tentou-se abranger uma gama bastante ampla de características, com o intuito de evitar eventuais polarizações na avaliação objetiva/subjetiva por especificidades dos sinais de áudio.

2.2 Base de fala

Para sinais de fala, normalmente o material deve ser formado por frases foneticamente balanceadas. A utilização de frases curtas dificulta a análise do ouvinte, enquanto que a utilização de frases longas pode ser mal avaliada, se o ouvinte passa a usar apenas sua memória recente. Uma frase típica contém duas ou três pequenas sentenças de 2 a 4 segundos cada, totalizando um estímulo entre 6 e 20 segundos.

Para a base de dados de fala, este trabalho emprega apenas frases em português [9], proferidas por falantes cariocas. As frases foram gravadas em um ambiente controlado (estúdio com baixo nível de ruído e RT_{60} de aproximadamente 200 ms), e armazenadas em formato

WAVE mono, com freqüência de amostragem de 48 kHz e 24 bits de resolução.

Tabela 3 – Base de sinais de áudio.

Nome	Gênero
A1	Hip-Hop – cantora e grupo vocal
A2	Vocal – cantora a capella
A3	Tango – cantora e piano
A4	Samba – instrumental com percussão
A5	Orquestra – frase contínua nas cordas
A6	Órgão – solo
A7	Balada – cantor e grupo
A8	Regional – cantor e grupo
A9	Harmônica-de-Vidro – solo
A10	Gafieira – instrumental de metais
A11	Jazz – grupo instrumental
A12	Fusion – Cantora em <i>scat</i> e grupo
A13	Carrilhão – solo
A14	Soft Rock – cantor e grupo
A15	Funky – instrumental
A16	Frevo – cantor e grupo
A17	Saxofone – solista e acompanhamento
A18	Flauta – solo
A19	Cravo – solo
A20	Castanholha – solo
A21	MPB – cantora e violões
A22	Blues – quarteto instrumental
A23	Violão – solo
A24	Aplauso – platéia
A25	Pop-Rock – cantores finlandeses e grupo

Cada arquivo de fala contém 3 frases aleatoriamente escolhidas entre 3 diferentes falantes com características de timbre distintas. Assim como na base de áudio, o objetivo é reduzir ao máximo a dependência de um falante específico no resultado da avaliação. Foram gerados 25 arquivos de fala, chamados de F1 a F25.

3 EXPERIMENTO PROPOSTO

Com o intuito de aferir possíveis diferenças entre as avaliações subjetivas de qualidade de sinais de áudio e fala que se poderia dizer num sentido amplo guardarem a mesma ‘fidelidade’ ao original, adotou-se a estratégia descrita a seguir.

3.1 Preparação dos sinais de teste e realização dos testes objetivos

Inicialmente, elegeu-se um avaliador objetivo de qualidade de áudio não-especializado: o PEAQ, método intrusivo para operação a 48 kHz. Reamostraram-se os 25 sinais da base de áudio para 48 kHz; então, associando-os aos sinais da base de fala, formou-se um conjunto de 50 sinais de referência.

Cada um dos 50 arquivos foi codificado em 12 taxas diferentes do padrão de codificação MPEG-1 Audio Layer III [10] e extensões (o popular MP3): 8, 16, 24, 32, 40, 48,

56, 64, 80, 96, 112 e 128 kbps. Posteriormente, os arquivos foram decodificados e rearmazenados no formato WAVE mono a 48 kHz. Ao final do processo, obtiveram-se 600 arquivos degradados.

Em seguida, realizou-se a medida de qualidade de cada um dos sinais de teste pelo PEAQ. O objetivo nesse ponto era gerar notas que cobrissem uniformemente toda a extensão da escala de saída do PEAQ, ou seja, de -4 a 0 na escala ODG. A Figura 2 mostra a dispersão das notas atribuídas a todos os sinais de entrada.

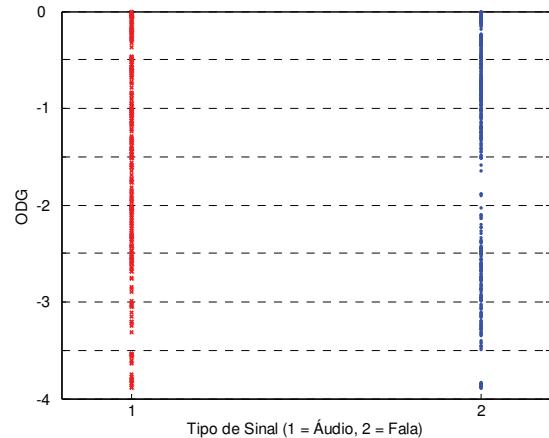


Figura 2 – ODGs atribuídas a todos os sinais da base de dados.

3.2 Redução dos sinais de teste

A etapa seguinte consistiu na escolha de pares de sinais fala/áudio que houvessem recebido do PEAQ a mesma nota na escala da Figura 2. Como o algoritmo PEAQ gera sinais com precisão de 4 casas decimais, considerou-se que um sinal de fala x_{fala} teria a mesma nota que o sinal de áudio $x_{áudio}$ se:

$$|PEAQ(x_{fala}) - PEAQ(x_{áudio})| \leq 10^{-2}. \quad (1)$$

Os pares que satisfizessem esta condição foram ordenados ao longo da escala ODG em passos de 0,125. Nota-se que, tanto para sinais de áudio quanto para sinais de fala, alguns trechos da escala ODG não foram mapeados, conforme ilustra em maior detalhe a Figura 3. Vale notar que mesmo utilizando 600 frases gravadas para multiplicar o número de sinais de fala disponíveis, não foi possível preencher as lacunas na escala para esta família.

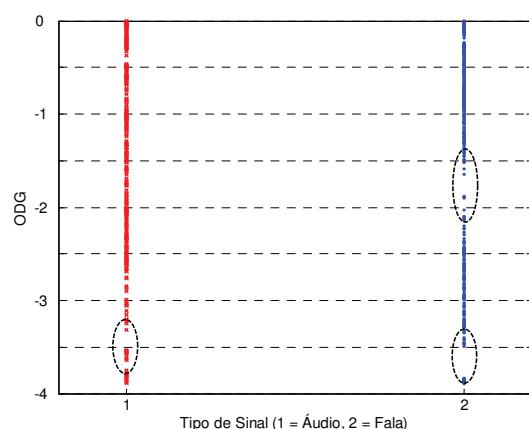


Figura 3 – Trechos não mapeados pelo PEAQ.

Este fato pode ser explicado por diversos motivos: número limitado de sinais da base de dados, no caso de áudio; sinais de características muito próximas gravados sob mesmas condições, no caso de fala; número limitado de taxas de compressão adotadas pelo padrão MPEG; aspectos internos ao cálculo do PEAQ.

Os pares não mapeados nos dois conjuntos foram omitidos. Ao final do processo, obteve-se um total de 21 pares, totalizando 42 arquivos que serão utilizados nos testes subjetivos. Estes pares conseguem representar significativamente os sinais da base de dados, conforme apresentado na Tabela 4 e ilustrado na Figura 4.

Tabela 4 – Pares de sinais de áudio e fala selecionados.

Par	ODG	Arquivo	Taxa (kbps)	Arquivo	Taxa (kbps)
1	-3,875	A3	8	F3	8
2	-3,250	A18	8	F5	16
3	-3,000	A6	32	F6	24
4	-2,875	A20	16	F21	16
5	-2,750	A2	16	F24	24
6	-2,625	A7	16	F6	32
7	-2,500	A7	32	F18	24
8	-2,375	A9	16	F21	24
9	-2,125	A10	24	F10	32
10	-1,500	A14	40	F5	40
11	-1,250	A19	40	F9	40
12	-1,125	A3	40	F10	40
13	-1,000	A6	48	F4	64
14	-0,875	A20	56	F17	48
15	-0,750	A23	56	F14	48
16	-0,625	A3	80	F20	56
17	-0,50	A12	56	F14	56
18	-0,375	A7	96	F11	80
19	-0,250	A2	96	F7	80
20	-0,125	A19	96	F4	96
21	0,000	A13	112	F1	112

A Tabela 4 expõe um aspecto importante: considerando a fundamentação perceptual tanto do PEAQ quanto do MP3, não se espera que as ODGs fornecidas pelo PEAQ sejam proporcionais à taxa de compressão.

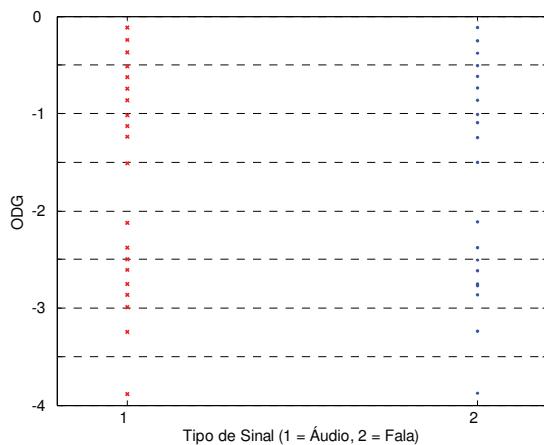


Figura 4 – Distribuição após seleção dos 21 pares de sinais de teste.

3.3 Realização dos testes subjetivos

A última etapa do experimento consistiu na realização de testes subjetivos. Conforme padronizado em [1], realizaram-se testes DCR, em que 24 indivíduos avaliaram a degradação percebida nos 42 sinais de teste (21 sinais de fala e 21 sinais de áudio), comparando-os com um sinal de referência.

A avaliação subjetiva foi realizada por ouvintes não-experientes (estudantes universitários, majoritariamente) equipados com *headphone* Sennheiser HD280®, em um ambiente com baixo nível de ruído. Utilizou-se o *software Subjective Test*, desenvolvido em MATLAB® pelo Grupo de Processamento de Áudio da UFRJ, para auxiliar a aquisição dos resultados. A nota subjetiva é dada pela média das notas atribuídas pelos ouvintes para cada um dos 42 sinais.

A Figura 5 ilustra o resultado dos testes subjetivos. O eixo das abscissas ordena o conjunto de arquivos utilizados no teste em qualidade crescente, segundo o PEAQ.

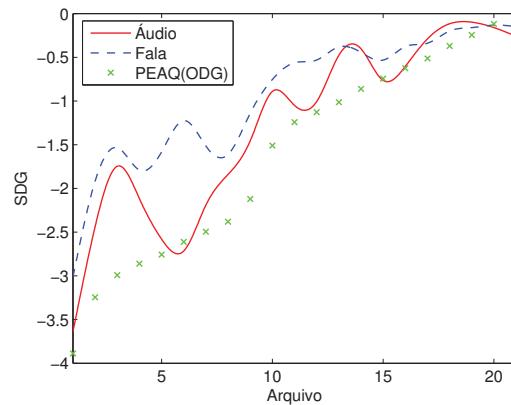


Figura 5 – Comparação áudio x fala: resultados dos testes subjetivos.

Para facilitar a comparação entre os resultados subjetivos, geraram-se duas visualizações alternativas. A Figura 6 ordena os sinais na ordem crescente de qualidade atribuída subjetivamente aos sinais de áudio; já a Figura 7 os ordena segundo as notas atribuídas aos sinais de fala.

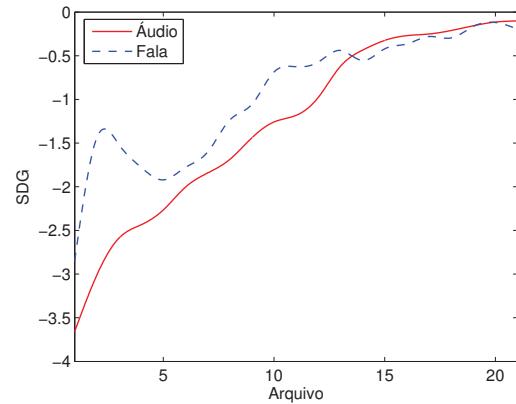


Figura 6 – Comparação áudio x fala: resultados dos testes subjetivos (áudio ordenado crescentemente).

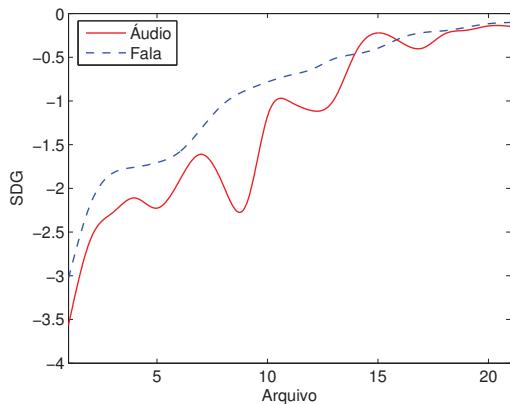


Figura 7 – Comparação áudio x fala: resultados dos testes subjetivos (fala ordenada crescentemente).

Como forma de aferir a coerência entre os resultados dos testes objetivos fornecidos pelo PEAQ e os dos testes subjetivos realizados neste trabalho, apresenta-se a seguir o coeficiente de correlação entre ambos, estimado pela fórmula de Pearson:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}}, \quad (2)$$

onde x_i e y_i são respectivamente os valores das medidas das variáveis aleatórias x e y na condição i ; e \bar{x} e \bar{y} são as médias de todos os x_i e y_i , respectivamente.

No caso em questão, calculam-se os coeficientes de correlação para cada subconjunto de arquivos de sinais de áudio e fala nas diferentes taxas de compressão, ou seja: para cada par sinal/taxa i , x_i representa a nota média atribuída pelo teste subjetivo e y_i representa a nota atribuída pelo teste objetivo. Quanto maior o valor de $|r| < 1$, melhor será a acurácia do avaliador objetivo.

A Tabela 5 apresenta os coeficientes de correlação encontrados no experimento, altos o suficiente para validarem o experimento.

Tabela 5 – Coeficientes de correlação entre os resultados obtidos

		Resultados Objetivos	
		Áudio	Fala
Resultados Subjetivos	Áudio	0,9180	0,9257
	Fala	0,9214	0,9263

4 CONCLUSÕES

A partir dos gráficos, é possível observar alguns pontos relevantes. O primeiro é a diferença entre as notas subjetivas dos sinais de áudio e fala para sinais que receberam a mesma nota objetiva. Fica claro que os indivíduos tendem a dar notas maiores para sinais de fala do que para sinais de áudio. Isto pode ser explicado pelo papel preponderante do fator ‘inteligibilidade’ mencionado anteriormente. Mesmo com alta taxa de compressão e consequente degradação da qualidade do sinal, se o ouvinte consegue compreender perfeitamente a informação, tende a dar-se por satisfeita com a qualidade do sinal de fala. Em contrapartida, na avaliação de qualidade de sinais musicais

entram em cena aspectos estéticos, suscitando uma avaliação mais criteriosa. Também se observa que a partir de um certo grau elevado de qualidade, os resultados se confundem.

O segundo ponto é o fato de o PEAQ, em geral, ter atribuído notas inferiores às dos testes subjetivos, tanto para sinais de fala quanto para sinais de áudio, e essa diferença se acentuar para degradações mais fortes. Como o PEAQ se destina a avaliar degradações leves, suas notas para sinais seriamente degradados podem ser pouco confiáveis. Ademais, seu treinamento foi realizado segundo testes subjetivos realizados por ouvintes especializados, possivelmente mais exigentes que os ouvintes que colaboraram no presente trabalho.

Por último, os vales e picos pronunciados em determinadas regiões das figuras podem causar certa surpresa. Embora se possa invocar novamente os argumentos do parágrafo precedente, é interessante notar pelo confronto da Figura 5 com a Tabela 4 que por vezes as ondulações das avaliações subjetivas parecem correlacionadas com a taxa de codificação. É como se a complexa estimativa do PEAQ por vezes acabasse por mascarar os aspectos mais simples da própria percepção que busca modelar.

Uma proposta de continuação deste trabalho seria a realização de novo teste subjetivo com a utilização de sinais de fala em outro idioma que não o nacional, preferencialmente um que todos os participantes do teste desconhecessem, a fim de suprimir o fator ‘inteligibilidade’. A expectativa é que, nesse caso, os resultados para sinais de fala convirjam para os resultados dos sinais de áudio, por apoiar-se a avaliação subjetiva primordialmente na ‘fidelidade’ do sinal degradado ao original.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] ITU-T Recommendation P.800, "Methods for Subjective Determination of Transmission Quality", 1996.
- [2] ITU-T Recommendation P.830, "Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs", 1996.
- [3] ITU-T Recommendation P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs", 2001.
- [4] ITU-R Recommendation BS.1387-1, "Method for objective measurements of perceived audio quality", 2001
- [5] ITU-T Recommendation P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications", 2004.
- [6] V. Grandcharov, D. Y. Zhao, J. Lindlbom, W.B. Kleijn, "Low Complexity, Non-Intrusive Speech Quality Assessment", IEEE Transactions on Speech and Audio Processing, vol.14, pp.1948-1956, 2006.
- [7] E. Zwicker, H. Fastl, "Psycho-acoustics, Facts and Models", Springer Verlag, 1990.
- [8] P. Kabal, "An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality", McGraw, 2002.
- [9] J. A. Moraes, A. Alcaim, J. A. Solewicks, "Freqüência de Ocorrência dos Fones e Listas de Frases

- Foneticamente Balanceadas no Português do Rio de Janeiro”, Revista da Sociedade Brasileira de Telecomunicações, vol. 7, n. 1, pp. 23-41, 1992.
- [10] ISO/IEC 11172-3, “Information technology - Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s - Part 3: Audio”, 1993.

Sessão 4

Análise, classificação e percepção do som
(Sound analysis, classification and perception)



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil

12ª Convenção Nacional da AES Brasil

5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Separação de instrumentos musicais com uma única mistura

Diego Barreto Haddad,¹ Mariane Rembold Petraglia² e Paulo Bulkool Batalheiro³

¹ CEFET, Unidade Descentralizada de Nova Iguaçu, Coordenadoria de Telecomunicações
Nova Iguaçu, RJ, 26041-271, Brasil

² Universidade Federal do Rio de Janeiro, COPPE, Del/Poli
Rio de Janeiro, RJ, 21941-947, Brasil

³ UERJ, Departamento de Engenharia Eletrônica e de Telecomunicações
Rio de Janeiro, RJ, 20559-900, Brasil

diego@pads.ufrj.br, mariane@pads.ufrj.br, bulkool@pads.ufrj.br

RESUMO

Atualmente, há um grande interesse em métodos de separação de fontes, devido as suas inúmeras aplicações. Este artigo trata do caso de apenas uma mistura de instrumentos musicais distintos, condição extrema na qual a maioria dos métodos de separação propostos não se aplica. Propomos um método que recorre a redes neurais e a um conhecimento *a priori* das características harmônicas de instrumentos musicais para efetuar a separação.

0 INTRODUÇÃO

Os métodos de separação de fontes, crescentemente sofisticados e contemplando casos a cada dia mais realistas, encontram em processamento de sinais de áudio aplicação imediata. Comumente, afirma-se que a relação entre o número M de misturas e o número N de fontes nos permite distinguir dois casos de separação que recorrem a técnicas bastante distintas. Estes casos seriam $M \geq N$ (determinado se $M = N$ ou sobredeterminado quando $M > N$, configuração normalmente contemplada pela análise de componentes in-

dependentes [1]) e o caso (mais difícil) indeterminado (quando $M < N$, normalmente tratado via análise de componentes esparsos [2]). Seria mais realista destacar da classificação “indeterminado” o caso extremo $M = 1$, pois mesmo as técnicas que tratam do caso indeterminado em geral não se aplicam à situação - bastante comum - de termos acesso a apenas uma mistura. Isto se deve, em última análise, ao fato de que exploram a diversidade espacial das fontes entre os sensores, a qual inexiste quando há apenas uma mistura. As técnicas atualmente dedicadas à condição $M = 1$ são as menos

bem-sucedidas dentre todas, devido à particular dificuldade desta configuração.

Estas técnicas efetuam uma decomposição das misturas por meio de uma das seguintes estratégias:

- (I) a partir de uma decomposição fixa, estimar a contribuição de cada uma das fontes em cada um dos componentes;
- (II) adaptativamente, encontrar uma decomposição cujos componentes sejam associados basicamente a apenas uma das fontes.

A estratégia (I) utiliza modelos das fontes (por exemplo, um modelo estatístico paramétrico devidamente treinado com trechos das fontes isoladas ou mesmo um recurso à estrutura harmônica dos sinais) para estimar a contribuição de cada fonte num dado componente [3]-[4]. Como é necessário um conhecimento das características das fontes, não podemos denominar estas técnicas de “cegas”.

Já a estratégia (II) não necessita de gravações isoladas das fontes, tampouco de um conhecimento *a priori* acerca de estruturas harmônicas, à medida em que pode recorrer a uma clusterização não supervisionada dos componentes de modo a associá-los, de forma cega (na realidade, com pressupostos bem gerais, como o da continuidade), a uma das fontes [5]. Porém, cabe ressaltar que os métodos para a clusterização supracitada [6],[7] costumam deteriorar bastante os resultados (vide [5]). Um outro exemplo promissor desta estratégia utiliza conhecimentos de psicoacústica e redes neurais de terceira geração (onde os neurônios são chamados de *spiking neurons*) [8].

O explicado acima nos permite concluir que os métodos de separação com apenas uma mistura baseiam-se principalmente na diversidade espectral das fontes (e não na diversidade espacial, como a maioria dos métodos para misturas indeterminadas com $N > M > 1$).

Este artigo apresenta um novo método, fundamentado na estrutura harmônica dos instrumentos musicais, que se utiliza da estratégia (I). Para efeitos de comparação, apresentaremos o método proposto em [3], aqui denominado FGMM, o qual utiliza a mesma estratégia. Contemplaremos aqui o caso com duas fontes.

Da mesma forma que o método proposto em [3], o nosso método necessita de trechos isolados das fontes, para treinamento. Neste aspecto, cumpre lembrar que há métodos de separação de fontes harmônicas que não recorrem a trechos das fontes em separado para treinamento [9],[10]. Estes métodos, não avaliados aqui, constituem uma alternativa promissora, em especial quando não possuímos um banco de dados disponível.

1 O MÉTODO FGMM

Utilizaremos a seguir a nomenclatura de [3]. Se as fontes s_1 e s_2 encontram-se misturadas de forma aditiva, podemos obviamente escrever a mistura x como:

$$x = s_1 + s_2 + n, \quad (1)$$

onde n é considerado ruído branco gaussiano de variância σ_n^2 . Dentro de um paradigma bayesiano, podemos escrever:

$$p(s_1, s_2|x) = \frac{p(s_1, s_2, x)}{p(x)}, \quad (2)$$

$$\frac{p(s_1, s_2, x)}{p(x)} = \frac{p(x|s_1, s_2)p(s_1, s_2)}{p(x)}, \quad (3)$$

$$p(s_1, s_2|x) \propto p(x|s_1, s_2)p_1(s_1)p_2(s_2). \quad (4)$$

A equação 4 foi obtida admitindo a independência das fontes e atentando para o fato de que $p(x)$ é fixo (a mistura é um dado constante). O modelo aditivo da mistura reflete-se na verossimilhança $p(x|s_1, s_2)$ e as informações disponíveis acerca das fontes são introduzidas em $p_1(s_1)$ e $p_2(s_2)$.

No caso de ausência de informações prévias acerca das fontes, uma estimativa de máxima verossimilhança poderia ser adotada. Porém, infelizmente, no contexto indeterminado a estimativa de máxima verossimilhança é inviável porque apresenta soluções múltiplas.

Logo, é necessário modelar as fontes no paradigma bayesiano adotado. O método FGMM admite que a distribuição das fontes (na realidade, como veremos mais adiante, do módulo das fontes no domínio da transformada de Fourier) é uma mistura de gaussianas de média zero (GMM, do inglês *Gaussian Mixture Models*), a qual pode ser expressa por:

$$\mathcal{G}\left(y, \left\{\varpi^{(i)}\right\}, \left\{\Sigma^{(i)}\right\}\right) = \sum_{i=1}^K \varpi^{(i)} g\left(y, \Sigma^{(i)}\right), \quad (5)$$

onde g é a distribuição normal (de média nula) e $\varpi^{(i)}$ é sempre não-negativo, com $\sum_{i=1}^K \varpi^{(i)} = 1$. Daí obtemos as distribuições das fontes:

$$p(s_1) = \sum_{i=1}^{K_1} \varpi_1^{(i)} \frac{e^{\left[-\frac{1}{2} s_1^T \Sigma_1^{(i)-1} s_1\right]}}{(2\pi)^{N/2} \left|\det(\Sigma_1^{(i)})\right|^{1/2}}, \quad (6)$$

$$p(s_2) = \sum_{j=1}^{K_2} \varpi_2^{(j)} \frac{e^{\left[-\frac{1}{2} s_2^T \Sigma_2^{(j)-1} s_2\right]}}{(2\pi)^{N/2} \left|\det(\Sigma_2^{(j)})\right|^{1/2}}, \quad (7)$$

onde os parâmetros ($\varpi_1^{(i)}$, $\varpi_2^{(j)}$, $\Sigma_1^{(i)}$ e $\Sigma_2^{(j)}$) devem ser estimados a partir dos dados da mistura x . Já para efetuar a estimativa das fontes, devemos adotar uma medida do custo da substituição dos parâmetros reais por estimativas dos mesmos, medida que deve, num momento posterior, ser minimizada.

As duas medidas de custo mais comuns são a do erro quadrático e a da distribuição de Dirac (a qual apresenta valor nulo, exceto quando as estimativas coincidem com os parâmetros reais). É possível provar que a primeira medida implica uma estimativa que equivale

à média posterior condicional, enquanto que a segunda implica a clássica estimativa MAP (*maximum a posteriori*).

Em termos matemáticos, supondo parâmetro θ , estimativa α e função custo $C(\alpha, \theta)$, a estimativa α_{opt} é calculada através de uma minimização do custo médio sobre todos os valores possível de θ :

$$\alpha_{opt} = \arg \min_{\alpha} \int_{\theta} C(\alpha, \theta) f(x|\theta) \pi(\theta) d\theta, \quad (8)$$

onde $\pi(\theta)$ representa o conhecimento que porventura tenhamos acerca do parâmetro θ , antes de observarmos a mistura x . Admitindo $C(\alpha, \theta) = |\alpha - \theta|^2$, podemos deduzir que $\alpha_{opt} = E(\theta|x)$ (estimativa de média posterior condicional - em inglês, abreviado por PM) [11]. Quando $K_1 = K_2 = 1$, as GMMs associadas a cada fonte se degeneram em meras distribuições gaussianas; nesta condição, podemos formular de forma bayesiana o filtro de Wiener (impondo a condição $\sigma_n = 0$) e facilmente demonstrar que tanto a estimativa MAP quanto a PM podem ser obtidas por:

$$\hat{s}_1 = \Sigma_1 x [\Sigma_1 + \Sigma_2 + \sigma_n^2 I]^{-1}, \quad (9)$$

$$\hat{s}_2 = \Sigma_2 x [\Sigma_1 + \Sigma_2 + \sigma_n^2 I]^{-1}. \quad (10)$$

No caso de distribuições GMMs das fontes, a dedução não é direta, sendo normalmente obtida através de dois passos, fundamentados numa interpretação de uma mistura de gaussianas por meio de um modelo gerador [12]:

- (1) seleção de uma das K gaussianas através de uma distribuição discreta que pode ser expressa por $\sum_{k=1}^K \varpi^{(k)} \delta(q-k)$; q é associado ao componente (ou estado) ativo;
- (2) geração de uma amostra seguindo a distribuição $g(y, \Sigma^{(q)})$.

Os estados ativos das fontes serão denominados q_1 e q_2 (um para cada fonte). Estes estados são normalmente desconhecidos, devendo ser estimados. A probabilidade posterior de que $q_1 = i$ e $q_2 = j$ é dada por:

$$\begin{aligned} \gamma_{i,j}(x) &= p(i, j|x) \propto p(x|i, j)p(i)p(j) \\ &\propto \varpi^{(i)} \varpi^{(j)} g\left(x, \Sigma_1^{(i)} + \Sigma_2^{(j)} + \sigma^2 I\right). \end{aligned} \quad (12)$$

Sendo $q_1 = i$ e $q_2 = j$, o estimador condicional de Wiener (o qual, num paradigma bayesiano, equivale tanto à estimativa MAP quanto à PM quando $\sigma_n = 0$) é dado por:

$$E(s_1|i, j) = \frac{\Sigma_1^{(i)} x}{\Sigma_1^{(i)} + \Sigma_2^{(j)} + \sigma^2 I} \quad (13)$$

$$E(s_2|i, j) = \frac{\Sigma_2^{(j)} x}{\Sigma_1^{(i)} + \Sigma_2^{(j)} + \sigma^2 I} \quad (14)$$

Dadas as distribuições GMM das fontes (com $K_1 > 1$ e $K_2 > 1$), as estimativas MAP e PM passam a ser distintas. A estimativa MAP estima o estado da GMM de cada fonte através do maior valor de $\gamma_{i,j}$, e então recorre às equações 13 e 14 para estimar as fontes (numa espécie de filtragem de Wiener adaptativa [3]); já as estimativas PM são obtidas por meio da ponderação a seguir mostrada [13]:

$$E(s_1|x) = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \gamma_{i,j}(x) \Sigma_1^{(i)} \left[\Sigma_1^{(i)} + \Sigma_2^{(j)} + \sigma^2 I \right]^{-1} \cdot x, \quad (15)$$

$$E(s_2|x) = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \gamma_{i,j}(x) \Sigma_2^{(j)} \left[\Sigma_1^{(i)} + \Sigma_2^{(j)} + \sigma^2 I \right]^{-1} \cdot x, \quad (16)$$

$$\text{onde } \gamma_{i,j} \propto \varpi_1^{(i)} \varpi_2^{(j)} g\left(x, \Sigma_1^{(i)} + \Sigma_2^{(j)} + \sigma^2 I\right).$$

Por fim, cabe ressaltar que todas as estimativas acima mostradas não se aplicam às amostras de sinais de áudio no domínio do tempo. Elas são aplicadas à cada raia (*bin*) da transformada discreta de Fourier (DFT, de *Discrete Fourier Transform*) dos sinais (na realidade, a transformada de Fourier de tempo curto, já que a DFT é aplicada em quadros, devidamente suavizados mediante uma janela de Hanning), o que acaba por gerar observações unidimensionais (portanto, a partir de agora, podemos substituir $\Sigma_1^{(i)}$ e $\Sigma_2^{(j)}$ por $\sigma_1^{2(i)}$ e $\sigma_2^{2(j)}$). À cada raia são associadas duas GMMs: uma para a fonte 1 e outra para a fonte 2. Na etapa de treinamento e na equação 12, utiliza-se o módulo da DFT, o que implica uma modelagem que não atenta para a fase do sinal (no domínio de Fourier).

Para o treinamento dos parâmetros das GMMs, a partir de valores iniciais arbitrários, recorremos ao clássico método iterativo EM (do inglês *Expectation Maximization*) [12],[14].

2 MÉTODO PROPOSTO: FGMM_H

Detalhado na seção anterior, o método FGMM almeja modelar (via GMMs) a distribuição dos módulos de cada raia da DFT dos sinais. Cada raia é modelada de forma independente. No entanto, é conhecida a alta dependência entre as raias quando um instrumento emite uma nota musical. Esta constatação inspirou nossa proposta (denominada, daqui em diante, FGMM_H), a qual apresenta duas diferenças em relação ao método FGMM, a saber:

- (1) contempla a dependência entre harmônicos;
- (2) utiliza uma outra técnica de reconhecimento de padrões: redes neurais de múltiplas camadas.

O recurso a outras técnicas de reconhecimento é importante devido a um grave problema na modelagem do método FGMM: a existência de um modelo gaussiano (o qual contempla todo o eixo dos conjuntos reais)

para sinais não-negativos (módulos das raias da DFT). Outra hipótese algo irreal do método é a de um modelo de média zero de sinais não-negativos (módulos da DFT).

2.1 Análise DFT para sinais harmônicos

A primeira dificuldade de recorrer à estrutura harmônica dos sinais reside na DFT. As freqüências por ela analisadas estão uniformemente distribuídas entre 0 e a metade da freqüência de amostragem. Isto provoca uma incompatibilidade com uma análise das freqüências associadas aos harmônicos, dado que estes se distribuem uniformemente numa escala logarítmica. Para contornar este problema, é possível utilizar outras transformadas (como *wavelets*).

Neste ponto, optamos por uma estratégia já utilizada em modelagem senoidal: a partir dos módulos das três raias (da DFT) mais próximas à freqüência desejada, interpolamos uma parábola. Nossa estimativa do módulo da DFT na freqüência desejada será o valor apresentado pela parábola nesta freqüência. Matematicamente, seja f a freqüência a analisar. As raias mais próximas a f apresentam freqüências f_1 , f_2 e f_3 , apresentando módulos X_1 , X_2 e X_3 , respectivamente. O módulo da DFT em f é estimado por $af^2 + bf + c$, onde a , b e c são calculados por:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} f_1^2 & f_1 & 1 \\ f_2^2 & f_2 & 1 \\ f_3^2 & f_3 & 1 \end{bmatrix}^{-1} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \quad (17)$$

2.2 Formação dos padrões

Para a tarefa de reconhecimento de padrões, são formados 12 diferentes conjuntos de padrões, cada qual associado a uma nota e aos seus respectivos harmônicos. O número exato de harmônicos depende da nota e da freqüência de amostragem. Por exemplo, a nota A0 possui freqüência de 27,5 Hz. Numa amostragem a 11 kHz (onde a freqüência do sinal deve ser limitada a 5,5 kHz), ela apresenta harmônicos em 55, 110, 220, 440, ..., 3520 Hz; o padrão associado a esta nota possui 8 componentes (A0 mais seus sete harmônicos). Os padrões são formados apenas pelos módulos da DFT nas freqüências de interesse. Os módulos são obtidos mediante a interpolação parabólica supracitada.

A idéia de se utilizar o módulo da DFT apenas de uma nota (e de seus harmônicos) para cada rede neural se inspira no fato de que as relações entre o módulo da DFT na freqüência da nota e de seus harmônicos costumam ser muito específicas para cada instrumento, possibilitando a distinção destes numa etapa de separação posterior.

2.3 Análise com Redes Neurais

Para treinamento, após a formação dos padrões, podemos utilizar redes neurais para classificar os

padrões associados à fonte 1 ou à fonte 2. Para este propósito, recorremos a redes neurais *feedforward* de duas camadas (a escondida e a de saída) com apenas um neurônio de saída, treinadas por meio do algoritmo rápido *resilient backpropagation* [15]. A função de ativação escolhida foi a tangente hiperbólica. No treinamento, associamos à saída do neurônio da camada de saída o valor 1 para a fonte 1 e o valor -1 para a outra fonte. Cabe ressaltar que há 12 redes neurais, uma para cada nota.

Na fase de teste, após o janelamento do sinal e posterior transformação para o domínio de Fourier, calculamos suas componentes nas freqüências desejadas (novamente, por meio da interpolação parabólica). Então, para cada um dos 12 padrões, obtemos a resposta da rede neural, a qual varia entre +1 e -1.

Seja y_o a resposta de uma das rede neurais (obtida a partir do neurônio de saída). Este valor é associado a um nível de *mascaramento* M_k , dado pela fórmula $M_k = (y_o + 1)/2$. O valor de mascaramento M_k é associado a todas as freqüências relacionadas à rede neural (a nota e seus harmônicos).

2.4 Mascaramento

As redes neurais geram, na fase de teste, um valor de mascaramento para as freqüências associadas a cada uma das notas. Então, deparamo-nos novamente com o problema de compatibilizar estas freqüências (uniformemente distribuídas por nota numa escala logarítmica) com as obtidas mediante uma DFT. Uma alternativa seria utilizar uma interpolação hiperbólica, como já visto acima. Escolhemos utilizar uma interpolação *spline* cúbica [16] (notamos que os resultados são praticamente os mesmos que os oriundos de uma interpolação parabólica). Todos os valores negativos obtidos por esta interpolação são considerados nulos; já os superiores a 1 são igualados à unidade. De todo modo, estas retificações são pouco freqüentes. A figura 1 apresenta um exemplo de mascaramento oriundo de redes neurais e sua respectiva interpolação (o eixo das abscissas está na escala logarítmica, mais adequada para a visualização).

A estimativa da fonte 1 é obtida mediante o mero produto (no domínio de Fourier) do sinal de teste pela interpolação do mascaramento ($M_k^{int}(f)$, visto na figura 1). Já a estimativa da fonte 2 é obtida pelo produto da DFT por $(1 - M_k^{int}(f))$. Dessa forma, o mascaramento é uma medida do grau em que determinada amostra (no espaço tempo × freqüência) pertence à fonte 1 (e consequentemente o grau em que esta não pertence à fonte 2). Uma outra interpretação conveniente seria a da probabilidade de classe *a posteriori* [17] (no nosso caso, as “classes” seriam as fontes 1 e 2). Uma forma de mascaramento alternativa (e menos suave) seria a do mascaramento binário, a qual associa cada ponto (no espaço tempo × freqüência) a apenas uma das fontes [18].

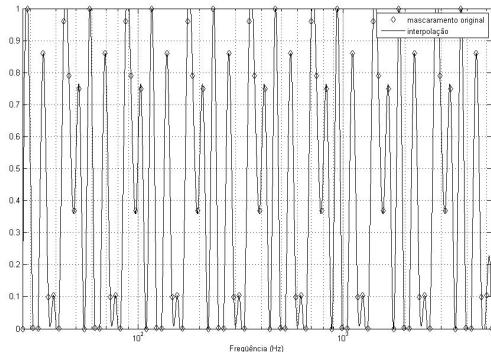


Figura 1: Exemplo de mascaramento via redes neurais e sua interpolação.

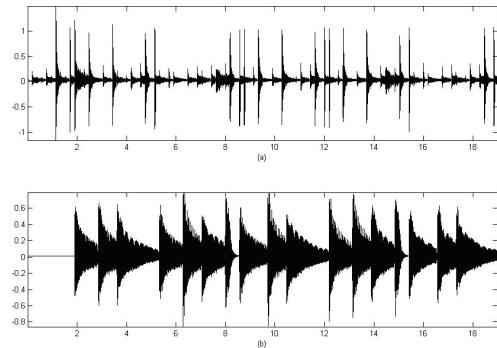


Figura 2: Fontes originais: (a) bateria, (b) piano.

3 RESULTADOS

Auditivamente, os resultados dos métodos FG-MM e FGMM_H são muito ruins quando as fontes não estão sincronizadas. Cremos que, em [3], as fontes utilizadas faziam parte de um trecho coerente de música por esta razão. Eis uma limitação de ambos os métodos, a qual comumente não é muito grave. Em [3], os testes foram efetuados com dois instrumentos: bateria e piano. Nas simulações por nós efetuadas, optamos também por estes dois instrumentos, amostrados em 11kHz, devidamente janelados (janela de Hanning) e com sobreposição de 75%. Eis uma configuração interessante, a partir do momento em que temos um instrumento (piano) que obedece ao “paradigma harmônico” (objeto de atenção especial do método proposto FGMM_H) e outro que não apresenta esta condição.

Duas medidas de qualidade de separação foram calculadas: SIR (razão sinal interferência, sigla oriunda do inglês *Source to Interference Ratio*) e SAR (razão artefato interferência, do inglês *Source to Artefact Ratio*). O SIR mede o resíduo da outra fonte na estimativa de cada fonte. Já o SAR reflete a quantidade de distorção em cada sinal estimado. O cálculo dessas medidas segue o formulado em [19].

Foi utilizado um trecho de cada fonte isolada com 1 minuto de duração para treinamento e validação. As redes neurais dedicaram 20% destes dados para validação. Foram inicializadas 10 diferentes redes neurais para cada nota, sendo escolhida para a fase de teste a que apresentou menor erro quadrático médio. Este procedimento, muito comum, reduz a probabilidade de uma rede neural se encontrar num mínimo local inadequado da superfície de erro.

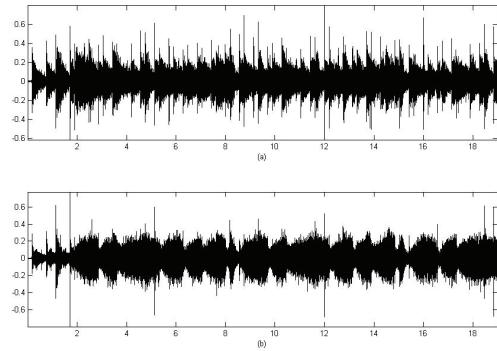


Figura 3: Fontes estimadas pelo método FGMM, com 4 gaussianas e janela de comprimento 512: (a) bateria, (b) piano.

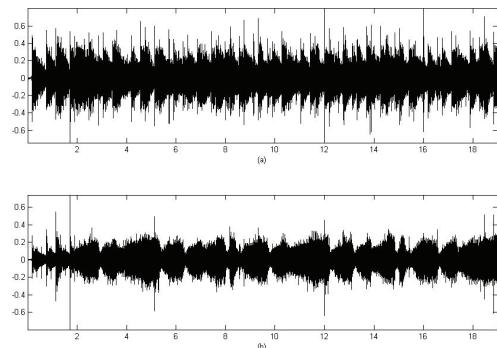


Figura 4: Fontes estimadas pelo método FGMM, com 16 gaussianas e janela de comprimento 512: (a) bateria, (b) piano.

Método	P	N	SIR_b	SIR_p	SAR_b	SAR_p
FGMM	4	512	13,42	3,56	-13,86	-15,06
FGMM	8	512	15,05	1,07	-14,12	-17,33
FGMM	16	512	17,56	0,23	-14,72	-18,27
FGMM	4	1024	27,73	-3,76	-13,54	-16,66
FGMM	8	1024	38,48	-7,02	-14,24	-17,72
$FGMM_H$	3	512	-7,14	10,03	2,75	1,76
$FGMM_H$	3	1024	-1,12	14,97	0,50	5,79
$FGMM_H$	4	1024	-3,46	12,51	0,71	5,22
$FGMM_H$	6	1024	-5,10	10,65	-0,09	4,86

Tabela I: Medidas SIR e SAR obtidas pelos métodos FGMM e $FGMM_H$.

A tabela acima exibe os resultados obtidos. O valor de P (parâmetro) indica o número de gaussianas (para o método FGMM) ou o número de neurônios na camada escondida (para o método $FGMM_H$) e N é o comprimento da janela de Hanning. Os subscritos “ p ” e “ b ” indicam piano e bateria, respectivamente.

Os resultados apresentados indicam que o SIR para a estimativa do piano obteve uma significativa melhora com o método proposto; já a da bateria sofre uma degradação muito grande, devido ao fato de esta não apresentar características harmônicas típicas. É interessante notar que o método proposto oferece um SAR superior, provavelmente devido ao fato de suas alterações no domínio da freqüência tenderem a ser suaves (o que não ocorre com o método FGMM, pois este modela cada raia da DFT de forma independente das outras).

Outro padrão a ressaltar consiste na melhora em SIR_b no método FGMM com o aumento do número de gaussianas, acompanhado de uma degradação em SIR_p . No método $FGMM_H$, notamos que um número de neurônios na camada escondida superior a 3 degrada os resultados (exceto SAR_b quando o número de neurônios é igual a 4; mas a diferença é muito pequena). Um tamanho de janela de 1024 se mostrou benéfico (em relação a um comprimento de 512) para o método proposto, enquanto que para o FGMM só o foi para a bateria. As figuras 2 a 7 mostram as fontes originais (amostras no domínio do tempo, num trecho de duração de aproximadamente 19s), bem como suas estimativas mediante várias configurações dos métodos FGMM e a melhor estimativa do método proposto $FGMM_H$. Embora a aparência da estimativa via $FGMM_H$ pareça muito superior, auditivamente a estimativa da bateria pelo método FGMM é significativamente melhor (o que é coerente com os valores de SIR_b apresentados). O método FGMM parece apresentar muito ruído (o que degrada o seu SAR). Provavelmente, tal fato se deve às descontinuidades no tratamento das raias (freqüências), já que cada raia é modelada de forma completamente independente das outras.

A figura 8 apresenta a evolução do mascaramento no espaço tempo \times freqüência em um pequeno trecho do sinal; observamos que a superfície de mascaramento é relativamente suave. Foram suprimidas as freqüências inferiores a 1000 Hz, para tornar o gráfico de mais fácil

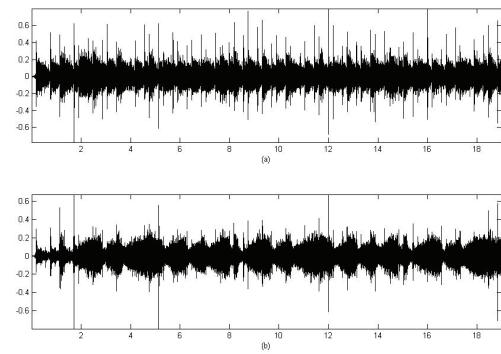


Figura 5: Fontes estimadas pelo método FGMM, com 4 gaussianas e janela de comprimento 1024: (a) bateria, (b) piano.

visualização (há uma grande densidade de picos nestas freqüências, já que nesta figura o gráfico se apresenta numa escala linear).

4 CONCLUSÕES

Neste artigo, apresentamos um método alternativo de separação de fontes mediante uma única mistura. Este método adota o mesmo paradigma do método apresentado em [3]. Duas características que distinguem o método proposto são o seu recurso à estrutura harmônica de sinais oriundos de instrumentos musicais e o uso de redes neurais *feedforward* para reconhecimento de padrões.

A modelagem da estrutura harmônica nos permitiu obter melhorias significativas na separação, apresentando distorção menor e maior razão sinal-interferência (para instrumentos harmônicos). A especificidade da nossa proposta implica uma degradação de desempenho quando o instrumento a separar não possui uma freqüência fundamental bem definida (como é o caso da bateria). Este fato nos leva a concluir que uma estratégia interessante seria empregar o método

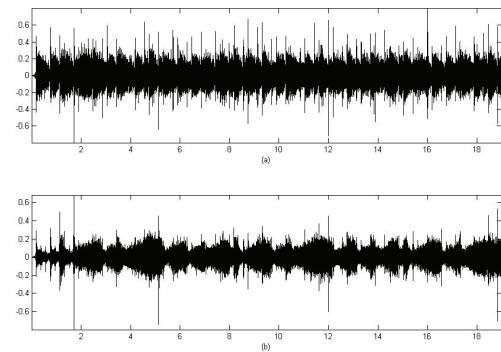


Figura 6: Fontes estimadas pelo método FGMM, com 8 gaussianas e janela de comprimento 1024: (a) bateria, (b) piano.

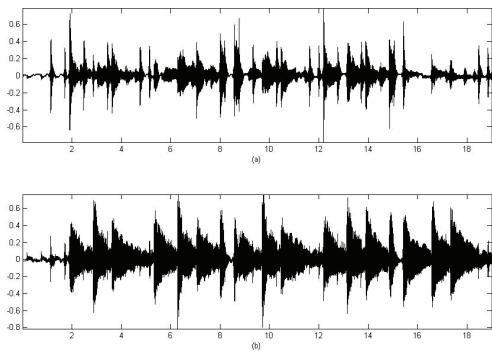


Figura 7: Fontes estimadas pelo método FGMM_H , com 3 neurônios na camada escondida e janela de comprimento 1024: (a) bateria, (b) piano.

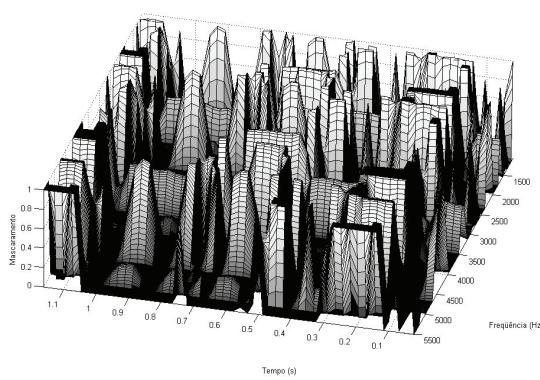


Figura 8: Padrão do mascaramento no espaço tempo × freqüência em um pequeno trecho de sinal.

proposto ou o método de [3] de forma dependente do instrumento. Uma alternativa para o mascaramento suave proposto seria utilizar um mascaramento binário ($M_k(f)$ podendo ser 0 ou 1 [18]). Pretendemos efetuar a comparação deste mascaramento com o por nós proposto no futuro.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] J. Karhunen e E. Oja A. Hyvärinen, *Independent Component Analysis*, New York: Wiley Interscience, 2001.
- [2] Paul Bofill e Michael Zibulevsky, “Underdetermined blind source separation using sparse representations,” *Signal Process.*, vol. 81, 2001.
- [3] Frédéric Bimbot e Rémi Gribonval Laurent Benaroya, “Audio source separation with a single sensor,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, 2006.
- [4] L. Deng e R. L. Brennan H. Sameti, H. Sheikhzadeh, “Hmm strategies for enhancement of speech signals embedded in nonstationary noise,” *IEEE Transactions Speech Audio Processing*, vol. 6, no. 5, pp. 445–455, 1998.
- [5] Tuomas Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.
- [6] M. A. Casey e A. Westner, “Separation of mixed audio sources by independent subspace analysis,” *Proc. Int. Comp. Music Conf.*, 2000 volume = 12, number = 1, pages = 157–165, month = March,.
- [7] S. Dubnov, “Extracting sound objects by independent subspace analysis,” *Proc. 22nd Int. Audio Eng. Soc. Conf.*, 2002.
- [8] Jean Rouat, “Source separation with one ear: Proposition for an anthropomorphic approach,” *EURASIP Journal on Applied Signal Processing*, pp. 1365–1373, 2005.
- [9] D. L. Wang e G. J. Brown, “Separation of speech from interfering sounds based on oscillatory correlation,” *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.
- [10] D. L. Wang e G. J. Brown, “Separation of stop consonants,” *ICASSP*, vol. 2, pp. 749–752, 2003.
- [11] Simon Haykin, *Adaptive Filter Theory*, Englewood Cliffs, NJ: Prentice Hall, 1996.

- [12] Carlo Tomasi, “Estimating gaussian mixture densities with em - a tutorial,” *Duke University*, data desconhecida.
- [13] Y. Ephraim e N. Merhav, “Hidden markov processes,” *IEEE Transactions on Information Theory*, vol. 41, no. 5, 1995.
- [14] J. A. Bilmes, “A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden mixture models,” *Tech. Rep.*, 1998.
- [15] M. Riedmiller e H. Braun, “A direct adaptive method for faster backpropagation learning: The rprop algorithm,” *Proceedings of the IEEE International Conference On Neural Networks*, 1993.
- [16] C. de Boor, *A Practical Guide to Splines*, Springer-Verlag, 1978.
- [17] Simon Haykin, *Neural Networks - A comprehensive foundation*, Prentice-Hall, 1999.
- [18] Ö. Yilmaz e S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Transaction on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.
- [19] E. Vincent e C. Févotte R. Gribonval, L. Benaroya, “Proposals for performance measurement in source separation,” *Proc. ICA*, pp. 715–720, 2003.



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Towards the evaluation of automatic transcription of music

Tiago Fernandes Tavares,¹ Jayme Garcia Arnal Barbedo,¹ and Amauri Lopes¹

¹ Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação,
Departamento de Telecomunicações
Av. Albert Einstein - 400, Campinas, São Paulo, CEP 13083-852, Brasil

ttavares@fee.unicamp.br, jgab@decom.fee.unicamp.br, amauri@decom.fee.unicamp.br

ABSTRACT

This article presents a method for automatic evaluation of music transcription algorithms. It is based on the comparison between a reference transcription, assumed as ground-truth, and the automatically transcribed file. The method presented here aims to provide an unified background to evaluate transcription algorithms, which, to the moment, have been assessed using different methods, obstructing direct comparison of their performance. The presented method is, also, compatible with monophonic and polyphonic transcription. The method allows the end-user to define what specific performance metrics will be calculated, thus being potentially applicable to any related work.

0 INTRODUCTION

Music transcription is the process of extracting a high-level symbolic representation, such as scores, tablatures or chord tables, from an audio signal. The information about musical structures contained in the final product of a transcription process allows a musician to entirely play the transcribed musical piece.

An automatic music transcriptor is a system that performs musical transcription without human interference. There are, today, several methods for automatic transcription of music. These methods are often designed to operate in a specific subset of all music, like monophonic sound [1, 2] (where only one note is

played at a given time), bass instruments (bass, eletric bass, etc.) [3], and voice over a polyphonic recording [4]. There exists, however, some attempts to design an automatic transcription process suitable for any kind of music [5, 6].

Although there is consensus about the desired output of an automatic transcriptor - a set of notes described by their pitch, start time and stop time - the methods used for performance evaluation in each work tend to differ. Some evaluation processes are subjective and qualitative [2, 7, 8], while others include the definition of performance rates based on the classification of detected notes in groups like *correct* and *incorrect*

[4, 3, 6, 1]. Since the exact definition of a correct note depends on a subjectively chosen error tolerance, these performance rates have different definitions depending on the author. This work presents a method that allows fast calculation of user-defined performance metrics, allowing fast evaluation and comparison between different transcription algorithms.

The presented method uses as input the final product of an automatic transcription process together with a manual reference transcription. By selecting, for each note in the reference transcription, a corresponding note from the automatic transcription according to a best match criterion, a data set of individual note performances is generated and may be clustered for further analysis. At the end of the evaluation process, the method returns feedback information about the performance of the automatic transcription algorithm.

This paper is divided as follows. In Section 1, the used mathematical notation and other formal aspects are addressed. The evaluation process is explained in Section 2. Section 3 contains conclusions and final considerations.

1 NOTATION

Before the evaluation algorithm is explained, it is important to make explicit what kind of data will be dealt with.

Automatic music transcription algorithms are designed to output a set of data structures that will be called, in this work, note descriptors. Each note descriptor contains information about time (when the note starts and stops playing) and frequency (which note, or notes, are being played). For simplicity, note descriptors will also be referred as notes.

The performance of an automatic music transcription is normally based on the similarity between its output and a reference transcription. This means that each audio file used in the evaluation process must be, first, manually transcribed. While manual transcription is often slower than manual evaluation of transcription processes, it only has to be performed once for each audio file.

The automatic evaluation system uses as inputs the note descriptor sets contained in both the manual and evaluation transcriptions - respectively, T_m (with N_m elements) and T_a (with N_a elements). Each characteristic of a note descriptor will be referred in this work in the form $T[n].characteristicName$, where:

- T is the set name and may be any defined note descriptor set, like T_m (for the manual transcription) or T_a (for the automatic transcription).
- n is the current element number and may be any integer value between 1 and the number of elements of the current set.
- $characteristicName$ is the referred characteristic name, and may be $start$ for the start time of the

note, $stop$ for its stop time, $length$ for its length or $number$ for its MIDI number.

The MIDI number related to a note is calculated according to its fundamental frequency F_0 through Equation 1.

$$N_{MIDI} = 69 + 12\log_2 \frac{F_0}{440}. \quad (1)$$

2 THE EVALUATION PROCESS

The evaluation process is divided into three parts. In the first one, the notes in T_a that best describe the notes in T_m are selected, in an one-to-one matching algorithm. Next, each match is analyzed to extract performance information regarding each individual note. Finally, statistical information about the overall transcription system performance is calculated.

2.1 The best match

In order to calculate the performance for each individual note, it is necessary to find out what is the note descriptor in T_a that best matches each note descriptor in T_m . Since only one note in T_a may be related to each note in T_m , the matching process may leave some elements unmatched. This will be relevant in later steps of the evaluation process.

2.1.1 Likelihood between note descriptors

In order to match note descriptors, their likelihood must be estimated. To do so, each note descriptor will be projected into two Euclidean spaces. The first one will be used for the distance calculation of note descriptors in time. The second one will be used for the distance calculation of note descriptors in frequency.

The time distance between two note descriptors N_1 and N_2 is calculated by Equation 2. It is deduced by using the start time and length of both notes as their horizontal and vertical coordinates in a Cartesian system and calculating the distance between them.

$$D_t(N_1, N_2) = \sqrt{(\Delta begin)^2 + (\Delta length)^2}, \quad (2)$$

where $\Delta begin = N_1.begin - N_2.begin$ and $\Delta length = N_1.length - N_2.length$.

In the same context, it is important to know the time overlap (T_{over}) between two notes. This is done through Equation 3, which calculates for how much time N_1 and N_2 are playing together.

$$T_{over}(N_1, N_2) = \max(0, end_{min} - begin_{max}), \quad (3)$$

where $end_{min} = \min(N_1.end, N_2.end)$ and $begin_{max} = \max(N_1.begin, N_2.begin)$.

The calculation of the frequency distance between two note descriptors requires a more carefull analysis. Most common frequency errors are octave errors (which means the transcription systems finds a note whose fundamental frequency is a multiple or submultiple of the reference note) or semitone errors (which

means the transcription system finds a note that is adjacent to the reference transcription). The frequency component of the note descriptor will be projected into a vector space in which the distance related to octave errors is close to the distance related to semitone errors, and both of them are smaller than the distances related to other errors. This makes these common errors less relevant than other errors during the note matching process.

These conditions may be obtained by projecting the frequency component of the note descriptor into cylindrical coordinates. The MIDI number N_{MIDI} of the note is used as basis for this. Increasing the MIDI number leads to a corresponding increment to the height (z-axis) coordinate and the angle (θ -coordinate), as seen in Equation 4. This projection makes the distance related to semitone and octave errors similar, while both of them are smaller than the distances related to other errors.

$$\begin{aligned} Z &= \frac{N_{MIDI}}{12} \\ \theta &= \text{mod}_{2\pi}\left(\frac{2\pi N_{MIDI}}{12}\right), \end{aligned} \quad (4)$$

where $\text{mod}_B(A) = A - kB$, where k is the greatest integer that satisfies $A - kB \geq 0$.

Equation 5 shows how to calculate the frequency distance D_f between two notes N_1 and N_2 , with projected coordinates Z_1, θ_1 and Z_2, θ_2 :

$$D_f(N_1, N_2) = \sqrt{(Z_1 - Z_2)^2 + (\psi^2(\theta))}, \quad (5)$$

where $\psi(\theta)$ is the smallest angular difference between θ_1 and θ_2 .

It is now possible to calculate a likelihood parameter as shown in Equation 6:

$$L(N_1, N_2) = \frac{T_{over}(N_1, N_2)}{\sqrt{D_f^2(N_1, N_2) + D_t^2(N_1, N_2)}}. \quad (6)$$

The likelihood parameter aims to be a measure of the similarity between two note descriptors, allowing the algorithm described in the next section to work.

2.1.2 Matching algorithm

The matching algorithm aims to define the best one-to-one match between notes in T_a and T_m , using as basis the likelihood parameter L (Equation 6). It is described in two steps.

The first step begins with the calculation of the likelihood L between each note $T_a[k]$ and each note $T_m[n]$, generating a likelihood matrix $M_{K \times N}$ where $m_{k,n} = L(T_a[k], T_m[n])$. Then, it selects, for each line k in M , the element with the greatest value, and keeps the related column number (n). This is equivalent to searching the note in T_m with greatest likelihood parameter L with respect to $T_a[k]$. The note $T_a[k]$ is, then, added to a set $C[n]$, that contains possible matches for $T_m[n]$.

The second step aims to ensure that a correct one-to-one match is chosen. In this process, each set $C[n]$ is analyzed. The note in $C[n]$ with the greatest likelihood parameter to $T_m[n]$ is chosen as its best match. The product of this step is an array $R[n]$ with N_m elements. $R[n]$ is the note descriptor in T_a that best matches $T_m[n]$.

This two-step algorithm ensures that a note in T_a will not be chosen as a match for more than one note in T_m , and, at the same time, that a note in T_m will not be matched by more than one note in T_a .

2.2 Data analysis

This stage aims to extract information from the stored data. It calculates performance metrics for each expected note descriptor using comparison between the note descriptor itself and its best match (calculated in the previous section).

The first parameter to be calculated is the detection lag. It is calculated for each note using Equation 7, which reveals the lag between the effective and the detected beginning of the note.

$$\text{Lag}[n] = T_m[n].start - R[n].start. \quad (7)$$

Second, the fraction of the expected note length that the automatic transcription system was able to detect is calculated using Equation 8.

$$\text{Frac}[n] = \frac{R[n].length}{T_m[n].length}. \quad (8)$$

Finally, the frequency error is calculated by Equation 9. This parameter shows the difference, in semitones, between the expected and the detected notes.

$$\text{Freq}[n] = |R[n].number - T_m[n].number|. \quad (9)$$

At the end of these calculations, there is, for each note descriptor in T_m , a corresponding performance description array. It is now necessary to divide this data set into groups, allowing statistical error analysis.

2.3 Error grouping

The simplest way to extract information from the performance description array data set is by defining rules that classify them as errors, hits, misses and other groups desired by the user. A possible set of rules for matching $T_m[n]$ and $T_a[k]$ is shown in Table 1.

Table 1: Set of rules for performance evaluation

Rules	Group
$\frac{T_a[k].length}{T_m[n].length} < \frac{3}{4}$	Miss
$\frac{T_a[k].length}{T_m[n].length} \geq \frac{3}{4}$	Found
$T_a[k].number = T_m[n].number$	Hit
$T_a[k].number \neq T_m[n].number$	Error

This table is not, by any means, definitive. Different criteria for error grouping can be found (e.g [4, 3]), especially regarding the classification of notes in groups “Miss” and “Found”. Since there is no restriction about overlapping error groups, rule sets regarding different metrics may be created.

In addition, there may be note descriptors in T_a without a corresponding best match in T_m . This means that the automatic transcription system has found notes that do not actually exist. The number of such notes, called *ghost notes*, should be counted as another performance evaluation criterion.

At the end of these calculations, the evaluation process returns a performance evaluation array consisted of:

- Mean and variation of the detection lag (Equation 7),
- Mean and variation of the detected fraction (Equation 8),
- Number of notes in each error group (Table 1),
- Number of ghost notes.

There are two relevant improvements, other than speed, that may be achieved by using the presented automatic evaluation process instead of simple manual evaluation. The first one is the possibility of statistical analysis over different projections of the whole result data set. Error grouping may be extended to involve rules regarding the related instrument, as done in [1], or specific pitch ranges. The second one is the use of the evaluation results as feedback for improving the performance of the automatic transcriptor. This can be done by many ways, like automatically modifying the execution parameters of the transcriptor until its performance is maximized or normalizing note start and stop times to have zero mean detection lag and 100% mean detected fraction.

3 CONCLUSION

In this paper, a method for automatic evaluation of music transcription algorithms was presented. It is based on the comparison between a reference transcription, assumed as ground-truth, and the results of an automatic transcription process. Through a matching algorithm, the performance of the automatic transcriptor is evaluated for each note, providing a performance data set that may be grouped according to user-defined criteria. This provides an unified background to evaluate automatic transcription methods, therefore allowing direct performance comparison.

This work is not concerned about the definition of the specific metrics that will be used for performance evaluation, which have to be defined by the end user. Once a set of metrics is defined, all calculation is done

automatically, allowing comparison between newly developed automatic transcription methods and its predecessors without requiring extensive manual evaluation work.

Also, the presented evaluation process allows performance feedback. This opens the possibility of designing transcription systems with self-adjusting behavior aiming performance improvement, which is expected to be done in future work.

REFERENCES

- [1] N. Trevilatto Jr., J. G. A. Barbedo, and A. Lopes, “Transcrição Automática de Sinais de Áudio Monofônico,” in *Anais do 10 Simpósio Brasileiro de Computação Musical*, 2005, vol. 1, pp. 291–294.
- [2] J. P. Bello, G. Monti, and M. Sandler, “Techniques for Automatic Music Transcription,” in *International Symposium on Music Information Retrieval*, 2000.
- [3] M. Ryynänen and A. Klapuri, “Automatic bass line transcription from streaming polyphonic audio,” in *Proc. 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawai’i, USA, Apr. 2007, pp. 1437–1440.
- [4] M. Ryynänen and A. Klapuri, “Transcription of the singing melody in polyphonic music,” in *Proc. 7th International Conference on Music Information Retrieval*, Victoria, BC, Canada, Oct. 2006, pp. 222–227.
- [5] K. D. Martin, “A Blackboard System for Automatic Transcription of Simple Polyphonic Music,” Tech. Rep. 385, M.I.T. Media Laboratory Perceptual Computing Section, 1996.
- [6] M. Ryynänen and A. Klapuri, “Polyphonic music transcription using note event modeling,” in *Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, Oct. 2005, pp. 319–322.
- [7] A. Klapuri, A. Eronen, J. Seppänen, and T. Virtanen, “Automatic transcription of music,” in *Symposium on Stochastic Modeling of Music*, Oct. 2001.
- [8] A. Klapuri, “Automatic transcription of music,” in *Proc. Stockholm Music Acoustics Conference*, Aug. 2003.



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Short-Term Classification of Musical Instruments: A Critical View

Jayme Garcia Arnal Barbedo¹, Amauri Lopes¹

¹Department of Communication, School of Electrical and Computer Engineering,
State University of Campinas
Campinas, São Paulo, 13083-852, Brazil
{jgab, amauri}@decom.fee.unicamp.br

ABSTRACT

This paper presents a discussion about the difficulties and limitations of automatically classifying musical instruments using short musical excerpts. The text focuses on two main constraining factors, which are closely intertwined: 1) the lack of homogeneous characteristics when comparing the sounds generated by similar instruments and 2) the insufficiency of the available databases in providing really representative sets of the instruments to be classified. To support the remarks stated along the paper, tests were performed using 14 well known features extracted from the signals, together with a pairwise classification procedure that has been successfully used in similar tasks. The final and most important objective of the paper is to promote a debate on the search for suitable solutions for such a difficult problem.

0 INTRODUCTION

The fast grow of multimedia databases has increased the demand for tools able to aid users searching and retrieving data and to improve the human-machine interaction. This is particularly true for audio databases, which have been growing since early 1990 decade. The automatic music genre classification is one of the most useful tools proposed in the last years, though the results have been inconsistent. One way to improve the music classification is to identify the instruments present in a given song, which would provide an important clue to the genre of such a song.

Instrument identification can also be very useful to automatically generate music scores, being in this case a complementary tool to automatic music transcribers. Finally, it can be an important part of future audio source separators, whose current performance is still far from ideal.

Several strategies to identify musical instruments have been proposed in the last decade, but their applicability is somehow limited by operational and/or mathematical constraints (see more details in Section 2). One of those limitations is that all methods require relatively long audio excerpts to acquire the temporal and spectral behaviors altogether, only then determining the instrument. Relying on long excerpts has some problems that can prevent such methods to be applied in practical conditions. First, the presence of a given instrument in a song may be intermittent, making it very difficult to determine in which excerpts it is actually present. Also, the instrument may be present only for a short period, in which case the method would not be able to perform the classification. Finally, in an actual song the temporal behavior, onsets, offsets, etc., of the concurrent instruments will be, in general, very correlated and aligned, making it nearly impossible to identify and classify them without some kind of a priori knowledge.

Therefore, it would be very useful to develop a technique able to classify the instruments in a short-term basis, taking small pieces of the signal (from now on called frames) and making an instantaneous classification of the instruments. However, such a task is extremely challenging, demanding strategies able to overcome the main difficulties.

Among the factors that difficult the design of an efficient instrument classifier, two deserve special attention. First, the spectrum of a given type of instrument will present a huge variability according to instrument manufacturer, musician, acoustics of the room, temperature, humidity, and a number of other factors. Also, in general a given frequency will not be precisely sustained for more than a few milliseconds. This prevents any frequency uniformity and makes it practically impossible to find a phase relation among all partials generated by a given note. Thus, in order to tune a strategy to deal with all variations expected for a given instrument, a huge database would be necessary. This is related to the fact that human beings need some education in order to refine their identification capabilities, which is, in some extent, equivalent to the training in computational methods. In practice, there are only a few databases available, being this the second major problem. The most used databases [1, 2, 3] are not even nearly enough to provide a robust training to the strategy, as will be seen along the paper. Finding ways to overcome these two problems is still an open challenge.

The main objective of this paper is to describe the difficulties and to promote a debate that can produce potential solutions for such a difficult problem.

To illustrate the effects of the cited problems, a method was implemented using 14 well known features and a classification procedure that uses a pairwise strategy previously used to classify audio signals into genres [4]. The results help to better understand the problem and provide a good starting point in the search for solutions.

The paper is organized as follows. Section 1 presents a description of some previous work. Section 2 presents a discussion about the problem of music instrument classification and the main difficulties involved. Section 3 presents the tests designed to gather more information on the subject. Section 4 shows some results. Finally, Section 5 presents the final remarks.

1 PREVIOUS WORK

The research on automatic classification of musical instruments is very recent, beginning in the second half of the 1990 decade. The complexity of the techniques and the relevance of the results have increased greatly since then, but a fundamental limitation has persisted: all proposals need long audio excerpts (at least one second) to perform the classification. Although several methods divide the signals into small frames (less than 50 ms), a statistical summarization of the collected data is always performed at some point due to the intrinsic characteristics of each strategy. Despite this, most proposals show very distinct approaches; the remaining of this section is dedicated to briefly describe some of the most important works in the area.

Early work started investigating very simple problems, usually considering only isolated notes (no simultaneous instruments). In 1997, a technique using Mel-based cepstral coefficients as features and a cluster-based probability model was used to distinguish between

saxophone and oboe sounds [5]. An evolution of this technique was presented in [6]. The technique presented in [7] also uses a statistical approach, but in this case a wider variety of acoustic features were extracted and an auditory model was applied; the proposal was able to distinguish between 15 instruments with an accuracy of 70%. The same author presents a wider analysis about sound source recognition in his doctoral thesis [8]. In [9] the authors use a template adaptation and music stream extraction, together with a statistical model, to distinguish between three instruments; the authors claim that the technique works for real music, but all tests were performed for isolated notes only. A critical review of the main instrument classification approaches proposed until 2000 is presented in [10]. A deep investigation about the effectiveness of different features to discriminate between woodwind instruments is performed in [11], showing that the results strongly depend on the training and test sets division. A deep study about instrument recognition, which resulted in a hierarchical classification approach, is presented in [12]. In [13], the authors use a number of spectral features to compare several different classification strategies in discriminating 27 different instruments; best results were achieved by support vector machines and quadratic discriminant classifiers, with accuracy around 70%.

In the last few years, a number of techniques able to deal with polyphonic music have been proposed, with relatively good results. In [14], the authors use a number of features and a k-nearest neighbors algorithm to perform the classification; the method was designed to deal with solo instruments, but tests revealed that it can also be applied to duet music. Independent subspace analysis is used in [15] to make possible to the method to deal with polyphonic music. Another technique able to identify instruments in duet music is presented in [16]; this technique uses a very complex decomposition scheme and neural networks to perform the classification. The technique presented in [17] explores some cues on the common structures of musical ensembles to recognize up to four instruments playing concurrently. Finally, in [18] the authors propose a feature weighting strategy to minimize the effects of sound overlaps in polyphonic music, making the instrument identification more suitable.

It is worth noting that most techniques use as reference to their performance the actual signal structure, meaning that the ultimate objective is to reach 100% of correct identifications. Some authors, on the other hand, compare their methods to the performance of expert human listeners in classifying instruments. In this case, the target performance can vary greatly according to the number and type of instruments being considered, but normally it is considerably lower than 100% (see [19] for a list of human recognition rates). This is a much more achievable objective, but the practical use of methods with such low identification capability is seriously limited.

2 THE PROBLEM OF MUSICAL INSTRUMENT CLASSIFICATION

2.1 Long-term versus short-term approaches

As stated before, all instrument identification systems proposed so far rely on relatively long audio excerpts (at least 1 second, normally more) to work properly. This kind of approach is usually preferred because features that describe the temporal behavior of the signal can be

extracted, and the corresponding spectral features can be statistically treated in order to result in more meaningful and homogeneous information. In contrast, instantaneous classification using short individual frames can rely only in the information contained in a small snapshot of the signal.

Despite the obvious advantages of classifying the signals using long excerpts, there are some problems associated that can clearly limit the use of such proposals in real situations. Some of the main problems:

1. The presence of a given instrument may be intermittent. This actually will happen for almost every instrument in any song. There are some ways to determine the occurrence of a new event in a song, however it is in general very difficult to determine if such an event is due to a new note played by an instrument that was already present, if it is due to the arise of a new instrument, or if it is due to both new and old instruments being played together. Therefore, it will be extremely difficult to the algorithm to determine which parts of the song should be considered in order to classify each instrument. At present, the only way to do that is to manually feed the algorithm with this knowledge prior to the classification itself.

2. A given instrument may be present only for a short time. In this case, the algorithm will not have enough information to perform a correct classification, leading to a drop in the accuracy.

3. The temporal behavior, onsets, offsets, etc., of concurrent instruments will be, in general, very correlated and aligned, making it nearly impossible to identify and classify them without some kind of a priori knowledge. It is worth noting that this will also occur for spectral characteristics, but in this case there are some efficient techniques that help in the identification of concurrent instruments (e.g. [20]).

4. The spectral and temporal characteristics of a given instrument type can vary greatly according to instrument manufacturer, musician, environmental variables, etc. This may cause the training set to be unmatched with the corresponding test set, reducing the robustness of the strategy. This problem is common to both long and short-term approaches. This point will be discussed in more depth later in the paper.

5. In real conditions, it will not be possible to a musician to keep the frequency and intensity of the sound constant for a long time. Sometimes those variations can be intense to the point that a statistic summarization becomes meaningless.

6. Real-time processing becomes impossible. Although this condition is not mandatory, the possibility of real-time processing may be desirable in a number of applications.

Therefore, there are several reasons to justify the search for a short-term classification system. However, the challenges involved in the task have made it difficult to such an approach thrive, as discussed next.

2.2 Problems of short-term approach

Among the factors that still restrain the arise of efficient short-term instrument classifiers, two closely interconnected problems are predominant: 1) the lack of homogeneous characteristics when comparing the sounds generated by similar instruments and 2) the insufficiency of the available databases in providing really representative sets of the instruments to be classified.

Similar instruments can generate very distinct signals both in time and frequency contexts, due to several

variables that influence the generation of the sound. Among the most important of such variables are:

- Instrument manufacturer: each manufacturer of a given instrument uses different assemble processes and different materials. This may cause that two similar instruments from different manufacturers have quite different acoustic attributes. Figure 1 shows the magnitude spectrum of two different violins played by the same musician. The differences are evident, as the partial ratios are quite different. Also, it can be seen that the partials tend to be much more spread around the desired frequency for the first violin.

- Musician: each musician has his own style and will play a given instrument in a particular way, imprinting particular acoustic attributes to the execution.

- Acoustics of the room: each environment will have its own acoustic characteristics, producing echoes that will interfere with certain frequency bands with a given delay. Those echoes can significantly change the spectral content of the signal.

- Environmental factors: variables like temperature, humidity, atmospheric pressure, etc., are important factors in the sound production.

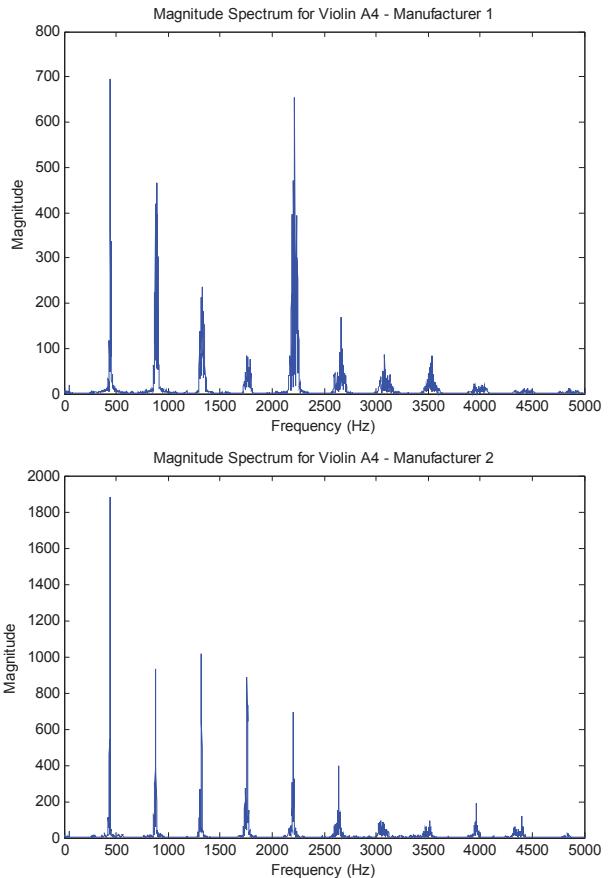


Figure 1 - Spectral differences between two violins playing the same note.

- Microphones: the position, sensitivity and manufacturer of the microphones also influence the sound attributes.

- Noise: noise will always be present in audio signals, and depending on the signal-to-noise ratio and the characteristics of the noise (white, pink, brown, etc.), the attributes of the signal can be significantly changed.

- Sound intensity: an instrument can be played with different intensities (e.g. forte, mezzo forte, piano, etc.). A change in the intensity will often imply in a change in the

way the instrument is being played. Additionally, lower intensities will be more sensitive to the effects of noise. Figure 2 shows the magnitude spectra of a violin played with two different intensities. It can be observed that not only the partial ratios have changed, but the noise is much more pronounced in the second case (pianissimo intensity).

- A/D conversion: the quality of the A/D converter will influence the acoustic attributes found in the digital signal.

- Musician inconsistency: no matter how good a musician is, there will always be important variations in the frequency and intensity along time. Therefore, even when one compares two consecutive frames, the instantaneous acoustical attributes will have changed. The plots in Figure 3 were generated taking two 21.3 ms frames from a continuous violin A4 excerpt. The frames are 1 second apart. It can be observed that the magnitude of the partials have varied greatly, showing a high degree of inconsistency along the excerpt.

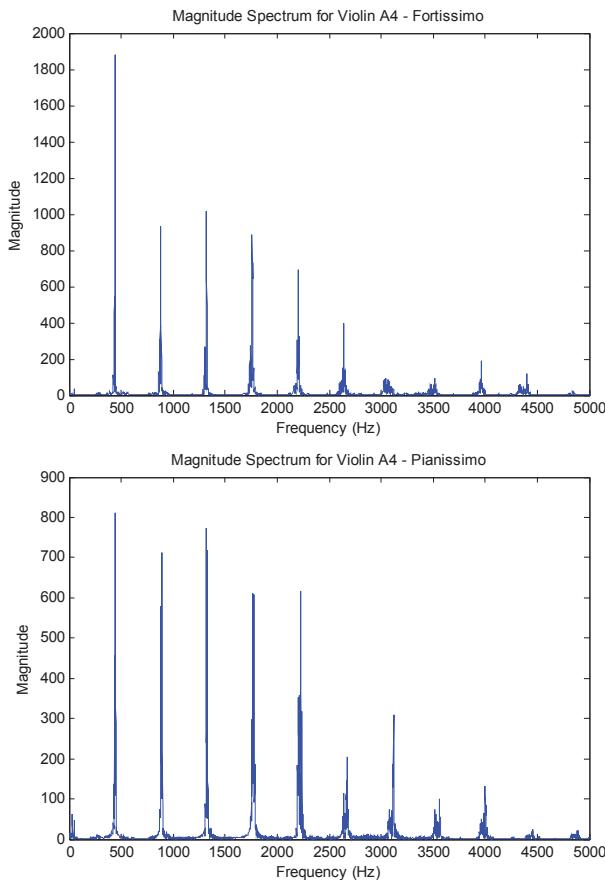


Figure 2 - Spectral differences of a violin played with different intensities.

As can be seen, there are lots of factors that contribute to the final acoustic attributes. Dealing with such a great variability is not an easy task. One possible solution would be using databases large enough to contain at least most of the different acoustical attributes expected for a given instrument. There are only a few carefully designed databases: the RWC database [1], the Iowa University database [2], and the McGill University database [3]. Unfortunately, their contents are not enough to train a short-term instrument classifier properly. However, the development of an instrument database able to provide a wide variety of acoustical attributes to each instrument would be extremely expensive both in terms of costs and

time demanded. Additionally, the number of variables involved in the sound production is so great that it would be impossible to guarantee that all possible behavior would be represented in the database, even in an approximate manner.

However, there are evidences that this problem could be solved without massive databases. A trained human being is able to identify an instrument even under difficult conditions, and to do that he does not need to be trained over a huge number, for example, of violins and different performers.

The most viable solution for this problem is to develop a feature (or a group of features) relatively invulnerable to the variations, a feature able to apprehend the "signature" of each instrument, no matter the surrounding conditions. The results presented next section can provide some clues toward this goal.

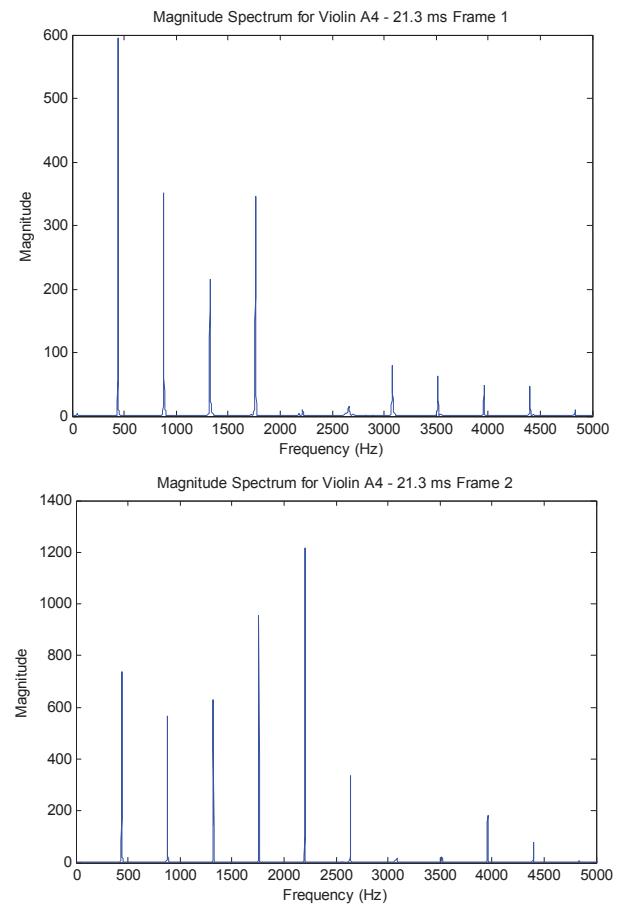


Figure 3 - Spectral differences for two frames of a same violin.

3 TEST DESIGN

In order to provide a better understanding and to gather more information about the problem, a short-term classifier was implemented. Such a classifier can be divided into two main parts: feature extraction and classification procedure.

3.1 Feature Extraction

The first stage of most musical instrument classifiers is the extraction of a number of features from the signals. The objective of such features is to provide the best characterization of the signal as possible. In the context of this work, 14 features are extracted from 21.3 ms frames.

Consecutive frames present a 50% overlap. The features, selected among the most used ones, are:

- *Spectral centroid*: it represents the “center of mass” of the spectral energy distribution of the signals, and is given by

$$ce_i = \frac{\sum_{k=1}^K k \cdot |X_i(k)|^2}{\sum_{k=1}^K |X_i(k)|^2}, \quad (1)$$

where $|X(k)|$ is the magnitude of the k^{th} spectral line resulting from an FFT (Fast Fourier Transform) applied to the frame i of the signal $x(n)$, and K is half the number of spectral bins. The spectral centroid is given in terms of spectral bins. To obtain the value in Hz, ce must be multiplied by the difference in Hz between two consecutive spectral bins.

- *Zero-crossing rate (ZCR)*: a zero crossing occurs whenever the amplitudes of two consecutive temporal samples have opposed signs, as indicated by the expression

$$zcr_i = 0.5 \cdot \sum_{n=1}^N |\operatorname{sgn}[x_i(n)] - \operatorname{sgn}[x_i(n-1)]|, \quad (2)$$

where $x_i(n)$ represents the samples of frame i of signal $x(n)$ and $\operatorname{sgn}(x)$ equals -1 or $+1$ as x is negative or positive, respectively.

- *Spectral roll-off*: this feature is defined as the frequency R_i below which 85% of the magnitude distribution is concentrated, as expressed by

$$\sum_{k=1}^{R_i} |X_i(k)| = 0.85 \cdot \sum_{k=1}^K |X_i(k)|, \quad (3)$$

- *Spectral flux*: this feature quantifies the changes in the spectral shape between consecutive frames. It is defined as

$$fe_i = \sum_{k=1}^K \left\{ \log_{10} |X_i(k)| - \log_{10} |X_{i-1}(k)| \right\}^2, \quad (4)$$

- *Bandwidth*: this feature determines the bandwidth of the signal, and is given by

$$lb_i = \sqrt{\frac{\sum_{k=1}^K (ce_i - k)^2 \cdot |X_i(k)|^2}{\sum_{k=1}^K |X_i(k)|^2}}, \quad (5)$$

Equation 5 gives the bandwidth in terms of spectral lines. To get the value in Hz, lb must be multiplied by the difference in Hz between two consecutive spectral lines.

- *Loudness*: this feature measures the signal intensity the way it is perceived by a human listener. The first step to calculate this feature is modeling the frequency response of outer and middle ears of an average person. Such response is given by [21]

$$W(k) = -0.6 \cdot 3.64 \cdot f(k)^{-0.8} - 6.5 \cdot e^{-0.6(f(k)-3.3)^2} + 10^{-3} \cdot f(k)^{3.6} \quad (6)$$

where $f(k)$ is the frequency in kHz, given by

$$f(k) = k \cdot d, \quad (7)$$

being d the difference between two consecutive spectral lines. The frequency response is used in the calculation of the loudness as a weighting function to emphasize spectral components for which the ear is more sensible and to attenuate the less audible ones:

$$L_i = \sum_{k=1}^K |X_i(k)|^2 \cdot 10^{\frac{W(k)}{20}}, \quad (8)$$

- *Inharmonicity*: this feature measures how much the partials deviate from the expected frequency, and is given by

$$in_i = \sum_{p=1}^P \left| \frac{f_{i,p} - ef_{i,p}}{ef_{i,p}} \right|, \quad (9)$$

where ef_p and f_p are, respectively, the expected and actual frequency of partial p . The number of partials was limited to 10.

- *Skewness*: this feature measures how asymmetric is the spectrum. It is the inharmonicity weighted by the spectrum; therefore, it is a different measure of the inharmonicity, where the relative importance of each component is taken into account, and is given by

$$sk_i = \sum_{p=1}^P \left| \frac{f_{i,p} - ef_{i,p}}{ef_{i,p}} \right| \cdot |X_{i,p}(k)|, \quad (10)$$

where $|X_p(k)|$ is the magnitude of the partial p .

- *Harmonic ratio*: this is actually a set of four features that measures the ratio between the magnitude of each one of the first four partials and the sum of all magnitudes:

$$h_{i,p} = \frac{|X_{i,p}(k)|}{\sum_{k=1}^K |X_i(k)|}, \quad (11)$$

where $p = 1, 2, 3, 4$.

- *Odd partials/even partials ratio*: this feature is the ratio between the three first odd partials and the three first even partials:

$$h_{i,p} = \frac{|X_{i,1}(k)| + |X_{i,3}(k)| + |X_{i,5}(k)|}{|X_{i,2}(k)| + |X_{i,4}(k)| + |X_{i,6}(k)|}, \quad (12)$$

- *Harmonic tendency*: this feature is given by the angular coefficient of the straight line that minimizes the squared distance to the partial magnitudes, as illustrated in Figure 4, where the asterisks indicate the partial magnitudes and ht is the harmonic tendency value according to the angle α .

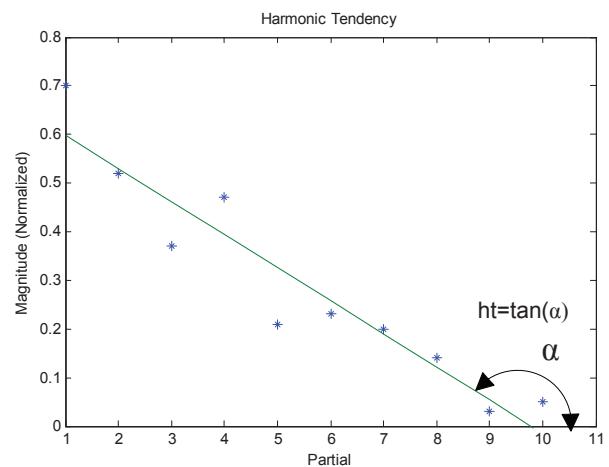


Figure 4 - Calculation of harmonic tendency.

3.2 Classification Procedure

The classification stage aims to use the information contained in the features as efficiently as possible. The scheme adopted here was successfully used before in [1].

where it was applied in the classification of musical signals into genres. It consists of a non-conventional pairwise classification procedure that compares all possible pairs of instruments and explores such information to improve the discrimination. A description of the strategy is presented next, and more details can be found in [4].

In order to the strategy to work properly, the four features that showed the best discrimination capabilities for a given pair of instruments were selected for that pair. This is because a great number of features cause, in general, a drop in the accuracy of the strategy. As a consequence, some features that showed a good global discriminating capability were more used than the less efficient ones. The proportion of times that each feature was used is presented in Section 4.

The selected features extracted for each frame are summarized along the duration of a given note by extracting three summary features: mean, variance and main peak prevalence, which is calculated according to

$$p_{fi} = \frac{\max[ft(i)]}{\frac{1}{I} \cdot \sum_{i=1}^I ft(i)}, \quad (13)$$

where $ft(i)$ corresponds to the value of feature ft in the frame i , and I is the number of frames into the note segment. This summary feature aims to infer the behavior of extreme peaks with relation to the mean values of the feature. High p_{fi} indicate the presence of sharp and dominant peaks, while small p_{fi} often means a smooth behavior of the feature and no presence of high peaks.

This procedure leads to 12 summary features, which are arranged into a vector. The reference vectors used to classify the instruments were determined through a training phase, as described next.

A. Training Phase

The algorithm was trained using 14 instruments present in the RWC database [1], which includes instruments from different manufacturers and played with different intensities. The result from the training is a set of reference vectors for each pair of instruments and for each possible note. The notes could be previously determined by some specialized algorithm, but in the tests presented here, the notes were manually determined by observing the magnitude spectra, in order to avoid error propagation from occasional algorithmic misestimates. In the test phase, the reference vectors that are closer to the feature vector extracted from a given frame in a Euclidian sense will determine the winner instrument, as described next.

B. Test Procedure

Figure 5 illustrates the final classification procedure of a signal. The figure was constructed considering a hypothetical division into 5 instruments (A, B, C, D and E). Nevertheless, all observations and conclusions drawn from Figure 5 are valid for the 14 instruments considered in this work.

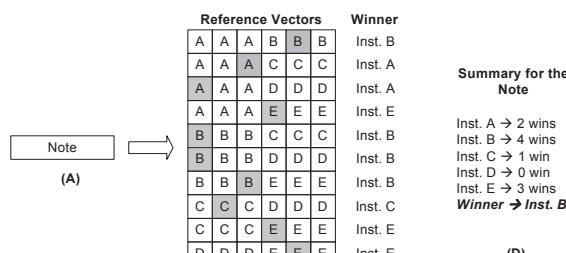


Figure 5 - Classification procedure.

As can be seen in Figure 5, the procedure begins with the extraction of the summary feature vector (Figure 5A). Such a vector is compared with the reference vectors corresponding to each pair of instruments, and the smallest Euclidean distance indicates the closest reference vector in each case (gray squares in Figure 5B). The labels of those vectors are taken as the winner instruments for each pair of instruments (C). In the following, the number of wins of each instrument is summarized, and the instrument with most victories is taken as the winner instrument (D); if there is a draw, all procedures illustrated in Figure 5 are repeated considering only the reference vectors of the drawn instruments; all other instruments are temporarily ignored.

Next section presents the results achieved using the features and classification procedure presented in this section.

4 RESULTS

The tests were performed using the 14 instruments used to train the algorithm. The database used in the tests was the University of Iowa Musical Instrument Samples [2], in order to avoid biased results. Table 1 shows the confusion matrix obtained in the tests, in percentages. The shaded cells indicate correct classifications.

Table 1 – Confusion matrix obtained in the tests.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	30	0	0	12	2	0	0	3	9	0	0	0	8	4
2	0	30	0	0	2	0	0	0	0	0	0	0	0	0
3	0	0	59	5	0	2	0	0	3	2	0	20	3	4
4	19	30	0	22	15	5	0	4	3	2	0	0	0	0
5	6	0	0	5	33	2	0	0	14	0	0	0	3	4
6	0	0	11	4	15	60	27	10	3	12	0	0	15	18
7	0	0	0	9	2	2	33	3	3	0	24	0	13	0
8	13	40	11	6	1	13	7	59	0	0	0	17	8	0
9	0	0	3	0	5	6	0	0	44	0	0	0	5	4
10	0	0	3	13	0	0	3	10	3	47	7	20	10	4
11	0	0	0	6	0	0	23	4	0	24	69	0	10	4
12	0	0	3	12	0	0	0	4	0	2	0	23	0	0
13	26	0	7	4	15	5	7	3	9	6	0	20	20	29
14	6	0	3	7	10	5	0	0	9	4	0	0	5	29

Legend:

- | | | |
|--------------------|--------------------|--------------|
| 1 - Alto Saxophone | 6 - Flute | 11 - Trumpet |
| 2 - Bass | 7 - Oboe | 12 - Tuba |
| 3 - Bassoon | 8 - Piano | 13 - Viola |
| 4 - Cello | 9 - Sop. Saxophone | 14 - Violin |
| 5 - Clarinet | 10 - Trombone | |

The overall accuracy of the algorithm was 40%. After the first tests, new ones were performed including some temporal information and a context-based classification correction, but the overall accuracy have reached only 45%, mostly thanks to piano, whose correct classifications have reached 84%. However, this performance is still too poor to an algorithm intended to be used for a wide variety of conditions. It is also worth noting that the wrong classifications are, in general, distributed over different classes of instruments (wind, string, etc.), meaning that the errors do not occur only among related instruments.

As commented before, only the four most effective features are used for each pair of instruments. Table 2 shows how many times each feature is used in the comparison between each pair of instruments.

Table 2 – Confusion matrix obtained in the tests.

Feature	Times Used
<i>Zero-Crossing Rate</i>	372
<i>Centroid</i>	186
<i>Bandwidth</i>	705
<i>Roll-Off</i>	778
<i>Spectral Flux</i>	14
<i>Loudness</i>	752
<i>Inharmonicity</i>	22
<i>Skewness</i>	181
<i>Harmonic Ratio 1</i>	1336
<i>Harmonic Ratio 2</i>	1038
<i>Harmonic Ratio 3</i>	745
<i>Harmonic Ratio 4</i>	426
<i>Odd/Even Partials Ratio</i>	11
<i>Harmonic Tendency</i>	786

As can be seen, most features are widely used, except spectral flux, inharmonicity and odd/even partials ratio. However, the data in Table 2 has to be taken into consideration in relative terms. One fact that draws immediate attention is how ineffective inharmonicity is. At a first analysis, this is very surprising, since in the study carried out in [13] this feature was by far the most successful. There are two factors that explain such a discrepancy: in [13], the authors used the mean of the inharmonicity instead of individual values for each frame; additionally, further analysis revealed that this feature would need far more training signals to become effective. This also happens with other features, always due to database limitations.

Those results confirm the previous statement that unfitting features, together with limited databases, reduce the effectiveness of short-term instrument classification strategies. Since a significant expansion of existing databases is too complicated from several points-of-view, the most promising solution would be developing new features capable to capture the essence of each instrument. Other possible direction would be finding alternatives to traditional spectral analysis that would emphasize the particular aspects of the instruments. A promising strategy would be applying an ear model to the signal in order to simulate the way humans perceive each sound, and then applying some kind of cognitive processing to extract the desired information (an approach of this kind can be found in [22]).

It is worth noting that even new features and new ways to treat the signals may not be enough to make the problem treatable. In this case, the conventional structure “feature extraction/classification procedure” may have to be replaced by some highly new and unconventional design.

The main objective of this paper was to provide a starting point to the process of thinking new ways to solve the problem of instrument identification in an efficient way.

5 FINAL REMARKS

This paper presented a critical view about the problem of instrument classification. It was shown that all proposals so far rely in relatively long excerpts of audio to perform the classification, and that this kind of approach has several problems that limit the use of those techniques in practical conditions. The solution to this problem would be to design

an algorithm able to classify small audio segments. The resulting high-resolution classification would make the algorithm robust to instrument intermittency and other problems associated to the use of longer excerpts.

Although promising, the feasibility of this approach is severely limited due to the extreme varying nature of the sounds generated by the instruments and also due to database limitations. Tests performed using 14 features and a pairwise-based classification scheme reveals how limited the current approaches are and provide clues to possible solutions.

There are promising ways to solve this problem, like the development of new features and the incorporation of new tools to emphasize particular signal characteristics (e.g. applying an ear model). The fuzzy nature of the problem may require that new unconventional strategies be developed to replace the widely used “feature extraction/classification procedure” scheme.

6 ACKNOWLEDGEMENTS

Special thanks are extended to FAPESP for supporting this work under Grants 04/08281-0 and 03/09858-6.

7 REFERENCES

- [1] M. Goto, *Development of the RWC Music Database*, Proceedings of the 18th International Congress on Acoustics, pp. 553-556, Granada, Spain, April 2004.
- [2] *University of Iowa Musical Instrument Samples*, <http://theremin.music.uiowa.edu/MIS.html>.
- [3] F. Opolko, J. Wapnick, *McGill University Master Samples*, <http://www.music.mcgill.ca/resources/mums/html>
- [4] J. G. A Barbedo and A. Lopes, *Automatic Genre Classification of Musical Signals*, EURASIP Journal on Advances in Signal Processing, vol. 2007, 12 pages, 2007.
- [5] J. C. Brown, *Cluster-based probability model for musical instrument identification*, Journal of the Acoustical Society of America, vol. 101, p. 3167, 1997.
- [6] J. C. Brown, *Computer identification of musical instruments using pattern recognition with cepstral coefficients as features*, Journal of the Acoustical Society of America, vol. 105, no. 3, pp. 1933-1941, 1999.
- [7] K. D. Martin and Y. E. Kim, *Musical instrument identification: a pattern-recognition approach*, Journal of the Acoustical Society of America, vol. 103, no. 3, pp. 1768-1779, 1998.
- [8] K. D. Martin, *Sound Source Recognition: A Theory and Computational Model*, Doctoral Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1999.
- [9] K. Kashino and H. Murase, *Sound source identification system for ensemble music based on template adaptation and music stream extraction*, Speech Communication, vol. 27, no. 3, pp. 337-349, 1999.
- [10] P. Herrera, X. Amatriain, E. Batlle, and X. Serra, *Towards instrument segmentation for music content description: a critical review of instrument classification techniques*, in International Symposium

- on Music Information Retrieval, 9 pages, Plymouth, MA, USA, October 2000.
- [11] J. C. Brown, O. Houix, S. McAdams, *Feature dependence in the automatic identification of musical woodwind instruments*, J. Acoust. Soc. Am., vol. 109, pp. 1064-1072, 2001.
- [12] A. Eronen, *Automatic musical instrument recognition*, Master thesis, Tampere University of Technology, Department of Information Technology, 69 pages.
- [13] G. Agostini, M. Longari, and E. Pollastri, *Musical instrument timbres classification with spectral features*, EURASIP Journal on Applied Signal Processing, vol. 1, no. 11, pp. 5-14, 2003.
- [14] A. Livshin and X. Rodet, *Musical instrument identification in continuous recordings*, in 7th International Conference on Digital Audio Effects (DAFX-4), Naples, Italy, October 2004.
- [15] P. Jincahitru, *Polyphonic instrument identification using independent subspace analysis*, in 2004 IEEE International Conference on Multimedia and Expo (ICME), pp. 1211-1214, Taipei, Taiwan, 2004.
- [16] B. Kostek, *Musical instrument classification and duet analysis employing music information retrieval techniques*, Proceedings of IEEE, vol. 92, no. 4, pp. 712-729, 2004.
- [17] S. Essid, G. Richard, and B. David, *Instrument recognition in polyphonic music*, in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'05), vol. 3, pp. 245-248, Philadelphia, Pa, USA, March 2005.
- [18] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, *Instrument Identification in Polyphonic Music: Feature Weighting to Minimize Influence of Sound Overlaps*, EURASIP Journal on Advances in Signal Processing, Vol. 2007, Article ID 51979, 15 pages, 2007.
- [19] A. Srinivasan, D. Sullivan, and I. Fujinaga, *Recognition of isolated instrument tones by conservatory students*, in Proc. International Conference on Music Perception and Cognition, pp. 17-21, Sidney, Australia, July 2002.
- [20] V.Y. Liepin'sh, *An Algorithm for Evaluating of a Discrete Fourier Transform for Incomplete Data*, Autom. Control and Comp. Sciences, vol. 30, no. 3, pp. 27-40, 1996.
- [21] E. Zwicker, H. Fastl, *Psychoacoustics, Facts and Models*, Berlin: Springer Verlag, 1990.
- [22] J. G. A. Barbedo, A. Lopes, *A New Cognitive Model for Objective Assessment of Audio Quality*, J. Audio Eng. Soc., Vol. 53, No. 1/2, January/February 2005.



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Automatic Estimation of Harmonic Complexity in Audio

José Fornari & Tuomas Eerola¹

¹Finnish Centre of Excellence in Interdisciplinary Music Research
Department of Music, University of Jyväskylä.
Jyväskylä, P.O. Box 35(M), Finland
{fornari,tuomas.eerola}@campus.jyu.fi

ABSTRACT

Music Complexity is here defined as the perceived entropy conveyed by music stimuli. Part of this information is given by musical chords, more specifically, by their structure (static) and progression (dynamic). We named this subset of music complexity as Harmonic Complexity (HC). A computational model able to automatically estimate HC in musical audio files can be useful in a broad range of applications for sound processing, computer music and music information retrieval. For instance, it could be used in large-scale music database searches or in adaptive audio processing models. It is rare to find in the literature studies about the automatic estimation of HC directly from audio files. This work studies several principles related to the perception of HC in order to propose an initial audio model for its automatic estimation. The prediction of these principles were compared with the data acquired in a behavioral experiment where thirty-three listeners rated one-hundred music excerpts. Lastly, a linear model with these principles is presented. The results are shown and described based on the comparison of the listeners' mean ratings with principles and model predictions.

0 INTRODUCTION

Harmonic Complexity (HC), as here studied, refers to the contextual musical feature that almost any listener can perceive and intuitively grade the overall amount of complexity conveyed by the harmonic structure of a musical stimuli. In [1], it is mentioned that, in music “subjective complexity reflects information content”, which agrees with the Information Theory perspective that relates complexity with the amount of entropy, or randomness, presented in the information source output. Therefore, in musical harmony, we assume that the amount of psychoacoustic entropy may correspond to the sensation of its complexity. In [2], it is suggested that four components seem to be related to HC: 1) Overall harmonic changing rate. 2) Harmonic changing on weak beats. 3)

Dissonant (ornamental) notes rate. 4) Harmonic distance of consecutive chords. We understand here that these principles can be roughly summarized as chord structures and chord progression. In [3] it is pointed out what it considered to be the two most important descriptors of music complexity 1) Coherence of spectral assignment to auditory streams and 2) Amount of variance in auditory streams. Once again, it seems to us that the first item is related to the amount of time-independent musical randomness (i.e. chord structure) and the second one, with its time-dependent information (i.e. chord progression).

Nevertheless, in our experiment collecting listeners' ratings for HC, it became clear how difficult it is for the listeners to have a common agreement upon their perception of HC. The listeners considered it related to a variety of musical features, such as: 1) chord composition

(from simple triads, major, minor, augmented, diminished, to complex clusters), 2) chord functions (fundamental, dominant, subdominant, etc.), 3) chord tensions (related to the presence of 7ths, 9ths, 13ths, etc.) and 4) chord progressions (diatonic, modal, chromatic, atonal). As these principles are based on the musical context, they lead to the development of a high-level descriptor, as described in [4], in order to properly predict the perception of HC in a computational model.

A high-level musical descriptor of HC is expected to deliver a scalar measurement that represents the overall HC of a musical excerpt. This perception is conveyed only in excerpts that are longer than the cognitive "now time" of music, around three to six seconds of duration [5].

The work here presented describes the study of several principles that seem to be related to the human perception of HC. We built a computational model with these principles that were then analyzed and improved, for the best prediction of HC that we could achieve, as shown in the experimental results.

1 COMPUTATIONAL MODEL DESIGN

The first step to create a computational model for HC was to investigate principles related to the amount of complexity found in chords structures and their progressions. Chords are related to the musical scale region where note events happen close enough in time to be perceived as simultaneous and therefore interpreted by the human audition as chords. In audio signal processing, this corresponds to the fundamental partials that coexist in a narrow time frame of approximately fifty milliseconds and are gathered in a region of frequency where music chords mostly occur. In musical terms this is normally located below the melody region and above the bass line. However, as the bass line also influences the interpretation of harmony, this one also had to be taken into consideration for the HC model.

The model starts by calculating the chromograms for these two regions (bass and harmony). A chromogram is a form of spectrogram whose partials are folded in one musical octave, with twelve bins, corresponding to the musical notes in a chromatic musical scale. We separate each region using pass-band digital filters in order to attenuate the effects of partials not related to the musical harmonic structure.

These two chromograms are calculated for each time frame of audio corresponding to the window-size of its lowest frequency. Therefore, each chromogram is presented in the form of a matrix with twelve lines, each one corresponding to one musical note from the chromatic scale, and several columns, corresponding to the number of time frames. We then tested three principles that initially seemed to be related to the HC perception, as they were created from the studies and evidences mentioned in the introduction of this work.

The first principle is what we named here as auto-similarity. This one measures how similar each chromogram column is with each other. This calculation returns an array whose size is the same as the number of columns in the chromogram. The variation of this array is inversely proportional to the auto-similarity. The rationale behind this principle is that auto-similarity may be proportional to the randomness of the chromogram and this one seems to be related to chord progression, one of the studied aspects of HC perception.

Next, these two chromogram matrixes are collapsed in time, which means that their columns were summated and their consequent twelve points arrays normalized. As the music excerpts rated in this experiment were all with five seconds of duration, collapsing the chromogram seemed reasonable once this is nearby the cognitive "now time" of music, however, for longer audio files, a more sophisticated approach should be developed. In this work, the collapsed chromogram array proved to efficiently represents the overall harmonic information in this particular small time duration. We then calculate two principals out of these two arrays, named here as: energy and density. Energy is the summation of the square of each array element. The idea was to investigate if the array energy was somehow related to the chord structure. The second principle, density, accounts for the array sparsity. The idea is that, for each array, the smaller the number of zeros between nonzero elements, the bigger is its density. The intention here was to investigate if the collapsed chromogram density is related to the chord structure. This came from the notion that, in terms of musical harmony, the closer the notes in a chord are, the more complex the harmony tends to be. Figure 1 shows the diagram for the calculation of these six principles.

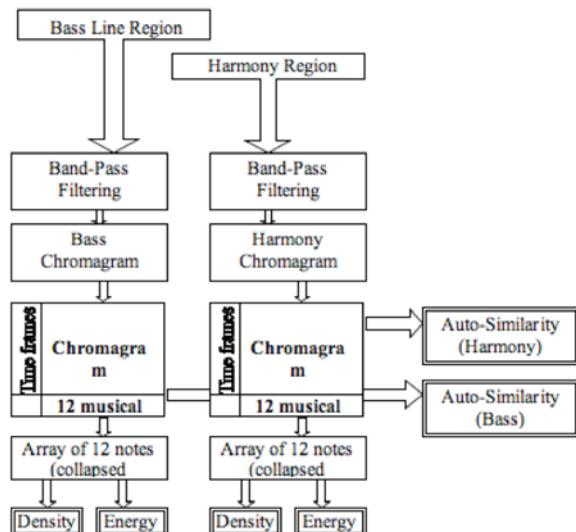


Figure 1 HC Principles Diagram.

This computational model was written and simulated in Matlab. As seen in Figure 1, this model calculated these three principles (auto-similarity, density and energy) for two chromograms representing the regions of frequency corresponding to the musical bass line and harmony. Results of our measurements are described in the next section.

2 HC LISTENERS RATING

For the evaluation of the computational model, a group of thirty-three students of music were asked to rate the harmonic complexity of one hundred music excerpts of five second of duration each. These excerpts were extracted from instrumental (without singing voice) movies soundtracks. The experiment was run individually using a computer and the order of the examples was randomized. The rating was done using a nine-point Likert scale [6] that goes from "no complexity" (at the leftmost side) to "extremely complex" (at the rightmost side).

The behavioral data was collected with a program that we developed, written in Pd (Pure Data) language. After the data being collected, it was analyzed, pruned, filtered, had outliers removed and its mean-rate given in the form of a function of time. This became the experiment ground-truth that was used to benchmark the HC principles and its model.

The listeners were also invited to share their comments about their ratings. Each one was asked to describe how difficult he/she had felt to rate HC and which musical features his/her rating was mostly based. Summarizing and grouping their opinions, they paid attention to: 1) harmonic change velocity and amount of change, 2) traditional chords structures versus altered chords, 3) predictability and sophistication of chord progression, 4) clarity of a tonal centre, 5) amount of instruments playing together and 6) dissonances. Overall, the listeners considered difficult to rate HC, especially for some excerpts with high amounts of musical activity, unclear chord structures and atonal music excerpts. As seen here, the listeners took into account different music features in order to rate HC. This was an important evidence for us to improve our initial computational algorithms for the principles. We then focused on the principles that are mostly related to the listeners' perception of HC.

3 HC PRINCIPLES EVALUATION

The model described in Figure 1 was used to calculate the principles of energy, density and auto-similarity for the chromograms of bass and harmony. Altogether, this resulted in six principles to be studied. Here, we calculated the prediction of these six principles for the same one hundred music excerpts rated by the listeners in the behavioral data acquisition experiment.

The correlation of each principle with the listeners mean-rating (the ground-truth) for the chromograms of Harmony and Bass are shown in the table below.

Table 1 Principles Correlation with Ground-Truth.

Principle	Harmony Chrom.	Bass Chromagram
density	0.56	0.51
auto-similarity	0.46	0.46
energy	0.27	0.35

As seeing in Table 1, the principle of Density for the Harmony chromagram was the one that presented the highest correlation with the ground-truth, followed by the principle of Auto-similarity, in the Bass chromagram.

We then created a multiple regression model with these six principles. This model presented a correlation coefficient with the ground-truth of $r = 0.61$. This yields to a coefficient of determination of $R^2 = 0.37$, thus explaining about 37% of data.

Figure 2 depicts the prediction of HC for this model (the circles) compared with the ground-truth rating (the bar-lines) for the same one hundred music excerpts. It is seeing in this figure that, with few exceptions, the predictions calculated by the model closely followed the human mean ratings.

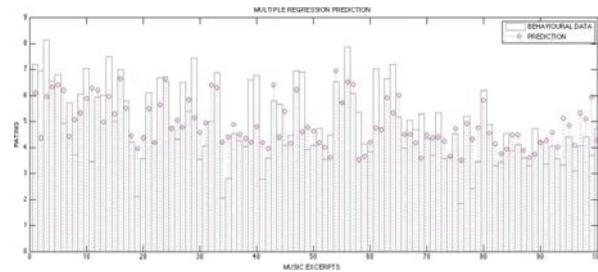


Figure 2 HC Model Prediction (dots) and Ground-Truth Rating (bars). Coefficient of Correlation: $r = 0.61$.

4 DISCUSSION

As seen in Table 1, the principle that presented the highest correlation in the harmony chromagram was density, followed by auto-similarity. In the bass chromagram, this was swapped. The highest descriptor was then auto-similarity, followed by density. This result seems to suggest that these principles are the best ones to describe harmonic complexity. These results support the initial premise that harmonic complexity is mostly related to chord structure (density) and chord progression (auto-similarity). It was interesting to observe, though, that auto-similarity presented the highest correlation in the bass chromagram, instead of in the harmony one, as we were initially expecting. This might be related to the particular set of music excerpts used in this experiment. Although this music excerpts were selected from a broad range of music styles, all are from sound-tracks without singing voice, which increase their similarity and may even place them into the same genre of music (i.e. movies soundtracks). If this is truth, further studies with a broader range of music genres may clarify this phenomenon.

The linear model, made with the multiple regression of these six principles, reached a correlation coefficient with ground-truth of $r = 0.61$. This is observed in Figure 2, on how close the prediction and the rating are. Although this is a fairly high correlation, further studies are needed in order to make sure that this multiple regression model is not over-fitted as a result of the large number of principles. Nevertheless, the high correlation achieved by the principles of density and auto-similarity in both chromagrams (harmony and bass), are sounding results that seem to point that these two principles are actually pinpointing meaningful aspects of HC perception.

5 CONCLUSIONS

This work presented an investigation of six principles that are related with HC perception. We compared the prediction of each principle and their linear model with the ground-truth obtained from the HC rating of one hundred music excerpts by thirty-three listeners. However, most of the music excerpts used in the experiment can be categorized as belonging to one musical genre; movies soundtracks. All excerpts were instrumental (without singing voices) and with five seconds of duration. We chose to work with short excerpts in order to be able to isolate static moments of harmonic complexity. As said before, this short-duration is slightly above the cognitive "now time" so was expected that the duration of these excerpts are enough to convey contextual features of music, such as HC, but without carrying the dynamic aspects of emotional prosody of music. For this introductory study this is desirable, as the prediction of

dynamic contextual aspects, such as HC, would necessarily have to consider the effects of memory (e.g. repeating similar patterns of music) and forgetfulness (e.g. the forgetting curve), that are enticing researches, but are beyond the scope of the present study.

As we mentioned in section 2, most of the listeners found difficult to rate HC. We believe that this is mostly related to the fact that HC can be given by static (chord structure) and dynamic (chord progression) features. Maybe a further study separating these two forms of harmonic complexity would be interesting. This would involve two sets of music excerpts, one with only static chords of different complexity and another with chords with similar structural HC but different degrees of progression HC.

Nevertheless, the results shown here are promising and we believe that this work can lead to better and more complete models for the automatic estimation of HC in audio files.

6 ACKNOWLEDGEMENTS

We would like to thank the BrainTuning project (www.braintuning.fi) FP6-2004-NEST-PATH-028570 and the Music Cognition Group of the University of Jyväskylä.

7 REFERENCES

- [1] Berlyne, D. E. (1971). *Aesthetics and psychobiology*. Appleton-Century-Crofts, New York.
- [2] Temperley, D. (2001). *The cognition of basic musical structures*. MIT Press, Cambridge, London.
- [3] Scheirer, E. D., Watson, R. B., and Vercoe, B. L. (2000). *On the perceived complexity of short musical segments*. In Proceedings of the 2000 International Conference on Music Perception and Cognition, Keele, UK.
- [4] Amatriain, X., Herrera, P. (2000). *Transmitting Audio Content as Sound Objects*. In Proceedings of the AES 22nd Conference on Virtual, Synthetic, and Entertainment Audio, Helsinki.
- [5] Leman et al., (2005). *Communicating Expressiveness and Affect in Multimodal Interactive Systems*. IEEE MultiMedia. vol. 12. pg 43 - 53. ISSN:1070-986X.
- [6] Likert, R. (1932). *A Technique for the Measurement of Attitudes*. Archives of Psychology 140: pp. 1-55.



Sociedade de Engenharia e Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Classificação Automática de Sons de Instrumentos Musicais usando Discriminantes Lineares

Jorge Costa Pires Filho¹, Paulo Antonio Andrade Esquef², Luiz Wagner Pereira Biscainho²

¹Instituto de Pesquisas da Marinha

Rio de Janeiro, RJ, Brasil

²PEE/COPPE & DEL/Poli, UFRJ

Rio de Janeiro, RJ, Brasil

jcpfilho@gmail.com, pesquef@yahoo.com, wagner@lps.ufrj.br

RESUMO

Este artigo apresenta um estudo da utilização de discriminantes lineares e transformações não-lineares do espaço de entrada num sistema para reconhecimento da assinatura sonora de instrumentos musicais. Utilizaram-se no trabalho amostras da base de dados MIS, da Universidade de Iowa. O sistema proposto é avaliado para uma seleção de escolhas de técnicas de pré-processamento de dados; formação do vetor de atributos; e classificação. Com o uso combinado de *Line Spectral Frequencies* (LSF) e Discriminantes Lineares, os desempenhos obtidos foram em torno de 92% e 86%, respectivamente, para a taxa de acerto da família do instrumento e do instrumento individualmente. Esses resultados são compatíveis com a faixa de taxas de acertos reportada na literatura.

0 INTRODUÇÃO

No atual contexto de reconhecimento de instrumentos musicais, ainda não há consenso quanto à melhor abordagem em sinais polifônicos (os quais apresentam simultaneamente sons de diversos instrumentos musicais). Atualmente, a maior parte dos estudos desta área contempla o caso monofônico, seja em notas isoladas, seja em trechos de música solo. Um breve levantamento em [6] destaca os seguintes trabalhos e resultados: Marques e Moreno [1] reportam taxas de acerto de 70% com um sistema que analisa segmentos de 0,2 s, utiliza *Linear Prediction Coding* (LPC), *Line Spectral Frequencies* (LSF), FFT e *Mel-Frequency Cepstrum Coefficients* (MFCC) para formar o vetor de características, e emprega modelos de misturas gaussianas (GMM) e *Support Vector Machines* (SVM) para classificar os trechos analisados de 9 instrumentos. Martin [2] usa um conjunto de características perceptuais derivadas de um correlograma lag-log para classificar notas isoladas de 27 instrumentos e reporta taxas de acerto de cerca de 86% para a família do

instrumento e cerca de 71% para o instrumento individualmente. Eronen e Klapuri [3] também usam um conjunto de características perceptuais para classificar notas isoladas de 30 instrumentos e reportam taxas de acerto de cerca de 94% e 85%, respectivamente, para família e instrumento individualmente. Agostini *et al.* [4] empregam somente características espetrais para classificar 27 instrumentos e reportam taxas de acerto de cerca de 96% e 92% para família e instrumento, respectivamente. Kitahara *et al.* [5] apresentam um classificador para tons de 19 instrumentos que utiliza uma distribuição normal de diversos parâmetros, dependente da frequência fundamental, obtendo taxas de acerto de cerca de 90% e 80% para família e instrumento, respectivamente. Finalmente, Krishna e Sreenivas [6] descrevem um sistema que usa LSF como características representativas de segmentos obtidos a partir de notas isoladas e modelos de misturas gaussianas para a classificação. Relatam ter obtido taxas de acerto de 95% e 90% para família e instrumento, respectivamente. Exceto para Brown *et al.* [7] e Marques e Moreno [1], todos os outros resultados

reportados se referem a sistemas classificadores que utilizam notas isoladas.

Uma das motivações dessa área de estudo são aplicações comerciais que visam a catalogar discotecas através de um processo automático, etiquetando cada música com a presença dos instrumentos musicais que a compõem, facilitando assim uma busca seletiva. Outras aplicações de interesse são a transcrição automática de música [8] e a codificação de áudio em alto nível, usando modelagem de fonte sonora [9].

A escolha de se abordar a classificação de instrumentos musicais a partir de notas isoladas nesse estudo pode ser justificada por diversos motivos. Primeiramente, ela pode ser adaptada tanto para classificar trechos de música monofônica quanto para outros sinais de áudio oriundos de uma única fonte. No mais, a identificação de instrumentos a partir de notas isoladas, apesar de não ser a mais adequada para resolver o problema na sua concepção mais geral (sinais de música contendo superposição no tempo e na frequência de vários instrumentos musicais), não é restritiva caso se queira identificar sinais que já tenham passado por um processo de separação de fontes. Pode-se assumir que o escopo do presente trabalho é identificar qual é o instrumento associado a um sinal que, tendo sido previamente separado, pertence a uma única fonte.

O fato de se usar nesse trabalho discriminante linear para a função de classificação tem como principal meta avaliar o desempenho de um algoritmo mais simples, mais robusto e mais rápido na fase de treinamento do que os classificadores mais elaborados propostos na literatura, tais como o SVM e o GMM.

1 METODOLOGIA

Uma das preocupações deste trabalho foi obter resultados que permitissem a avaliação comparativa da eficiência do uso da técnica de discriminação linear na identificação de instrumentos frente a soluções concorrentes. Assim, para se traçar uma avaliação de desempenho utilizaram-se como paradigmas os resultados apresentados por diversos autores, summarizados em [6]. Isso permite avaliar quão bom é o desempenho que se obtém com o uso do discriminante linear combinado com a forma de obtenção do vetor de características proposta neste artigo.

Para atacar o problema de reconhecimento da assinatura sonora de notas isoladas de instrumentos musicais, subdividiu-se o problema em módulos funcionais independentes. O sistema que os integra para formar o classificador é ilustrado na Figura 1.

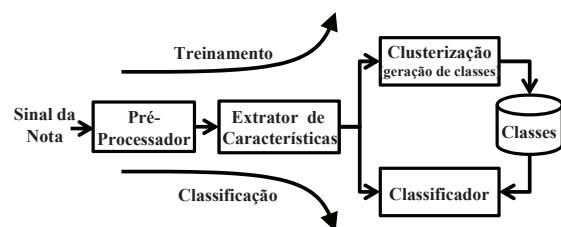


Figura 1: Sistema de Identificação de Instrumentos Musicais.

Como pode ser visto, o procedimento se divide em 4 blocos, de forma bem próxima ao modelo em [9]:

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

O Pré-processador é responsável pelo escalonamento dinâmico dos sinais e sua segmentação em trechos de interesse. O Extrator de Características é responsável pela obtenção do vetor de características que representa o sinal entregue pelo Pré-processador. O Gerador de Classes é responsável pela definição das classes em que serão agrupados os sinais da base de dados, tanto no estágio de treinamento como no de testes do classificador. E o Classificador é o módulo responsável por decidir sobre qual a classe a que pertence um dado sinal de teste, representado por seu vetor de características.

Durante a etapa de classificação, dois conjuntos distintos de sinais (amostras) foram utilizados: um conjunto de treinamento para o classificador e outro conjunto de teste para a avaliação da taxa de acerto.

2 BANCO DE DADOS

As amostras de sons de instrumentos musicais (*Musical Instrument Samples - MIS*) da Universidade de Iowa foram gravadas em uma câmara anecóica no "*Wendell Johnson Speech and Hearing Center*" na Universidade de Iowa [10] com os seguintes equipamentos:

1. Microfone Neumann KM 84;
2. Mixer Mackie 1402-VLZ;
3. Gravador DAT Panasonic SV-3800.

As únicas exceções foram os sons de piano, cuja gravação não-anecóica ocorreu em um pequeno estúdio.

Os sinais são monofônicos (com exceção do piano, gravado em estéreo), amostrados a 44,1 kHz e representados em 16 bits, e foram armazenados em formato AIFF.

Para cada instrumento, os sinais gravados englobam a tessitura usual do instrumento em escalas cromáticas tocadas em três níveis dinâmicos não normalizados: *pp*, *mf* e *ff*, ou seja, *pianissimo*, *mezzo forte* e *fortissimo*. Quando cabível, foram gravados estilos diferentes de execução, por exemplo, com e sem *vibrato*, com arco e *pizzicato*. Cada nota tem aproximadamente 2 segundos de duração, sendo imediatamente precedida e seguida de silêncio.

Dos sinais disponíveis na base de dados, este trabalho utilizou somente os listados na Tabela 1, apresentada na Seção 6.

3 PRÉ-PROCESSAMENTO DO SINAL

O módulo Pré-processador recebe o sinal de áudio correspondente a uma nota isolada e devolve um conjunto de trechos do sinal já pré-processados.

O objetivo principal do estágio de pré-processamento é segmentar o sinal em três trechos que correspondem ao ataque, sustentação e decaimento da nota. Para isso foram utilizados 2 limiares sobre um sinal de detecção, escolhido como a potência instantânea do sinal da nota gravada: o primeiro localizado a 10% e o segundo a 90% da média do sinal de detecção. Portanto, o ataque corresponderá ao trecho compreendido entre o instante de tempo em que o sinal de detecção ultrapassa pela primeira vez o primeiro limiar (10%) e o instante em que o sinal de detecção ultrapassa pela primeira vez o segundo limiar (90%). Já para obtenção do trecho de decaimento faz-se um

procedimento análogo, porém começando do final do sinal de detecção para seu inicio. O trecho de quase-estacionariedade, ou seja, de sustentação, é considerado como o intervalo de tempo entre o final do ataque e o início do decaimento. Na Figura 2, observa-se a envoltória do sinal de detecção correspondente a uma nota, com os dois limiares usados para a segmentação.

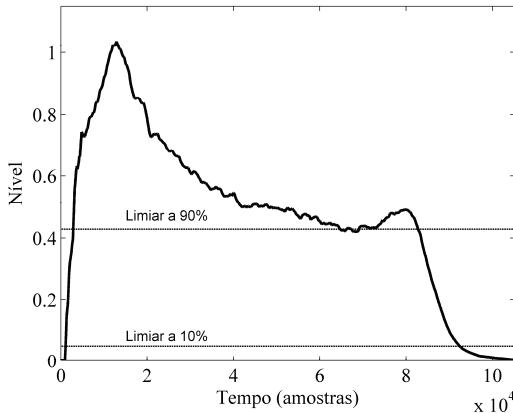


Figura 2: Envoltória do sinal de detecção (potência instantânea do sinal) contra os limiares de segmentação. Para melhor legibilidade, optou-se por substituir o sinal de detecção por sua envoltória.

4 EXTRAÇÃO E PROCESSAMENTO DE CARACTERÍSTICAS

O módulo Extrator de Características é responsável pela obtenção de um conjunto de características representativas do trecho de sustentação do sinal segmentado pelo módulo anterior. Seguindo a estratégia utilizada em [11] e [6], no presente estudo avaliaram-se características estatísticas (os momentos de segunda e terceira ordem) e parâmetros relacionados à análise espectral paramétrica do sinal (os coeficientes de predição linear ou as LSFs, para ordens predefinidas). O número de elementos do vetor de características será variado numa faixa arbitrariamente predefinida.

Antes da extração do vetor de características, o trecho de sustentação sofre um escalamento [11] de forma a assumir média zero e desvio-padrão unitário. Para esse procedimento, o momento de segunda ordem já teve de ser calculado. Calcula-se depois o momento de terceira ordem.

Em seguida, são estimados os coeficientes de predição linear (LPC) e aqueles associados aos polinômios das LSFs. Para um LPC de ordem N , o procedimento consiste em encontrar um conjunto de coeficientes a_k que minimiza o erro quadrático médio do seguinte preditor *forward*, aplicado em uma seqüência s_n :

$$\hat{s}_n = \sum_{k=1}^N a_k s_{n-k} + e_n. \quad (2)$$

O preditor da Equação (2) pode ser visto como a saída de um filtro gerador só-polos $H(z) = 1/A(z)$, com

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_N z^{-N}, \quad (3)$$

excitado por e_n . Dois polinômios, simétrico e assimétrico, associados às LSFs podem ser definidos a partir de $A(z)$, respectivamente, por

$$P(z) = A(z) + z^{-(N+1)} A(z^{-1}) \quad (4)$$

$$Q(z) = A(z) - z^{-(N+1)} A(z^{-1}). \quad (5)$$

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

Por fim, o vetor de características é formado pelo conjunto de coeficientes LSF ou LPC mais os momentos de ordens 2 e 3 do trecho de sustentação.

Vale ressaltar que redundâncias presentes no vetor de características definido acima poderiam ser eliminadas com a aplicação de técnicas de redução de dimensionalidade. Entretanto, isso fica além do escopo deste trabalho.

5 CLASSIFICADOR

Foram avaliados três tipos distintos de classificadores, a saber:

1. Vizinho mais próximo (VMP)
2. Máquina de Vetor Suporte (SVM)
3. Discriminante Linear Generalizado (DLG)

Como os classificadores SVM e DLG somente conseguem separar duas classes, optou-se por avaliar apenas uma forma específica de se obter o resultado no caso multiclasse.

A generalização para discriminação multiclasse adotada neste trabalho utilizou o procedimento um-contra-um (*one-against-one*) [2,6]: calculam-se P discriminantes, onde P representa o número de duplas possíveis no total de classes que estão sendo avaliadas. Uma dada amostra é testada segundo todos os P discriminantes, e o número de atribuições da amostra a cada classe é contabilizado. A amostra é classificada como pertencente à classe que recebeu mais votos.

5.1 Vizinho mais Próximo (VMP)

Este método estima a classe mais provável de uma dada amostra a ser classificada pela sua distância (segundo alguma métrica) a um conjunto de treinamento formado por amostras cujas classes são previamente conhecidas.. Percorre-se o conjunto de treinamento, calculando a distância de cada uma de suas amostras à amostra a classificar, em busca da que apresenta a menor distância. A classe atribuída à amostra sob teste será a desse ‘vizinho mais próximo’.

Neste trabalho arbitrou-se como métrica de distância a distância Euclidiana de ordem 2 entre a amostra \mathbf{X} e a amostra \mathbf{M}^j do conjunto de treinamento:

$$D_x^j = \sqrt{\sum_{i=1}^n (x_i - M_i^j)^2}, \quad (6)$$

onde:

x_i = elemento i do vetor de características da amostra \mathbf{X} .

M_i^j = elemento i do vetor de características da amostra \mathbf{M}^j do conjunto de treinamento \mathcal{M} .

5.2 Máquina de Vetor Suporte

Concisamente, uma SVM implementa discriminantes lineares (hiperplanos) no espaço obtido por uma

transformação não-linear do espaço de entrada. A Figura 3 ilustra a separação por discriminantes lineares.

Na sua forma tradicional, uma SVM diferencia uma classe, a positiva, de outra, a negativa, em um esquema binário de classificação. Para isso a SVM constrói um hiperplano que maximiza a margem de separação entre os exemplares positivos e os negativos. Esse objetivo é atingido através de uma abordagem baseada na Teoria Estatística de Aprendizagem [12,13], implementando aproximadamente o método de minimização do risco estrutural [12].

Apesar de utilizar discriminantes lineares, uma SVM não necessita, para efeitos de generalização, de classes linearmente separáveis. Tal propriedade se deve ao fato de a discriminação ser empregada num espaço de características já submetido a uma transformação não-linear. Esta operação pode ser justificada invocando-se o célebre Teorema de Cover [12], que afirma que padrões não-linearmente separáveis pertencentes a um dado espaço de características são, com alta probabilidade, linearmente separáveis num espaço de características transformado, desde que: (a) a transformação seja não-linear; (b) a dimensão do espaço transformado seja alta o suficiente.

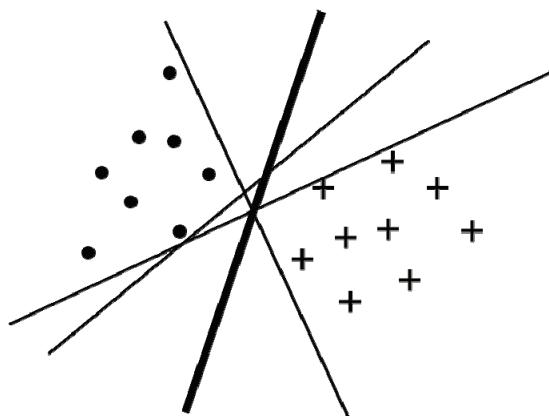


Figura 3: Hiperplano separador ótimo (linha grossa), de duas classes indicadas pelos marcadores • e +.

Para se obter a transformação não-linear do espaço de características requerida pelo SVM, utiliza-se o conceito de *kernel*. A idéia dá função *kernel* é aplicar operações no espaço de características ao invés de no espaço transformado, cuja dimensão é potencialmente maior. Assim, torna-se desnecessário calcular explicitamente o produto interno no espaço transformado para realizar as projeções, na tentativa de contornar o problema da dimensionalidade.

A SVM resolve um problema de programação não-linear que maximiza a margem entre os vetores transformados e o hiperplano separador, de modo a posicioná-lo de forma equidistante em relação aos chamados vetores-suporte.

Neste trabalho utilizou-se um *toolbox* SVM para o MATLAB (em dll) denominado SVM-KM e disponibilizado na Internet via licença geral de uso público da GNU. Em particular, foram utilizadas as funções *svmclasslib* e *svmlib*. O *kernel* utilizado nesse trabalho para o SVM foi o gaussiano.

5.3 Discriminante Linear Generalizado (DLG)

Analogamente a uma SVM, o DLG tenta encontrar um hiperplano que separe duas classes. O objetivo é achar a partir de um conjunto de treinamento o vetor \vec{w} que define um hiperplano separador, pela minimização do quadrado do erro de classificação

$$\varepsilon = t_{\vec{x}} - \tilde{y}(\vec{x}), \quad (10)$$

onde $t_{\vec{x}}$ (que pode assumir os valores $\{-1,1\}$) é a classe da amostra \vec{x} , e \tilde{y} é uma função estimadora da classe.

Assim, espera-se que se $\vec{w}^T \vec{x} > 0$, sendo T o operador transposição, a amostra \vec{x} pertença à classe 1; caso contrário, pertencerá à classe -1. Portanto, a classe da amostra \vec{x} é determinada por

$$y(\vec{x}) = \text{sign}(\vec{w}^T \vec{x}). \quad (11)$$

Para viabilizar a minimização por métodos que utilizam a direção do gradiente, substituiu-se a função sinal em (11) pela função tangente hiperbólica. A mudança se justifica, uma vez que esta, assim como a função sinal, possui sua imagem limitada pelos valores $\{-1,1\}$, sendo contudo totalmente diferenciável em seu domínio. Redefine-se, então, a classe da amostra \vec{x} como

$$\tilde{y}(\vec{x}) = \tanh(\vec{w}^T \vec{x}). \quad (12)$$

5.4 Transformação do Espaço de Características

Também foi investigado o efeito de uma extensão do espaço de características, consistindo na incorporação das potências, até um inteiro k , de cada parâmetro do vetor de características. Desta forma, se n é a dimensão do vetor de características associado a uma amostra, após a extensão kn será a nova dimensão tanto deste vetor de características transformado, agora definido por

$$\vec{x}_p = \begin{cases} [\vec{x}], & \text{se } k=1 \\ [\vec{x}^T \vec{x}^{2^T}]^T, & \text{se } k=2 \\ \vdots & \vdots \\ [\vec{x}^T \vec{x}^{2^T} \cdots \vec{x}^{\Psi^T}]^T, & \text{se } k=\Psi \end{cases}, \quad (13)$$

quanto do hiperplano separador, agora dado por

$$\vec{w} = [\vec{w}_1^T \quad \vec{w}_2^T \quad \cdots \quad \vec{w}_k^T]^T. \quad (14)$$

Tomou-se na Eq. (13) a liberdade de notar por \vec{v}^i a versão de \vec{v} com todos os seus elementos elevados à potência i .

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

$$\tilde{y}'(\vec{x}) = \tanh(\vec{w}^T \vec{x}_p). \quad (15)$$

Esta transformação não-linear foi usada em particular com o classificador DLG. Como se verá mais adiante, ela provocou um aumento na taxa de acerto das classes.

6 CLUSTERIZAÇÃO E FORMAÇÃO DAS CLASSES

Os tipos mais tradicionais de instrumentos musicais podem ser classificados de diversas formas, sendo uma das mais comuns a que se baseia no processo de produção de

som. O estudo dos instrumentos musicais designa-se por organologia, que foge ao escopo deste artigo.

Nesse trabalho, seguindo o procedimento de [6], quatro famílias de instrumentos foram definidas: flautas, palhetas, metais e cordas (FRBS – *Flutes, Reeds, Brass, and Strings*) agregando os instrumentos da Tabela 1 conforme a Tabela 2.

Tabela 1: Classe *Default* (Instrumento).

Instrumento	# Notas
Clarinete em Mi Bemol	119
Clarinete em Si Bemol	139
Fagote	122
Flauta	227
Flauta Baixo	102
Flauta Contralto	99
Oboé	104
Saxofone Contralto	192
Saxofone Soprano	192
Trombone Baixo	131
Trombone Tenor	99
Trompa	96
Violino	601
Violoncelo	668

Tabela 2: Classe FRBS (Família).

Família	Instrumentos
Flautas	Flauta, Flauta Baixo, Flauta Contralto
Palhetas	Clarinete em Mi Bemol, Clarinete em Si Bemol, Fagote, Oboé
Metais	Saxofone Contralto, Saxofone Soprano, Trombone tenor e Trompa
Cordas	Violino, Violoncelo

7 EXPERIMENTOS E RESULTADOS

Para realizar as simulações com o sistema classificador que utiliza o DLG, variou-se a quantidade dos coeficientes da parametrização LPC e LSF. Mais especificamente, foram predefinidas parametrizações com 8, 16 e 24 coeficientes.

Para cada uma das seis combinações possíveis, modificou-se também o grau de potenciação (conforme definido na Seção 5.4) para a transformação do espaço de entrada entre $k = 1$ e $k = 4$. Tais configurações de processamento foram adotadas para ambas as classes (*Default* e FRBS), perfazendo assim, um total de 48 simulações usando o DLG.

As demais simulações usando VMP e SVM não fizeram uso de transformação direta do espaço de características, conforme a Eq. (13). No entanto, para o classificador SVM foi usado um *kernel* gaussiano.

Todas as simulações, independentemente do classificador empregado, foram feitas usando o mesmo conjunto de treinamento e teste. Mais especificamente, aproximadamente 90% do total das amostras foram usadas para treinamento dos classificadores. O conjunto complementar restante foi usado para os testes de aferição de desempenho.

A seguir são apresentados os resultados das simulações realizadas, conforme as configurações experimentais supracitadas. Cada tabela contém uma descrição sumária

destes resultados, tanto para a classe FRBS quanto para a classe *Default*.

A Tabela 3 e a Tabela 4 apresentam os resultados associados ao desempenho do classificador DLG, para discriminar os instrumentos musicais nas classes FRBS e *Default*, respectivamente. São mostradas as taxas médias de acerto conforme a combinação do valor da potenciação k , usado na transformação no espaço de características (ver Eq. (13)), com o tipo e o número de coeficientes da representação paramétrica (LPC ou LSF) do sinal no trecho de sustentação.

Tabela 3: Classe FRBS - DLG. As maiores taxas de acerto em cada coluna são indicadas em negrito. A maior taxa de acerto é indicada na célula com fundo cinza.

DLG	$k=1$	$k=2$	$k=3$	$k=4$
LSF-8	57,84%	73,87%	78,40%	86,06%
LPC-8	63,41%	70,38%	71,78%	78,05%
LSF-16	64,11%	89,20%	88,15%	75,26%
LPC-16	66,90%	68,64%	75,96%	80,49%
LSF-24	71,78%	88,85%	91,99%	74,22%
LPC-24	74,91%	76,66%	81,53%	80,49%

Tabela 4: Classe *Default* – DLG. As maiores taxas de acerto em cada coluna são indicadas em negrito. A maior taxa de acerto é indicada na célula com fundo cinza.

DLG	$k=1$	$k=2$	$k=3$	$k=4$
LSF-8	73,40%	80,85%	81,91%	82,98%
LPC-8	67,02%	72,70%	78,01%	77,30%
LSF-16	78,01%	85,11%	82,62%	77,30%
LPC-16	72,70%	80,50%	79,08%	80,14%
LSF-24	80,50%	85,46%	85,82%	85,11%
LPC-24	77,30%	83,33%	80,14%	83,33%

Os melhores resultados alcançados pelos classificadores DLG, SVM e VMP para discriminação na classe FRBS, em um sistema que usou a parametrização LSF com 8, 16 e 24 coeficientes são mostrados, respectivamente, na Tabela 5, na Tabela 6 e na Tabela 7. Note que $k = 3$ foi adotado para o classificador DLG.

Tabela 5: Classe FRBS – LSF com 8 coeficientes.

	Flautas	Palhetas	Metais	Cordas	Taxa
DLG	78,57%	64,58%	92,96%	92,86%	86,06%
SVM	88,10%	85,42%	94,37%	92,86%	91,29%
VMP	85,71%	81,25%	88,73%	88,89%	87,11%

Tabela 6: Classe FRBS – LSF com 16 coeficientes.

	Flautas	Palhetas	Metais	Cordas	Taxa
DLG	75,87%	79,17%	90,14%	96,03%	89,20%
SVM	78,57%	89,58%	85,92%	95,24%	89,55%
VMP	88,10%	77,08%	95,77%	88,89%	88,50%

Tabela 7: Classe FRBS – LSF com 24 coeficientes.

LSF-24	Flautas	Palhetas	Metais	Cordas	Taxa
DLG	88,10%	68,75%	98,59%	98,41%	91,99%
SVM	78,57%	79,17%	87,32%	96,83%	88,85%
VMP	85,71%	75,00%	98,59%	81,75%	85,37%

A Tabela 8 apresenta as taxas médias de acerto alcançadas pelos classificadores avaliados, em um cenário de teste no qual a parametrização LSF com 24 coeficientes foi adotada para discriminar os instrumentos presentes na classe *Default*.

Tabela 8: Classe *Default* – LSF com 24 coeficientes.

LSF-24	DLG	SVM	VMP
Flauta Contralto	100,00%	100,00%	77,78%
Flauta Baixo	90,00%	100,00%	70,00%
Flauta	100,00%	86,36%	81,82%
Clarinete em Si Bemol	46,15%	53,85%	30,77%
Clarinete em Mi Bemol	36,36%	45,45%	45,45%
Oboé	40,00%	70,00%	20,00%
Saxofone Contralto	94,74%	100,00%	100,00%
Trombone Baixo	69,23%	53,85%	69,23%
Trompa	77,78%	77,78%	44,44%
Saxofone Soprano	63,16%	84,21%	89,47%
Trombone Tenor	100,00%	100,00%	88,89%
Violoncelo	98,48%	96,97%	98,48%
Violino	95,00%	95,00%	83,33%
Fagote	91,67%	91,67%	75,00%
Taxa Global	85,82%	87,59%	79,43%

8 DISCUSSÃO

Como se pode observar pela Tabela 3 e pela Tabela 4, a utilização de parametrização por LSF para formar o vetor de características no sistema de classificação com DLG apresentou um desempenho superior melhor ao observado quando do uso de LPC. Isto corrobora os resultados obtidos por Krishna e Sreenivas [6].

Também pela Tabela 3 e pela Tabela 4 constata-se que uso da transformação não-linear sobre o espaço de características ($k > 1$) tende a favorecer o aumento das taxas de acerto. Os melhores resultados para ambas as classes ocorreram para a combinação $k = 3$ e LSF-24. Isso sugere ser desnecessário usar potenciação de grau maior que 3 em associação com 24 LSFs. Contudo, esse ponto merece melhor investigação.

Ainda pela Tabela 3 e pela Tabela 4 verifica-se que, tanto com LPC quanto com LSF, as taxas de acerto tendem a crescer com o número de coeficientes. Portanto, pode-se especular que taxas ainda maiores de acerto possam ser alcançadas com a utilização de parametrização de mais alta ordem.

Com base nos dados mostrados nas Tabelas 5, 6 e 7, a análise comparativa dos desempenhos dos classificadores para discriminar famílias de instrumentos (classe FRBS) revela que o uso de LSF-16 implica nas mais altas taxas médias de acerto para o SVM e o VMP. Já o melhor

desempenho do DLG, que representa o melhor desempenho global, é alcançado quando se utiliza parametrização LSF-24. Nota-se que a diferença de desempenho entre o primeiro e o segundo (SVM com LSF-8) colocados é menor que 1%.

Ao se repetir a análise anterior para a discriminação de instrumentos (classe *Default*) observou-se que o uso de LSF-24 implica a obtenção dos melhores desempenhos de classificação para o DLG e o SVM. Como visto na Tabela 8, o SVM alcança a maior taxa média de acerto, sendo seguido pelo DLG com uma diferença de menos de 2%.

Ainda com referência à Tabela 8, verifica-se que as taxas de acerto obtidas pelo DLG (com $k = 3$) mostraram-se iguais ou superiores às do VMP e do SMV, respectivamente, para 78% e 57% dos instrumentos. Logo, pode-se alegar que, para a base de dados e às condições de teste descritas, o desempenho do DLG com $k = 3$ é similar ao do SVM com *kernel* gaussiano.

Na tarefa de discriminação de família de instrumentos, a mais alta taxa de acerto alcançada pelo DLG foi de aproximadamente 92%. Entretanto, o desempenho cai para aproximadamente 86% quando o objetivo é identificar os instrumentos. Esses resultados ficaram bem próximos daqueles obtidos por Krishna e Sreenivas [6], em que as taxas de acerto são de 95% e 90%, respectivamente, para a identificação das mesmas família e classes de instrumentos individuais.

Uma vantagem do emprego de DLG em relação ao SVM é a menor complexidade computacional daquele, fator que se manifesta principalmente na maior rapidez para a obtenção da convergência no estágio de treinamento do classificador.

9 CONCLUSÕES

Este artigo abordou o problema de identificação automática de sons de instrumentos musicais a partir de sinais acústicos correspondentes a gravações anecóicas de notas isoladas. Em particular, o foco de interesse foi avaliar o desempenho de um sistema classificador que utiliza transformações não-lineares do espaço de entrada que alimenta um classificador do tipo Discriminante Linear Generalizado (DLG). Para fins comparativos, os resultados obtidos são confrontados com os alcançados por sistemas classificadores do tipo SVM e Vizinho mais Próximo.

Na tarefa de discriminação da família de instrumentos definida no trabalho, a mais alta taxa de acerto alcançada pelo DLG foi de aproximadamente 92%. Entretanto, o desempenho cai para aproximadamente 86% quando o objetivo é identificar os instrumentos individualmente.

Os resultados acima ficaram bem próximos daqueles obtidos por Krishna e Sreenivas [6], que obtêm com um classificador GMM taxas de acerto de 95% e 90%, respectivamente, para a identificação das mesmas classes de família e instrumentos individuais. Assim pode-se concluir que o classificador DLG apresentou um desempenho satisfatório, visto que, dentro das configurações adotadas neste estudo, seus resultados ficaram próximos ou melhores que os classificadores concorrentes avaliados.

Por fim, uma vantagem do emprego do DLG em relação ao SVM é a menor complexidade computacional daquele, fator que se manifesta principalmente na maior rapidez para a obtenção da convergência no estágio de treinamento do classificador.

Aspectos associados às outras etapas do sistema completo de classificação, como a operação ao longo do tempo, a formação das classes etc., estão sendo investigados em conjunto com a classificação propriamente dita.

10 AGRADECIMENTOS

Os trabalhos de pesquisa de Paulo Esquef e Luiz Biscainho são financiados pelo CNPq, respectivamente, através de bolsas de Pós-Doutorado Júnior (Processo 152042/2007-5) e de Produtividade em Pesquisa.

11 REFERÊNCIAS

- [1] J. Marques and P. Moreno, *A Study of Musical Instrument Classification Using Gaussian Mixture Models and Support Vector Machines*, Cambridge Research Labs Technical Report Series CRL/4, 1999
- [2] K. D. Martin, *Sound Source Recognition: A Theory and Computational Model*, PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1999.
- [3] A. Eronen and A. Klapuri, “Music instrument recognition using cepstral coefficients and temporal features,” in Proc. of ICASSP, pp. 753-756, 2000.
- [4] G. Agostini, M. Longari and E. Pollastri, “Music instrument timbres classification with spectral features,” in Proc. of ICME, pp. 97-102, 2001.
- [5] T. Kitahara, M. Goto and H. G. Okuno, “Music instrument identification based on F0-dependent multivariate normal distribution,” in Proc. of ICASSP, pp. 421-424, 2003.
- [6] A. G. Krishna and T. V. Sreenivas, “Music instrument recognition: from isolated notes to solo phrases,” in Proc. of ICASSP, pp. 265-268, 2004.
- [7] J. C. Brown, O. Houix and S. McAdams, “Feature dependence in the automatic identification of musical woodwind instruments,” J. Acoust. Soc. Am., Vol. 109, No. 3, pp. 1064-1072, 2001.
- [8] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*, Springer, pp. 3-17, 2006
- [9] Kim, H.-G., *Introduction to MPEG-7 Audio*, John Wiley & Sons, Inc., New York, 2005.
- [10] MIS – Musical Instruments Samples of IOWA University.
- [11] J. C. P. Filho, D. B. Haddad, L. P. Calôba, “Classificação de padrões de varredura de radares,” in Anais do VIII Congresso de Redes Neurais, 2007.
- [12] S. Haykin, *Neural Networks: a Comprehensive Foundation*, Prentice Hall, 2nd. Ed., 1999.
- [13] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
- [14] Steve R. Gunn, *Support Vector Machines for Classification and Regression*, Technical Report - Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, Southampton University, 1998.



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Um modelo para a transcrição automática de melodias para partitura

Gabriel Simões Gonçalves da Silva¹ e Antonio Cezar de Castro Lima¹

¹Departamento de Engenharia Elétrica – Universidade Federal da Bahia
Salvador, Bahia, (40210-630), Brasil
gsimoes@gmail.com, acdcl@ufba.br

RESUMO

Técnicas de transcrição automática são hoje peças-chave no funcionamento de diversos sistemas de computação musical e processamento digital de áudio. Este artigo apresenta uma solução para a transcrição automática de melodias, não obrigatoriamente provenientes da execução de instrumentos monódicos ou gravadas em ambientes anecóicos, utilizando a partitura como notação musical. Além de detalhar o funcionamento dos módulos de análise e processamento, serão apresentados testes de transcrição com sinais sintéticos e reais e uma conclusão sobre os resultados obtidos.

1 INTRODUÇÃO

Mesmo já sendo estudada por mais de 35 anos, a transcrição automática de melodias continua atraindo o interesse de músicos, engenheiros e cientistas da computação. Em muito, esse interesse se deve a inexistência, até a presente data, de uma solução única, capaz de transcrever com perfeição todas as informações referentes à execução de uma melodia; e também a evolução tecnológica das últimas décadas que firmou a plataforma de áudio digital como o novo padrão de mercado.

O processo de transcrição musical pode ser definido como o ato de ouvir um trecho de música e extrair deste as informações necessárias para que se possa reproduzir o que foi ouvido [1], o que requer a identificação das notas tocadas, suas respectivas alturas, durações e intensidades, e a classificação dos instrumentos utilizados. Dessa forma, a transcrição automática é o processo computacional de análise de sinais de áudio digitalizados e consequente extração de informações simbólicas que possam ser relacionadas com estruturas musicais de mais alto nível [2].

Do surgimento das primeiras pesquisas da área até os dias de hoje, o leque de aplicações amparadas em sistemas de transcrição automática se expandiu, transcendendo o cunho puramente acadêmico. Exemplo disso é o surgimento freqüente de soluções baseadas em transcrição por computador como sistemas de acompanhamento musical, auto-afinação, performance interativa e *e-learning*, além dos sistemas de auxílio a transcrição e escrita musical.

Este artigo tem como objetivo apresentar um modelo computacional para a transcrição automática de melodias, não obrigatoriamente provenientes da execução de instrumentos monódicos, utilizando a partitura como notação musical. Através de técnicas de DSP (*Digital Signal Processing* – Processamento Digital de Sinais), o sistema proposto busca identificar o momento do início e fim da execução de cada nota presente nos sinais analisados, além de suas respectivas alturas, combinando ao final as diferentes informações obtidas para efetuar a segmentação da melodia. Já o processo de transcrição para a partitura é realizado através de heurísticas de análise, aproximação e conversão dos dados resultantes desta segmentação.

O conteúdo deste artigo está estruturado da seguinte forma: a Seção 2 apresenta um *overview* sobre o modelo proposto; as Seções 3 a 6 detalham cada um dos módulos que compõem o sistema e o funcionamento dos seus algoritmos; a Seção 7 descreve os testes efetuados e os resultados obtidos; e, por fim, a Seção 8 apresenta as conclusões sobre o projeto.

2 VISÃO GERAL

O processo de transcrição musical envolve mais do que a segmentação do áudio em notas e silêncios. Existem também detalhes de execução, como dinâmica, técnicas e efeitos, que precisam ser levados em consideração para que a reprodução do material transscrito seja o mais fiel possível em relação à melodia original.

Devido à subjetividade e também a dificuldade de estimativa precisa de tais informações, o escopo da proposta apresentada neste artigo se restringe a transcrição da altura, posição e duração das notas e pausas presentes nos sinais analisados, ignorando informações relativas à dinâmica e demais sinais de execução.

Por escolher a partitura como notação musical padrão, o sistema precisa ser capaz de converter os valores de posição e duração transcritos, de real para musical. Para isso, é necessário conhecer e tomar como base o andamento relativo à execução da melodia a ser transcrita, o que também foge ao escopo atual deste projeto. Sendo assim, cabe ao usuário definir o andamento a ser utilizado como referência para cada transcrição, além da clave, das unidades de tempo e compasso e de um valor de duração mínimo para notas e pausas.

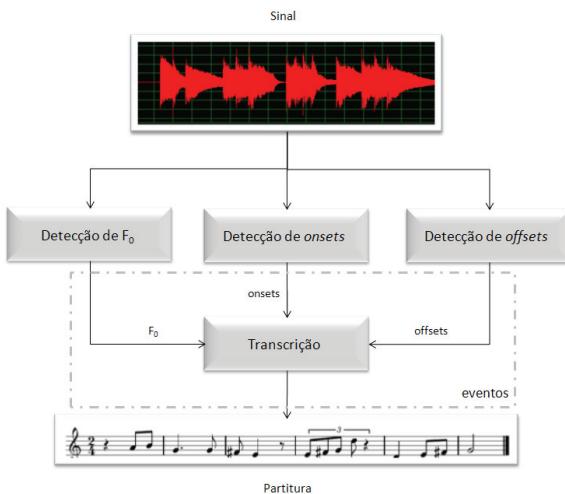


Figura 1 Descrição da arquitetura geral do sistema.

O sistema proposto foi implementado com base em uma arquitetura modular e desacoplada (figura 1). Três dos módulos exercem funções de análise e extração de informações, como descritos abaixo:

- Detecção de freqüências fundamentais (F_0): módulo responsável por identificar a freqüência fundamental em diferentes trechos do sinal, agrupando-as em eventos;
- Detecção de *onsets*: módulo responsável pelo levantamento do momento em que cada nota da melodia é executada;

- Detecção de *offsets*: módulo responsável por identificar o momento em que cada nota que precede uma pausa deixa de ser executada.

Um quarto módulo efetua a transcrição propriamente dita, tomando como base as informações levantadas pelos três módulos de análise acima apresentados e realizando as conversões e aproximações devidas. As próximas seções deste artigo detalham melhor o funcionamento destes.

Por fim, para possibilitar a visualização da partitura, o sistema formata os dados já transcritos de acordo com o padrão definido pelo *LilyPond* [3]; um software sob licença GNU, capaz de desenhar partituras a partir de descrições textuais.

3 DETECÇÃO DE F_0

A função básica do módulo de detecção de F_0 é identificar, em diferentes trechos do sinal, a altura das notas que compõem a melodia em análise. Devido à natureza do problema a ser resolvido, optou-se por trabalhar no domínio da freqüência através da subdivisão do sinal em janelas, aplicação da função de suavização de Hann e, por fim, cálculo da FFT. A partir deste ponto, o processo de detecção pode ser subdividido em 3 etapas: escolha e estimativa precisa de parciais, análise de evolução de parciais e detecção da freqüência fundamental.

Com o sinal já no domínio da freqüência, o espectro da janela passa por um processo de triagem. Visando minimizar a influência de ruídos e também de vazamento espectral nas etapas posteriores da detecção, apenas as magnitudes dos escaninhos relativos aos picos do espectro (máximos locais) com energia proeminente são mantidas, enquanto as demais são igualadas a zero.

A estimativa precisa de parciais está diretamente associada aos problemas de resolução espectral decorrentes do uso da FFT e à necessidade de análise de informações referentes às baixas freqüências do espectro. Espectros com baixa resolução espectral dificultam a diferenciação entre parciais de baixa freqüência enquanto as técnicas para o aumento da resolução espectral podem levar a problemas como aumento do custo computacional ou a análise do espectro de duas ou mais notas em apenas uma janela.

Como solução para o problema de resolução espectral, foi utilizado o algoritmo de interpolação quadrática proposto em [4], capaz de precisar a freqüência relativa a um máximo local através da sua magnitude e da maior magnitude entre os seus escaninhos adjacentes. Por meio dessa heurística aliada ao uso da janela de Hann é teoricamente possível detectar, com margem de precisão de até 3%, qualquer freqüência F em uma janela com duração maior ou igual ao dobro do período de F . Outras vantagens do uso desse algoritmo são o seu baixo custo computacional, com comportamento assintótico igual a $\theta(1)$, e a necessidade do cômputo da FFT de apenas uma janela para a sua aplicação.

Considerando casos perfeitos, onde todas as melodias seriam estritamente monofônicas, seria possível aplicar apenas a heurística de detecção de freqüência fundamental logo após a estimativa precisa de parciais para que a F_0 da janela fosse detectada. Porém, devido a fatores como reverberação e uso de instrumentos não monódicos para a execução de melodias, os espectros de notas seqüenciais podem, por um determinado período, se sobrepor criando trechos de polifonia. Como resultado, é possível gerar um

novo espectro composto por um conjunto de parcias referente às duas notas, porém semelhante à série harmônica de uma terceira nota mais grave, o que pode “confundir” os algoritmos de detecção de F_0 .

Para ilustrar o problema, imagine a execução de duas notas em seqüência: um G4 (392,00 Hz) e um D5 (587,36 Hz). Caso haja um pequeno momento de polifonia no inicio da execução da segunda nota, o conjunto de parcias formado pela sobreposição dos espectros poderia levar identificação de um G3 (196,00 Hz) como freqüência fundamental, já que todos os componentes das séries harmônicas das notas sobrepostas também fazem parte da sua série.

Buscando aumentar a robustez do módulo de detecção de F_0 , possibilitando também a análise correta de melodias não estritamente monofônicas, foi proposto neste projeto um sistema de acompanhamento de evolução de parcias no tempo (SAEPT). A função deste é identificar a sobreposição de notas executadas em seqüência e bloquear o efeito dos componentes espetrais remanescentes, permitindo a detecção correta da F_0 referente à nova nota.

O funcionamento do sistema de acompanhamento de evolução de parcias está diretamente ligado a heurística de detecção de freqüências fundamentais. Dessa forma, a descrição do funcionamento de ambos será feita em conjunto na seqüência desta seção.

Existem diferentes métodos para se identificar freqüências fundamentais através da análise espectral de janelas de um sinal. Nesta pesquisa, escolheu-se utilizar a heurística proposta em [5] devido aos ótimos resultados obtidos durante testes iniciais e também em projetos de pesquisa anteriores [6].

Visando diminuir o custo computacional e a complexidade do algoritmo de detecção de F_0 escolhido, foi proposta uma alteração na regra de cálculo da função Fração de Banda Crítica (ψ) [5]. Sendo i o i -ésimo parcial da série harmônica em análise, a fração de banda crítica será calculada de acordo com (1):

$$\psi(i) = \begin{cases} 1.000, & i = 1 \\ 0.999, & i = 2 \\ 0.998, & i = 3 \\ 0.997, & i = 4 \\ \Gamma[i] - \Gamma[i-1], & i > 4 \end{cases} \quad (1)$$

$$\Gamma(n) = \log_{2^3} \left(n * \sqrt{\frac{n+1}{n}} \right) \quad (2)$$

O intuito desta alteração é, através da diferenciação numérica dos valores de ψ (pesos) aplicados aos 4 primeiros parcias da série harmônica de cada F_0 candidata, permitir a correta identificação das freqüências fundamentais nos espectros analisados sem a necessidade de implementar a seleção final entre os candidatos de maior proeminência harmônica (eq. 12 e eq. 13 em [5]).

O módulo inicia o processo de identificação da freqüência fundamental de uma janela analisando o espectro composto pelos parcias já escolhidos e re-estimados, através da execução do algoritmo de detecção de F_0 e aproximação segundo o padrão da ISO ($A_4 = 440$ Hz). A partir do resultado obtido, o sistema de acompanhamento de evolução de parcias procura identificar vestígios de polifonia decorrentes da possível

sobreposição de duas notas executadas em seqüência. Freqüências fundamentais maiores ou iguais a da janela anterior são automaticamente consideradas corretas. Já freqüências mais baixas podem indicar uma sobreposição espectral com formação de um novo conjunto de parcias, semelhante ao da série harmônica de uma nota mais grave.

Para distinguir entre uma nova nota com F_0 mais baixa e uma detecção incorreta em um trecho de sobreposição espectral, o sistema precisa ser capaz de identificar a presença de componentes da série harmônica da nota anterior em meio ao espetro da janela atual, analisando a sua influência. Tomando como base o modelo de envelope ADSR [5], a pesquisa partiu do pressuposto que, a partir da etapa de estabilização (*sustain*), a energia de uma nota, e por consequência de seus componentes espetrais, só poderá se manter constante ou diminuir com o passar tempo. Porém, devido a influências externas como a reverberação, as magnitudes dos parcias podem apresentar pequenas variações positivas no tempo. Para diferenciar entre componentes da série harmônica da janela atual e parcias da série harmônica da janela anterior prolongados no tempo, foi definido um parâmetro chamado fator de oscilação γ . Sendo k o índice da janela, i o i -ésimo parcial do espetro da mesma, SH a série harmônica relativa à F_0 de cada janela já analisada e j o j -ésimo parcial da série, o método de seleção do sistema de acompanhamento de evolução de parcias foi definido de acordo com (3).

$$\left\{ \begin{array}{l} Janela[k][i]_{magnitude} = 0 \forall i, j | \\ ((SH[k-1][j]_{freq} * 0.7) < Janela[k][i]_{freq} < (SH[k-1][j]_{freq} * 1.3)) \\ \frac{Janela[k][i]_{magnitude}}{SH[k-1][j]_{magnitude}} < \gamma \end{array} \right. \quad (3)$$

A escolha do valor de γ deve ser de tal forma que os efeitos decorrentes de sustentação ou reverberação sejam removidos, porém todos os parcias que compõem a série harmônica da F_0 da janela atual, coincidentes com os da janela anterior, sejam mantidos.

Ao final do processo de seleção, o algoritmo de detecção de F_0 é novamente executado, analisando apenas o novo conjunto de parcias. Caso o resultado obtido seja menor ou igual ao da primeira detecção, estará constatado que não houve sobreposição espectral e o sistema manterá a primeira F_0 obtida como correta. Caso contrário, estará comprovada a influência da série harmônica da janela anterior sobre a nota atual, o que resultará na utilização da segunda freqüência fundamental detectada e no “bloqueio” dos parcias eliminados pelo sistema de acompanhamento de evolução de parcias.

O conceito de bloqueio de parcias tem como intuito anular o efeito decorrente da sobreposição de notas não apenas na janela em que foi detectada, mas até que a influência da nota já executada se dissipe. Quando um parcial é marcado como bloqueado, este fica impedido de compor o espetro de qualquer janela subsequente até que uma nova nota, de cuja série harmônica este faça parte, seja executada. Ao final de toda etapa de escolha e estimativa precisa, cada parcial selecionado é confrontado com a lista de parcias bloqueados. Para todo caso de identificação positiva, o sistema analisa a evolução da magnitude do parcial em questão ainda de acordo com (2). Caso a relação entre as magnitudes seja menor do que γ o parcial será mantido como bloqueado, tendo o valor da sua

magnitude atualizado na lista e sendo removido do espectro da janela. Caso contrário, o sistema entenderá que uma nova nota foi executada e que o parcial em análise faz parte da sua série harmônica, devendo assim ser removido da lista de bloqueados.

Ao finalizar a análise de todo o sinal e já de posse da lista de freqüências fundamentais de todas as janelas, o módulo de detecção de F_0 agrupa as seqüências de resultados de mesmo valor em eventos com início e duração definidos. A combinação destes eventos com os resultados dos demais módulos possibilitará a transcrição do sinal, como será apresentado na seção 6.

4 DETECÇÃO DE ONSETS

A detecção de *onsets* é por si só um tema de pesquisa de grande relevância na computação musical. Como resultado, existem hoje um grande número de trabalhos publicados na área, os quais apresentam soluções distintas para o problema com base em variadas heurísticas.

Infelizmente, devido às diferentes metodologias de avaliação utilizadas e a inexistência de bases de dados padrão para testes, não é incomum encontrar inconsistências ao analisar resultados apresentados em diferentes artigos de pesquisa sobre o assunto. Em [7] e [8], por exemplo, os autores buscam analisar e avaliar diversas heurísticas já propostas para solucionar o problema da detecção de *onsets*, ilustrando cada uma das abordagens estudadas, porém alcançando resultados divergentes em alguns testes.

Nesta pesquisa foram implementadas quatro diferentes abordagens para o cálculo da função de detecção (FD) de *onsets*, a serem escolhidas livremente pelo usuário no momento da transcrição: *Equal Loudness Contours* (eq. 13 em [8]), *Log Spectral Power* (eq. 6 em [8]), *Hi Frequency Content* e *Phase Deviation*. A decisão de não escolher apenas uma abordagem visa aumentar a robustez da solução proposta e também diminuir o risco de escolha de uma solução não tão apropriada, baseada apenas nos resultados apresentados nos artigos pesquisados.

Depois de calculada, a função de detecção (independente da heurística utilizada) passa por um *Envelope Follower* (EF) com um filtro IIR passa baixa de primeira ordem, ativado sempre que o sinal apresenta decaimento. O objetivo da utilização do *Envelope Follower* é suavizar a FD em trechos de decaimento pós *onsets*, reduzindo o número de máximos locais de pequena expressão com o intuito evitar a detecção de falsos positivos. A vantagem da aplicação do EF em detrimento a um filtro passa baixa comum é a preservação das informações de energia e de posição no tempo dos picos relativos aos *onsets*.

Por fim, foi implementado um sistema de *peak picking* baseado numa função de *thresholding* estática (valor fixo) ou dinâmica (eq. 21 em [7]), ficando também a escolha a cargo do usuário.

5 DETECÇÃO DE OFFSETS

No processo de transcrição musical, a percepção e a escrita correta das pausas são tão importantes quanto a das notas. Sendo assim, no caso dos sistemas de transcrição melódica é necessário identificar com precisão, além dos *onsets*, os *offsets* que assinalarão o instante em que cada pausa começou a ser executada.

A partir das pesquisas efetuadas para este projeto foram encontradas na literatura duas abordagens distintas para resolver o problema em questão: a aplicação de um *Silence*

Gate [9]; algoritmo que, ao analisar o envelope do sinal, busca identificar momentos em que a energia deste cai abaixo de um nível considerado silêncio; ou a inferência de que trechos do sinal sem F_0 determinada ocorrem devido à execução de pausas (silêncios) [10].

Como já apontado em [6], a abordagem do *Silence Gate* sozinha não se mostra adequada para sistemas de transcrição para partitura. Devido a fatores como sustentação natural dos instrumentos musicais (decaimento não instantâneo) e reverberação, os momentos assinalados como inicio das pausas tendem a ser posteriores aos *offsets* reais, o que culmina com o encurtamento da duração das mesmas e, consequentemente, com transcrições incorretas. Também por motivos de imprecisão temporal (fatores acima apresentados somados aos tamanhos da janela e do salto de análise), a identificação de *offsets* através de detectores de freqüência fundamental não se mostra como uma solução confiável para atingir o objetivo desta pesquisa.

Tentando solucionar a questão, foi proposta em [6] uma nova heurística para a detecção precisa do início de pausas, porém esta se mostrou ingênua, apresentando boa taxa de acerto apenas quando aplicada a sinais sintéticos (com envelope “perfeito”).

Partindo do pressuposto que um *offset* marca também o início do estágio de dissipação no envelope *ADSR* [6], foi proposto neste projeto um algoritmo capaz de identificar, com maior precisão do que o seu antecessor, o momento em que notas que precedem pausas deixam de ser executadas (equivalentes aos *offsets* no caso de melodias). Devido à natureza do seu funcionamento, este pode ser aplicado tanto a sinais sintéticos quanto reais, desde que o comportamento do envelope das notas siga o modelo *ADSR*.

Com base em estudos de comportamento de energia de sinais envelope com *offsets* assinalados manualmente, foram identificados três diferentes padrões de variação de energia característicos do início do estágio de dissipação (*offset*):

- 1) Energia em queda com forte acentuação devido ao fim da execução da nota, sendo o *offset* o ponto de início da mudança de tendência;
- 2) Energia estabilizada com mudança de tendência para queda abrupta, sendo o *offset* o ponto de início da mudança de tendência;
- 3) Energia em crescimento alternando rapidamente para queda, sendo o *offset* o máximo local que antecede o decaimento.

Inicialmente, calcula-se o envelope do sinal analisado através da retificação de onda completa, com posterior processamento por um filtro digital passa-baixa. O modelo de filtro escolhido foi o *Bessel* (IIR) de ordem 2, por apresentar *overshooting* mínimo no sinal resultante e atender as necessidades de atenuação. A freqüência de corte foi definida em 12,5 Hz através de testes empíricos.

Em seguida o sistema calcula a diferença de primeira ordem (FOD) do envelope com o intuito de ressaltar as variações de energia deste. O sinal resultante é novamente filtrado por um passa-baixa do tipo *Bessel* também de ordem dois, mas com a freqüência de corte um pouco mais baixa (10,0 Hz), suavizando ainda mais o sinal com o intuito de minimizar possíveis erros de detecção.

Na seqüência, a heurística do *Silence Gate* é aplicada ao sinal envelope apenas para possibilitar a identificação dos momentos de silêncio da melodia, base para identificação dos *offsets*. Para cada trecho de silêncio encontrado, o sistema inicia uma busca pelo *offset* associado através de uma análise de comportamento de energia no trecho equivalente do sinal diferença de primeira ordem.

Partindo do fim do silêncio (ponto de aumento de energia em que o *threshold* e o envelope se cruzam) e **retrocedendo** no tempo, o sistema busca identificar uma seqüência de estágios de energia característicos que possibilitem a identificação do *offset*. O primeiro deles é o mínimo local que antecede o ponto do inicio da busca, equivalente a maior variação negativa de energia durante o estágio de dissipação. Em seguida, o sistema “caminha” pelo sinal calculando a diferença de segunda ordem (SOD) ponto a ponto, buscando encontrar outros dois diferentes estágios: diminuição da aceleração de decaimento (SOD apresentando diminuição contínua) e crescimento da aceleração de decaimento (SOD apresentando aumento contínuo). O *offset* então pode ser identificado em dois diferentes momentos do processo de análise:

1. Sendo semelhante aos comportamentos 1 e 2, o *offset* será assinalado no ponto referente ao início do crescimento da aceleração de decaimento (quando a SOD parar de aumentar).
2. Sendo semelhante ao comportamento 3, o *offset* será assinalado no ponto onde o sinal envelope inicia a dissipação de energia (após a identificação do mínimo local, o ponto onde a FOD alcançar um valor maior ou igual a zero).

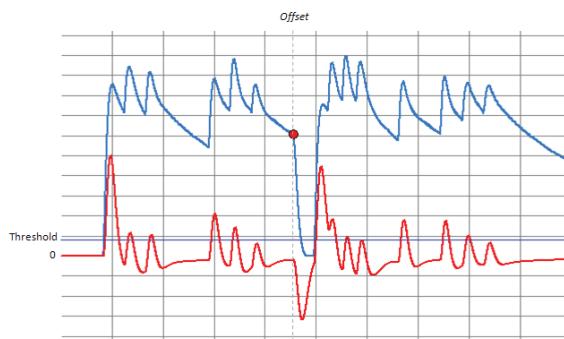


Figura 2 Exemplo de *offset* semelhante ao comportamento 1. Sinais envelope e diferença de primeira ordem. *Offset* marcado no ponto indicado.

6 TRANSCRIÇÃO

Para efetuar a transcrição do sinal analisado, foi implementado um módulo com o intuito de compilar todos os dados obtidos através das detecções de F_0 , *onset* e *offset*, produzindo ao final uma lista contendo os valores reais de altura (em Hertz) e a duração (em segundos) das notas transcritas.

Como ilustrado na figura 3, o sistema busca alinhar todos os eventos levantados pelos módulos de análise e, através da combinação entre eles, segmentar o sinal no tempo. Essa segmentação por si só é capaz de fornecer as durações das notas. As informações referentes à altura são retiradas apenas dos dados resultantes da detecção de F_0 .

Para efetuar a transcrição para partitura, o sistema precisa ser capaz de converter o valor real das informações de tempo e altura para estruturas musicais de mais alto

nível. Para as informações de altura, foi implementado um algoritmo de conversão, tomando como base a tabela padrão definida pela ISO.

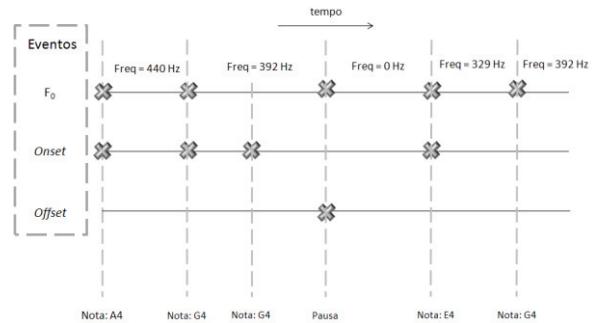


Figura 3 Segmentação do sinal baseado nos eventos levantados pelos módulos de análise.

O processo para a associação da duração de cada uma das notas às figuras a serem utilizadas na partitura se baseia no andamento e nas unidades de tempo e compasso, fornecidos pelo usuário. Para cada período equivalente a duração de um compasso, o sistema busca encontrar, de forma recursiva, a melhor relação entre as figuras e as durações.

Para cada nota da melodia pertencente ao compasso em análise, o sistema identifica as duas melhores aproximações entre a sua duração real e as durações fixas das figuras, guardando os valores referentes ao módulo das diferenças “real x aproximação”. Em seguida, para cada figura associada o sistema calcula todas as possibilidades de combinação com as aproximações das demais notas que a sucedem no mesmo compasso. Ao final, será escolhido como compasso transcrita corretamente aquele com a soma das durações das figuras mais próxima a duração real do compasso, sendo o critério de escolha para casos de empate a menor soma dos módulos das diferenças de aproximação.

7 RESULTADOS

Para avaliar o desempenho da proposta apresentada neste artigo cada um dos diferentes módulos de análise que compõem o sistema foi, inicialmente, submetido a testes preliminares, apenas com o intuito de garantir que o seu funcionamento estivesse de acordo com o comportamento esperado. Devido à natureza qualitativa dos testes e também ao foco desta seção, basta afirmar que os resultados obtidos apresentaram consistência quando confrontados com gabaritos de transcrição criados manualmente por músicos profissionais.

Após confirmar o bom funcionamento de cada módulo, o sistema de transcrição como um todo foi avaliado através da análise das transcrições de diferentes melodias. Devido à inexistência de um conjunto de critérios padrão para a avaliação de sistemas de transcrição automática, em especial para sistemas que utilizem a partitura como notação musical, os autores deste artigo decidiram utilizar os critérios listados abaixo para ilustrar os resultados alcançados durante os testes efetuados:

- Notas corretas: número de notas transcritas com altura, posição e figura de duração corretas;
- Notas com erro de altura: número de notas transcritas com altura incorreta;

- Notas com erro de duração: número de notas transcritas com altura correta, porém com figura de duração incorreta;
- Notas inexistentes: número de notas transcritas, porém inexistentes na melodia original;
- Notas não transcritas: número de notas presentes na melodia original, porém não transcritas;
- Notas com duração unida: número de notas cuja duração foi somada a da nota anterior;
- Pausas corretas: número de pausas transcritas com posição e figura de duração corretas;
- Pausas com erro de duração: número de pausas transcritas com figura de duração errada;
- Pausas inexistentes: número de pausas transcritas, porém inexistentes na melodia original;
- Pausas não transcritas: número de pausas presentes na melodia original, porém não transcritas.

Devido a não existência de uma base de dados padrão para a análise de sistemas de transcrição monofônica [5], e em especial composta apenas por melodias gravadas tomando como base andamentos fixos, foi criado um conjunto proprietário com o intuito de possibilitar a formalização e quantificação dos testes realizados. Este é composto por 308 notas e 54 pausas subdivididas em 22 melodias, todas gravadas com taxa de amostragem igual a 44100 Hz, podendo ser agrupadas da seguinte maneira:

- 12 melodias sintetizadas a partir de amostras de sinais reais de diferentes instrumentos (baixo e guitarra elétricos, violino, violoncelo, flauta, oboé, piano, órgão, piano Rhodes, saxofone, trombone e trompete);
- 10 melodias gravadas a partir da execução de instrumentos reais (guitarra e baixo elétricos);

Para a realização dos testes, os diferentes parâmetros de configuração do sistema foram ajustados da seguinte forma:

- Tamanho da janela do módulo detecção de F_0 : 2048 amostras;
- Tamanho do salto da janela do módulo detecção de F_0 : 1024 amostras;
- Fator de oscilação (γ): 2.5;
- Tamanho da janela do módulo detecção de *onsets*: 2048 amostras;
- Tamanho do salto da janela do módulo detecção de *onsets*: 256 amostras;
- *Threshold* do módulo de detecção de *onsets*: dinâmico, com piso fixo igual a 19% do valor máximo da função de detecção;
- *Threshold* do módulo de detecção de *offsets* (*Silence Gate*): estático em 8% do valor máximo do envelope.

O algoritmo escolhido para efetuar a detecção de *onsets* foi o *Equal Loudness Contours* com *threshold* dinâmico, devido à taxa de acerto apresentada nas análises iniciais, superior a dos demais algoritmos implementados.

Visando possibilitar uma melhor mensuração dos resultados obtidos através de comparações diretas, o mesmo conjunto de melodias e critérios de avaliação foi aplicado ao sistema *AudioScore Profesional 3* [11], software comercial especializado na transcrição de melodias para partitura. Para igualar as condições de execução dos testes e evitar distorções no processo de comparação entre os resultados, a função de autodetectação de andamento do *AudioScore* foi desabilitada, tendo sido a entrada de tais valores efetuada manualmente.

Seguem abaixo as informações obtidas através da análise e comparação entre transcrições geradas pelos sistemas e transcrições gabarito, criadas por um músico profissional devidamente treinado. A tabela 1 apresenta os resultados obtidos com as transcrições das melodias sintetizadas; a tabela 2 ilustra os resultados obtidos com as transcrições das melodias gravadas com instrumentos reais.

Tabela 1 Consolidação dos resultados das transcrições da base de dados de melodias sintetizadas

Melodias Sintetizadas		
Critérios de Avaliação	Modelo Proposto	AudioScore
Número de melodias	12	
Número total de notas	168	
Número total de pausas	24	
Total de notas transcritas corretamente	162	102
Total de notas transcritas com erro de F_0	0	0
Total de notas transcritas com erro de duração	4	57
Total de notas inexistentes transcritas	0	8
Total de notas não transcritas	1	3
Total de notas com duração unida	1	6
Total de pausas transcritas corretamente	24	12
Total de pausas transcritas com erro de duração	0	0
Total de pausas inexistentes transcritas	0	1
Total de pausas não transcritas	0	12

Tabela 2 Consolidação dos resultados das transcrições da base de dados de melodias provenientes da execução de instrumentos reais

Melodias provenientes da execução de instrumentos reais		
Critérios de avaliação	Modelo Proposto	AudioScore
Número de melodias	10	
Número total de notas	140	
Número total de pausas	30	
Total de notas transcritas corretamente	133	85
Total de notas transcritas com erro de F_0	0	1
Total de notas transcritas com erro de duração	7	53
Total de notas inexistentes transcritas	0	6
Total de notas não transcritas	0	1
Total de notas com duração unida	0	0
Total de pausas transcritas corretamente	29	10
Total de pausas transcritas com erro de duração	1	10
Total de pausas inexistentes transcritas	0	0
Total de pausas não transcritas	0	10

Um terceiro teste foi efetuado com o intuito de analisar a capacidade do sistema de transcrever melodias não estritamente monofônicas. Para isso, foram escolhidas duas das melodias sintetizadas (piano e flauta), transcritas pelo modelo proposto com perfeição no teste anterior, que posteriormente foram processadas por um algoritmo de reverberação “leve”, com o intuito de criar curtos trechos de polifonia no início da execução de cada nova nota. É importante ressaltar que a intenção deste teste não é provar que o sistema é capaz de transcrever sinais polifônicos ou gravados em ambiente com forte reverberação, mas sim analisar a sua robustez ao transcrever melodias que apresentem trechos de polifonia em decorrência da execução de instrumentos não monódicos ou de suave e curta reflexão sonora do ambiente.

As figuras 4 e 5 ilustram o desempenho do módulo de detecção de F_0 ao analisar primeiramente os sinais sem a adição de reverberação e, em seguida, com a adição de reverberação utilizando ou não o sistema de acompanhamento de evolução de parciais no tempo. Já a tabela 3 apresenta um comparativo entre os resultados das transcrições em ambos os casos.

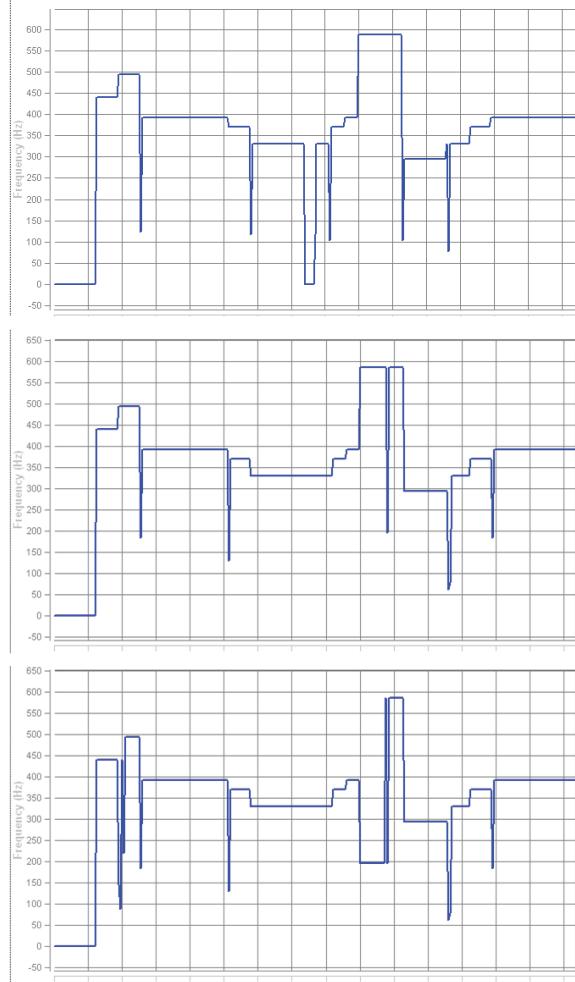


Figura 4 Resultado da análise da melodia de piano pelo módulo de detecção de F_0 : em cima a análise da melodia sem reverberação e utilizando o SAEPT; no meio a análise da melodia com reverberação e utilizando SAEPT; em baixo a análise da melodia com reverberação e não utilizando o SAEPT.

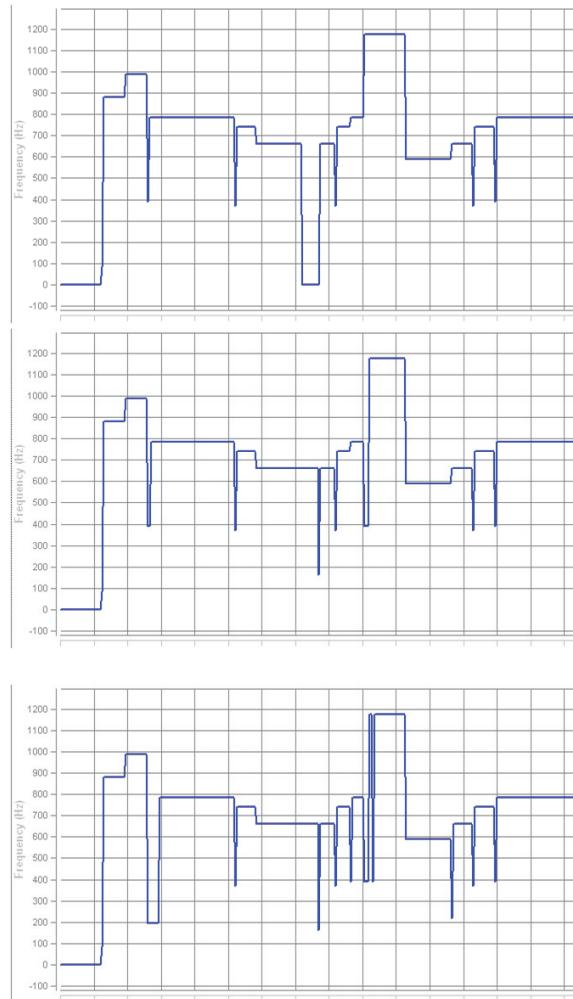


Figura 5 Resultado da análise da melodia de flauta pelo módulo de detecção de F_0 : em cima a análise da melodia sem reverberação e utilizando o SAEPT; no meio a análise da melodia com reverberação e utilizando SAEPT; em baixo a análise da melodia com reverberação e não utilizando o SAEPT.

Tabela 3 Consolidação dos resultados das transcrições da base de dados de melodias com reverberação sintética

Melodias com adição sintética de reverberação		
Critérios de avaliação	com SAEPT	sem SAEPT
Número de melodias	2	
Número total de notas	28	
Número total de pausas	4	
Total de notas transcritas corretamente	28	25
Total de notas transcritas com erro de F_0	0	0
Total de notas transcritas com erro de duração	0	3
Total de notas inexistentes transcritas	0	2
Total de notas não transcritas	0	0
Total de notas com duração unida	0	0
Total de pausas transcritas corretamente	4	4
Total de pausas transcritas com erro de duração	0	0
Total de pausas inexistentes transcritas	0	0
Total de pausas não transcritas	0	0

Ao comparar os primeiro e último quadros das figuras 4 e 5, é possível perceber que a adição de reverberação foi capaz de criar vales (erros de detecção) em alguns trechos referentes ao início de notas devido a polifonia formada pela sobreposição temporal destas em relação a suas respectivas antecessoras. Considerando esses trechos, a aplicação do SAEPT não se mostrou suficiente para eliminar por completo os efeitos provocados na capacidade de análise do módulo detector de F_0 , mas conseguiu atenuá-los melhorando a taxa de acerto das transcrições e, consequentemente, aumentando a robustez da solução proposta (vide tabela 3). Os autores deste artigo compreendem a necessidade de se efetuar testes mais conclusivos em relação aos efeitos do SAEPT na detecção de freqüências fundamentais em janelas seqüenciais, porém acreditam com base nos resultados das análises preliminares que, no geral, o seu uso no pior caso leva a resultados semelhantes aos da sua não utilização.

Ainda analisando as figuras 4 e 5, é possível perceber também que após a adição de reverberação os silêncios presentes nas melodias originais, quando precedidos por notas, foram sobrepostos pela reverberação das mesmas. Para esses casos é importante destacar que, mesmo neste cenário, o algoritmo de detecção de *offsets* proposto neste projeto se mostrou capaz de identificar o momento de início das pausas com precisão suficiente para permitir a transcrição correta dos novos sinais.

8 CONCLUSÃO

Através deste artigo, um modelo para a transcrição automática de melodias (não obrigatoriamente estritamente monofônicas) para partitura foi apresentado. Com base nos resultados obtidos durante os testes relatados na seção 7, pode-se afirmar que a proposta descrita alcançou o objetivo inicial da pesquisa, atingindo um índice de acerto que superou as expectativas dos autores. Em comparação direta com o sistema de transcrição *AudioScore*, o modelo proposto apresentou resultados iguais ou superiores em todos os critérios de avaliação aos quais estes foram submetidos.

Em consequência do uso da heurística proposta em [5] no módulo detector de F_0 , o escopo de transcrição do sistema foi limitado a instrumentos cujas notas apresentem afinação estável, o que a princípio exclui fontes sonoras como a voz humana. Porém, devido à arquitetura modular deste, outros algoritmos podem ser implementados em paralelo ao já existente, permitindo a ampliação do escopo de aplicação da solução e o consequente aumento da sua robustez.

As novas heurísticas propostas neste projeto, tanto para a detecção de *offsets* como para a transcrição para a partitura, se mostraram adequadas e eficientes. Em comparação direta com [6], a heurística de detecção de *offsets* apresenta como melhoria a sua alta taxa de acerto tanto para sinais sintéticos como para sinais reais, apresentando uma precisão muito superior a das demais propostas estudadas em todos os testes realizados. Já a heurística de transcrição para partitura apresentou ganho principalmente nas transcrições de sinais reais onde as durações das notas executadas não são perfeitas, o que muitas vezes leva a transcrições incorretas quando utilizados sistemas de aproximação direta.

A necessidade de ajuste manual de parâmetros a cada transcrição pode ser considerada como o ponto fraco do modelo proposto. Dessa forma, os autores deste artigo já

estão trabalhando em soluções para a automatização da definição de andamento e de níveis de *threshold*, deixando para o usuário apenas a entrada de informações referentes aos demais parâmetros musicais.

9 AGRADECIMENTOS

Os autores deste artigo agradecem ao CNPQ – Conselho Nacional de Desenvolvimento Científico e Tecnológico - pelo apoio dado por este no desenvolvimento desta pesquisa.

10 REFERÊNCIAS

- [1] Keerthi Nagaraj. *Toward automatic transcription – pitch tracking in polyphonic*. EE381K– Multidimensional digital signal processing, 2003.
- [2] Eric Scheirer. *Extracting expressive musical performance information from recorded music*. Master thesis, MIT, 1995.
- [3] LilyPond. *Music Notation for Everyone*. Disponível em <www.lilypond.org>. Último acesso em 27/01/2008.
- [4] Thomas Grandke. *Interpolation algorithms for discrete Fourier transforms of weighted signals*. IEEE Transaction on Instruments and Measurements. [S.I.], v.32, nº 2, p. 350-353, 1983.
- [5] Adriano Mitre, Marcelo Queiroz, Regis Faria. *Accurate and Efficient Fundamental Frequency Determination from Precise Partial Estimates*. AES Brasil, São Paulo, 2006.
- [6] Gabriel Simões, Allan Freitas, Hercules Souza. *Desenvolvimento de um sistema computacional de transcrição de melodias monofônicas para partitura*. WCOMPA – XXVI Congresso da SBC, Campo Grande, 2006.
- [7] Juan Pablo Bello, et all. *A Tutorial on Onset Detection in Music Signals*. IEEE Transaction on speech and audio processing, vol. 13, nº 5, 2005.
- [8] Nick Collins. *A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions*. 118th AES Convention, Barcelona, 2005.
- [9] Paul Brossier, Juan Bello, Mark Plumley. *Fast labelling of notes in musical signals*. Centre for Digital Music, Queen Mary College, Londres.
- [10] Giuliano Monti, Mark Sandler. *Monophonic transcription with autocorrelation*. Department of Electronic Engineering, King's College, Londres.
- [11] Neuratron. *AudioScore Professional 3*. Disponível em <www.neuratron.com/audioscore.htm>. Último acesso em 17/04/2008.

Sessão 5

Análise, síntese e sistemas para computação musical
(Analysis, synthesis and computer music systems)



Sociedade de Engenharia de Áudio Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

AURAL: ambiente interativo aplicado à sonificação de trajetórias robóticas

Artemis Moroni¹, Josué Ramos¹, Sidney Cunha¹ e Jônatas Manzolli²

¹Divisão de Robótica e Visão Computacional do Centro de Pesquisas Renato Archer
(DRVC/CenPRA)

Campinas, São Paulo, 13069-901, Brasil

²Núcleo Interdisciplinar de Comunicação Sonora – NICS/UNICAMP
Campinas, São Paulo, 13091-970, Brasil

Artemis.Moroni, Josue.Ramos, Sidney.Cunha@cenpra.gov.br, Jonatas@nics.unicamp.br

RESUMO

Este artigo descreve a pesquisa desenvolvida no projeto AURAL e sua implementação, que propõe a construção de um ambiente computacional para controlar a interação de informação sonora, visual e robótica. Em síntese, o processo culmina com a criação de trajetórias para robôs móveis que se integram com o controle visual e sonoro. As trajetórias, aqui denominadas de coreografias, poderão ser experimentadas e apresentadas *in loco* ou à distância, através de laboratório de acesso remoto pela internet. O resultado é um ambiente robótico interativo com as características funcionais do ambiente JaVOX, sobre o qual o AURAL está sendo construído.

0 INTRODUÇÃO

O desenvolvimento de sistemas que associam o comportamento de robôs móveis ou também o movimento humano com eventos sonoros tem sido estudado nos últimos anos. Camurri et al. [1] apresenta um sistema vinculado ao WebEye onde o processo de sonificação é associado ao movimento corporal de um bailarino. Nesta pesquisa associa-se a Teoria do Movimento de Laban com atributos sonoros. Manzolli et al. [2] criaram o sistema Roboser, desenvolvido com o pequeno robô Khepera, onde eventos sonoros foram associados ao controle adaptativo denominado de DAC (Distributed Adaptive Control). Posteriormente, o Roboser foi utilizado na composição da paisagem sonora interativa para a instalação “Ada: intelligent space”. Nesta pesquisa, o comportamento humano e um conjunto de sistemas robóticos foram

utilizados para criar um espaço que respondia ao comportamento humano e cuja expressão sonora visual foi denominada de “emoções sintéticas” [3].

Os sistemas robóticos, além de fornecerem trajetórias ou comportamentos para serem mapeados em sons, podem ser guiados por sinais sonoros. Neste caso o objetivo é estudar o controle do posicionamento do robô através de estímulos sonoros como apresentado por Murray [4].

Semelhante aos sistemas desenvolvidos por Manzolli & Verschure e Murray, o AURAL organiza uma sequência de eventos sonoros a partir do comportamento do robô móvel no espaço físico. Diferente destes dois sistemas, o mecanismo de sonificação é feito através da associação da trajetória do robô com curvas na *pad* de controle de um ambiente de computação evolutiva, denominado JaVOX. A informação recebida do robô modifica o comportamento da

função de fitness, redimensionando a adequação de eventos sonoros à trajetória em tempo real.

O segundo aspecto é a relação de interação entre o robô móvel *Nomad* e outros robôs que será usada para modificar os controles de performance sonora do JaVOX [5, 6]. A interação de parâmetros físicos e a presença dos corpos mecânicos dos robôs tem o potencial de gerar uma complexa cadeia interacional. Neste sentido os processos estudados pelo AURAL são vinculados ao conceito de auto-organização, utilizado como paradigma composicional, como discutido em Manzolli [7].

O ambiente AURAL, cujo diagrama é apresentado na Figura 1, reúne: a) o uso do ambiente JaVOX como mecanismo de interação entre as trajetórias e o processo de sonificação; b) um sistema de visão omnidirecional que se utiliza de um espelho esférico e câmera para localizar o robô no espaço; c) um módulo supervisor que recebe a trajetória e supervisiona o robô, para que este a percorra satisfatoriamente.

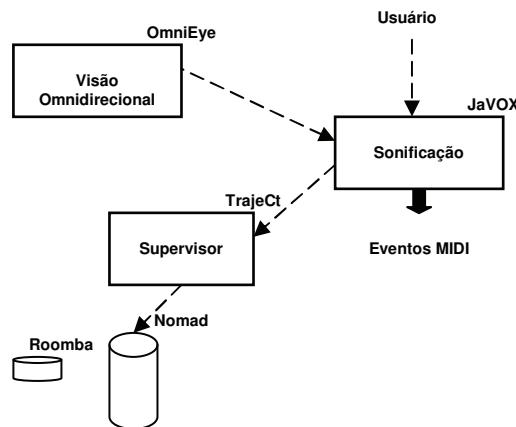


Figura 1 Diagrama de processos do ambiente AURAL

1 SONIFICANDO TRAJETÓRIAS

O sistema de sonificação do AURAL está estruturado na aplicação de Computação Evolutiva para geração de eventos sonoros. Neste contexto, a representação do protocolo MIDI foi tomada como informação genotípica como no desenvolvimento original do VOX POPULI [8, 9]. Este ambiente, inicialmente desenvolvido em Visual Basic foi posteriormente traduzido para a linguagem Java, resultando no sistema JaVOX. As funcionalidades descritas nesse trabalho estão presentes em ambos os ambientes, VOX POPULI e JaVOX.

No AURAL, as trajetórias geradas pelos robôs são associadas às estruturas sonoras do JaVOX. A característica sonora de ambos os ambientes, o JaVOX e o VOX POPULI, é descrever o processo de sonificação através de populações de *clusters* e acordes criando uma nova sonoridade a cada passo do processo.

Em ambos os ambientes, VOX POPULI e JaVOX, uma área de controle (pad) da interface interativa habilita o usuário a desenhar curvas num espaço de fase, associando a cada uma delas trajetórias que guiam a produção sonora. Na Figura 2 são mostradas curvas desenhadas pelo usuário na interface gráfica do ambiente VOX POPULI e a sequência sonora resultante.

As curvas vermelhas estão associadas aos parâmetros melódico (**mel**), na componente *x*, e octave (**oct**), ou

intervalo de vozes, na componente *y*. A Figura 3 apresenta, no detalhe, a associação da curva vermelha com os parâmetros *melódico* e *octave*. Já as curvas azuis estão associadas com os parâmetros biológico (**bio**) na componente *x* e ritmo (**rhy**), na componente *y*.

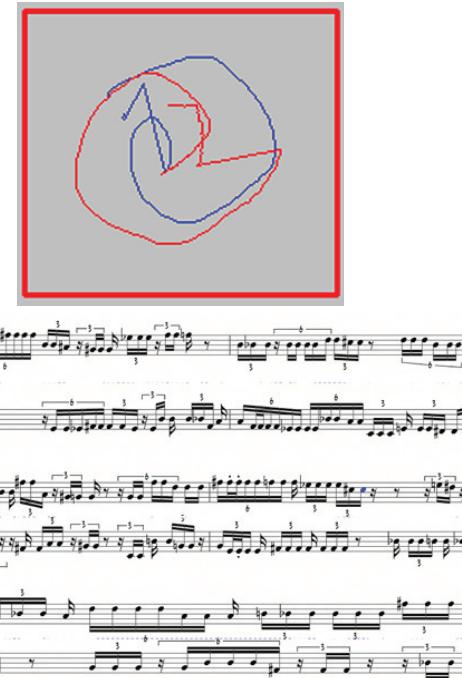


Figura 2 Sequência sonora resultante das curvas desenhadas acima no pad interativo do ambiente VOX POPULI

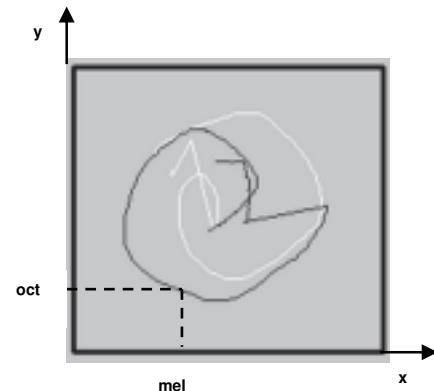


Figura 3 Parâmetros de controle (*mel*, *oct*) associados à curva vermelha, em destaque.

A Figura 4 apresenta a interface do ambiente JaVOX., onde as linhas desenhadas pelo usuário na interface interativa direcionam a produção em tempo real de uma seqüência sonora. Assim como no VOX POPULI, o JaVOX associa a cada uma delas os parâmetros de controle da interface.

No AURAL, através de uma facilidade similar ao pad interativo, as trajetórias são desenhadas e transmitidas para um robô móvel. O robô móvel percorre um espaço estruturado que é associado, através de uma projeção bidimensional, ao espaço que aproxima as coordenadas com eventos MIDI. O robô é observado por um sistema de visão omnidirecional, que observa e informa a localização espacial do robô.

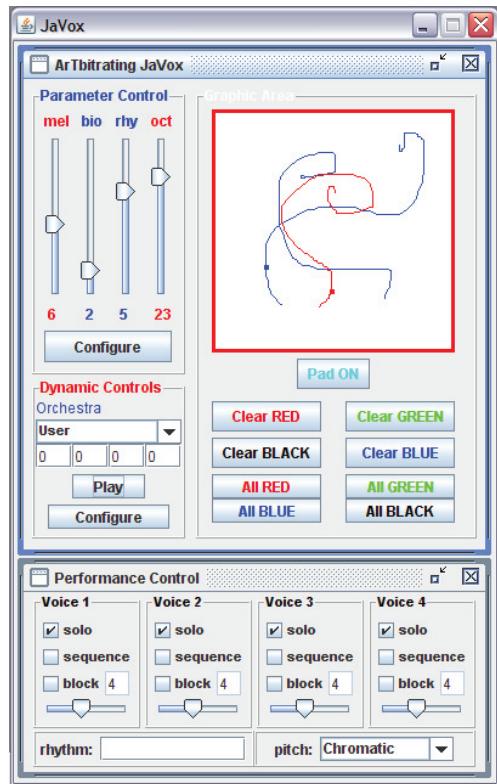


Figura 4 A interface do ambiente JaVOX, núcleo do ambiente AURAL

1.1 Trajetórias Convexas

A interação entre curvas na interface de controle e o robô se faz através do envio de pontos discretos consecutivos da curva desenhada na pad interativa para o robô, sob a forma de trajetória. Ao receber comandos para executar a sequência, o robô gera uma nova trajetória – a trajetória efetivamente percorrida – que é observada pelo sistema de visão, enviada para o sistema JaVOX e atualizada na pad de controle.

A interação entre a trajetória enviada e a executada pelo robô é descrita através de uma combinação convexa conforme apresentado a seguir:

$$\begin{aligned} X(k+1, i) &= \alpha.X(k,i) + (1-\alpha)Rx(i) \\ Y(k+1, i) &= \beta.Y(k,i) + (1-\beta)Ry(i) \end{aligned} \quad (1)$$

onde $i=1, 2..N$. N é o número de pontos que compõem a curva. Os pontos são tomados a cada 5 pixels de distância da curva na tela do monitor. $X(k, i)$ e $Y(k, i)$ são as coordenadas (x, y) para cada ponto do pad de controle do JaVOX tomadas no instante k. $Rx(i)$ e $Ry(i)$ são as coordenadas (x, y) da curva observada pelo sistema de visão e os coeficientes da combinação convexa são $0 \leq \alpha, \beta \leq 1$.

Desta forma, a informação obtida a partir da observação do comportamento do robô é utilizada para re-alimentar a informação na pad, que é usada na produção sonora, gerando um ciclo de interação.

1.2 Controles de Performance Musical

Além da trajetória, o JaVOX possui outro meio para controle em tempo-real da sonificação (ver parte inferior da Figura 3). Para cada uma das quatro vozes do JaVOX há

três controles denominados de 1) *solo*, 2) *sequence* e 3) *block*. No primeiro, os eventos sonoros são enviados para porta MIDI através do controle direto do pad e do processo de computação evolutiva, que fornece uma sequência de eventos MIDI a cada interação do JaVOX. Portanto, para o primeiro controle o resultado sonoro é dependente da interação entre a curva do pad e a trajetória observada pelo sistema de visão.

No segundo controle (*sequence*) de performance sonora do JaVOX, são executadas eventos MIDI tocados sequencialmente. Desta forma, é enfatizado o caráter horizontal do processo de sonificação, gerando uma textura sonora com caráter melódico. No terceiro (*block*), os eventos MIDI são enviados tão próximos quanto possível, quase que simultaneamente, para a placa MIDI, gerando uma superposição de blocos de notas. Neste caso a ênfase é na verticalidade dos eventos sonoros, gerando uma textura complexa de acordes.

Estes três modos de operar sobre a sonificação geram modificações significativas no resultado sonoro e podem ser utilizados como estratégia composicional. A interação destes controles com o comportamento dinâmico do Nomad (o robô móvel), o sistema de visão e, eventualmente, a presença de outros robôs no espaço, pode gerar uma organização sonora complexa.

O processo de vínculo entre o comportamento dos robôs no espaço e a sonificação foi desenvolvido com o objetivo de verificar a capacidade do AURAL de criar texturas sonoras auto-organizadas a partir de interações simples entre os agentes do sistema, os robôs móveis. O módulo supervisor (TrajeCt) recebe a seqüência de pontos enviada pelo JaVOX e envia os comandos para o Nomad. Outro robô móvel, o Roomba, se movimenta livremente pelo espaço, utilizando um sistema de navegação autônomo. Quando há colisão, o Roomba se afasta. A interação entre a navegação livre do(s) Roomba(s) e a trajetória do Nomad gera um comportamento coletivo entre os robôs que é utilizado como controle de performance do JaVOX. No caso limite, haverá 4 robôs no ambiente, cada um associado a uma das vozes do JaVOX.

A Tabela 1 apresenta as relações entre os controles de performance do JaVOX e o comportamento dos robôs móveis. Esta tabela descreve uma das possibilidades de associação de proximidade entre um Roomba e o Nomad.

O fluxo de informação parte e retorna para o JaVOX no processo de sonificação. A trajetória do Nomad e dos outros robôs é capturada pelo sistema de visão que fornece as coordenadas (Equação 1) e os critérios de comportamento (Tabela 1) que controlam o JaVOX. A comunicação entre cada parte do sistema é feita através do protocolo TCP-IP.

Tabela 1 Proximidade e Controle de Performance no JaVOX

Interação entre Nomad e Roombas			
	Solo	Sequence	Bloco
Distante	X		
Médio		X	
Próximo			X

O sistema de amplificação sonora é realizado no entorno do espaço onde os robôs navegam e para isto são utilizadas duas caixas acústicas e uma placa de som conectada ao computador que controla o JaVOX. Numa etapa posterior, alternativas de difusão sonoras deverão ser testadas. A posição física dos robôs poderá ser usada para mapear a

especialização sonora num modelo de disposição de fontes sonoras como o *surround 5.1*.

2 CONTROLE ROBÓTICO

Na interface gráfica do ambiente JaVOX, a área do pad interativo está conceitualmente associada a um espaço estruturado de dimensões 4x4 m. Uma vez desenhada uma curva no pad interativo da interface gráfica, pontos consecutivos desta curva são transmitidos ao TrajeCt, que aplica o algoritmo de controle de trajetória em conjunto com uma camada de decisão para que o Nomad percorra a trajetória associada no espaço estruturado de forma segura [10]. É este o sistema chamado de supervisor, pois ele recebe os pontos do JaVox e envia comandos de movimentos calculados pelo algoritmo de controle de trajetórias. O deslocamento do(s) robô(s) no ambiente é observado pelo OmniEye, o sistema de visão omnidirecional. A trajetória observada é enviada para o JaVOX, apresentada na interface. Ambas as trajetórias, a enviada e a observada, passam então a ser consideradas na produção sonora (Equação 1).

2.1 TrajeCt: Controle de Trajetórias

Em linhas gerais, o algoritmo funciona da seguinte maneira: dada uma seqüência de pontos, o robô percorre cada um destes através de retas traçadas entre pontos adjacentes na seqüência. A velocidade à frente do robô é mantida constante, sendo alterada somente a velocidade angular de cada roda para o ajuste da direção que ele deve seguir.

Caso o robô não esteja na posição desejada, é calculada a distância de sua posição atual até a reta que liga os pontos da trajetória. Essa distância é então minimizada através da correção da direção atual do robô pela velocidade angular.

Quando o robô se aproxima a um raio R do próximo ponto da trajetória, um novo ponto da seqüência é obtido e o algoritmo de correção de trajetória é executado novamente, até que a lista de pontos da trajetória se esgote. Esse raio R foi determinado de forma empírica. Os melhores resultados foram apresentados quando ele é igual a 20% do tamanho do segmento de reta que liga os próximos pontos da trajetória.

O sistema supervisor foi escrito em TDL/TCM (Task Description Language/Task Control Manager) [11, 12]. A linguagem TDL é uma extensão de C++, permitindo a sincronização e execução de tarefas bem como o tratamento de exceções geradas por estas, que impeçam que as mesmas sejam concluídas satisfatoriamente. A interligação do módulo TrajeCt com o JaVOX, escrito em Java, foi realizada utilizando a Java Native Interface (JNI).

2.2 OmniEye: Visão Omnidirecional

Sistemas de visão omnidirecional provêem imagens com campo de visão de 360°. Tais sistemas podem ser construídos segundo diversos modelos, que englobam desde o uso de múltiplas câmeras apontadas em diferentes direções até o uso de uma única câmera livre para girar em torno de um eixo fixo [13].

Diferentes espelhos convexos podem ser utilizados para conseguir um sistema omnidirecional. Nesta fase da pesquisa foi empregado um espelho meia-bola 360° fixado no teto do ambiente (o laboratório da DRVC/CenPRA)

com fios de nylon. O OmniEye foi construído com uma *webcam* montada na cabeça de um tripé pequeno, adaptado para o experimento [14]. Um prumo em forma de prisma foi usado para dar estabilidade ao sistema. A Figura 5 apresenta um foto do dispositivo e a Figura 6 mostra uma imagem adquirida com ele.

A construção de um sistema de visão omnidirecional empregando câmeras com espelhos parabólicos, hiperbólicos ou elípticos assegura a propriedade de um único centro de projeção, o que permite o cálculo preciso de funções de calibragem e retificação. Já para espelhos esféricos tal propriedade pode ser apenas localmente aproximada na parte central do espelho.

Para a calibragem do espelho foi aplicada uma toolbox desenvolvida por Scaramuzza [15, 16]. Esta técnica assume apenas que a função de imageamento pode ser descrita por uma *Série de Taylor* cujos coeficientes são estimados em minimização linear de quatro passos aplicando o método dos Mínimos Quadrados, seguido de um refinamento não-linear baseado no critério de vizinhança máxima. Apesar de o espelho esférico não apresentar a propriedade de ter um centro único de projeção, resultados consistentes foram obtidos com a Toolbox. Posteriormente, um algoritmo genético [17, 18] foi aplicado para otimizar os coeficientes da função de imageamento, aproximada por uma Série de Taylor de ordem 4. Para tanto, foi desenhada uma grade no chão do ambiente, e foram usados os pontos da grade (24 pontos) como critério de avaliação, obtendo-se um erro de 1% em relação à menor distância máxima entre os pontos estimados e os pontos reais. Antes da otimização com o algoritmo genético o erro era de 8%.



Figura 5 O sistema omnidirecional composto por espelho meia-bola, câmera e peso em forma de prisma.

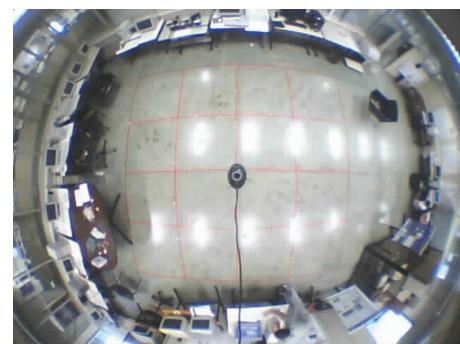


Figura 6 Uma imagem do laboratório da DRVC/CenPRA obtida com o sistema omnidirecional.

Para o monitoramento do robô, foi colocada sobre ele uma lâmpada (led). Após a captura das imagens em tempo real usando o OmniEye, são realizadas uma série de operações tais como: subtração de imagens, operações morfológicas e *thresholding* [19], com o objetivo de isolar na imagem apenas o objeto de interesse, nesse caso a luz emitida pelo led. A partir daí basta transformar as coordenadas do robô (em pixel, na imagem) para coordenadas no mundo real, de acordo com a função encontrada no processo de calibração da câmera. As coordenadas são enviadas então para o sistema JaVOX, e a trajetória observada é mostrada na interface gráfica.

3 RESULTADOS

São apresentados a seguir alguns resultados obtidos com o ambiente AURAL. A Figura 7 apresenta uma imagem da trajetória observada (azul) pelo robô Nomad, processada pelo sistema de visão omnidirecional. Os pontos da trajetória são enviados por TCP-IP para o ambiente JaVOX e desenhados na área gráfica, na Figura 8. A outra curva (vermelha) que aparece na interface é a trajetória enviada. Também na interface aparece o botão com o rótulo “T. Convexa”. Ao pressioná-lo, o resultado é uma curva azul, combinação convexa de todas as curvas vermelhas e azuis, que poderá ser enviada como nova trajetória para o robô percorrer.



Figura 7 Acima, uma imagem da trajetória percorrida pelo robô Nomad.
Abaixo, o negativo em preto e branco da imagem.

A Figura 9 mostra a trajetória observada pelo robô Roomba em movimento. A interação deste robô com o Nomad, utilizada como estratégia composicional e, eventualmente, a presença de mais robôs no espaço, geram modificações significativas e de complexa organização no resultado sonoro.

4 CONCLUSÃO

Conceitualmente, esta proposta se coloca dentro da área de Criatividade Computacional, sub-área da Inteligência Artificial, aplicada à Computação Musical. Dado o caráter interdisciplinar da proposta, insere-se também nas áreas de Arte, Aplicações à Distância, Robótica e Visão Computacional.

Nessa área interdisciplinar e emergente ligada à interatividade, Arte e Ciência influenciam e fertilizam uma à outra. O ambiente AURAL representa no Brasil uma rara oportunidade de reunir música e alta tecnologia numa único ambiente. Construído sobre o ambiente JaVOX, o AURAL disponibilizará recursos para a programação de coreografias para robôs móveis, *in loco* ou à distância.

Tendo como contexto o tratamento computacional da criatividade, no AURAL o processo criativo do compositor, do artista, do engenheiro, da audiência ou dos intérpretes - no caso, também os robôs - estará apto a examinar os mecanismos das diferentes áreas envolvidas, com resultados onde o material e a idéia são induzidos a coincidir na operação final. Muitas são as possibilidades de interação a serem exploradas, como por exemplo a relação entre a velocidade do Nomad e o andamento (tempo musical) da sonificação ou o uso de padrões visuais para identificação dos robôs.

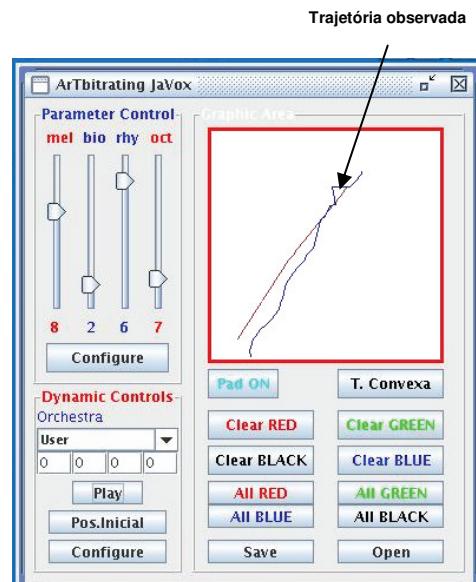


Figura 8 Trajetórias enviada para o robô Nomad e observada pelo sistema omnidirecional. Note que a trajetória enviada é muito mais simples que a observada. Isso se deve aos ajustes de direção que o robô teve que fazer para cumprir a trajetória.



Figura 9 Trajetória do robô Roomba observada pelo OmniEye
Abaixo, o negativo em preto e branco da imagem.

Numerosos são os desdobramentos que poderão advir desta parceria entre o Núcleo Interdisciplinar de Comunicação Sonora da Unicamp (NICS/Unicamp) e a Divisão de Robótica e Visão Computacional do Centro de Pesquisas Renato Archer (DRVC/CenPRA), propiciando o surgimento de resultados interdisciplinares de pesquisa em novas mídias.

5 AGRADECIMENTOS

Aos alunos Thiago E. D. Spina, Lucas Soares, Rafael Maiolla, Igor Martins, Igor Dias, Felipe Augusto, Luiz Fernando Faria Pereira, Eduardo Camargo, Gustavo

Solaira, Leonardo Laface e Daniel Domingues. Ao programa PIBIC/CNPq. O projeto Aural é viabilizado pela colaboração da DRVC/Cenpra e o NICS/Unicamp, que se integram no desenvolvimento desta pesquisa com o apoio da FAPESP através do projeto Jovens Pesquisadores 05/56186-9. Manzolli tem o apoio de projeto Pq do CNPq.

6 REFERÊNCIAS

- [1] Camurri, A., Hashimoto, S., Ricchetti, M., Ricci, A., Suzuki, K., Trocca, R., Volpe, G. EyesWeb: Toward Gesture and Affect Recognition in Interactive Dance and Music Systems. *Computer Music Journal*, 24:1:57-69, Spring 2000.
- [2] Manzolli, J., Verschure, P. F. M. J. Roboser: a Real-world Musical Composition System. *Computer Music Journal*, 2005.
- [3] Wassermann, K. C., Eng, K., Verschure, P. F. M. J., Manzolli, J. Live Soundscape Composition Based on Synthetic Emotions. *IEEE Multimedia*, Out-Dec, 2003, pg. 82-90, 2003.
- [4] Murray, J., Wermter, S., Erwin, H. Auditory robotic tracking of sound sources using hybrid cross-correlation and recurrent networks. In Proceedings of the IEEE/RSJ Intelligent Robots and Systems (IROS 2005), 3554-3559 D.O.I. 10.1109/IROS.2005.1545093, 2005.
- [5] Moroni, A. S., Manzolli, J., Von Zuben, F. ArTbitrating JaVox: Evolution Applied to Visual and Sound Composition. Em Brunet, P., Correia, N., Baranowski, G. (eds.) Ibero-American Symposium in Computer Graphics 2006, pp. 97 – 108. Santiago de Compostela, Eurographics Chapter Proceedings, 2006.
- [6] Moroni, A.; Maiolla, R., Manzolli, J. "Seeing and Hearing" Evolutionary Compositions. 3IA-2007: 10th International Conference in Computer Graphics and Artificial Intelligence, Athens – Greece, 2007.
- [7] Manzolli, J. Auto-organização: um paradigma Composicional. In Auto-organização: Estudos Interdisciplinares em Filosofia, Ciências Naturais e Humanas, e Artes, (orgs.) M. Debrun, Gonzales, Pessoa Jr.. Coleção CLE, Vol. 18, pg. 417-435, 1996.
- [8] Moroni, A., Manzolli, J., Von Zuben, F. J. & Gudwin, R. "Vox Populi: An Interactive Evolutionary System for Algorithmic Music Composition", *Leonardo Music Journal*, 10, pp. 49-54, 2000.
- [9] Moroni, A., Manzolli, J., Von Zuben, F.J. & Gudwin, R. "Vox Populi: Evolutionary Computation for Music Evolution" em Bentley, P. & Corne, D. (eds.) Creative Evolutionary Systems, San Francisco, USA: Morgan Kaufmann, pp. 205 - 221, 2002a.
- [10] "Nomad 200 User's Manual". Nomadic Technologies, Mountain View, CA, EUA, Janeiro de 1996.
- [11] Apfelbaum, D. "The Task Description Language Technical Report". Carnegie Mellon University, Pittsburg, PA, EUA, 1999.
- [12] Simmons R. & Apfelbaum D. "The Task Description Language – TDL". Carnegie Mellon University, <http://www.cs.cmu.edu/~tdl/>.
- [13] Yagi, Y.: Omnidirectional sensing and its applications. IEICE Transactions on Information and Systems, vol. E82-D, No. 3, pop. 568—579, 1999.
- [14] Moroni, A.; Cunha, S. OmniEye: A Spherical Omnidirectional Vision System for Tracking Robots in the AURAL Environment. 3IA-2008: 11th International Conference in Computer Graphics and Artificial Intelligence, Athens – Greece, 2008 (*a ser impresso*).
- [15] Scaramuzza, D., Martinelli, A. and Siegwart, R.: A Flexible Technique for Accurate Omnidirectional Camera Calibration and Structure from Motion, Fourth IEEE International Conference on Computer Vision Systems. (ICVS 2006), New York, 2006.
- [16] Scaramuzza, D. Omnidirectional Camera Calibration Toolbox for Matlab. http://asl.epfl.ch/~scaramuz/research/Davide_Scaramuzza_files/Research/OcamCalib_Tutorial.html
- [17] Holland, J. H. Adaptation in Natural and Artificial Systems. University of Michigan Press, 1975
- [18] Michalewicz, Z. Genetic Algorithms, Numerical optimization, and Constraints. In L. J. Eshelman (Ed.), Proceedings of the 6th International Conference on Genetic Algorithms, pp. 151—158 (1995)
- [19] Gonzalez, R. C., Woods, R. E., Eddins, S. L. Digital Image Processing. Prentice-Hall, 2002.



Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Applications of Group Theory on Sequencing and Spatialization of Granular Sounds

Renato Fabbri¹, Adolfo Maia Jr.¹

¹Núcleo Interdisciplinar de Comunicação Sonora – UNICAMP,
Cx. P. 6166, Campinas, SP, 13 091-970, Brazil

²Instituto de Matemática, Estatística e Computação Científica (IMECC) – UNICAMP,
Cx.P. 6065, Campinas, SP, 13091-970, Brazil

{renato, adolfo}@nics.unicamp.br

ABSTRACT

We present an application of the theory of finite (permutation) groups on sequencing and spatialization of granular sounds. We show how to apply finite groups to an ordered set of grains in order to get sound streams with cyclical characteristics. Combining different layers of time sequencing with spatial addressing based on permutation groups we get a rich “polyphony” of spatialized granular sound streams. We present a computer implementation of our model named FIGGS (Finite Groups in Granular Synthesis), a flexible interface developed in the Python programming language and SAGE, a software for algebra and geometry experimentation.

0 INTRODUCTION

The composer I. Xenakis [1] dedicated a chapter in his book *Formalized Music* to applications of groups in algorithmic composition. On the other hand, due to the exponential growth of digital industry, in the second half of the twentieth century, electroacoustic music had its major developments, and now early tedious procedures can be relied on fast computer calculations.

Roughly speaking, composition in electroacoustic music incorporates sound construction methods as well temporal organization of these sounds. So, after Xenakis, it is natural to argue whether group theory can have a role in Electroacoustic Music exploring symmetries and cyclic time organization of the sound material. This work is our contribution in this direction. In order to have a focus on this problem we are most interested in the well known technique named *Granular synthesis* (see [5, 6] for an

overview). So, in this work we propose an exploration of granular synthesis through the point of view of applications of symmetries of finite groups acting on sonic structures which on their turn are also built taking into account symmetry of an “internal finite group”. In the next section, we present some theoretical preliminaries and some simple examples of finite groups and also some comments on Granular Synthesis. In Section 2, we show our model, that is, an application of finite groups in granular synthesis. We stress the fact that a number of models are possible and ours is just a demonstration of our method’s potential. Nevertheless, as far as we know, our model is the first one that makes use of Finite Groups in Granular Synthesis. In Section 3 we describe shortly the computer implementation of our model based on language Python with external packages for numeric manipulation and SAGE, a known software for algebraic manipulation. In the same section, we will make an analysis of some examples of outputs. In section 4 we present our model for spatialization of grain

streams. In section 5, we make a conclusion and state some perspectives of the method we called FIGGS (Finite Groups in Granular Synthesis). Section 6 we present our bibliography.

1 GRANULAR SYNTHESIS

Granular synthesis is commonly known as a technique that works by generating a rapid succession of tiny sounds, metaphorically referred to as sound grains or yet as microsounds. Granular synthesis is widely used by musicians to compose electronic or computer music because it can produce a wide range of different sounds, but it also has been used in speech synthesis. Clearly, discussions about musical aesthetics arise from these developments and although it is a very interesting topic by itself we will not deal with these matters in this paper. A good account of the aesthetics of microsound can be found in [7]. Granular synthesis is largely based upon D. Gabor idea of representing a sound using hundreds or thousands of elementary sound particles [3]. His approach to "elementarity" was inspired by the Uncertainty Principle of Quantum Mechanics. He proposed the basis for representing sounds combining frequency and time domains information. D. Gabor's point of departure was to acknowledge the fact that the ear has a time threshold for discerning sound properties. Below this threshold, different sounds are heard as clicks, no matter how different their spectra might be. The length and shape of a wavecycle define frequency and spectrum properties, but the ear needs several cycles to discern these properties. D. Gabor referred to this minimum sound quantity as an acoustic quantum and estimated that it usually falls between 10 and 30 milliseconds.

The first studies using D. Gabor's sound representation (*time x frequency* space) in music were, probably, initiated by composer I. Xenakis, but the first computer-based granular synthesis system did not appear, however, until C. Roads [4] and B. Truax, for real-time, began systematically to investigate the potential of the technique. As far as the idea of sound grains is concerned, any synthesizer capable of producing rapid sequences of short sounds may be considered as a granular synthesizer. However, it is important to stress that the very concept of grain is not the same for D. Gabor and I. Xenakis. The latter describes a grain as the instantaneous measured pair of a Fourier Partial with a particular frequency and amplitude, which is, for him, the most elementary characteristic of a sound. In this case, a cloud of grains can be thought as a cluster of points relatively close to each other in the *frequency x amplitude* space. And indeed, I. Xenakis proposed a density parameter to measure the compactness of a cloud and the duration of a grain is an external parameter to control the sound as a whole. For D. Gabor, however, duration is an internal parameter of the grain itself, which can have a complex content in terms of Fourier Partials. D. Gabor borrows the concept of quantum of action from Quantum Mechanics to define a quantum of sound. The quantum action of sound A is of order of unity, that is

$$A = \Delta\omega \cdot \Delta T = 1 \quad (1)$$

More recently, C. Roads suggested the following definition: "*A grain is a signal with an amplitude envelope in the shape of a quasi-gaussian bell curve*" [5]. This is

close to D. Gabor's original definition in the sense that it is implicit that a short time signal may have complex content. However, the concept of "elementary" (or quantum) is not strictly taken into account in C. Roads' definition. The interpretation of "quasi-gaussian bell curve" can be very general. In addition, his concept of grain density is actually a measure of the number of grains occurring within a given time interval. C. Roads also seems to suggest that granular synthesis can be classified as a form of additive synthesis. We therefore prefer to consider this definition as being for a specific type of granular synthesis, which we refer to as Short-Time Additive Synthesis (STAS). A key problem in granular synthesis is the control of the evolution of the sound grains in time. Most granular synthesis systems have used stochastic methods to control the time evolution of hundreds or even thousands of grains. A handful of alternative methods have also been proposed. For instance, E. Miranda devised Chaosynth, a granular synthesizer of the STAS type that uses cellular automata to manage the spectrum of the sound grains [8]. Chaosynth explores the emergent behavior of cellular automata to produce coherent grain sequences with highly dynamic spectra. The states of the cellular automata define frequency and amplitude values for an additive synthesis engine that produced the grains.

In this work we take C. Roads' definition of sound grain as a point of departure to develop a formal but flexible granular synthesis model. A critical review and present status of Granular Synthesis can be found in [7]. Once the sonic grains are discrete entities, it is possible to think them as objects that in which we can impose internal and external symmetries, the last one related to time organization of grains.

2 THEORETICAL MODEL: GROUP-SEQUENCING OF GRAINS STREAMS

Formally, a group G is a set with a binary rule (which we will denote by '•' in this work) that together satisfy the four fundamental properties:

- 1) If g_1, g_2 are in G then $g_1 \bullet g_2 \in G$ (**Closure**)
- 2) $g_1 \bullet (g_2 \bullet g_3) = (g_1 \bullet g_2) \bullet g_3$ (**Associativity**)
- 3) There exists an element e in G such that $g \bullet e = e \bullet g$ (**Identity**)
- 4) For each g in G there exists an element g^{-1} in G such that $g \bullet g^{-1} = e$ (**Inverse Element**)

We denote as (G, \bullet) a Group with an operation \bullet . A Group is finite if it has a finite number of elements, otherwise it is called infinite. A Group is called *commutative* or *abelian* if the commutative property is satisfied for all its elements, that is:

for g_1, g_2 in G $g_1 \bullet g_2 = g_2 \bullet g_1$ (**Comutativity**)

The groups we are most interested in this work are the symmetry groups and permutations groups of a number n of objects (in our case, these objects are sound grains). More on Group Theory see, for example, [2].

Here we present our model and the next we show its computer implementation. The idea is just to find a way to construct an application from the group on the set of grains.

Let be G_n a group with n elements and $T: G_n \rightarrow G_n$ an endomorphism of G_n , that is, T is an application of G_n onto itself, which we denote as:

$$p \xrightarrow{\quad} T_p(q) = p \bullet q \quad (2)$$

For a fixed q in G_n , consider the set $\mathcal{O} = \{T_q(p)\}$, with p in $G_n\}$. This is named a *co-set* or an *orbit* in G_n . If two elements are in the same orbit (defined by q) they are related in the sense that any element can be obtained from another one through repeated applications of the transformation T . That is, $p_1 \sim p_2$ if there exists a natural number k such that

$$T^k(q) = p_2 \quad (3)$$

With this application the group G_n is sliced in a number of co-sets.

We have considered, for our applications, three permutations groups.

a) *The Symmetric Group of degree n*: the group of all the permutations on an ordered set of n elements.

b) *The Alternating Group of degree n*: the group of even permutations on a set of n elements, that is, the set of permutations obtainable from an even number of two-element swaps.

c) *Cyclic Group*: a group which can be generated by a single element and the group operator. In our case we use, for simplicity, Permutations Cyclic Groups.

In this first and simple model we take the group as one of the three above mentioned. We can also take an *ordered set of grains* and consider the application (2) defined above with the permutations (belonging to one of the above mentioned groups) applied to the ordered set of grains. The overall effect is a sequence of sound segments which evolves cyclically in time. Each new choice for the q element will lead to another orbit with some new characteristics. Of course this also depends on the sound contents of set of grains. The point here is that we have previously ordered initial set of, say, n grains fixed by the user. Now that group transformations are, for example, permutations of a set of n elements. These permutations are then sequenced in time resulting in a sound stream whose main psychoacoustic characteristic is the cyclical sound structure in time. In addition we can also apply group transformations inside the grain itself. This can be done, for example, using some set of transformations on the spectral parameters of the grain (frequencies, duration, amplitude, etc) and imposing on this set of transformations a Group Structure. This approach leads a construction of grainy sound structures with a strong internal correlation.

3 INTERFACE FIGGS

We have developed a GUI (Graphical User Interface) in Python named FIGGS (Finite Groups in Granular Synthesis), see Figures 1 and 2. This allows the user/composer have some facilities in order to concentrate in his musical creation. This interface has two tabs: one for grain parameters specifications and another for the sound parameters on which act symmetry group, each with its proper specifications.

	frequency(Hz)	duration(ms)	amplitude(Peak*100)	fade(ms)	separation(ms)
Grain 1	1127	98	92	26	-19
Grain 2	439	33	10	7	379
Grain 3	439	58	11	28	69
Grain 4	925	97	88	24	-50
Grain 5	294	29	21	9	268
Grain 6	213	14	91	7	163
Grain 7	1093	11	65	5	-6
Grain 8	265	74	53	11	0

Figure 1: FIGGS Panel for Grain Specifications

Figure 1: FIGGS Panel for Groups - Choosing Grain Parameters

Each parameter of each grain can be inputted independently. For experimentation, we can input random parameters with a specific random distribution which can include Gaussian, Bernoulli, Binomial, among others (in future development). There is a parameter controlling the grains time separation. This could be also randomized and negative separations are understood as superposition. This is another way to construct asynchronous sequencing.

In this panel we can choose independently those parameters we want to permute. The number box on the top is the number of times the specified grains are going to be played, the number of cycles. This is the only necessary input before the sound file can be done, the user can choose to play the grains in the exact sequence inputed in the grain panel.

Figure 3 shows some group usage. The number boxes besides each group list are the number of elements in which the group acts and the number of cycles that repeats before the group acts again. There are several ways in which a group can act on a parameter set, and options are going to be available in future developments. In this version's screen shot, the action is performed by applying a random element in the group (i.e. a permutation) on the set. The user can freely move back and forth from grain panel and group panel, and drive the sound file to be written any number of times. This allows one to make a number of sounds with a controlled behavior.

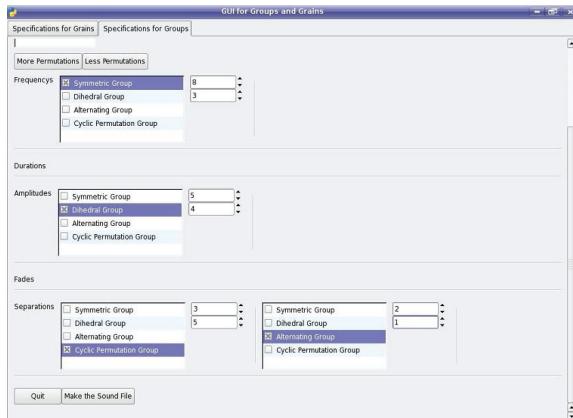


Figure 2: FIGGS Panel for Groups - Specifying Groups

4 SPATIALIZATION ADDRESSING VIA GROUPS

Today there are a number of algorithms designed to spatial control of granular sound streams [13, 5]. Most of these algorithms assembly all the grains together in a time sequence whose spatialization is designed to this “global sound” through pan parameters. Another approach is to control the spatialization of the stream, say, *grain by grain*. Unfortunately, for streams with thousands of grains, of course, there is no hope for this kind of control. The remedy for this problem relies again on a “coarse control”, that is, we must use an algorithm in order to choose automatically the spatial addressing of grains.

Now, a large number of electroacoustic music pieces, as well soundscapes for a specific ambient, present cyclicity as major characteristics. As mentioned above, group theory is well suitable to construct large sequencing of cyclical sound structures. The point here is that we can use the group approach to control also the spatial positioning of stream, *grain by grain*. The cyclical characteristics of finite groups are then reflected on the spatial distribution of the grain stream among the channels. We present, below, a simple method which relies on the cyclical structure of finite groups to get a direct spatialization of grain streams. This is as follows.

Suppose we have a set of m channels outputs $\mathbf{Ch} = \{ch_1, ch_2, \dots, ch_m\}$ and a ordered set of n grains. Consider a partition of m the channels in subsets, say, $m = 5$ and the partition $S_1 = \{ch_2, ch_3\}$, $S_2 = \{ch_1\}$, $S_3 = \{ch_4, ch_5\}$. The idea now is t

We can apply on these sets several kinds of finite groups. For simplicity, consider, for example, the groups of permutation G_k with $1 \leq k \leq n$. So, on the set $S_1 = \{ch_2, ch_3\}$ will act the group G_2 , on $S_2 = \{ch_1\}$ will act G_1 and on $S_3 = \{ch_4, ch_5\}$ will act G_2 .

As mentioned above groups slice the ordered set o n grains in orbits (or cosets), say, O_1, O_2, \dots, O_q . Now we assign the grains in O_1 to the S_1 channels set with action of a G_2 . The effect is the alternating spatialization of the grains of the coset O_1 among the two channels $\{ch_2, ch_3\}$. We can also address, for example, O_2 to $S_2 = \{ch_1\}$. Of, course, in this case, all grains of this coset have the same local sound output. This kind of operation we named *Spatialization Addressing*. Of course, our algorithm can be generalized to any finite group.

It is interesting to note that in order to construct musical structures we can use superposition of sequences of grains. In other words we can construct layers which can be controlled independently (or interdependently) in terms

of content and duration of the grains as well as their time sequencing. This conception can be envisaged as a tool for granular synthesis composition. In addition, new rhythmic aspects can emerge from this kind of sound design. Combining different layers of time sequencing with spatial addressing based on permutation groups we get a rich polyphony of granular sound streams.

5 CONCLUSION AND PERSPECTIVES

As mentioned in the introduction section, group theory can be valuable tool for algorithmic composition. Nevertheless, as far as we know, its application in sound synthesis was not pursued seriously until now. Our model is a preliminary study in this direction in the area of sound synthesis. We have used, for the sake of simplicity, the well known finite permutation and symmetrical groups which can demonstrate the potential of the model for more complex applications. In addition we can point out some directions and perspectives for future work:

- A) The method can be generalized to include other non granular sounds. This kind of approach could be interesting in order to create soundscapes which evolve cyclically and whose elements merge one into another.
- B) In an actual composition, different groups can be used for each sound layer. This includes different sets of sounds in which these groups act.
- C) In order to generate an arbitrary and great quantity of grains we can make use of probabilistic distributions such as Gaussian, Binomial, Bernoulli, among others. In addition we can also define random sequencing for the time ordering of the grains such as random walks, Markov chains, and other stochastic processes.

6 REFERENCES

- [1] Xenakis, I., *Formalized Music*, Bloomington: Indiana University Press (1971); also, *Formalized Music*, 2d ed., New York: Pendragon Press (1991).
- [2] Budden, F.J., *The Fascination of groups*, CUP,(1972).
- [3] Gabor, D., *Acoustical Quanta and the Theory of Hearing*, Nature **159** (4044), pp. 591-594,(1947).
- [4] Roads, C., *Introduction to Granular Synthesis*, Comp. Mus. Jour., **12** (2), pp. 11-13 (1988).
- [5] Roads, C., *Microsound*, MIT Press, Cambridge,MA, (2001).
- [6] Roads, C., *Computer Music Tutorial*, MIT Press, Cambridge, MA (1996).
- [7] Thomson, P., *Atoms and errors: towards a history and aesthetics of microsound*, Organized Sound, **9** (2), pp. 207-218, (2004).
- [8] Miranda, E. R., *Computer Sound Design: Synthesis Techniques and Programming*, Oxford: Focal Press (2002).
- [9] Python Software Foundation, “Python Documentation”, <http://www.python.org/doc/>, as accessed in 02/06/2007.

- [10] Stein, W., *SAGE – System for Algebra and Geometry Experimentation*,
<http://www.sagemath.org/sage/documentation.html>,
accessed in 02/05/2007.
- [11] Scipy and Numpy documentation at
<http://www.scipy.org/Documentation>, accessed in
02/06/2007.
- [12] Cournapeau, D., “Pyaudiolab homepage”, at
<http://www.ar.media.kyoto.ac.jp/members/david/software/pyaudiolab/> as accessed in 02/06/2007
- [13] López, D., Martí, F., Resina, E., *Vocem: An Application for Real-Time Granular Synthesis*, dafx98 (1998).



Sociedade de Engenharia de Áudio Artigo de Congresso

Apresentado no 6º Congresso da AES Brasil
12ª Convenção Nacional da AES Brasil
5 a 7 de Maio de 2008, São Paulo, SP

Este artigo foi reproduzido do original final entregue pelo autor, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, www.aes.org. Informações sobre a seção Brasileira podem ser obtidas em www.aesbrasil.org. Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Mecanismo de argumentação e controle de versão para um ambiente cooperativo de prototipação musical

Aurélio Faustino Hoppe, Evandro Manara Miletto, Marcelo Soares Pimenta
Instituto de Informática – Laboratório de Computação Musical
Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre, RS, 91501-970, Brasil
{afhoppe, miletto, mpimenta}@inf.ufrgs.br

RESUMO

Este artigo se insere na área de Computação Musical e propõe a integração multidisciplinar de conceitos da computação - Interação Homem-Computador (IHC) e Trabalho Cooperativo Suportado por Computador (CSCW) - no desenvolvimento de um ambiente cooperativo de prototipação musical baseado na web. Discute também o conceito de prototipação musical e introduz os principais aspectos da prototipação cooperativa de peças musicais feitas no ambiente por usuários que não possuem necessariamente conhecimentos musicais, focando em particular nos aspectos relativos aos mecanismos de argumentação e controle de versão, responsáveis pela compreensão e percepção (*awareness*) de um usuário das ações realizadas pelos outros membros do grupo.

0 INTRODUÇÃO

Hoje ainda são poucos os aplicativos e ambientes – especialmente aqueles para a *Web* – que aproveitam potencialmente os aspectos interativos e cooperativos do fazer musical. Estes aspectos – mesmo de natureza exploratória – podem ser determinantes na realização e no entendimento do modo pelo qual usuários se envolvem em processos de experimentação musical e na concepção de ambientes computacionais que o auxiliem.

Assim como [1], vemos a música como uma atividade social, que nos leva a compartilhar nossas experiências musicais. A tecnologia atual tem claramente contribuído para o surgimento de novas modalidades sociais de apreciação musical, porém estamos convencidos de que ela igualmente oferece também muitas contribuições para formas sociais de “criação” musical. Paralelamente, apesar de músicos experientes e amadores estarem habituados a usar tecnologia para fazer música, este uso não é tão óbvio

quando consideramos pessoas leigas em música e por isto estamos especialmente interessados, como [2], em facilitar para qualquer usuário (leigo ou não) o acesso a experiências musicais envolventes e significativas. Nos últimos anos temos investigado a idéia de criar um ambiente de composição musical para leigos, surgindo assim, o CODES – “*C*Ooperative *M*usic *P*rototype *D*Esign”, um ambiente baseado na *Web* para prototipação musical cooperativa. O propósito do CODES é permitir que seus usuários (tanto leigos em música quanto músicos experientes) façam experimentos musicais e interajam entre si combinando peças musicais simples, aqui denominadas protótipos musicais. Definimos prototipação musical como sendo um processo interativo (usuários interagem com os mecanismos oferecidos pelo CODES), iterativo (permite a natural repetição e o refinamento gradual dos resultados deste processo criativo) e cooperativo (permitindo que grupos de usuários interajam entre si) para a construção de um protótipo musical.

No CODES, qualquer pessoa pode criar seu esboço musical (protótipo), que pode então ser repetidamente ouvido, testado e modificado, tanto pelo autor original quanto por seus parceiros remotos, que estarão cooperando no refinamento desse protótipo. Para que isso seja possível, o CODES foi projetado de modo a satisfazer aspectos relacionados à flexibilidade de interação e a usabilidade, além de fornecer apoio adequado para interações com informação musical complexa, atividades de cooperação e percepção de grupo (ou “*group awareness*”), provendo mecanismos de apoio ao entendimento das ações e decisões dos membros de um grupo que compartilha e coopera em protótipos musicais.

Outros conceitos e características principais do CODES são encontrados em [3], [4] e [5], bem como idéias iniciais sobre mecanismos de argumentação aqui descritos.

1 PROTOTIPAÇÃO MUSICAL NA WEB

Prototipação é um processo cíclico normalmente adotado pela indústria para criação da versão simplificada de um produto a fim de compreender suas características e processos de concepção e produção. Assim, sucessivas versões do produto são criadas de modo incremental, incluindo melhorias de uma versão para a próxima.

Contudo, a composição musical é uma atividade complexa onde não existe concordância sobre quais atividades devem ser executadas e em que ordem: cada pessoa tem seu estilo único e modo de trabalhar e, além disso, a maioria dos compositores atuam sós, talvez por ainda não terem a tradição de compartilhar suas idéias musicais e de colaborar durante as atividades compostoriais.

“Prototipação” não é uma expressão comum na literatura musical. De fato, a atividade realizada normalmente por compositores é denominada “composição”. Mas, em princípio, leigos não são compositores e os resultados de suas experiências criativas são deliberadamente chamados de “protótipos musicais” neste artigo, para ressaltar essa diferença.

Em nossa opinião, música é um produto artístico que pode ser concebido pelo processo de prototipação. Uma idéia musical (nota, seqüência de acordes, ritmo, estrutura ou pausa) é criada por alguém (tipicamente para execução em um instrumento musical) e a seguir cíclica e sucessivamente modificada e refinada de acordo com sua intenção inicial ou com idéias que surgem durante o processo de prototipação. Além dos músicos, os leigos provavelmente também possuem interesse em criar e participar de experiências musicais, mas carecem de ambientes orientados ao seu perfil de usuário.

2 SUPORTE A COOPERAÇÃO

Prototipação musical cooperativa é definida aqui como sendo uma atividade que envolve pessoas trabalhando em conjunto num protótipo musical. A cooperação no CODES é assíncrona, já que não é necessário gerenciar a complexidade de eventos de tempo real para o desenvolvimento de protótipos musicais. Os usuários podem acessar o protótipo, fazer seus experimentos e escrever comentários em tempos diferentes.

Autenticado no ambiente CODES, o usuário poderá elaborar um protótipo musical inicial e solicitar a colaboração de outros “parceiros” através do envio de convites explícitos (normalmente usando recursos de correio eletrônico). Os parceiros que aceitam o convite

podem participar da manipulação e do refinamento musical cooperativo. Deste modo, o grupo de parceiros pode evoluir para uma Comunidade Virtual.

A coordenação das atividades em um contexto musical deste tipo pode ocorrer naturalmente quando o grupo reconhece um dos membros como alguém que possui mais habilidades musicais ou experiência. No entanto, acreditamos que não é necessário fazer distinção e representação explícita do papel de coordenador, pois a hierarquização das ações e comunicações do grupo não é nossa intenção. Geralmente as opiniões e ações de usuários mais experientes em um grupo com coordenação explícita podem inibir a participação dos demais usuários.

3 MECANISMOS DE PERCEPÇÃO DO CODES

Para apoiar os aspectos cooperativos da prototipação musical, propõe-se no CODES o uso de três tipos de mecanismos de percepção ou *awareness*. Este conceito recebe bastante atenção na literatura de CSCW (*Computer Supported Cooperative Work*), mas não possui uma definição única ou precisa [6]. No contexto do CODES, a noção adotada de *awareness* é a percepção e compreensão das ações dos outros usuários, o que fornece a um determinado usuário um contexto para as suas próprias ações. Os três mecanismos de percepção no CODES são: a) *Music Prototyping Rationale*: que permite usuários associarem explicações às ações nos protótipos musicais. b) *Action Logging*: para manter explicitamente registrado o histórico dos passos e das decisões que conduziram o protótipo ao estado atual. c) *Modification marks*: para indicar a um usuário que o protótipo foi alterado por outros.

Os mecanismos de *awareness* oferecem diversas vantagens para a prototipação musical, tais como:

- registrar a evolução das decisões;
- recuperar o progresso na prototipação musical e identificar conflitos, que podem iniciar um processo de negociação entre diversos pontos de vista;
- apoiar a construção de conhecimento cumulativo da prototipação;
- ajudar na integração de perspectivas de vários membros de um grupo;

A percepção das ações realizadas pelo grupo, desempenha um papel crucial para apoiar atividades cooperativas e multidisciplinares do CODES. Os principais aspectos desse mecanismo de percepção serão discutidos a seguir.

A capacidade para associar argumentações a passos em um projeto é um processo originário da área de IHC (Interação Homem-Computador) e é chamado de *Design Rationale* [7].

Design Rationale é um mecanismo de comunicação da equipe para documentar as decisões críticas tomadas durante um projeto, quais alternativas foram investigadas e a justificativa para a alternativa escolhida. É um meio para auxiliar um membro do grupo a entender melhor as decisões e ações dos outros integrantes do grupo. Ações e decisões musicais são em geral subjetivas, e daí a importância de se ter um mecanismo de comunicação específico para a argumentação das ações, de modo a informar as razões de cada ação tomada aos demais membros do grupo, como selecionar ou adicionar padrões sonoros, instrumento, pausas, etc. ou a decisão de fazer combinações ou excluir elementos.

Quando os usuários estão prototipando no CODES, eles combinam padrões sonoros a partir da percepção das alterações feitas por outro usuário. Sendo assim, escolhas, seleções, inclusões, remoções e audições são tarefas realizadas constantemente em processo cíclico até que se atinja um consenso sobre o resultado da prototipação. Todas essas ações podem ser argumentadas no sistema por usuários para que possam ser informadas aos outros os motivos que levaram a estas ações. Esta é de fato uma maneira segura de garantir a percepção (*awareness*) em ambientes colaborativos assíncronos.

Os elementos básicos de *Music Prototyping Rationale* do CODES basicamente são Tópicos e Comentários. Tópicos correspondem a decisões, ações e estados alcançados durante a criação de um protótipo musical colaborativo e seu refinamento. Por exemplo, um tópico pode ser “trocar um padrão sonoro, inserir uma pausa, misturar diferentes ritmos, etc.”. Tópicos são motivados por escolhas consensuais e alternativas relacionadas das ações em curso.

Comentários são declarações feitas para apoiar uma ação (comentários a favor) ou advertir o interesse de outros usuários através de uma expressão de objeção (comentários contra).

Além disso, comentários podem expressar sugestões, perguntas ou observações genéricas sobre um tópico. Toda decisão ou ação pode ser associada a comentários. Não há, entretanto, uma estrutura rígida nem um tipo específico de mensagem.

Um exemplo prático de *Music Prototyping Rationale*, após uma sessão de experimento rítmico no ambiente do CODES, é descrita como se segue. Três usuários chamados Robert, Jimmy e John participam do mesmo protótipo musical. Entretanto, Robert tem a idéia de misturar diferentes ritmos e decide adicionar padrões sonoros de Jazz dentro do estilo Pop. Para obter uma opinião dos outros participantes do protótipo, Robert escreve um comentário sobre este tópico.

Na janela de comentário, a edição de um comentário inclui o assunto e um corpo de texto livre para o usuário comentar. O usuário pode marcar uma característica opcional que é tornar o comentário privado (para controle próprio) ou público (para conhecimento dos demais).

CODES salva o comentário e o associa aos tópicos correspondentes do protótipo, informando aos outros usuários através de um ícone do tipo *post-it* que algum comentário foi feito por algum usuário relacionado a alguma ação executada no protótipo. Clicando neste ícone, o usuário recupera o comentário. Nossa abordagem de *Music Prototyping Rationale* usa uma estrutura hierárquica para representar as razões dos usuários. Cada entrada em “*Descrição*” corresponde a um elemento de argumentação. Nesta janela cada elemento é acompanhado por três ícones, um (+/-) que serve para propósitos de apresentação, outro (olhos) indica que o comentário foi lido e o último (fone de ouvido) para indicar que o evento sonoro foi ouvido pelo usuário correspondente, conforme exemplo da Figura 1.

Histórico de Alterações	Ver Comentários	Padrões Sonoros
Descrição		
Usuário	Evento	
<input checked="" type="checkbox"/> Misturar diferentes ritmos		<input checked="" type="checkbox"/> Robert
<input type="checkbox"/> Re: Misturar diferentes ritmos Concordo com você. Ficou legal!		<input type="checkbox"/> John

Figura 1 Tela Ver Comentários - CODES

CODES tem um mecanismo de *log de ações* (*Action Logging*) - para registrar informações (como data, autor, ação, elemento do protótipo afetado, etc) de todas as ações tornando-as disponíveis para todos os usuários para advertir o que foi realizado em qual sequência, bem como apresentar textualmente o histórico destas realizações, funcionando como uma Memória de Grupo.

CODES suporta sessões longas de prototipação e atividades cooperativas. As modificações sobre um protótipo podem perdurar de poucos dias a anos e as pessoas que cooperaram em um protótipo musical devem ter acesso fácil às modificações. O mecanismo de Marcas de Modificação (*Modification Marks*) permite a persistência destas modificações e as notificações da sua existência são explicitamente mostradas para todos os usuários.

CODES possui mecanismos de compartilhamento de informações iguais ou semelhantes aos encontrados em outros ambientes cooperativos, como por exemplo, o *Google Docs*, mas os mecanismos de argumentação que usamos diferem-se ou não existem neste tipo de ambiente.

4 CONTROLE DE ACESSO E VERSÃO

O controle de acesso e versão é uma ferramenta que representa de forma visual ou textual o histórico de todas alterações realizadas pelo usuário. É uma base de dados comum para usuários que compartilham o mesmo protótipo musical. Permite não apenas o entendimento das suas próprias atividades (consultando o histórico de versões), mas também, atividades realizadas por outros membros do grupo, através dos mecanismos de “*Music Prototype Rationale*”.

No CODES, o controle de acesso e versão está dividido em três níveis de edição: a) *Nível Sessão*: neste nível o usuário trabalha localmente, realiza seus experimentos ou anotações na forma de rascunhos, sem a participação dos demais colaboradores do protótipo. b) *Nível Versão*: ao iniciar a sessão (nível sessão), o usuário é alertado sobre modificações ocorridas no protótipo (inclusão de novos comentários ou versões). Estas notificações ocorrem quando algum usuário compartilha seus rascunhos com o grupo, através da ação “salvar versão”. Suas contribuições serão visualizadas, analisadas e discutidas pelos membros daquele protótipo musical. c) *Nível Publicação*: chegando a um consenso sobre o protótipo, o grupo poderá disponibilizar o resultado final para acesso público na Web (em formato MP3). A Figura 2 mostra exemplos do controle de versões, onde é possível ver o registro de algumas ações realizadas no protótipo, organizadas em ordem cronológica. O usuário pode navegar pelos registros para acessar, ver e ouvir as diferenças das ações (versões) anteriores.

5 AVALIANDO O USO DO CODES

CODES foi disponibilizado para uso em contexto acadêmico restrito. Levando-se em conta alguns métodos de avaliação da literatura de IHC e Ergonomia [8], temos aplicado entrevistas orientadas à satisfação dos usuários como procedimento de avaliação simples. Para capturar críticas e comentários relevantes (a também para evitar opiniões não-gravadas) os usuários são convidados a responder todas as questões *in loco*, imediatamente após o uso do CODES. As entrevistas possuem questões objetivas que tornam o processo de análise mais rápido, fácil e de baixo custo.

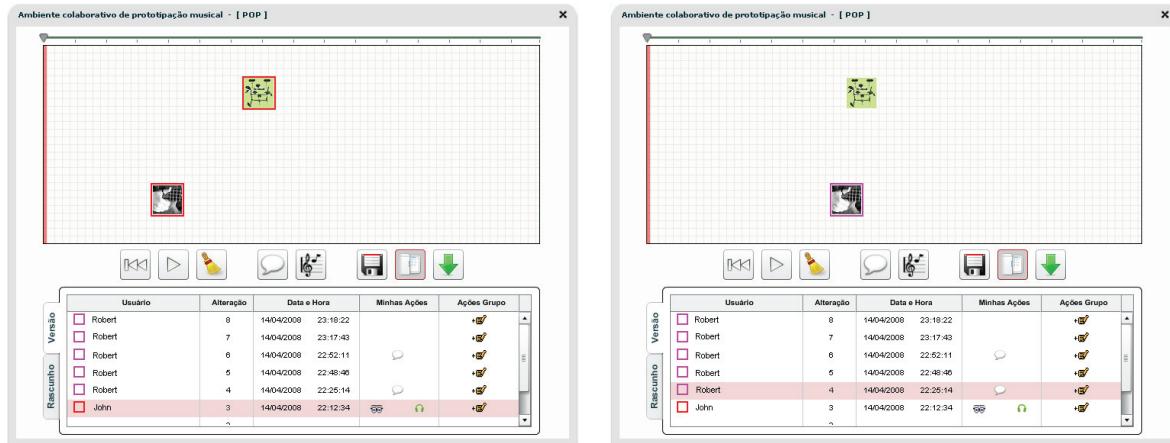


Figura 2 Exemplos do Controle de Versões - CODES

As perguntas foram concebidas para identificar conveniência da ordem e estrutura da informação para as atividades do usuário durante o processo de criação ou edição do protótipo, bem como da colaboração, incluindo itens específicos sobre usabilidade, acessibilidade, complexidade de navegação e também se o usuário está satisfeito com as funcionalidades implementadas. Além das respostas objetivas, a entrevista dá ao usuário oportunidade de fornecer comentários e para nós uma boa oportunidade para analisar os comentários.

Os resultados preliminares apontam para um relativo sucesso do nosso trabalho, mas, surpreendentemente o que mais chamou atenção foi o conjunto de alternativas possíveis que os usuários encontram para o uso do CODES. De fato, além dos procedimentos tradicionais para os quais desenvolvemos as suas funcionalidades, alguns usuários encontraram outras aplicações para o ambiente, como suporte efetivo ao aprendizado musical, ferramenta de entretenimento (DJ's) e sistema de acompanhamento para performance instrumental.

6 CONSIDERAÇÕES FINAIS

A principal motivação do CODES é propiciar que leigos – autodidatas, músicos amadores ou mesmo pessoas sem educação formal em música – possam usar um ambiente computacional para fazer a própria música (como dizia John Cage, “o som nosso de cada dia”) e compartilhar suas idéias com outras pessoas. Evidentemente, não está em discussão a qualidade musical do trabalho final, e sim a possibilidade de “criá-lo”, no lugar de apenas consumir o que já vem pronto.

Este desafio nos conduz a algumas questões importantes, em particular em relação a formas e mecanismos de suporte a esta cooperação.

A abordagem CODES para cooperação entre usuários na criação de protótipos musicais coletivos é um exemplo de ferramenta bastante promissora, que permite compartilhar conhecimento através de uma interação rica e de mecanismos de argumentação associados a cada modificação no protótipo. Conseqüentemente, cada participante pode compreender os princípios e regras envolvidas no complexo processo da experimentação e criação musical.

Nosso objetivo inicial foi desenvolver mecanismos úteis e ativos que não apenas estruturassem a informação envolvida no processo de prototipação musical, mas também ajudassem usuários durante este processo.

Acreditamos que a integração dos mecanismos de *awareness* e controle de acesso e versão aqui discutidos é uma forma razoável para facilitar a cooperação entre usuários e principalmente para estimular o surgimento de interações não planejadas, e assim consequentemente ampliar as possibilidades de prototipação musical.

7 REFERÊNCIAS

- [1] Gurevich, M., *JamSpace: Designing a Collaborative Networked Music Space for Novices*. In: Proceedings of the NIME06 - International Conference on New Interfaces for Musical Expression, Paris. p. 118-123.
- [2] Weinberg, G., *The Aesthetics, History, and Future Challenges of Interconnected Music Networks*. In: Proceedings of the ICMC 2002 - International Computer Music Conference, Gotemburgo, Suécia. p. 349-356.
- [3] Miletto, E. M.; Pimenta, M. S.; Vicari, R. M.; Flores, L. V., *CODES: a web-based environment for cooperative music prototyping*. Organised sound (Print), Cambridge University Press, v. 10, n. 3, p. 243-253, 2005.
- [4] Miletto, E. M.; Flores, L. V.; Rutily, J.; Pimenta, M. S., *CODES: Supporting Awareness in a Web-based Environment for Collective Music Prototyping*. In: Simpósio de Fatores Humanos em Sistemas Computacionais, 2006, Natal. Anais do IHC2006, 2006.
- [5] Miletto, E. M.; Flores, L. V.; Pimenta, M. S.; Santagada, L. *Interfaces for Music Activities and Interfaces for Musicians are not the same: the Case for CODES a Web-based Environment for Colaborative Music Prototyping*. In: The Ninth International Conference on Multimodal Interfaces (ICMI 2007), 2007, Nagoya. Proceedings do IX ICMI, 2007. p. 201-207.
- [6] Liechti, O., *Awareness and the WWW: An Overview*. ACM SIGGROUP Bulletin, v.21, n.3, p. 3-12, 2000.
- [7] Lee, J.; Lai, K.-Y., *What's in Design Rationale?* In: *Human-Computer Interaction Special Issue on Design Rationale*. Mahwah, NJ: Lawrence Erlbaum Associates, p. 251-280, 1991.
- [8] Dix, A. J.; Finlay, J. E.; Abowd, G. D.; Beale, R., *Human-Computer Interaction*. 2.ed. [S.l.]: Prentice Hall, 1998.

Índice de Autores

Author Index

Barbedo, J. G. A.	96, 100
Batalheiro, P. B.	88
Biscainho, L. W. P.	58, 72, 80, 120
Cunha, S. P.	128
Dietrich, P.	46
Eerola, T.	108
Espiga, M. L.	46
Esquef, P. A. A.	112
Fabbri, R.	134
Fagundes, R. D. R.	24
Filho, J. C. P.	112
Fornari, J.	108
Gerges, S.	46
Gerscovich, D. S.	80
Goldemberg, R.	20
Gomes, L. C. T.	65
Haddad, D. B.	88
Hoppe, A. F.	139
Lima, A. C. C.	119
Lopes, A.	96, 100
Maia Jr., A.	134
Mannis, J. A.	30
Manzolli, J.	15, 128
Martins, F. C. V.	72
Menezes, F.	20
Miletto, E. M.	139
Moroni, A.	128
Neto, M. U.	65
Netto, S. L.	58
Nunes, L. O.	72

Oliveira, L. C.	15
Paul, S.	38, 46
Pepe, I.	20
Petraglia, M. R	52, 88
Pimenta, M. S.	139
Prado, C. B.	58
Ramos, J.	128
Romano, J. M. T.	65
Sato, C. T	46
Scolari, D.	24
Silva, F. J. F.	11
Silva, G. S. G. S.	119
Siola, F. B.	11
Tavares, T. F.	96
Tenenbaum, R. A.	52
Torres, J. C. B.	52
Tovo, F. C.	52
Tygel, A. F.	72
Vasconcelos, L. G. L. B. M.	58
Zwetsch, I. C.	24



Audio Engineering Society - Seção Brasil

Anais do 6º Congresso de Engenharia de Áudio

12ª Convenção Nacional da AES Brasil

Proceedings of the 6th AES Brazil Conference

12th AES Brazil National Convention

Patrocinadores:



Expositores:

Attack Áudio System	Meteoro Amplifier	Santo Angelo
Audicare	Pride Music	Selenium
Beyma/Audiobrands	Produção Profissional	Sennheiser
Decomac Brasil	Quanta AV-Pro	Snake Pro
Electrovoice	Quanta Music	Sotex
FZ	Revista Áudio Música & Tecnologia	Spectral Balance Pro Áudio
Habro Music	Revista Backstage	Staner Eletrônica
Hinor Alto Falantes	Revista Igreja Equipar	Strike Music
Hot Sound	Revista Música & Mercado	Studio R
IATEC	RAG Consultores Associados	Superlux
IAV	Roland Brasil	Yamaha
Libor	Royal Music	