Anais do 4º Congresso de Engenharia de Áudio
10ª Convenção Nacional da AES Brasil

Proceedings of the 4th AES Brazil Conference - 10th AES Brazil National Convention

O áudio na era da comunicação
The audio in the communication era

08 a 10 de Maio de 2006 - 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
Regis Rossi A. Faria e Marcelo K. Zuffo

Audio Engineering Society - Seção Brasil

Coordenador Geral Convenção: Joel Brito (Presidente AES Brasil)

Coordenador do Congresso e do Comitê de Programa Técnico: Regis Rossi Alves Faria (LSI-EPUSP)

Coordenador Editorial: Marcelo K. Zuffo (LSI-EPUSP)

Comitê de Programa Técnico: Aníbal Ferreira (Univ. do Porto, Portugal)
Eduardo R. Miranda (Univ. Plymouth, UK)
Fábio Kon (IME-USP)
Fernando Iazzetta (ECA-USP)
Francisco J. Fraga (LSI-EPUSP)
João Antônio Zuffo (LSI-EPUSP)
João Benedito dos Santos Junior (PUC-MG)
Jônatas Manzolli (IA-UNICAMP)
Luiz Wagner Pereira Biscainho (EP-UFRJ)
Marcelo Gomes Queiroz (IME-USP)
Marcelo Knörich Zuffo (LSI-EPUSP)
Maurício Loureiro (EM-UFMG)
Miguel Arjona Ramirez (EPUSP)
Paulo Esquef (FPF-AM)
Pedro Donoso Garcia (EE-UFMG)
Phillip Burt (EPUSP)
Regis Rossi Alves Faria (LSI-EPUSP)
Rubem Dutra R. Fagundes (PUC-RS)
Sidnei Noceti Filho (EEL-UFSC)
Sylvio R. Bistafa (EP&FAU-USP)

Apoio logístico: Aurélio Antônio Mendes Nogueira
Elena Saggio
Leandro Ferrari Thomaz
Simone Carvalho
Maria Francesca Neglia

Agradecimentos: Thereza Leonard (AES Past President)
AES Board of Governors
Luiz Wagner P. Biscainho
Sidnei Noceti Filho
Silvia Regina Saran Della Torre

Editoração e arte: Totum Marketing e Comunicação

Realização / Promoção: AUDIO ENGINEERING SOCIETY - SEÇÃO BRASIL

Organização: Laboratório de Sistemas Integráveis da Escola Politécnica da USP

Apoio: SBC Sociedade Brasileira de Computação, Francal Feiras

Congresso de Engenharia de Áudio (4.: São Paulo: 2006); Convenção Nacional AES Brasil (10.: São Paulo: 2006)
Anais 4. Congresso de Engenharia de Áudio; 10. Convenção Nacional AES Brasil / ed. R.R.A. Faria, M.K. Zuffo - São Paulo: AES Brasil, 2006.
133 p.

ISBN 85

1. Engenharia de áudio (Congressos) 2. Computação musical (Congressos) 3. Processamento de sinais (Congressos)
I. Convenção Nacional AES Brasil (10.: São Paulo, 2006) II. Áudio Engineering Society. Seção Brasil III. Faria, Regis Rossi Alves IV. Zuffo, Marcelo Knörich V.t

CDD621.3828

* Anais em CD-Rom: ISBN 85-99997-01-7 (Anais em CD-Rom)

Impresso no Brasil.
Printed in Brazil.

Sociedade de Engenharia de Áudio
AES – Audio Engineering Society – Brazil Section
Endereço de correspondência: 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

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

Sessão 2 - Processamento Digital de Áudio, Voz e Sistemas Eletrônicos de Áudio

(Digital Audio and Speech Processing, and Audio Electronic Systems)

Sessão 3 - Sonorização Espacial, Som 3D, Acústica de Salas e Ambientes II

(Spatial sound systems, 3D Sound, Environmental and Room Acoustics II)

Sessão 4 - Síntese, Modelagem de Instrumentos e Computação Musical
(Synthesis, Instrument modelling and Computer Music)

Sessão 5 - Psicoacústica, Percepção Auditiva, Análise e Audição Automática
(Psychoacoustics, Auditory Perception, Analysis and Automatic Listening)

Prefácio dos Organizadores

É com grande prazer que escrevo esta introdução aos Anais do 4º Congresso da AES Brasil. Este ano experimentamos um crescimento substancial não só em quantidade mas também na infra-estrutura, divulgação e participação no Congresso. Para isso contribuiu de forma excepcional o apoio da Sociedade Brasileira de Computação que nos cedeu acesso ao sistema de submissões de artigos, facilitando enormemente nosso trabalho.

O Congresso ocupa um espaço especial em nosso encontro pois representa o ponto fundamental da sociedade, cujo objetivo é claro: estimular o estudo e o desenvolvimento do áudio. Foi pensando em como poderíamos apoiar esse avanço que empreendemos o esforço de organizar o Congresso há três anos.

Os verdadeiros heróis de um Congresso são o Coordenador do Programa (Papers Chair) e o Comitê. Eles convidam, imploram, mandam, chantageiam, cobram favores, bajulam, enfim fazem tudo para conseguir que autores apresentem trabalhos, com isso fazendo com que o todo seja muito maior do que a soma das partes. A esses dedicados colaboradores, meu mais sincero agradecimento.

O que eu posso escrever sobre esses Anais? Eles cobrem um amplo espectro de áreas extremamente especializadas. Seus autores são pesquisadores acadêmicos, fabricantes e profissionais do mais alto quilate. Os autores são nossos Bandeirantes do Século 21. Assim como seus antecessores de séculos atrás, os trabalhos que os autores nos trazem abrem novas trilhas que nos levam à fontes de sabedoria e conhecimento (o equivalente às minas de diamantes do passado).

Os trabalhos vão desde o teórico até aplicações que já encontram-se no mercado (ou quase). Eles representam o estado da arte em suas respectivas especializações.

Tenho a certeza de que o conhecimento aqui compartilhado será de muita utilidade a todos e que ano que vem teremos ainda mais trabalhos para apresentar. Aos Congressistas de 2006, meus votos de que aproveitem esses dias de intensa sinergia.

Joel Brito
Presidente AES Brasil
Coordenador Geral da Convenção

Sejam benvindos ao 4º Congresso da AES Brasil 2006 para três dias de uma programação rica e diversificada sobre as atualidades e avanços que nos aguardam num futuro próximo da engenharia de áudio e disciplinas afins. O tema da convenção este ano é “o áudio na era da comunicação” em linha com as mudanças e desafios trazidos pela digitalização dos nossos maiores meios de comunicação: o rádio e a televisão.

Vinte artigos distribuídos por 5 sessões foram publicados este ano, cobrindo novidades e contribuições inéditas principalmente nas áreas de processamento de áudio, áudio espacial, sonorização, acústica ambiental e computação musical. Para enriquecer ainda mais o evento, organizamos três workshops especiais: um sobre saúde auditiva (audiologia e questões relacionadas à preservação da audição), um voltado para a prática de medições acústicas, e um cobrindo o processo de implantação do rádio e da TV digital no Brasil, contando com especialistas, pesquisadores, representantes de agências governamentais, associações comerciais e convidados internacionais.

Este ano fizemos um esforço considerável para aumentar os números do congresso em termos de artigos e de participação, ampliando sua divulgação e construindo uma programação diversificada, que fosse ao mesmo tempo atraente para a academia, para os engenheiros e para os profissionais do áudio. Ampliamos o comitê de programa, convidando também membros da comunidade científica internacional, e buscamos apoio à divulgação junto à AES Internacional e Região Latino-Americana.

Juntamente com a convenção nacional da AES Brasil, os congressistas ainda terão acesso a uma intensa programação de palestras nacionais e internacionais abordando diversos tópicos em tecnologias e sistemas para áudio, bem como acesso à feira de exposições, demonstrações e atividades especiais espalhadas pelo centro de convenções.

São Paulo é uma metrópole plena de diversidade cultural e gastronômica, e a localização central do centro de convenções Rebouças facilita ainda a visita a museus, restaurantes e sofisticados centros de compras nos arredores. Finalmente queremos agradecer à AES Internacional e à SBC pelo apoio, e especialmente agradecer toda a colaboração e disposição dos membros do comitê técnico, dos revisores, secretários e demais profissionais envolvidos na realização deste evento.

Regis Rossi A. Faria
Coordenador do Congresso
Coordenador do Comitê de Programa Técnico

Organization Greetings

It is with pleasure that I write this introduction to the Proceedings of the 4th AES Brazil Conference. This year we experienced a substantial increase not only in quantity but also in infrastructure, spreading and participation in the conference. The institutional support from the Brazilian Computer Society contributed exceptionally to this, making available the access to its paper submission system, greatly easing the organization work.

The conference takes a special part in our meeting while representing the fundamental key of the society, which of course aims to foster the study and development of audio. It was thinking in how we could support these advances that we undertook the effort to organize this conference three years ago.

The actual heroes of a conference are the technical program chairman and the committee. They invite, beg, order, blackmail, charge favors, at last make everything to get that authors present their works, this way making the whole a lot larger than the sum of the parts. To these dedicated collaborators, my very sincere thanks.

What can I write about the proceedings? They cover a wide spectrum of extreme specialized areas. Their authors are academic researchers, manufacturers and professional of highest esteem. The authors are our pioneers of XXI century. As well as their antecessors centuries ago, their works take us to new trails to the source of knowledge and wisdom (equivalent to the diamond mines in the past).

The works go from theoretical to the applications already found in the market (or nearly). They represent the state-of-the-art in their respective specializations.
I am sure that all the knowledge here shared will be of great utility to all and that next year we will have yet more works to present. To the 2006 conferencees my votes that they enjoy these days of intense synergy.

Joel Brito
AES Brazil President,
Convention General Coordinator

Welcome to the 4th AES Brazil Conference 2006 for three days of a rich and diversified program over several novelties and forecoming advances in the audio engineering and related disciplines. This year's theme is "the audio in the communication era" in line with the changes and challenges brought by the digitalization of our most important communication media: the radio and the television.

Twenty papers distributed over 5 sessions were published this year, covering novel contributions mainly in the areas of audio processing, spatial audio, sound systems, environmental acoustics and computer music. To further enrich the event, we organized three special workshops: one about auditory health (audiology and issues related to auditory loss prevention), one turned to the practice of acoustic measurements, and one addressing the process of implantation of digital radio and TV in Brazil, counting with experts, researchers, representatives from government agencies and commercial associations, and international guests.

This year we made a considerable effort to increase the conference numbers both in terms of papers and participation, amplifying its spreading and building a diversified program at the same time interesting for the academia, engineers and the audio professionals. We enlarged the technical program committee, inviting also members from the international scientific community, and got the support from AES International and Latin America Region to spread the event.

Jointly with the AES Brazil National Convention, the conferencees will also have access to an intense program of national and international lectures approaching several topics in audio technologies and systems, as well as access to the exhibition, demos and special activities all over the convention center.

São Paulo is a metropolis full of gastronomic and cultural diversity, and the convention center localization is strategic for accessing museums, restaurants and sophisticated shopping spots around. Finally we want to thank the AES International and the SBC (Brazilian Computer Society) for their institutional support, and specially thank all the collaboration and disposition of the technical program committee members, reviewers, secretaries and other professionals involved in the realization of this event.

Regis Rossi A. Faria
Conference Coordinator
Technical Program Committee Chairman

Revisores

Reviewers

Aníbal Ferreira
Eduardo R. Miranda
Fábio Kon
Fernando Iazzetta
Fernando Pacheco
Francisco J. Fraga
João Antônio Zuffo
João Benedito dos Santos Junior
Jônatas Manzolli
Leandro F. Thomaz
Luiz Wagner Pereira Biscainho
Marcelo Gomes Queiroz
Marcelo Knörich Zuffo
Mário Minami
Maurício Loureiro
Miguel Arjona Ramirez
Monique Nicodem
Paulo Esquef
Pedro Donoso Garcia
Phillip Burt
Regis Rossi Alves Faria
Rubem Dutra R. Fagundes
Sergio Rodriguez Soria
Sidnei Noceti Filho
Sylvio R. Bistafa

Sessões de Artigos

Papers Sessions

Sessão 1

Sonorização Espacial, Som 3D, Acústica de Salas e Ambientes I

(Spatial sound systems, 3D Sound, Environmental and Room Acoustics I)

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Análise comparativa dos resultados dos parâmetros objetivos de avaliação da qualidade acústica de um auditório multifuncional, obtidos por meio de medições, simulações, e cálculos matemáticos.

Lineu Passeri Jr. (i), Sandra R. Moscati (ii), Paulo Pinhal (iii), Heloísa Helena Afonseca Silva (iv), e Sylvio R. Bistafa (v)

Faculdade de Arquitetura e Urbanismo da USP, Departamento de Tecnologia da Arquitetura, Cidade Universitária, 05424-970, São Paulo, SP.

(i) lineupasseri@uol.com.br
(ii) smoscati@uol.com.br
(iii) paulo@pinhalarquitetura.com.br
(iv) heloisahasarq@ig.com.br
(v) sbistafa@usp.br

RESUMO

Serão apresentados os resultados de medições de diversos parâmetros objetivos de avaliação da qualidade acústica de salas – obtidos *in-loco* – de um auditório multifuncional na Grande São Paulo. Em seguida, serão apresentados os resultados dos mesmos parâmetros obtidos por intermédio de um programa de simulação acústica por traçado de raios. Por fim, os resultados do tempo de reverberação obtidos nos dois experimentos serão comparados com aqueles que se obtém a partir da aplicação direta da fórmula de *Sabine*. As semelhanças e as diferenças entre os resultados dos mesmos parâmetros, obtidos de maneiras diferentes, serão analisadas e discutidas. De posse desses resultados, também serão analisadas algumas soluções de projeto do ambiente.

INTRODUÇÃO

Salas para usos específicos (concerto, ópera, teatro e música de câmara, por exemplo) não são comuns no Brasil, uma vez que tais especificidades não seriam condizentes com a demanda por espaços tão particulares. Assim, a grande maioria das salas, construídas ou em construção, no Brasil, são salas multifuncionais. Tais salas não se prestam a uma atividade específica, mas procuram oferecer características – acústicas e funcionais – capazes da abrigar o maior número possível de espetáculos dos mais diversos tipos.

Pode-se definir como "qualidade sonora" de uma sala o conjunto de atributos acústicos subjetivos que atendam às expectativas acústicas dos ouvintes. Para cada finalidade de sala, há atributos acústicos subjetivos correspondentes que devem ser atendidos. Em auditórios multifuncionais, espera-se que esses atributos sejam atendidos da forma mais ampla possível, dentro das limitações que salas desse tipo, via de regra, impõem.

Diversos fatores influenciam o resultado daquilo que ouvimos no interior de uma sala. Controlar esses fatores é, portanto, fundamental na determinação do resultado sonoro que se espera em seu interior. D'ANTONIO *et al* [1] descreve esses fatores como sendo: (i) as dimensões da sala, (ii) a geometria da sala, (iii) a localização do ouvinte e sua habilidade de escuta, (iv) a localização da(s) fonte(s) sonora(s), (v) os materiais de revestimento das superfícies internas da sala, e sua disposição no ambiente, (vi) e a qualidade dos equipamentos de reprodução do som – se houverem.

As características acústicas de uma determinada sala, também referidas como "atributos subjetivos de qualidade acústica e musical" foram descritas pela primeira vez por BERANEK [2] como sendo as seguintes: (i) presença, (ii) calor, (iii) intimidade, (iv) claridade, (v) difusão, e (vi) brilho ou textura. BARRON [3] relacionou as características arquitetônicas de salas de diversos tipos, tamanhos e finalidades com suas características acústicas.

Os parâmetros acústicos mais conhecidos – o tempo de reverberação e o nível de ruído de fundo – não se têm mostrado suficientes no sentido de atender aos atributos subjetivos julgados mais relevantes. Alguns índices objetivos, por sua vez, não se encontram ainda totalmente validados no sentido de estabelecerem correlações confiáveis com as impressões subjetivas que se espera atender nos diversos tipos de salas.

Nesse contexto, uma série de ferramentas digitais (programas computacionais) se propõe a fornecer dados confiáveis, tanto de predição quanto de análise e emissão dos resultados de parâmetros objetivos da qualidade acústica de um determinado ambiente. Faz-se, portanto, necessário investigar o desempenho desse tipo de ferramenta em um ambiente construído, comparando seus resultados com aqueles normalmente obtidos a partir do cálculo do tempo de reverberação com a aplicação da fórmula de Sabine.

RESUMO DOS PARÂMETROS OBJETIVOS E SUA CORRELAÇÃO COM ATRIBUTOS SUBJETIVOS

De acordo com SIEBEIN *et al* [4], diversos indicadores da qualidade acústica de salas de grandes dimensões podem ser calculados a partir de sua resposta impulsiva. Todos os indicadores são derivados de *p(t)*, ou seja, a pressão sonora ao longo do tempo, medida em diversos pontos de um mesmo ambiente, por intermédio de uma fonte sonora e um microfone.

Os indicadores mais comumente utilizados na avaliação acústica de salas são os seguintes:

Tempo de Reverberação (RT_{60})

É mais antigo e, ainda, o parâmetro objetivo mais importante na avaliação acústica de uma sala. Pode ser definido como o tempo necessário para que o nível de um som diminua de 60 dB, a partir do instante de sua interrupção, num determinado ambiente, expresso em segundos.

Early Decay Time (EDT_{10})

É o tempo necessário para que o som decaia de 10dB, multiplicado por seis, cujo resultado é extrapolado para uma curva representando o seu decaimento de 60dB, expresso em segundos.

Initial Time Delay Gap (ITDG)

Também chamado de "Retardo Inicial", é o tempo decorrido entre o som direto e a primeira reflexão num determinado ponto da sala. Este índice tem sido correlacionado com a impressão subjetiva de "intimidade".

Definition (D_{50})

Ou "Definição", baseia-se na característica da audição humana, na qual reflexões sonoras que cheguem ao ouvinte em até 50ms após a chegada do som direto, são consideradas benéficas, melhorando sua audibilidade. Seu cálculo é feito a partir da razão entre (1) a somatória das energias contidas no som direto e no som proveniente das reflexões até 50ms, e (2) a energia total da resposta impulsiva medida num determinado ponto da sala. É comumente correlacionada com a inteligibilidade da fala.

Clarity (C_{80})

De cálculo similar ao da Definição, com a diferença de que, neste caso, consideram-se como benéficas aquelas reflexões que chegam ao ouvinte em até 80ms após a chegada do som direto. Por esse motivo, tem sido usada para caracterizar a "clareza" ou a "transparência" da música em salas de concerto.

Early-to-late Energy Ratios (El_t)

É uma proporção logarítmica obtida a partir da resposta impulsiva da sala, entre a energia inicial (som direto) medida no intervalo de tempo $t_{[0,t]}$, e a energia final (som reverberante) medida no intervalo de tempo $t_{[t,\infty]}$.

Tempo central (t_s)

Trata-se do "centro de gravidade temporal" da resposta impulsiva ao quadrado. Caracteriza a duração da resposta impulsiva e, portanto, trata-se de uma medida do grau de interferência da sala no sinal.

Relative Loudness (L) ou Relative Strenght (G)

Definido como o nível de energia sonora num determinado ponto (em geral, uma poltrona) de uma sala, é medido a partir da energia sonora produzida por uma fonte no palco, em relação ao nível de energia sonora obtido a 10m da mesma fonte instalada em um ambiente anecóico. Este índice mede a contribuição efetiva das primeiras reflexões e da reverberação, à potência do som em um ambiente.

Bass Ratio based on EDT

Este índice foi proposto pela primeira vez por BERANEK [2], e utilizava as informações do tempo de reverberação por banda de freqüências, para avaliar o timbre (ou balanço tonal) de um ambiente, especialmente o seu "calor". Em 1994 propôs-se a substituição de RT_{60} por EDT_{10} e, atualmente, o índice é obtido por intermédio da relação entre (1) a soma dos *EDT*s em 125Hz e 250Hz dividida pela (2) soma dos *EDT*s em 500Hz e 1000Hz.

Treble Ratio based on EDT

Proposto pela primeira vez por CHIANG [5] para avaliar o timbre (ou balanço tonal) de um ambiente, especialmente o seu "brilho", este índice é obtido por intermédio da relação entre (1) a soma dos *EDT*s em 2000Hz e 4000Hz dividida pela (2) soma dos *EDT*s em 500Hz e 1000Hz.

Inter-Aural Cross Correlation Coeficient ($IACC_{80}$)

O índice *IACC* está diretamente relacionado à sensação de "espacialidade" da sala, uma vez que mede a diferença relativa entre mesmos sons percebidos pelos ouvidos direito e esquerdo do ser humano, num ponto determinado. Este índice é chamado de *Early Inter-Aural Cross Correlation Coeficient* ($IACC_E$ ou $IACC_{80}$) se o intervalo de tempo utilizado na apropriação dessa diferença estiver compreendido entre 0s e 80ms.

Lateral Energy Fraction (LEF)

Calculado por meio da proporção obtida entre (1) a energia sonora integrada nos primeiros 80ms após o som direto, em ambos os lados (ouvidos) de um espectador hipotético, dividida pelo (2) nível total de energia sonora nos mesmos 80ms, medido no mesmo ponto, este índice está supostamente correlacionado à sensação da "impressão espacial" por parte dos espectadores, sendo que valores mais elevados de *LEF* corresponderiam a uma maior sensação de "espacialidade" do ambiente.

Support (ST_1)

Proposto para medir o "apoio" ou o "suporte" que o som refletido pelas superfícies do palco dá aos músicos que lá estão se apresentando, porquanto está diretamente relacionado à sensação de "conjunto" e "balanço" dos músicos no palco.

De acordo com SIEBEIN *et al* [6], tais parâmetros têm sido cada vez mais utilizados no processo de projeto de salas de espetáculos, auditórios e teatros. Entretanto, ainda há muito a ser pesquisado, com o intuito de estabelecer de uma forma mais precisa quais as decisões do projeto de arquitetura que, realmente, interferem na resposta impulsiva em pontos diferentes de uma sala, e o quanto a resposta impulsiva da sala efetivamente contribui para o resultado da qualidade acústica percebida pelos espectadores.

BISTAFA [7] conduziu um trabalho em que oito teatros da cidade de São Paulo foram medidos segundo quatro dos treze parâmetros objetivos descritos acima – RT_{60}, EDT_{10}, C_{50}, e ST_1 – além de um quinto parâmetro *S*, denominado *speech sound level* (em português: nível sonoro da palavra falada). A principal conclusão desse trabalho é que os resultados obtidos nos oito teatros reiteram as recomendações de BARRON [3] para o projeto de salas com proscênio.

OBJETIVOS DO TRABALHO

Os objetivos do presente trabalho são (1) comparar os resultados de determinados parâmetros de avaliação da qualidade acústica de uma sala multifuncional, obtidos *in-loco*, por intermédio de medições, e obtidos por intermédio da utilização de um programa de simulação acústica por traçado de raios, (2) comparar alguns resultados anteriores com aqueles obtidos a partir da aplicação direta da fórmula de *Sabine*, (3) analisar e discutir as semelhanças e diferenças entre os resultados obtidos, e (4) analisar a influência das soluções de projeto do ambiente nos resultados obtidos.

BREVE DESCRIÇÃO DA SALA OBJETO DE ANÁLISE

A sala escolhida para ser objeto deste trabalho foi o Teatro Municipal Clara Nunes, localizado na cidade de Diadema, na Grande São Paulo.

Trata-se de uma sala de múltiplo uso, com capacidade para 434 espectadores, construída em 1983 e reformada ao longo do ano de 2004 (Fig. 1).

Figura 1: Vista parcial da platéia do Teatro Clara Nunes.

Seu palco original foi ampliado para permitir a apresentação de espetáculos de diversos tipos, incluindo grupos de música de câmara e orquestras (Fig. 2).

Figura 2: Vista parcial do palco do Teatro Clara Nunes.

O piso da platéia é de concreto revestido com borracha tipo Plurigoma™. As paredes laterais são revestidas em lambris de madeira e placas vibrantes. A parede dos fundos é revestida por painel absorvente em lã de rocha. O forro é constituído por painéis difusores policilíndricos, construídos em compensado de madeira.

DESCRIÇÃO DOS PROCEDIMENTOS

Os parâmetros objetivos analisados neste trabalho foram: Tempo de reverberação ($T_{30+}T_{60}$), Early Decay Time (EDT_{10}), Definição (D_{50}) e Clareza (C_{80}).

A partir da conclusão das obras de reforma da sala objeto deste trabalho, e da adequação dos desenhos de projeto *as built*, as seguintes atividades foram desenvolvidas:

Medições *in-loco*

Os parâmetros objeto deste trabalho foram medidos em 9 (nove) pontos na platéia, sendo três na 3ª fila de poltronas (um à direita, um no centro e um à esquerda), três na 8ª fila (um à direita, um no centro e um à esquerda), e três na 13ª fila (um à direita, um no centro e um à esquerda).

As medições foram feitas com a sala sem ocupação. Em todas as situações, a sala foi excitada a partir do estouro de balões de borracha, colocados no palco, a 1,50m de altura do piso (Fig. 3).

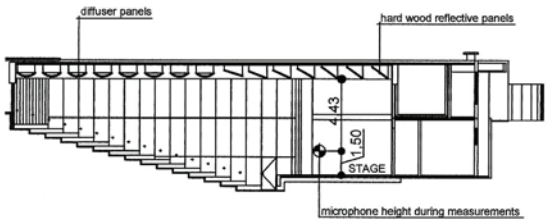


Figura 3: Indicação da localização da fonte sonora no palco do Teatro Clara Nunes

A captação foi feita por meio de um microfone omnidirecional Shure™ Beta 58, e o sinal foi processado por intermédio do programa computacional Aurora™ (8).

Simulação acústica da sala

Os parâmetros objeto deste trabalho foram então calculados nos mesmos 9 (nove) pontos na platéia, sendo três na 3ª fila de poltronas, três na 8ª fila, e três na 13ª, considerando a sala sem ocupação.

Neste experimento, após a modelagem em AutoCAD™, a exata localização da fonte e dos nove receptores, passou-se à simulação acústica da sala, por intermédio do programa de traçado de raios Catt Acoustic™, versão 7.2 (9).

Cálculo do tempo de reverberação utilizando a fórmula de Sabine

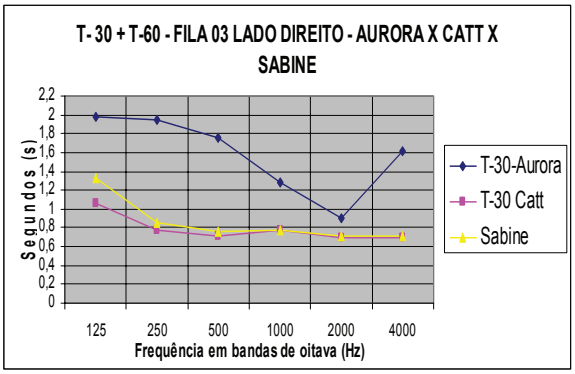
Por fim, calculamos o Tempo de reverberação (T_{60}) da sala a partir da fórmula de Sabine (10).

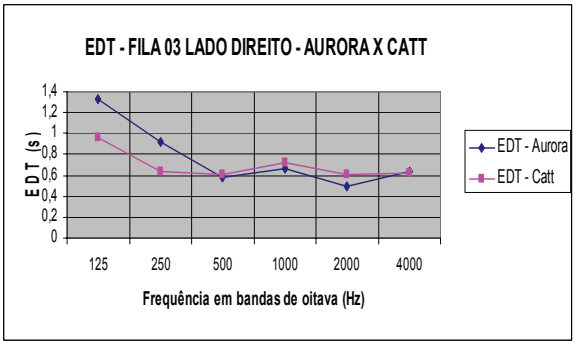
RESUMO DOS RESULTADOS OBTIDOS

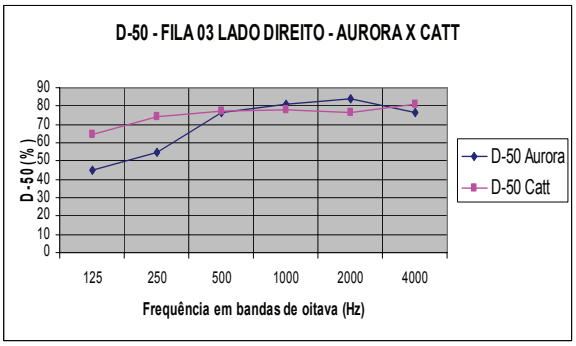
Dos nove pontos de medição e simulação, apresentaremos os resultados comparativos de três deles (3ª fila, à direita; 8ª fila, ao centro; e 13ª fila, à esquerda), resultados estes que foram impressos nos gráficos mostrados a seguir, para melhor visualização de suas semelhanças e diferenças:

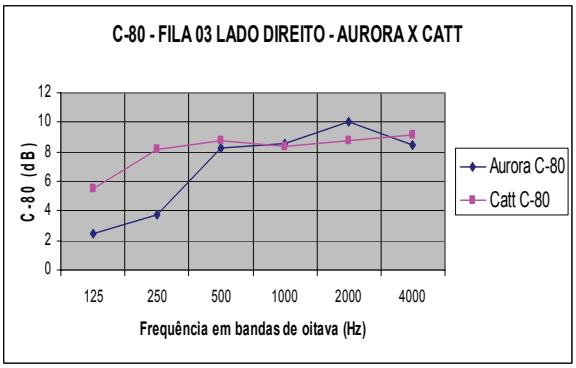
Resultados obtidos na 3ª fila, à direita

Os resultados de $T_{30+}T_{60}$, EDT_{10}, D_{50} e C_{80} foram os seguintes:



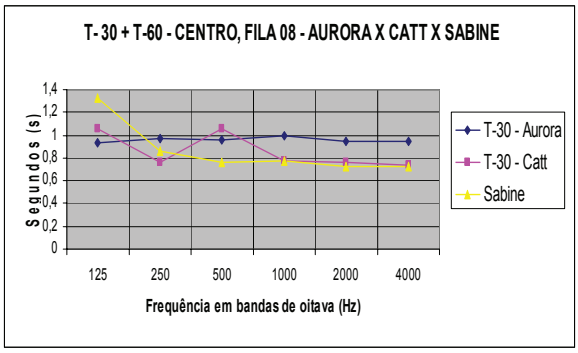


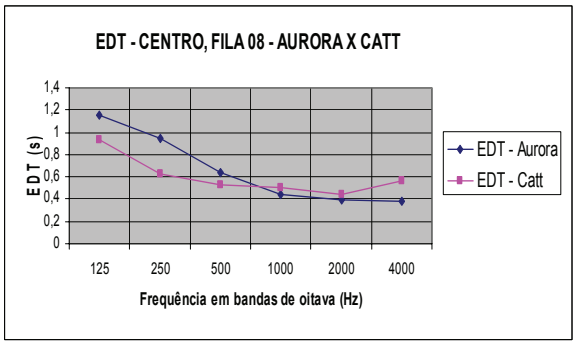


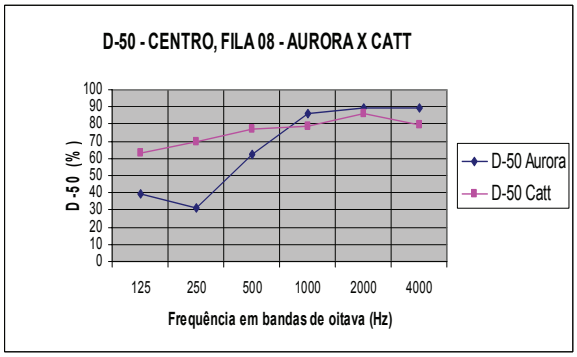


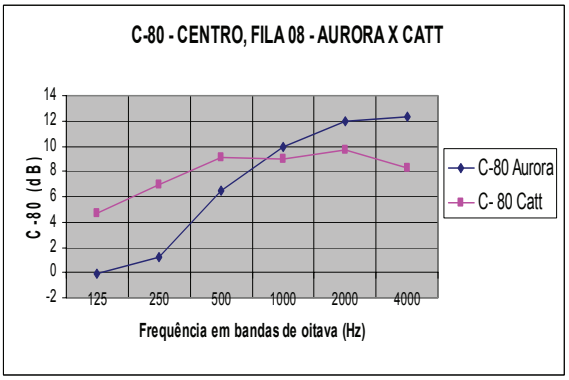
Resultados obtidos na 8ª fila, centro da sala

Os resultados de $T_{30+}T_{60}$, EDT_{10}, D_{50} e C_{80} foram os seguintes:



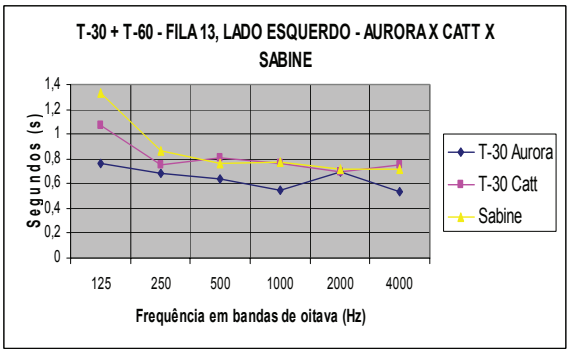


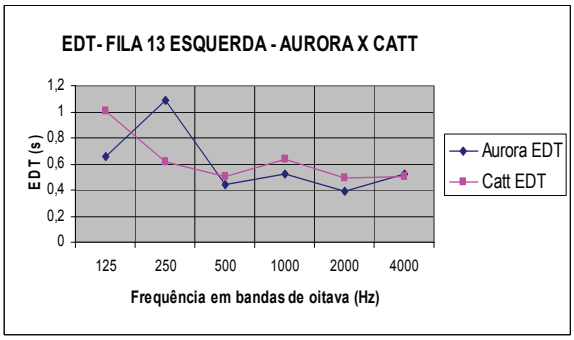


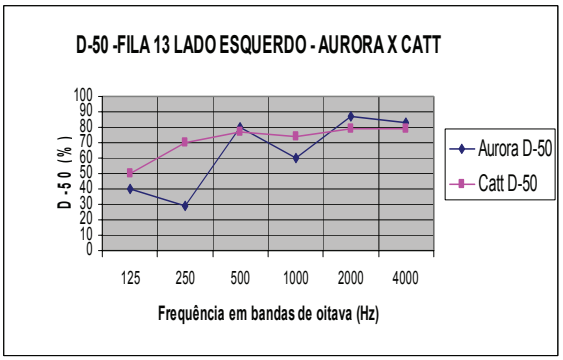


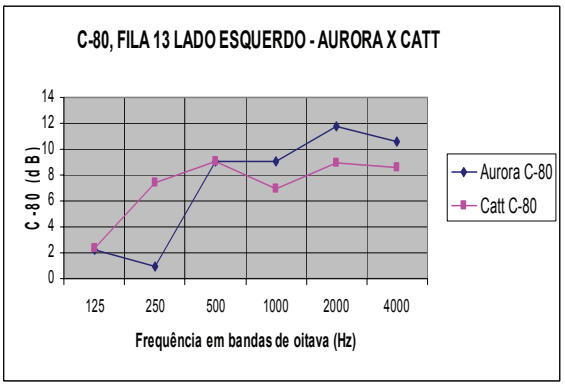
Resultados obtidos na 13ª fila, à esquerda

Os resultados de $T_{30+}T_{60}$, EDT_{10}, D_{50} e C_{80} foram os seguintes:









CONCLUSÕES

Em relação ao Tempo de reverberação

Os resultados de $T_{30+}T_{60}$, na 3ª, 8ª e 13ª fileiras revelam que as curvas relativas à simulação da sala obtidas por intermédio do programa de traçado de raios utilizado neste trabalho, e aquelas obtidas a partir da fórmula de Sabine, são muito semelhantes em seu comportamento, ainda que o resultado obtido por meio do programa de simulação, na 8ª fila, tenha apresentado um "pico" em 250Hz. Já as curvas obtidas a partir das medições *in-loco* distanciam-se das demais, nos três pontos.

Portanto, em relação ao Tempo de reverberação, podemos concluir que é possível obter resultados expeditos bastante seguros a partir da utilização da fórmula de Sabine, ao longo do desenvolvimento de projetos de ambientes de audição, permitindo que sua simulação, por meio de um programa de traçado de raios, seja feita na fase final do projeto, servindo para "afinar" a sala.

Em relação ao EDT

Os resultados de EDT_{10} na 3ª, 8ª e 13ª fileiras, por sua vez, mostraram que as curvas relativas à simulação da sala obtida por intermédio do programa de traçado de raios utilizado neste trabalho, e aquelas obtidas a partir das medições in-loco, apresentam comportamento e resultados bastante próximos, a partir de 500Hz. Abaixo disso, os resultados obtidos *in-loco* são superiores, nas três situações, provavelmente em decorrência da qualidade do microfone utilizado.

Em relação ao EDT_{10}, podemos concluir que é possível obter resultados seguros com um programa de traçado de raios como o que foi utilizado neste experimento, atentando para a necessidade de se fazer uso de um microfone com bom desempenho, principalmente no que se refere à captação dos sons de baixas freqüências.

Em relação ao D_{50}

Somente a partir de 1000Hz, ainda que na 13ª fila verificaram-se algumas discrepâncias. Porém, de um modo geral, os resultados de D_{50} revelam que as curvas obtidas por intermédio do programa de traçado de raios e aquelas obtidas a partir das medições *in-loco* apresentam comportamento e resultados bastante próximos. Abaixo de 1000Hz, os resultados obtidos *in-loco* apresentam distorções que não permitem avaliar o funcionamento do programa.

É provável que tais distorções nos resultados abaixo de 1000Hz, obtidos por meio de medições in-loco, seja igualmente decorrente da qualidade do microfone utilizado, o que aponta para a necessidade de se fazer uso de um microfone com bom desempenho, principalmente no que se refere à captação dos sons de baixas freqüências.

Em relação ao C_{80}

Já os resultados de $C_{80,}$ nos mesmos três pontos, demonstram que as curvas obtidas por intermédio do programa de traçado de raios utilizado neste trabalho, e aquelas obtidas a partir das medições, apresentam comportamento relativamente próximo a partir de 1000Hz, porém com resultados distintos. Abaixo dessa freqüência, ambas as curvas apresentam comportamento e resultados que, a exemplo de D_{50}, não permitem avaliar o seu desempenho.

Possivelmente, tais distorções sejam decorrentes da qualidade do microfone utilizado, o que aponta para a necessidade de se fazer uso de um microfone com bom desempenho, principalmente no que se refere à captação dos sons de baixas freqüências. No entanto, tendo em vista os resultados dos demais índices, talvez seja necessário refazer o procedimento para medição deste parâmetro.

Em relação ao projeto da sala

A conclusão mais significativa, em relação ao projeto da sala, pode ser obtida a partir da observação dos gráficos de EDT_{10} na 3ª, 8ª e 13ª fileiras. Nota-se que os resultados do comportamento da sala medidos *in-loco*, nas três situações, apresentam valores superiores àqueles obtidos por meio do programa de simulação.

É possível que tais diferenças sejam decorrentes dos coeficientes de absorção considerados para as placas vibrantes instaladas no ambiente, cujo desempenho real seja inferior àquele levado em conta no cálculo computacional.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] D'ANTONIO, P. & Cox, T. J. "Room optimiser: a computer program to optimise the placement of listener, loudspeakers, acoustical surface treatment, and room dimensions in critical listening rooms". 103rd AES Convention, preprint 4555, paper H-6, New York (1997).

[2] BERANEK, Leo Leroy. "Music, acoustics and architecture". John Willey & Sons, Inc., USA (1962).

[3] BARRON, M. "Auditorium acoustics and architectural design". E & Fn Spon, London, UK (1993).

[4] SIEBEIN, G. W. & Gold, M. A. "The concert hall of the 21st century: historic precedent and virtual reality". Architecture: material and imagined, Proceedings of the 85th ACSA Annual Meeting., Washington, DC, pp 52-61 (1997).

[5] CHIANG, W. "Effects on architectural parameters on six acoustical measures in auditoria". Ph.D. Dissertation, University of Florida, Gainesville, FL (1994).

[6] SIEBEIN, G. W. & Kinzey Jr., B. Y. "Recent innovations in acoustical design and research". In: Architectural acoustics: principles and practice (edited by William Cavanaugh & Joseph Wilkes), John Wiley & Sons, Inc., New York, NY (1999).

[7] BISTAFA, Sylvio R. "The acoustics for speech of eight auditoriums in the city of São Paulo". First Pan-American/Iberian meeting on acoustics, Cancún, MX (2002).

[8] FARINA, Angelo. In http://www.ramsete.com/aurora.

[9] DALENBÄCK, Bengt-Inge. In http://www.catt.se.

[10] SABINE, Wallace C. "Collected papers on acoustics", 1993, Peninsula Publishing, Los Altos, US.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Sistema Eficiente para Auralização Usando Agrupamento e Modelagem de HRTFs por Wavelets

Julio C. B. Torres[1], Mariane R. Petraglia[1], Roberto A. Tenenbaum[2]

[1]Escola Politécnica - Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ, Brasil
juliotorres@ufrj.br
mariane@pads.ufrj.br

[2]IPRJ - Universidade do Estado do Rio de Janeiro
Nova Friburgo, RJ, Brasil
tenenbaum@iprj.uerj.br

RESUMO

Os sistemas de realidade virtual acústica requerem uma complexidade computacional muito elevada para reproduzir as características tridimensionais do som. Uma forma de reduzir a complexidade de tais sistemas é modelar de forma eficiente a propagação do som. Utilizando um modelo baseado na decomposição das funções de transferência relacionadas ao receptor (HRTFs) por uma transformada wavelet, este artigo apresenta um sistema de auralização eficiente, que explora a similaridade dos coeficientes do modelo correspondentes às baixas freqüências das HRTFs provenientes de direções próximas.

INTRODUÇÃO

Nos últimos anos, tem-se observado um crescimento considerável dos sistemas de áudio imersivo, seja em sistemas com diversos alto-falantes ou através de fones de ouvido. Tal crescimento deve-se principalmente ao desenvolvimento de novas tecnologias e da necessidade do ser humano sentir-se imerso no programa áudio-visual. Um exemplo disso é a recente inclusão de faixas de áudio em DVDs, gravadas com cabeças artificiais, que possibilitam ao ouvinte perceber as características tridimensionais do som no momento da gravação. Porém, esse tipo de gravação não permite ao ouvinte modificar sua posição dentro do campo sonoro.

A fim de permitir que o ouvinte interaja com o sistema de áudio, modificando sua posição, orientação e até características do campo sonoro, foram criados os sistemas de realidade virtual acústica (SRVAs). Estes sistemas exigem um elevado grau de complexidade para que o som produzido seja equivalente ao gravado com cabeças artificiais e, mesmo com o desenvolvimento tecnológico atual, não é possível a utilização desses sistemas em tempo real. A utilização em tempo real só se torna possível caso sejam aceitas simplicações no sistema. Contudo tais simplificações implicam, geralmente, na redução da qualidade e da fi-

delidade do áudio produzido, quando comparado com um sistema não simplificado.

Uma forma de reduzir a complexidade dos sistemas de realidade virtual acústica é modelar de forma mais eficiente a propagação do som. A modelagem do receptor se dá através das funções de transferência relacionadas à cabeça (*Head-Related Transfer Functions* – HRTFs) [1, 2], que correspondem a pares respostas impulsivas (HRIRs) medidas para diversas direções ao redor do receptor.

Quando se deseja fazer com que um indivíduo perceba que uma fonte sonora encontra-se em um determinado ponto ou direção do espaço ao seu redor, deve-se então convoluir o sinal anecóico produzido pela fonte com as HRIRs relativas a essa direção. Removendo-se a influência do sistema de reprodução, como por exemplo realizando uma equalização de fones de ouvido, o som percebido deverá ser idêntico ao ouvido em um ambiente real livre de reverberação (sala anecóica).

Um sistema de realidade virtual acústica pode possuir diversas fontes; mesmo com apenas uma fonte, as ondas sonoras por ela emitidas podem sofrer múltiplas reflexões nas superfícies da sala. Assim, para cada direção possível de chegada de uma frente de onda no receptor, o sinal da fonte deverá ser convoluído com a HRIR da respectiva direção. Observa-se, portanto, que quanto mais reverberante for um ambiente, maior será o número de direções necessárias para gerar o sinal de áudio tridimensional.

Entretanto, o ser humano possui uma capacidade limitada em reconhecer a direção exata de uma fonte sonora [3]. A capacidade média do ser humano varia entre 5° e 20° [1] e, portanto, um conjunto discreto de direções pode ser utilizado para medir as HRTFs sem perda da capacidade de reconhecimento de direção. Geralmente utilizam-se aproximadamente 700 direções ao redor da cabeça, com a fonte situada entre 1 e 1,2 metros, resultando em um conjunto de 1400 HRTFs [4, 2].

O custo computacional de um sistema com processamento simultâneo de diversas direções pode ser reduzido de duas formas: diminuindo o número de direções e/ou reduzindo o comprimento das HRIRs. Reduzir o número de direções pode levar à degradação da "espacialidade" do áudio, uma vez que nem todas as direções nas quais o som poderia atingir o receptor seriam utilizadas na simulação. A redução do comprimento das HRIRs também poderá interferir na percepção da direção. Porém, se as características espectrais de cada direção forem mantidas, será possível reduzir seu comprimento sem interferir na qualidade da auralização.

Essa redução foi realizada com sucesso através da modelagem das HRTFs por transformadas wavelets e filtros esparsos [5, 6, 7], onde obteve-se uma redução de aproximadamente 70% em relação a sua implementação tradicional. Assim, uma HRIR que originalmente possuia 100 coeficientes no tempo pôde ser implementada por uma transformada wavelet acrescida de um conjunto de 30 coeficientes.

Apesar desse ganho computacional considerável, obtido com a modelagem por wavelets, a grande redundância de informação do conjunto de HRTFs pode ser utilizada para reduzir ainda mais o custo computacional. Nesse sentido, verificou-se que, na faixa de baixas freqüências, as HRTFs de direções próximas possuem um comportamento similar. Essa similaridade existe pois sons de baixa freqüência possuem grandes comprimentos de onda, maiores até que um torso humano, o que dificulta ao ser humano definir a direcionalidade da fonte, principalmente devido ao efeito de difração. Essa dificuldade em reconhecer a direção dos sons de baixa freqüência se traduz em uma característica praticamente plana do módulo das HRTFs até aproximadamente 1kHz.

Com base nesse modelo de HRTFs com wavelets, este artigo apresenta uma análise de como o processamento do som proveniente de direções próximas pode ser reduzido. Esse ganho de desempenho é obtido considerando-se a similaridade dos coeficientes da wavelet responsáveis pelas freqüências baixas das HRTFs.

CARACTERÍSTICAS DAS HRTFs

As HRTFs são funções cujas respostas em freqüência variam conforme a direção da fonte sonora. A Fig. 1 apresenta os módulos das respostas em freqüência de um conjunto de HRTFs pertencentes ao plano horizontal situado na altura das entradas dos canais auditivos. Este plano é equivalente a uma elevação de 0° em um sistema de coordenadas esféricas.

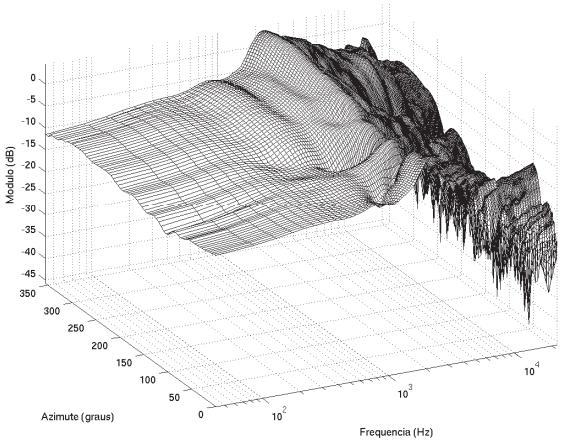


Figura 1: Módulo da resposta em freqüência das HRTFs com elevação de 0°.

Da Fig. 1 pode-se observar que na área de baixas freqüências (de 20 Hz a 1kHz) não há praticamente variação no módulo das HRTFs em função do ângulo de azimute. Este padrão se mantém para as

demais elevações onde se tem medição das HRTFs. As variações no módulo e na fase das HRTFs e as diferenças entre as HRTFs de direções diferentes auxiliam na identificação da localização da fonte sonora. Como em baixas freqüências não há praticamente diferenças, nessa faixa as HRTFs não fornecem informação necessária para o reconhecimento da direção. Neste caso, prevalecem as diferenças interaurais de tempo e de nível de pressão sonora na discriminação da direção [8, 9].

MODELAGEM DAS HRTFS COM A TRANSFORMADA WAVELET

Nessa abordagem a HRIR é vista como um sistema de resposta impulsional finita (FIR) e a modelagem é realizada com base na decomposição polifásica da sua função de transferência [10, 11, 12], como mostrado na Fig. 2.

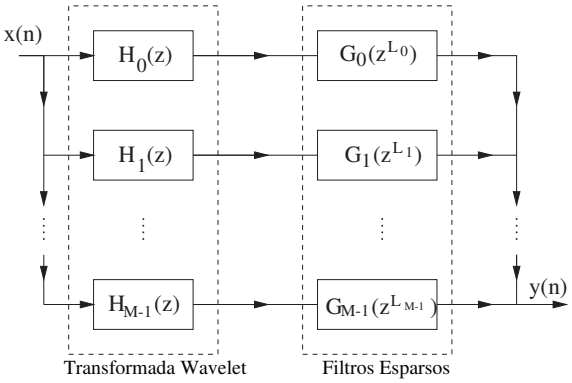


Figura 2: Sistema que utiliza a transformada wavelet para modelagem de uma HRTF.

Na Fig. 2 o banco de filtros de análise $H_m(z)$ implementa uma transformada wavelet discreta e os filtros esparsos $G_m(z^{L_m})$ são filtros cujos coeficientes proporcionam uma resposta impulsiva igual à HRIR da direção que está sendo modelada [13]. Os filtros base utilizados na implementação da transformada wavelet foram selecionados por apresentarem a melhor relação custo/benefício entre a seletividade e o comprimento [7]. Após diversos testes com diferentes filtros, inclusive biortogonais, os filtros protótipos Daubechies de comprimento 8 (daub8) [14] foram empregados em quatro estágios em uma estrutura de decomposição em oitavas.

Como exemplo, na Fig. 3 estão apresentados os coeficientes dos filtros esparsos $G_m(z^{L_m})$ que modelam as HRTFs de cada ouvido para a direção definida pela elevação $\phi = 0°$ e o azimute $\theta = 90°$ (fonte situada a 90° à direita do ouvinte).

REDUÇÃO DO CUSTO COMPUTACIONAL

Nesta seção são apresentadas duas técnicas baseadas nas características espectrais das HRTFs e dos coeficientes obtidos com a modelagem através de wavelets para reduzir o custo computacional e tornar um sistema de realidade virtual acústica mais eficiente. Primeiro será utilizado um procedimento para reduzir o número total de coeficientes esparsos, considerando um critério de perda de energia das HRTFs. Em seguida, o custo de implementação das HRTFs de direções próximas será reduzido, considerando a similaridade dos coeficientes.

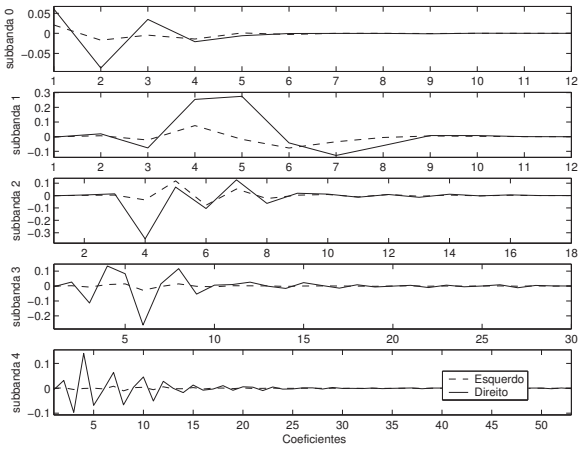


Figura 3: Coeficientes dos filtros esparsos de cada subbanda e de cada ouvido para a direção (0°, 90°).

Redução do Número de Coeficientes

A redução do número de coeficientes é obtida através de uma análise da energia acumulada dos coeficientes em cada subbanda. Contudo, a energia de cada HRTF varia conforme a direção. Os valores máximo e mínimo de energia ocorrem para os ângulos de azimute de 90° e 270°, respectivamente. Dessa forma, um critério de energia não deve ser definido em termos absolutos, mas sim em percentuais de energia em cada subbanda, para cada direção.

A energia da HRIR $E(\phi, \theta)$ é dada por

$$E(\phi, \theta) = \sum_{n=0}^{N-1} p^2_{\phi,\theta}(n), \tag{1}$$

onde N é o comprimento da HRIR $p_{\phi,\theta}(n)$.

A energia por subbanda $E_m(\phi, \theta)$ é dada por

$$E_m(\phi, \theta) = \sum_{k=0}^{K_m-1} g^2_{m,k}(\phi, \theta), \tag{2}$$

onde K_m é o número de coeficientes esparsos da subbanda m.

A contribuição cumulativa de cada coeficiente esparso, em cada subbanda, pode ser observada na Fig. 4, para o ouvido direito e direção $\phi = 0°$ e $\theta = 90°$. A soma das energias acumuladas em cada subbanda fornece a energia total da HRIR.

Conforme pode ser observado na Fig. 4, a energia cumulativa na terceira banda, por exemplo, atinge

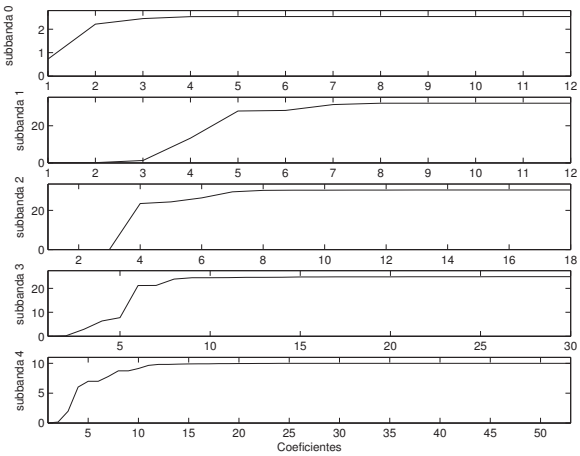


Figura 4: Energia cumulativa dos coeficientes esparsos para a direção $\phi = 0°$ e $\theta = 90°$, para o ouvido direito.

valor considerável somente após o terceiro coeficiente e tem praticamente toda energia acumulada até o sétimo coeficiente. Assim, se desprezarmos os coeficientes anteriores ao terceiro e posteriores ao sétimo nesta banda restarão apenas cinco coeficientes. Esta mesma análise é aplicada a todas as subbandas, porém definindo-se limites de tal forma que a energia total perdida com os coeficientes não-significativos seja no máximo 10% da energia da HRIR original. Aplicando o critério descrito em [7] para todas as direções, obtém-se os intervalos (janelas) descritos na Tab. 1. Esses intervalos garantem que haverá uma perda máxima de 10% de energia em cada HRTF. Entretanto, para diversas direções a perda não é máxima. Como mostrado em [7], a perda de 10% da energia total da HRTF através da redução dos coeficientes esparsos produz menos erros em frequência do que a perda direta de coeficientes das HRIRs. Uma análise do erro devido à redução dos coeficientes é apresentada em [15]

Protótipo	subbanda					total
Daub8	0	1	2	3	4	$\tilde{K}$
Intervalos	1-6	3-7	4-7	3-9	3-8	
No. coefs.	6	5	4	7	6	28

Tabela 1: Intervalos e número de coeficientes significantes dos filtros esparsos para cada subbanda.

Dessa forma, o número de coeficientes pode ser reduzido para aproximadamente 30% do total se considerarmos em cada subbanda apenas os coeficientes de maior significância. A energia perdida com o descarte de coeficientes é de no máximo 10% da energia total da HRTF e não altera significativamente o conteúdo espectral das mesmas. No exemplo da Fig. 4, a energia perdida é de apenas 4%, pois esses intervalos foram obtidos com uma média para todas as direções.

Redução do Número de Direções

Os coeficientes de cada subbanda são responsáveis por uma região do espectro da HRTF e a influência desses coeficientes nas demais bandas depende da seletividade dos filtros protótipos utilizados na estrutura em oitavas. Considerando que o protótipo utilizado (daub8) possui uma relação satisfatória entre seletividade e custo de implementação (comprimentos dos filtros $H_m(z)$ e atrasos produzidos), pequenas variações nos valores dos coeficientes das bandas 0 e 1 (freqüências mais baixas) não produzem alterações significativas nas demais bandas. O erro médio quadrático para as demais bandas é da ordem de -40 dB. Por outro lado, variações nos coeficientes da última banda provocam alterações em todo o espectro, devido à baixa seletividade do filtro de análise nessa banda.

Se considerarmos uma região do espaço ao redor do receptor (definida por um intervalo de valores de elevação e azimute) [16], dentro dessa região haverá diversas HRTFs que por sua vez serão substituídas pelas funções reduzidas, conforme a modelagem proposta. Analisando os coeficientes obtidos em uma determinada banda para todas as direções pertencentes a essa região do espaço, observa-se que os coeficientes relativos às baixas e médias freqüências possuem pouca variação. Para bandas mais altas, a variação dos coeficientes é mais acentuada. Isto é esperado por dois motivos: a baixa seletividade dos filtros das bandas mais altas e a grande variação existente entre os espectros das HRTFs em alta freqüência.

Considerando a direção $\phi = 0°$ e $\theta = 90°$ como principal e utilizando um ângulo de abertura de 40° tanto na elevação quanto no azimute, tem-se uma região cujas extremidades são $-20° < \phi < 20°$ e $70° < \theta < 110°$. A Fig. 5 apresenta na primeira coluna os coeficientes de todas as HRTFs percententes a esta região, por subbanda. Nessa figura pode-se observar a variação dos valores dos coeficientes devida à variação de direção. Na segunda coluna são apresentadas, por subbanda, as curvas correspondentes à média e à media mais o desvio padrão dos coeficientes.

Analisando as variações dos valores dos coeficientes, verifica-se que os maiores desvios ocorrem nas duas últimas bandas. Se não há praticamente variação nos coeficientes das bandas mais baixas, e uma pequena variação não é capaz de introduzir distorções consideráveis na resposta em freqüência, devido à seletividade dos filtros da wavelet, então é possível utilizar um conjunto comum de coeficientes para a mesma banda de todas as HRTFs da região.

Substituindo-se os coeficientes originais da primeira subbanda de uma dada HRTF da região pela média dos coeficientes da primera subbanda de todas as HRTFs da mesma região, verifica-se que essa modificação realmente não introduz variação que afete a percepção da direção do som processado. Isto pode ser observado na Fig. 6, onde o módulo e a fase da resposta em

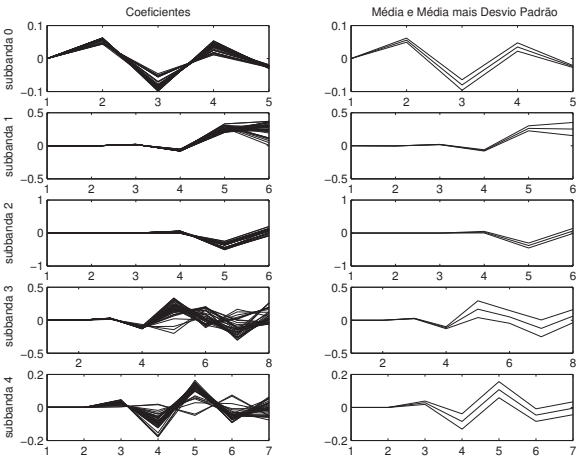


Figura 5: (a) Variação dos coeficientes de cada banda para as HRTFs de uma região e (b) média e média mais desvio padrão dos coeficientes.

freqüência da HRTF original (coeficientes originais) são comparados com os da HRTF onde os coeficientes da primeira banda foram substituídos pela média dos coeficientes de todas as primeiras bandas. A Fig. 6 apresenta o resultado obtido para a direção $(0°, 90°)$, para ambos ouvidos. Este comportamento é similar ao das demais direções dessa região.

Utilizando a média dos coeficientes das duas primeiras bandas obtém-se o resultado apresentado na Fig. 7. A Fig. 8 apresenta o resultado obtido utilizando-se os coeficientes médios das três primeiras bandas.

A partir dos gráficos apresentados nas Figs. 6 a 8 pode-se verificar que a substituição dos filtros esparsos responsáveis pelas baixas e médias freqüências não afetam significativamente as resposta em freqüência das HRTFs pertencentes a essa região do espaço.

Dessa forma um considerável ganho computacional pode ser obtido se, ao invés de processamos todas subbandas de todas as direções da região, realizarmos o processamento individual apenas das últimas subbandas de cada direção (HRTF) e apenas uma vez as primeiras subbandas, visto que estas serão iguais para todas as direções da região. Tomemos como exemplo uma região com 25 direções e cada direção com 28 coeficientes esparsos, conforme a Tab. 1. Sem a utilização do método proposto, seriam necessárias $25 \times 28 = 700$ operações de soma e multiplicação. Utilizando-se a média das bandas 0 e 1 em substituição dos coeficientes originais, serão necessárias apenas $11 + 25 \times 17 = 436$ operações, proporcionando uma redução de 37,7% na carga computacional.

Fica evidente que quanto maior for a região (maiores ângulos de abertura) maior será o ganho computacional. A análise apresentada neste artigo refere-se a regiões com ângulo de abertura de aproximadamente 40° ao redor de uma direção principal. É importante ressaltar que há uma relação de compromisso entre

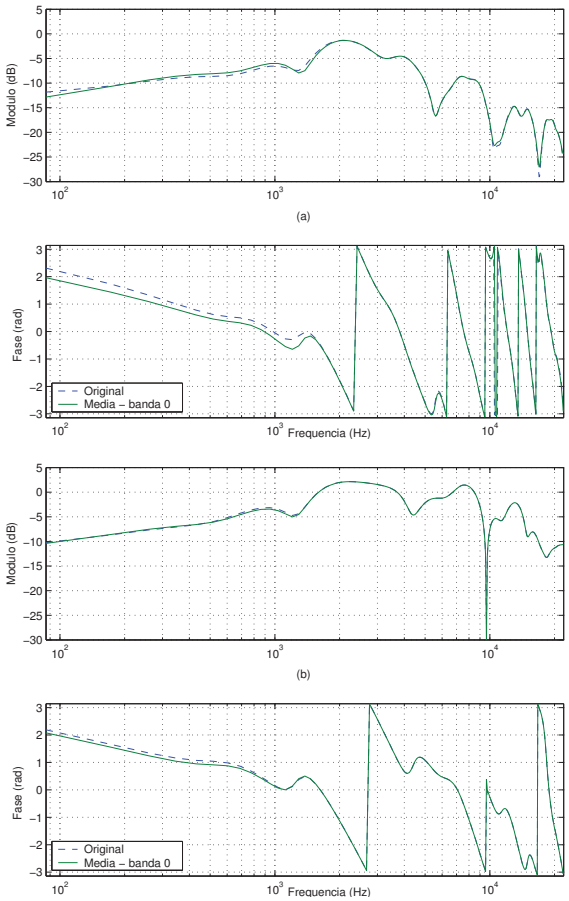


Figura 6: Comparação de módulo e fase entre as respostas em freqüência para a direção $(0°, 90°)$, substituindo os coeficientes da primeira banda pelos coeficientes médios: (a) ouvido esquerdo e (b) ouvido direito.

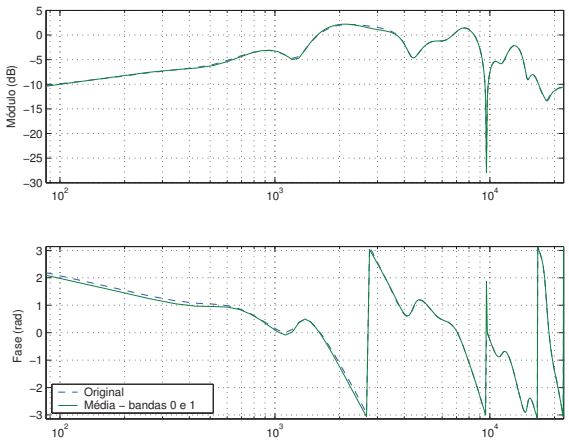


Figura 7: Comparação entre as respostas em freqüência para a direção $(0°, 90°)$, substituindo os coeficientes das duas primeiras bandas pelos respectivos coeficientes médios.

o ganho computacional e a qualidade de auralização, que será influenciada pelos desvios nas respostas em freqüência das HRTFs em função do número de

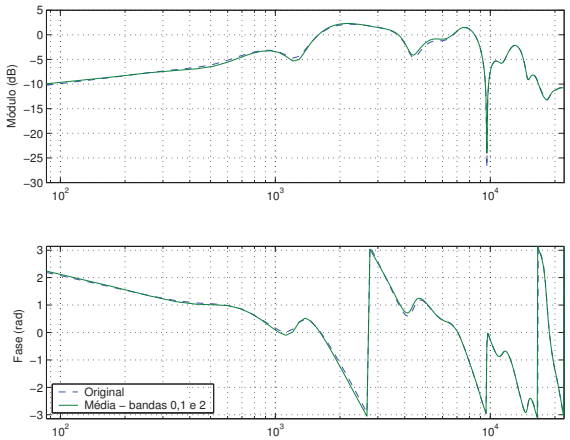


Figura 8: Comparação entre as respostas em freqüência para a direção (0°, 90°), substituindo os coeficientes das três primeiras bandas pelos respectivos coeficientes médios.

direções englobadas em um região do espaço. Assim, diversos testes subjetivos serão ainda necessários a fim de avaliar, sob o aspecto psico-acústico, quais são os ângulos de abertura e as direções principais que fornecem a melhor relação qualidade/ganho computacional.

CONCLUSÕES

Neste artigo foi apresentado um sistema para auralização com complexidade computacional reduzida, baseado em um modelo eficiente para as HRTFs e no agrupamento destas funções para direções próximas. Este agrupamento é possível devido à similaridade dos coeficientes do modelo correspondentes às freqüências baixas das HRTFs. Através da análise do erro gerado pela simplificação proposta, podem ser definidos os ângulos de abertura (azimute e elevação) e o número de direções agrupadas, sem que a qualidade do sistema de áudio 3D seja prejudicada, considerando sua aplicação em um sistema de realidade virtual acústica (acústica de salas).

REFERÊNCIAS BIBLIOGRÁFICAS

[1] J. Blauert, *Spatial Hearing*, The MIT Press, Cambridge, 1997.

[2] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *WASPAA '01 (2001 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics)*, Oct. 2001, CIPIC website: http://interface.cipic.ucdavis.edu/.

[3] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2841–2853, May 1999.

[4] W. G. Gardner and K. D. Martin, "HRTF measurements of a kemar," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908, 1995, MIT website: http://sound.media.mit.edu/KEMAR.html.

[5] J. C. B. Torres, M. R. Petraglia, and R. A. Tenenbaum, "Auralização de salas utilizando wavelets para modelagem das HRTFs," *Seminário de Engenharia de Áudio*, 2002.

[6] J. C. B. Torres and M. R. Petraglia, "Performance analysis of an adaptive filter employing wavelets and sparse subfilters," in *EUSIPCO 2000*, Sep 2000, vol. II, pp. 997–1001.

[7] J. C. B. Torres, M. R. Petraglia, and R. A. Tenenbaum, "An efficient wavelet-based HRTF model for auralization," *Acustica/Acta Acustica*, vol. 90, no. 1, Jan 2004.

[8] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1648–1661, Mar. 1992.

[9] F. L. Wightman and D. J. Kistler, "Monaural sound localization revisited," *J. Acoust. Soc. Am.*, vol. 101, no. 2, pp. 1050–1063, Feb. 1997.

[10] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[11] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambrigde-Press, Cambrigde, 1997.

[12] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice-Hall, Englewood Cliffs, New Jersey, 1995.

[13] J. C. B. Torres, M. R. Petraglia, and R. A. Tenenbaum, "HRTF modeling using wavelet decomposition," *XIV Congresso Brasileiro de Automática*, pp. 2208–2213, Sep 2002.

[14] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inform. Theory*, vol. 36, pp. 961–1005, Sept. 1990.

[15] J. C. B. Torres, M. R. Petraglia, and R. A. Tenenbaum, "Low-order modelling of head-related transfer functions using wavelet transform," *ISCAS 2004*, 2004.

[16] J. C. B. Torres, M. R. Petraglia, and R. A. Tenenbaum, "Low-order modeling and grouping of hrtfs for auralization using wavelet transforms," *ICASSP 2004*, 2004.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

AVALIAÇÃO OBJETIVA DE PARÂMETROS SONOROS EM SALAS: DIAGNÓSTICO DE QUALIDADE ACÚSTICA EM IGREJA LUTERANA - SP

Bianca Carla Dantas de Araújo (1), Maria Luiza Belderrain (2), Thaís Helena Luz Palazzo (3), Sylvio Reynaldo Bistafa (4)
FAU-USP, Pós-graduação em Tecnologia da Arquitetura
01240-001, São Paulo, SP, Brasil
(1) dantasbianca@gmail.com
(2) mlacustica@hotmail.com
(3) thaispalazzo@yahoo.com.br
(4) sbistafa@usp.br

RESUMO

A qualidade acústica das salas tem por objetivo otimizar a geração e recepção de informações, visando o uso a que são destinadas. Os requisitos para se alcançar uma boa qualidade sonora estão diretamente relacionados a geometria do local e suas dimensões, características das superfícies internas e materiais de acabamento, entre outras. O presente trabalho pretende avaliar a qualidade acústica de uma igreja, a partir dos parâmetros objetivos e subjetivos de análise, obtidos a partir do software de medições AURORA e da simulação computacional no software CATT-ACOUSTICS. Os resultados indicam baixa inteligibilidade da fala, mostrando que o espaço construído não corresponde ao propósito para o qual foi idealizado.

INTRODUÇÃO

Os esforços técnicos para reduzir o nível de ruído num dado local procedente de um recinto contíguo exterior, ou deste local para os recintos adjacentes, constituem o que se convencionou chamar de "acústica destrutiva". Já a "acústica construtiva" seria aquela com os esforços dirigidos a aperfeiçoar os níveis sonoros que se deseja conceber num local com um mínimo de interferência (SANCHO, 1982).

Referindo-se a esta "acústica construtiva", o aperfeiçoamento acústico define as condições sonoras internas nos recintos, que se baseiam no objetivo fundamental de se conseguir otimizar a geração e recepção de informações, ou seja a comunicação. Os recintos referidos são aqueles em que o comportamento do som é definido pelo uso destinado ao espaço, e são comumente denominados salas.

Os requisitos exigidos a um recinto para se conseguir uma qualidade acústica satisfatória variam segundo o uso a que é estabelecido. Alguns destes requisitos estão diretamente relacionados com a geometria do local, outros com suas dimensões, características das superfícies interiores,

e até com a implantação do recinto dentro do edifício e deste em relação à outra área exterior.

Cada sala exige critérios e condições particulares tanto para a comunicação como para o conforto acústico (SANCHO, 1982). Os critérios gerais de definição de acústica de salas estabelecem a qualidade sonora das mesmas, como o tempo de reverberação, por exemplo, porém são especificados em relação ao seu uso. Podem ser critérios objetivos e subjetivos, estando sempre relacionados entre eles e o uso a que se referem, conforme mencionado.

O tempo de reverberação era o único parâmetro acústico que relacionava o fenômeno físico com as impressões produzidas nas pessoas. Hoje, parâmetros diferentes podem relacionar o comportamento físico da sala com diferentes tipos de sensações auditivas. Essas sensações podem ser descritas como, por exemplo: intensidade, impressão espacial, clareza, brilho, presença, dentre outros (GERGES, 2000).

A garantia de níveis de ruído compatível com as atividades humanas tem sido a principal componente do conforto acústico em ambientes. No entanto, a acústica arquitetônica vem se desenvolvendo no sentido de propiciar algo mais aos usuários de ambientes diversos – a qualidade sonora.

> "Entende-se por qualidade sonora, um conjunto de atributos acústicos subjetivos que venham de encontro às expectativas da experiência acústica do ouvinte. Conscientemente ou não, a expectativa do usuário de uma sala de conferências, é que esta propicie condições acústicas para uma adequada inteligibilidade da fala. Isto irá requerer baixos níveis de ruído com certeza, porém algo mais é necessário para a adequada comunicação oral neste ambiente." (BISTAFA, 2005, p. 3)

Para cada finalidade da sala, há atributos acústicos subjetivos que devem ser atendidos. Diferentemente da sala onde o uso é a palavra falada, ou seja, uma sala de conferência, onde a reverberação deve ser reduzida, numa sala destinada à música, certa reverberação é necessária, no sentido de garantir a experiência acústica que o ouvinte espera ao escutar música (BISTAFA, 2005).

Os atributos não se encontram ainda totalmente definidos para a maioria das salas de audição crítica, sendo muitos dos existentes, alvo de considerável debate e controvérsia, e por este motivo objeto de pesquisa e desenvolvimento. Os atributos de uma sala de conferências são diferentes daqueles de uma sala destinada à música; envolvem muitas vezes várias dimensões subjetivas. Na sala destinada à música, um atributo subjetivo relevante é sentir-se "envolvido" pela música – uma outra dimensão subjetiva (BISTAFA, 2005).

Para tanto, é necessário dispor-se de um índice que quantifique objetivamente esta impressão subjetiva. Neste sentido, existem alguns índices mensuráveis que se correlacionam com algumas das dimensões subjetivas, que são os parâmetros objetivos, ainda, também, sujeitos a discussões e pesquisas.

De forma a contribuir com o contexto apresentado, o presente trabalho busca avaliar, por métodos de medições e simulações, a qualidade acústica de uma sala com audição crítica, no caso uma igreja, a partir da interpretação e registro de parâmetros sonoros subjetivos e objetivos, com vistas a adequação do espaço ao uso concebido; além de permitir uma comparação dos métodos propostos para análise.

PROCEDIMENTOS METODOLÓGICOS

Características gerais da edificação

A sala selecionada é uma Igreja Luterana – Igreja da Paz, localizada na Rua Verbo Divino, 392, Granja Julieta, São Paulo/SP. O uso predominante é para a palavra falada (cultos) e, eventualmente, música (apresentações de corais e orquestra de câmara); possui uma área em planta de 250 m^2 e um pé-direito médio de 9,0 m perfazendo um volume aproximado de 2.250 m^3. A forma hexagonal da planta da edificação possui como programa de necessidades um altar, platéia e balcão. Os acessos são: entrada principal pela parede da frente; acesso alternativo pela parede lateral esquerda; acesso ao balcão por escada estruturada em parte da parede lateral esquerda.

As superfícies são constituídas por piso altar em mármore; piso platéia em granito; escada em mármore; piso balcão em madeira (taco); paredes em alvenaria rebocada e pintada; janelas em vitrais; portas e bancos em madeira; teto abobadado em laje maciça pintada. (Figuras 1 e 2).

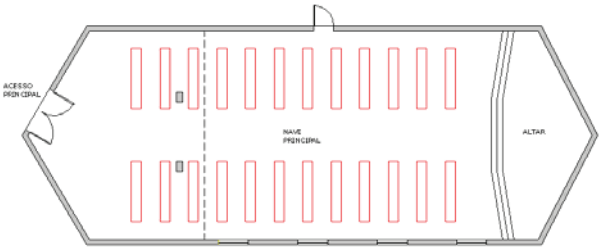


Figura 1 – Planta Baixa da Igreja analisada

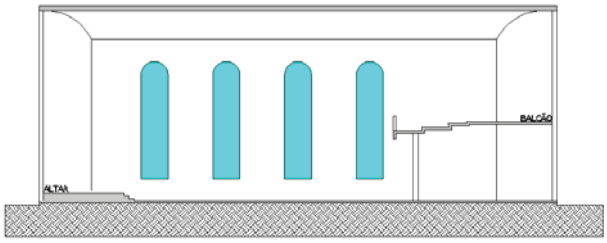


Figura 2 – Corte esquemático da igreja analisada

Medições dos parâmetros acústicos

O princípio das medições é identificar os parâmetros objetivos de qualidade acústica da sala real, a partir da Resposta Impulsiva (RI). As medições foram viabilizadas com o uso do software *Aurora,* desenvolvido pelo prof. Angelo Farina (Itália). A obtenção da Resposta Impulsiva (RI) foi realizada a partir de três sinais: Balão estourando; Multi MLS Signal; Sine Sweep (estes dois últimos emitidos pelo próprio programa de medição). A fonte sonora foi posicionada no centro do altar e a captação dos sinais foi feita em três locais da Igreja: na frente da audiência (P1), no fundo da audiência (P2) e no balcão (P3), conforme Figura 3.

Os sinais foram emitidos e captados com tréplica, ou seja, em cada ponto três vezes, e a partir daí retirada a média aritmética dos valores dos parâmetros objetivos da resposta impulsiva encontrada. Foi um total de 27 medições (9 para cada ponto).

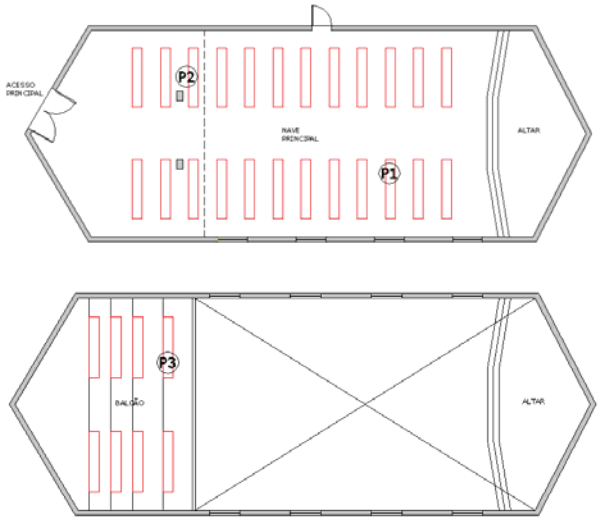


Figura 3 – Planta baixa da igreja analisada com destaque para localização dos pontos de medição

Os equipamentos e materiais utilizados nas medições foram:

Computador portátil (Sager 2850);
Microfone sem fio (Gemini UHF 1610);
Caixa de som (SP 5000);
Potência (Crown 460 CSL);
Pré-amplificador (Gemini PH 700);
Cabos de conexão;
Tripé RMW para caixa PA;
Softwares: Adobe Audition; Aurora; Excel;
Balões de festa (bexigas).

As medições refletem a condição de "sala vazia" ou sem público. Os dados obtidos com o sinal MLS (maximum length sequence) apresentaram distorções, em função da igreja em questão ser muito reverberante (devido às suas superfícies lisas e refletoras), o que foi agravado pela condição de ambiente vazio.

Com relação às medições executadas com estouro de balão, observou-se muita discrepância entre algumas frequências, em certos parâmetros. Por esse motivo, na análise dos resultados, optou-se por desprezar tanto as avaliações realizadas com o sinal MLS, como com o estouro de balão. Esse procedimento procurou aumentar a confiabilidade nos valores dos parâmetros em geral.

Simulações sonoras

As simulações do desempenho acústico da Igreja analisada foram desenvolvidas no *software Catt-Acoustics*.

Foi necessário adequar o modelo geométrico 3D (sistema *Autocad),* de modo a definir todas as superfícies como planos formados por pontos no sistema ortogonal. O trabalho gráfico exigiu que os planos ficassem totalmente fechados, tornando o modelo da igreja estanque ou sem vazamentos.

Após essa etapa foi preciso fornecer ao software informações a respeito dos materiais de acabamento das superfícies (descritos anteriormente), através de coeficientes de absorção sonora e coeficientes de difusão sonora, nas frequências de 125 Hz a 4 kHz, disponíveis na literatura. A variação desses coeficientes tem o intuito de "calibrar" o modelo, de modo a se obter resultados mais próximos da realidade.

O arquivo *master.geo* sintetiza todos esses dados, enquanto os arquivos *source* e *receiver* referem-se ao posicionamento da fonte sonora (centro do altar) e dos receptores (pontos P1, P2 e P3).

ANÁLISE DA QUALIDADE ACÚSTICA

Escolha dos parâmetros

O software *Aurora* fornece inúmeros parâmetros acústicos que qualificam uma sala, tais como: tempos de reverberação (T20, T30, Tuser), "early decay time" (EDT), tempo central (Ts), definição (D50), clareza (C80), força ("strength"), etc.

A fim de comparar os mesmos parâmetros que também o software de simulação fornece, são apresentados cinco deles: T30 (s), EDT (s), C80 (dB), D50 (%) e Ts (s). A seguir são apresentadas as definições dos parâmetros selecionados, conforme Barron (2000). Tem-se:

- T30 (s) – tempo de reverberação: tempo que a energia acústica dentro de um recinto leva para decair 30 dB (usualmente de – 5 dB a – 35 dB), depois que a fonte sonora é cessada. O parâmetro mais conhecido é o T60, ou tempo de decaimento para a energia sonora diminuir 60 dB, o qual foi desenvolvido por Sabine (1922), através da relação inversamente proporcional entre o volume da sala (m³) e a quantidade de absorção total da sala (m² sabine). Os valores de T60 para salas destinadas à fala variam entre 0,8 e 1,2 s.
- EDT (s) – "early decay time" ou tempo do decaimento inicial é uma medida da taxa de decaimento sonoro, baseada na primeira porção de 10 dB do decaimento. Em espaços altamente difusos, onde o decaimento é linear, as duas quantidades: EDT e T60 serão idênticas. O parâmetro EDT mostrou ser mais bem relacionado à sensação subjetiva de reverberação, do que o próprio tempo de reverberação (SCHROEDER, 1965).
- C80 (dB) ou clareza objetiva está relacionada ao equilíbrio entre a clareza percebida e a reverberância, o que é particularmente delicado no caso de audição musical. Pode ser expressa por (Equação 1):

$$C80\ (dB) = 10 \log \frac{[\text{energia chega} < 80\ \text{ms do som direto}]}{[\text{energia chega depois de 80 ms do som direto}]} \quad (1)$$

Este parâmetro tem equivalência direta com a fala. Os valores da clareza devem estar compreendidos entre $-3 < C80 < 0$; quanto mais próximo a zero, melhor.

- D50 (%) ou definição está diretamente relacionada ao entendimento da fala. Corresponde à razão direta entre a energia que chega aos primeiros 50 ms e a energia total. Assim, D50 é sempre um número entre 0,0 e 1,0. D50 > 70% representa uma inteligibilidade de 95% da fala.

- Ts (s) ou tempo central representa o centro de gravidade da área da resposta impulsiva integrada [equivalente a um triângulo, no gráfico: nível de pressão sonora (dB) x tempo (ms)]. O tempo central indicado para a fala corresponde a 70 ms.

Valores obtidos com os softwares *Aurora* e *Catt-Acoustics.*

As médias obtidas em cada ponto, para cada parâmetro, relativas aos resultados do *Aurora* e do *Catt-Acoustics*, comparados aos valores ideais ao local, lembrando que seu uso principal é para a fala, estão registradas nas tabelas 1 e 2. Os resultados obtidos são bem distintos para cada ponto, devido à sua localização, principalmente em relação à fonte sonora.

Tabela 1 – Valores obtidos no AURORA x critérios de qualidade

Param	V. Ideal	**P1**	Comp	**P2**	Comp	**P3**	Comp
T30 (s)	1,0 s	2,87	>>	2,85	>>	2,66	>>
EDT (s)	1,0 s	3,42	>>	3,26	>>	3,09	>>
C80 (dB)	-3 a 0 dB	- 4,2	<	- 7,0	<<	- 5,2	<
D50 (%)	70%	17,7	<<	8,2	<<	7,8	<<
Ts (s)	70 ms	246,5	>>	251,1	>>	260,6	>>

A tabela 1 mostra que todos os parâmetros analisados: T30, EDT, C80, D50 e Ts estão desfavoráveis, ou seja, a Igreja em questão é muito reverberante, o que implica na baixa inteligibilidade da fala e falta de clareza. Entre os pontos analisados, o ponto P1 – localizado na parte frontal da igreja – apresenta condições acústicas um pouco melhores do que os pontos P2 e P3, em função da proximidade em relação à referida fonte.

Tabela 2 – Valores obtidos no CATT-ACOUSTICS x critérios de qualidade

Param.	V. Ideal	**P1**	Comp	**P2**	Comp	**P3**	Comp
T30 (s)	1,0 s	2,78	>>	3,08	>>	3,08	>>
EDT (s)	1,0 s	3,07	>>	2,97	>>	2,92	>>
C80 (dB)	-3 a 0 dB	- 0,8	ok	- 1,4	ok	- 2,1	ok
D50 (%)	70%	34,6	<	30,5	<	26,0	<
Ts (s)	70 ms	175,1	>>	193,2	>>	197,2	>>

A tabela 2 mostra que os parâmetros analisados: T30, EDT, D50 e Ts estão desfavoráveis, definindo falta de clareza e entendimento da palavra falada. Entretanto, os valores de C80 (dB) clareza - estão dentro da faixa ideal, Isso mostra que a relação entre a energia sonora inicial (até 80 ms) e a energia sonora tardia (após 80 ms) é boa.

Apesar disto, não define que a sala esteja adequada, pois quanto mais próximo a zero o valor melhor; além disso, foi o único parâmetro cujo valor está dentro do considerado ideal pela literatura, não sendo suficiente para caracterizar a sala.

De uma forma geral, as ordens de grandezas dos valores encontrados foram coerentes nos dois métodos utilizados, no entanto, pode-se perceber que há um distanciamento bastante evidente dos valores ideais quando se considera a escala de variação, principalmente dos parâmetros D50 e Ts. Observa-se que estes parâmetros, obtidos no software *Catt-Acoustics*, apesar de fora dos valores recomendados, são melhores do que os obtidos com o software de medição *Aurora*, porém ainda muito longe dos valores ideais para o uso da fala.

A fim de permitir a comparação direta entre os dois métodos, os gráficos de cada parâmetro são apresentados com os valores médios dos seguintes parâmetros analisados: T30, EDT, C80, D50 e Ts; com os resultados do *Aurora* e do *Catt Acoustics* para os pontos P1, P2 e P3 (Figuras 4 a 8).

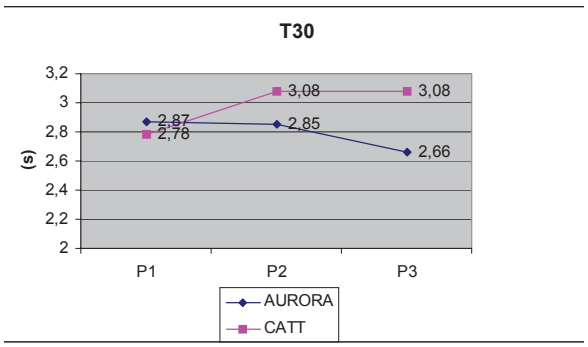


Figura 4 - Gráfico dos valores de T30

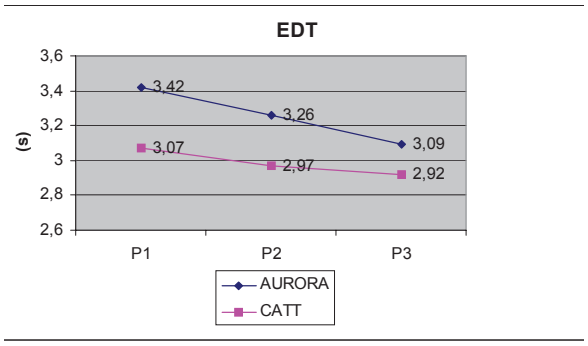


Figura 5 – Gráfico dos valores de EDT

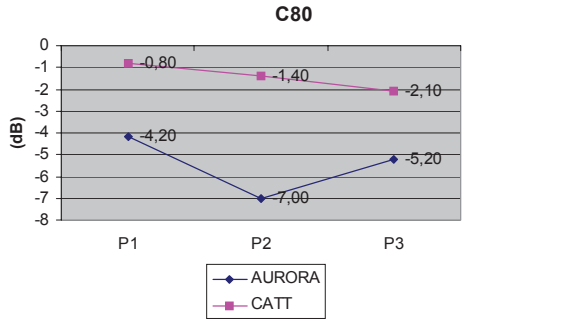


Figura 6 – Gráfico dos valores de C80

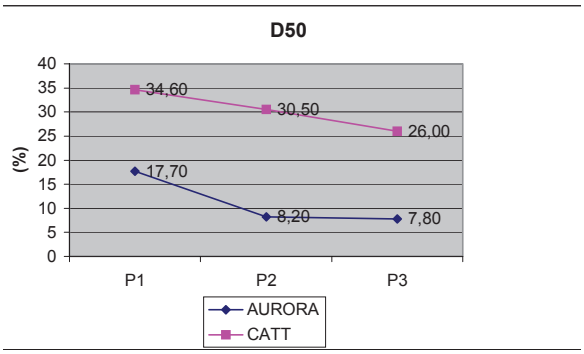


Figura 7 – Gráfico dos valores de D50

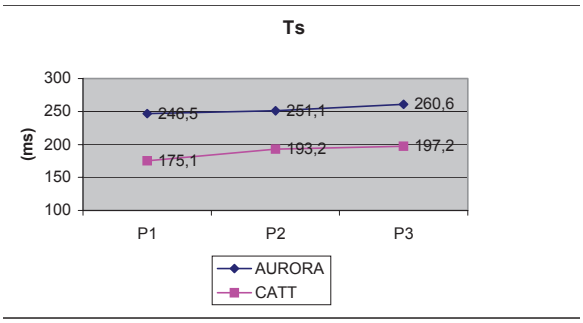


Figura 8 – Gráfico dos valores de Ts

Os gráficos mostram que as curvas em geral têm a mesma tendência, com exceção do parâmetro T30 no ponto P1, que apresentam valores próximos medidos e simulados. Em dois dos parâmetros – C80 e D50 – os valores obtidos no *Catt Acoustics* são maiores do que os obtidos no *Aurora*, o que representa resultados menos distantes dos valores ideais, porém ainda não satisfatórios.

O mesmo acontece com o T30, que apresenta valores simulados maiores do que medidos, com exceção feita ao ponto P1 que está mais próximo à fonte sonora; fato que pode ter interferido no resultado, já que este parâmetro está diretamente relacionado ao tempo de reverberação, ou seja, tempo que a energia acústica dentro de um recinto leva para decair 30 dB.

Pode-se observar nos gráficos que há uma tendência de oposição dos resultados dos pontos P1 e P3. Quando o primeiro apresenta resultados maiores, o terceiro apresenta resultados menores em relação a este, e vice versa. Este fato confirma o posicionamento mais desfavorável em relação à fonte sonora, que prejudica a comunicação, no caso da fala.

Outra constatação é a de que as curvas obtidas para os 3 pontos: P1, P2 e P3, na simulação acústica, são muito próximas entre si, com uma tendência linear, como pode-se observar nos gráficos apresentados, com exceção do parâmetro T30. No caso do parâmetro EDT, as curvas são quase coincidentes – formando uma reta. Esses resultados diferem daqueles obtidos na medição, a qual não apresentou similaridade entre as curvas para os diversos pontos.

CONCLUSÕES

A análise da qualidade acústica da Igreja da Paz, feita através de medições acústicas, com o uso do software *Aurora* e também da simulação computacional, com o uso do software *Catt-Acoustics*, apresentaram conclusões esperadas, quando confirmaram tanto a percepção subjetiva tida "in loco" pelos autores, quanto à opinião emitida pelo pastor da referida Igreja, de que a mesma não é apropriada à fala (pregação), por ser muito reverberante, mesmo com público.

Apesar de ter sido realizado o estudo da sala vazia, os valores identificados do tempo de reverberação estão muito superiores ao ideal para fala, constatando-se que mesmo a audiência de pessoas não é capaz de absorver o som a ponto de baixar um mínimo de aproximadamente 1,66 s, considerando o valor menor de T30 (2,66 s) encontrado independente do método.

Outro resultado constatado fora a falta de correlação entre os resultados obtidos com os métodos de medição e simulação. Acredita-se que a diferença confirmada nos resultados do *Aurora* e do *Catt Acoustics* deve-se às seguintes questões:

- Imprecisão na definição dos coeficientes de difusão sonora e, em menor escala, dos coeficientes de absorção sonora das superfícies da sala, na simulação;
- Necessidade de simplificação do modelo geométrico 3D da sala, para a simulação computacional, distanciando-o do modelo real;
- Realização das medições e simulação com a sala vazia, o que realça a condição reverberante do espaço (pode-se supor que na presença de audiência, parcial ou completa, devido à absorção oferecida pelo público, a qualidade acústica da igreja seja um pouco melhorada).

Este trabalho ressalva a necessidade de mais estudos neste contexto, a fim de subsidiar a "apuração", ou seja, a melhoria dos métodos utilizados para avaliar salas com audição crítica, além de revisão e adaptação das normas existentes, e criação de outras mais específicas.

Em função do distanciamento dos valores obtidos em relação aos valores ideais, para os cinco parâmetros pesquisados, nos dois métodos analisados, indica-se a necessidade de correção acústica à sala considerada, Igreja Luterana da Paz.

REFERÊNCIAS BIBLIOGRÁFICAS

BARRON, M. (2000). **Auditorium Acoustics and Architectural Design**. E&FN SPON. 2000.

BISTAFA, S. R. (2005). **Acústica Arquitetônica: Qualidade Sonora em Salas de Audição Crítica. Descrição detalhada.** Acesso em out. 2005. Disponível em: www.poli.usp.br/p/sylvio.bistafa/ACUSARQ

GERGES, S.H.Y. (1992). **Ruído: Fundamentos e Controle**. Departamento de Engenharia Mecânica da Universidade Federal de Santa Catarina. 1ª Edição, Florianópolis.

SANCHO, V.M., SENCHERMES A.G. (1982). **Curso de Acustica en Arquitectura**. Colegio Oficial de Arquitectos de Madrid, Madrid, 1982.

Sociedade de Engenharia de Áudio
Artigo de Congresso
Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Avaliação de Dois Novos Métodos para Geração de Som 3D

Fábio P. Freeland[1], Luiz W. P. Biscainho[1], Paulo S. R. Diniz[1]
[1]LPS – PEE/COPPE & DEL/Poli, UFRJ
Caixa Postal 68504, Rio de Janeiro, RJ, 21941-972, Brasil
[freeland,wagner,diniz]@lps.ufrj.br

RESUMO

Este trabalho trata da geração de som tridimensional reproduzido através de fones de ouvido. Nesse contexto, os autores desenvolveram recentemente duas novas técnicas para interpolação de HRFTs (*Head-Related Transfer Functions*) medidas para um conjunto finito de pontos ao redor de um ouvinte. Essas funções modelam o caminho do som da fonte sonora virtual às orelhas e, interpoladas, geram a ilusão do som em movimento. Neste artigo, realizam-se avaliações subjetivas daquelas técnicas, comparando-as ao método bilinear triangular.

INTRODUÇÃO

A geração de som tridimensional com fones de ouvido tem sido bastante investigada nos últimos anos [1, 2, 3, 4, 5]. Uma das técnicas empregadas para se criar esse efeito é a que utiliza as chamadas Funções de Transferência Relativas à Cabeça (HRTFs—*Head-Related Transfer Functions*). Essas funções modelam o caminho entre a posição da fonte virtual e as orelhas e, como são medidas para um conjunto finito de posições ao redor do ouvinte, devem ser interpoladas para se poder posicionar a fonte em qualquer outra posição. Essa interpolação normalmente é feita sobre as respostas ao impulso correspondentes a cada HRTF, chamadas de Respostas ao Impulso Relativas à Cabeça (HRIRs—*Head-Related Impulse Responses*) [6].

Recentemente, os autores do presente artigo desenvolveram duas técnicas de interpolação: uma baseada em uma função auxiliar chamada de Função de Transferência Interpositional (IPTF—*Interpositional Transfer Function*) que reduz a complexidade computacional do procedimento de interpolação [7]; e outra que interpola incrementalmente os coeficientes da transformada Karhunen-Loève (KLT—Karhunen-Loève *Transform*) relativos às HRIRs [8]. Naqueles trabalhos, foram realizadas comparações através de medidas objetivas que indicaram que o desempenho dos métodos propostos equivalem ao de um método clássico de interpolação chamado de bilinear [3, 9, 10].

No presente artigo, realiza-se a avaliação subjetiva desses dois métodos e compara-se o resultado ao atingido com o método bilinear. Na próxima seção, faz-se uma breve explanação sobre os métodos propostos em [7, 8]. Na seção seguinte, são mostradas as configurações dos testes subjetivos e os resultados obtidos. Por fim, apresentam-se as conclusões do trabalho.

MÉTODOS DE INTERPOLAÇÃO

Nesse trabalho, são comparados três métodos de interpolação: o método bilinear (chamado aqui de clássico), o método com IPTFs e o método incremental sobre os coeficientes da KLT (KLT incremental). Esses três métodos consideram que são conhecidas as HRIRs de determinadas posições sobre uma casca esférica ao redor do ouvinte, e obtêm a função interpolada como uma combinação linear de três HRIRs relativas aos pontos que formam uma região triangular que contém a posição desejada.

A diferença básica entre esses métodos está no tipo de função ao qual são aplicados os ponderadores calculados. Para uma dada posição, o valor dos ponderadores nos três casos são os mesmos, calculados através das distâncias angulares entre as posições que formam a região triangular onde se encontra a posição desejada, como no método clássico [10].

No caso do método KLT incremental, esses ponderadores são utilizados somente para se interpolar a HRIR da posição desejada na primeira vez que se entra em uma determinada região triangular. A partir dessa primeira interpolação, se não houver mudança de região, a interpolação incremental apenas corrige o valor da função de acordo com a diferença entre as posições anterior e atual [8].

Método Clássico

Na Fig. 1, pode-se ver um setor de uma esfera, sobre a qual foram medidas as HRIRs dos pontos A, B, C e D. Nesse caso, os

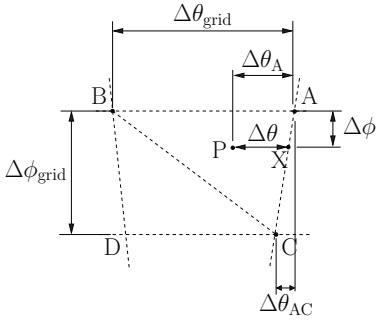


Figura 1: Detalhe das regiões triangulares sobre a esfera de referência.

ponderadores utilizados pelos métodos de interpolação mencionados para um determinado ponto P podem ser obtidos a partir das coordenadas de elevação ϕ e de azimute θ de acordo com

$$w_{\text{C}} = \frac{\Delta\phi}{\Delta\phi_{\text{grid}}}, \qquad w_{\text{B}} = \frac{\Delta\theta}{\Delta\theta_{\text{grid}}}, \tag{1}$$

$$w_{\text{A}} + w_{\text{B}} + w_{\text{C}} = 1, \tag{2}$$

sendo as distâncias angulares definidas como

$$\Delta\phi = \phi_{\text{P}} - \phi_{\text{A}}, \qquad \Delta\theta = \theta_{\text{P}} - \theta_{\text{X}}, \tag{3}$$

$$\Delta\theta_{\text{grid}} = \theta_{\text{B}} - \theta_{\text{A}} \quad \text{e} \quad \Delta\phi_{\text{grid}} = \phi_{\text{C}} - \phi_{\text{A}}. \tag{4}$$

Como pode ser visto na Figura 1, deve-se calcular a distância $\Delta\theta$ em função das coordenadas dos pontos envolvidos na interpolação. Assim, como

$$\frac{\Delta\phi}{\Delta\phi_{\text{grid}}} = \frac{\Delta\theta_{\text{A}} - \Delta\theta}{\Delta\theta_{\text{AC}}}, \tag{5}$$

pode-se obter

$$\Delta\theta = \Delta\theta_{\text{A}} - \frac{\Delta\phi}{\Delta\phi_{\text{grid}}}\Delta\theta_{\text{AC}}, \tag{6}$$

onde $\Delta\theta_{\text{A}} = \theta_{\text{P}} - \theta_{\text{A}}$ e $\Delta\theta_{\text{AC}} = \theta_{\text{C}} - \theta_{\text{A}}$.

Deve-se notar que $\Delta\theta$ é a distância do ponto P até o lado do triângulo que liga as duas elevações a partir do ponto A. Na prática, assume-se, sem perda de generalidade, que os pontos A e B têm a mesma elevação.

De uma forma ou de outra, os métodos de interpolação partem das HRIRs referentes a cada um dos pontos (A, B e C) e, com os ponderadores, geram a HRIR do ponto P. Tendo-se as HRIRs medidas ou aproximadas[1], o resultado final da interpolação é descrito por

$$\hat{h}_{\text{P}}(k) = w_{\text{A}} h_{\text{A}}(k) + w_{\text{B}} h_{\text{B}}(k) + w_{\text{C}} h_{\text{C}}(k), \tag{7}$$

onde $h_{(\cdot)}(k)$ é a HRIR do ponto $(\cdot)$ e $\hat{h}_{\text{P}}(k)$ é a HRIR do ponto P.

Deve-se notar que a interpolação é realizada sobre as funções de fase mínima [11]. Para se obter a aproximação final o atraso δ da HRIR desejada deve ser incluído na estrutura de interpolação. Para isso, calcula-se o excesso de fase de cada HRIR com relação à sua versão de fase mínima, que se aproxima muito de um atraso puro [9], e calcula-se δ através da ponderação dos atrasos estimados das três HRIRs dos pontos A, B e C.

A Fig. 2 mostra o diagrama em blocos do procedimento de interpolação descrito para um dos canais (esquerdo ou direito) do sistema binaural.

Método IPTF

O método IPTF [1, 7] se aproveita da redução de ordem conseguida para o modelo de IPTFs para diminuir a complexidade computacional da interpolação clássica. Esse método

[1] O método clássico utiliza as HRIRs medidas. Não é necessário estimá-las.

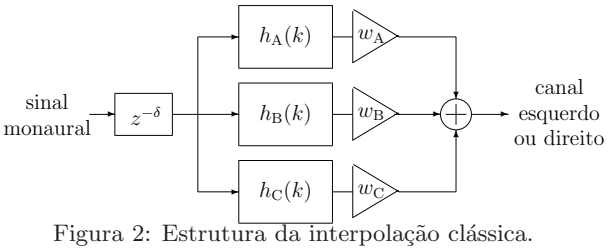


Figura 2: Estrutura da interpolação clássica.

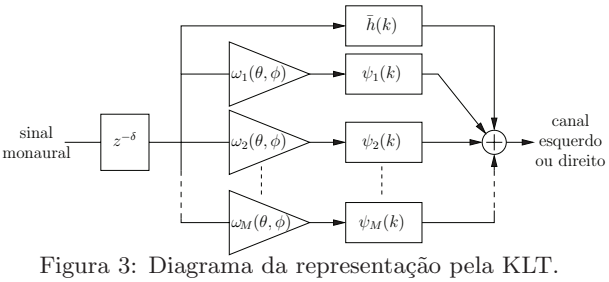


Figura 3: Diagrama da representação pela KLT.

realiza a interpolação através da Eq. (7) utilizando a HRIR medida relativa ao ponto mais próximo do ponto P e duas HRIRs aproximadas pela cascata desta HRIR medida e duas IPTFs (uma para cada aproximação).

A IPTF pode ser definida como

$$\text{IPTF}_{i,f} = \frac{\text{HRTF}_f}{\text{HRTF}_i}, \tag{8}$$

onde HRTF_i e HRTF_f são as HRTFs associadas com os pontos inicial e final, respectivamente.

Seguindo a nomeação de vértices explicada anteriormente, as HRTFs relativas ao ponto P, como na Eq. (7), podem ser descritas por

$$\text{HRTF}_{\text{P}} = \text{HRTF}_{\text{A}}(w_{\text{A}} + w_{\text{B}}\text{IPTF}_{\text{A,B}} + w_{\text{C}}\text{IPTF}_{\text{A,C}}), \tag{9}$$

onde os pesos w_{A}, w_{B} e w_{C} são calculados através das Eqs. (1) e (2). Nesse caso, o ponto mais próximo ao ponto P é o ponto A. Note que, para se obter redução da complexidade computacional, deve-se utilizar o modelo de ordem reduzida para as IPTFs obtidas pela Equação (8).

Método KLT Incremental

Os coeficientes da interpolação clássica podem ser utilizados também sobre os coeficientes de uma transformada cujas funções da base representem as HRIRs [4, 12]. Em [8], foi proposta uma forma incremental de se realizar a interpolação no domínio da transformada KLT.

Com as funções-base $\psi_j(k)$ da KLT do conjunto de HRIRs medidas, torna-se possível calcular a HRIR associada a cada ponto (θ, ϕ) sobre a esfera de referência fazendo-se

$$\hat{h}(\theta, \phi, k) = \overline{h}(k) + \sum_{j=1}^{N} \omega_j(\theta, \phi)\psi_j(k), \tag{10}$$

onde $\omega_j(\theta, \phi)$ são as funções de coeficientes a serem interpoladas, $\overline{h}(k)$ é a HRIR média do conjunto medido e N é o número de funções-base utilizadas na representação. A KLT consegue com um número reduzido de funções-base concentrar quase toda a energia do conjunto representado. Com isso, pode-se utilizar um numero $M < N$ de funções-base na representação. A Fig. 3 mostra o diagrama em blocos que aproxima uma das HRIRs (canal direito ou esquerdo) de um sistema binaural através da KLT. A grande vantagem dos métodos de interpolação no domínio da transformada está no fato de que ao acrescentar-se mais uma fonte sonora virtual, o número de multiplicações é acrescido apenas de M, já que são os coeficientes que contêm a informação de direção.

Partindo de um valor inicial, que pode ser interpolado fazendo-se

$$\omega_j(\theta, \phi) = w_{\text{A}}\omega_j(\theta_{\text{A}}, \phi_{\text{A}}) + w_{\text{B}}\omega_j(\theta_{\text{B}}, \phi_{\text{B}}) + w_{\text{C}}\omega_j(\theta_{\text{C}}, \phi_{\text{C}}), \tag{11}$$

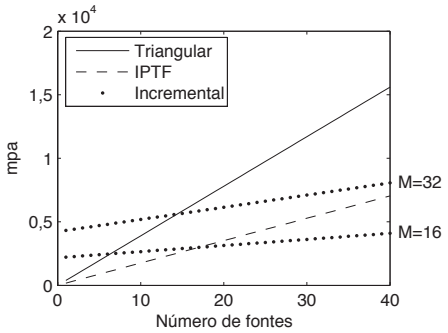


Figura 4: Comparação da complexidade computacional.

onde w_{A}, w_{B} e w_{C} são calculados pelas Eqs. (1) e (2) e as funções $\omega_j(\theta_{\mathrm{A}}, \phi_{\mathrm{A}})$, $\omega_j(\theta_{\mathrm{B}}, \phi_{\mathrm{B}})$ e $\omega_j(\theta_{\mathrm{C}}, \phi_{\mathrm{C}})$ são os coeficientes da KLT para as HRIRs das posições A, B e C, respectivamente, pode-se aproximar por

$$\hat{\omega}_j(\theta_l, \phi_l) = \hat{\omega}_j(\theta_{l-1}, \phi_{l-1}) + \Delta\omega_{j,l-1} \tag{12}$$

os coeficientes da l-ésima posição angular partindo da posição anterior $(l-1)$. O incremento $\Delta\omega_{j,l-1}$ aplicado aos pesos $\hat{\omega}_j$ da posição $l-1$ para a posição l pode ser calculado como

$$\begin{aligned}\Delta\omega_{j,l-1} &= (\theta_l - \theta_{l-1}) \left.\frac{\partial\omega_j(\theta,\phi)}{\partial\theta}\right|_{\substack{\theta=\theta_{l-1}\\ \phi=\phi_{l-1}}} + \\ &+ (\phi_l - \phi_{l-1}) \left.\frac{\partial\omega_j(\theta,\phi)}{\partial\phi}\right|_{\substack{\theta=\theta_{l-1}\\ \phi=\phi_{l-1}}}.\end{aligned} \tag{13}$$

Ao se trocar de região triangular, deve-se utilizar novamente a interpolação dada pela Eq. (11).

COMPLEXIDADE COMPUTACIONAL

A complexidade computacional de cada um dos métodos mencionados acima pode ser obtida em função do número de fontes F, fazendo-se

$$C_C = (3N+6)F \qquad C_{IPTF} = (2Q+N+6)F \tag{14}$$

$$C_{KLT} = 3MF + (M+1)N, \tag{15}$$

onde C_C, C_{IPTF} e C_{KLT} são os números de multiplicações necessárias aos métodos clássico, IPTF e KLT incremental. N e Q são os números de multiplicações associadas às HRIRs e IPTFs, respectivamente, e M é o número de funções-base da KLT utilizadas na representação das HRIRs.

Na Fig. 4, tem-se um gráfico do número de multiplicações em função do número de fontes simultâneas. Pode-se notar que com pouco mais de dez fontes o método KLT incremental já é mais eficiente que o triangular. Comparando o KLT com o IPTF, vê-se que isso ainda é verdade para $F > 17$, no caso de $M = 16$. Com isso, pode-se dizer que os mais eficientes, dendendo do número de fontes desejado, são os métodos KLT incremental e IPTF.

TESTES SUBJETIVOS

Os métodos de interpolação tratados neste artigo já foram confrontados de forma objetiva contra o método clássico em [7, 8], onde foram comparadas as respostas em freqüência interpoladas ao longo das posições. Para uma efetiva validação dessas técnicas, faz-se necessário algum tipo de avaliação subjetiva.

Na presente seção, esses métodos são comparados através de três testes subjetivos. Primeiramente, realiza-se a descrição dos testes aplicados, indicando-se o seu objetivo. É realizada, então, a análise dos resultados desses testes, obtendo-se dela algumas conclusões.

Descrição dos Testes

De maneira geral, os testes têm como princípio comparar direta ou indiretamente os resultados dos métodos de interpolação. Em cada teste, apresenta-se aos avaliadores o som pré-gravado, gerado segundo cada tipo de interpolação, a fim de que eles julguem o efeito percebido. A característica a ser julgada deve ser bem esclarecida aos avaliadores, e a forma de resposta deve ser a mais simples possível para que a resposta seja quase imediata.

Para que a influência de qualquer diferença seja facilmente percebida, o tipo de sinal a ser apresentado também é importante. O que se faz normalmente é utilizar algum tipo de ruído que excite todos os modos do sistema auditivo. Um tipo de ruído bastante utilizado é o chamado ruído rosa. Esse tipo de ruído tem espectro de potência com decaimento de 3 dB por oitava (10 dB por década) com a freqüência. Como a percepção de energia ao longo da freqüência é aproximadamente logarítmica, esse decaimento com a freqüência resulta em uma percepção mais uniforme da energia. Em todos os testes realizados utilizou-se ruído rosa obtido de [13].

Trinta e três pessoas com idades entre 20 e 40 anos e sem problemas auditivos diagnosticados foram submetidas aos mesmos testes. Nenhuma delas tinha conhecimento específico de som tridimensional, sendo a maioria leiga nesse assunto. Os testes foram realizados em grupos de 3 a 6 pessoas e o controle de apresentação de cada seqüência foi feito pelos autores do presente artigo, sendo possível a reapresentação de qualquer seqüência de acordo com a necessidade de algum usuário. A intensidade dos sinais foi regulada previamente, mas aos avaliadores era permitida a alteração do nível de volume. Utilizaram-se fones de ouvido fechados[2] com amplificação fornecida por equipamento dedicado de 8 canais[3], permitindo que se fizessem até 8 avaliações simultaneamente. Não foi realizada medição do ruído de fundo no interior da sala, mas com os fones de ouvido do tipo fechado utilizados, o efeito do já bem reduzido ruído ambiente pôde ser desconsiderado. Foi ainda sugerido que as pessoas fechassem os olhos a cada seqüência. Para que a avaliação levasse em conta apenas o efeito dos métodos de interpolação, escolheram-se apenas posições onde a interpolação é necessária.

Verificação de Mudança de Posição e/ou Timbre

Para avaliar se os métodos de interpolação são equivalentes, o primeiro teste aplicado foi o de simples comparação entre os sinais gerados em uma mesma posição. Nesse teste, cada comparação foi feita entre dois trechos de sinal de 1 segundo de duração, exibidos em seqüência, com um intervalo entre eles também de 1 segundo. As posições foram escolhidas de forma aleatória e independente, segundo uma distribuição uniforme nos intervalos $-180^o < \theta < 180^o$ e $-40^o < \phi < 90^o$.

Foram geradas 35 seqüências em posições distintas, das quais 20 contêm uma comparação entre a interpolação clássica realizada diretamente com as HRIRs e uma das outras desenvolvidas em [7, 8]: a com IPTFs de ordem reduzida ou a KLT incremental. As outras 15 seqüências são formadas por sinais idênticos gerados com a mesma forma de interpolação, sendo 10 com a interpolação clássica e as outras 5 divididas de maneira aleatória entre os outros métodos interpolação. As seqüências foram apresentadas em uma ordem aleatória.

O julgamento foi realizado pedindo-se que os avaliadores dessem uma nota de 1 a 4 que indicasse quão perceptível era a diferença entre os sinais da mesma seqüência quanto à mudança na posição e no timbre do ruído (distorção e perda de fidelidade). Da maior para a menor, os significados das notas eram "Diferença imperceptível", "Quase imperceptível", "Bem evidente" e "Muito acentuada", respectivamente.

Na Fig. 5, podem-se ver as notas médias atribuídas a cada um dos métodos e os limites de $\pm$ um desvio-padrão (linhas horizontais acima e abaixo da média). Da esquerda para a direita, vêem-se as médias para os métodos: clássico sobre as HRIRs (considerado o padrão), de IPTFs de ordem reduzida e KLT incremental. Pode-se notar que, apesar de haver um decrescimento da média, ela ainda está dentro da faixa do desvio da nota para o método clássico.

O método de Análise de Variância (ANOVA—*Analysis*

[2] HD265, marca registrada da Sennheiser.

[3] Powerplay Pro-8 HA8000, marca registrada da Behringer.

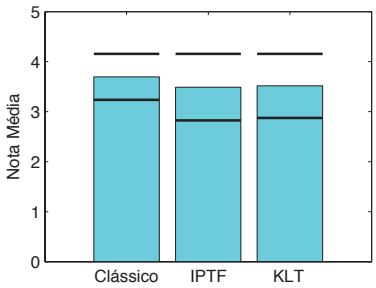


Figura 5: Médias das notas de diferença entre o método clássico e todos os outros.

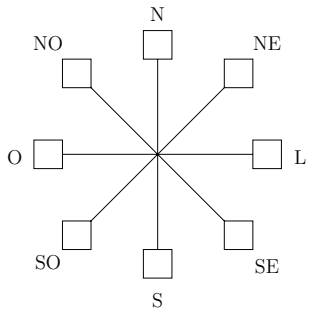


Figura 6: Sentidos testados na avaliação de percepção de movimento.

of Variance) [14, 15] indicou que as estimativas encontradas para as médias tinham significância estatística maior que $99,99\%$. Assim, as pequenas diferenças encontradas indicam grande similaridade entre os métodos na comparação direta. Pode-se, ainda, confirmar que os resultados são bastante próximos pelo fato de a mesma diferença percebida entre o método clássico (padrão) e os outros ter sido "percebida" entre o método clássico e ele mesmo (primeira barra na figura).

Com relação à dispersão das notas em torno da média, nota-se que houve um aumento aproximadamente igual para todos os métodos, comparados ao clássico. Isso indica uma certa diferença entre cada método testado e o clássico, mas insuficiente para alterar significativamente a média para os 33 avaliadores.

Verificação da Percepção do Movimento

O segundo teste aplicado procurou avaliar como é percebido o sentido do movimento. Para tanto, foram gerados 24 sinais, 8 para cada tipo de interpolação. Cada um desses 8 partia da posição $(0, 0)$ (frente do ouvinte) e seguia por um arco na superfície da esfera em direção a um dos oito pontos cardeais mostrados na Fig. 6 que estão posicionados 20° acima (N), abaixo (S), à esquerda (L) ou à direita (O), ou estão na direção diagonal, com 20° para cima e à direita (NE), para baixo e à direita (SE), para cima e à esquerda (NO) e para baixo e à esquerda (SO). Cada sinal tinha duração de 5 segundos, sendo que no primeiro e no último segundo a fonte virtual permanecia parada nas posições inicial e final, respectivamente. Aos avaliadores perguntou-se para qual das 8 posições a fonte havia se deslocado.

Na Fig. 7, podem-se ver os resultados das taxas de acerto para cada um dos métodos de interpolação. Nota-se que os métodos de interpolação IPTF e KLT incremental conseguem ser pouco melhores que o clássico, podendo, portanto, substituí-lo com alguma vantagem.

Esse teste indica que o método KLT incremental é preferível na substituição do clássico, já que, além de ser estruturalmente mais eficiente para o caso de múltiplas fontes, apresentou uma taxa de acertos mais elevada.

Nas Tabelas 1–3, pode-se observar o percentual das respostas dadas pelos avaliadores para cada sentido gerado. Como

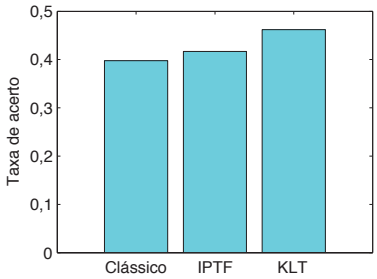


Figura 7: Taxas de acerto para o teste de sentido do movimento.

Tabela 1: Tabela de confusão. Percepção do movimento para o método clássico. Valores percentuais.

		Sentido Gerado							
		N	**NE**	**L**	**SE**	**S**	**SO**	**O**	**NO**
Percebido	**N**	**51,5**	0	0	0	57,6	0	0	6,0
	NE	0	**48,5**	39,4	42,4	0	0	0	0
	L	0	33,3	**42,4**	42,4	0	0	0	0
	SE	0	18,2	18,2	**15,2**	0	0	0	0
	S	45,5	0	0	0	**33,3**	0	0	0
	SO	3,0	0	0	0	3,0	**27,3**	30,3	15,2
	O	0	0	0	0	0	36,4	**36,4**	15,2
	NO	0	0	0	0	6,1	36,3	33,3	**63,6**

Tabela 2: Tabela de confusão. Percepção do movimento para o método IPTF. Valores Percentuais.

		Sentido Gerado							
		N	**NE**	**L**	**SE**	**S**	**SO**	**O**	**NO**
Percebido	**N**	**42,4**	0	0	0	36,4	3,0	0	3,0
	NE	6,1	**57,6**	21,2	33,3	6,0	0	0	0
	L	0	30,3	**60,6**	39,4	6,1	0	0	0
	SE	12,1	9,1	18,2	**27,3**	6,0	0	0	0
	S	36,4	3,0	0	0	**45,5**	0	0	0
	SO	0	0	0	0	0	**24,3**	30,3	24,3
	O	3,0	0	0	0	0	39,4	**36,4**	33,3
	NO	0	0	0	0	0	33,3	33,3	**39,4**

Tabela 3: Tabela de confusão. Percepção do movimento para o método KLT incremental. Valores percentuais.

		Sentido Gerado							
		N	**NE**	**L**	**SE**	**S**	**SO**	**O**	**NO**
Percebido	**N**	**54,5**	0	0	0	51,5	3,0	0	3,0
	NE	6,1	**48,5**	27,3	27,3	0	0	0	0
	L	3,0	33,3	**57,6**	27,3	6,0	0	0	0
	SE	6,1	18,2	12,1	**45,4**	6,1	0	0	0
	S	30,3	0	3,0	0	**36,4**	0	0	0
	SO	0	0	0	0	0	**30,3**	15,1	24,3
	O	0	0	0	0	0	39,4	**45,5**	21,2
	NO	0	0	0	0	0	27,3	39,4	**51,5**

mostrado na Fig. 7, nota-se que há uma pequena melhora nas taxas de acerto (diagonal nas tabelas) para os métodos KLT incremental e IPTF, em relação ao método clássico. Isso fica mais evidente para o método KLT incremental. Pode-se perceber, também, que as maiores confusões são entre os sentidos N e S, entre os sentidos NO, SO e O e entre os sentidos

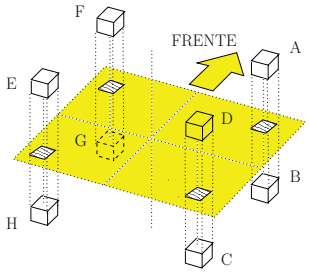


Figura 8: Posições testadas na avaliação de percepção da posição estática.

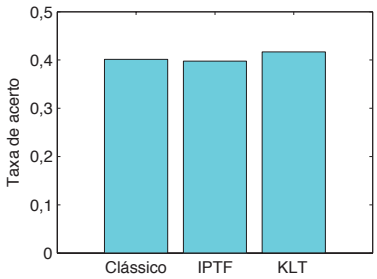


Figura 9: Taxas de acerto para o teste de posição estática.

NE, SE e L. De fato, a confusão entre cima e baixo é a mais evidente. A distinção lateral é feita em quase todos os casos. Deve-se chamar atenção também para o fato de as taxas de acerto serem todas em torno de 40%, o que é justificavel [16] pela simplicidade do modelo, que só leva em conta a posição angular da fonte.

Verificação da Percepção da Posição Estática

O terceiro e último teste de avaliação aplicado foi para avaliar a eficácia de cada método com relação à percepção da posição estática. Para esse teste, como no anterior, geraram-se 24 sinais, 8 para cada método de interpolação. Para cada método, posicionou-se a fonte virtual nas localizações referentes a cada um dos cubos mostrados na Fig. 8. Foi pedido às pessoas que respondessem em qual dos cubos a fonte sonora estava posicionada, considerando que a posição do ouvinte na figura era representada pelo cruzamento dos eixos, que ele estaria olhando na direção da seta e que o plano sombreado passava na altura das orelhas.

Na Fig. 9, vêem-se as taxas de acerto para cada um dos métodos de interpolação. Nota-se novamente que os métodos IPTF e KLT incremental conseguem praticamente a mesma taxa de acertos que o clássico, com ligeira vantagem para o KLT incremental.

Nas Tabelas 4–6 pode-se ver o percentual das respostas dadas pelos avaliadores para cada uma das posições geradas. Pode-se notar que para nenhum método houve confusão lateral (nenhum sinal pareceu estar vindo do lado oposto àquele em que foi gerado). O que realmente acontece é a confusão frente/trás e cima/baixo. Esse tipo de confusão é considerada normal, já que a fonte foi posicionada em pontos do mesmo cone de confusão. Apesar disso, o método incremental com a KLT obteve um número maior de acertos para a maioria das posições.

Geralmente, retira-se do cálculo de erros o efeito da confusão frente/trás, comum a todos os métodos, para se conseguir uma comparação mais clara entre os métodos quanto à identificação da posição [2]. Nesse caso, somando-se os valores percentuais de mesma elevação e azimutes de mesmo sinal (mesmo lado), o método IPTF obtém um número maior de acertos para a maioria das posições, ficando com uma média

Tabela 4: Tabela de confusão. Percepção da posição para o método clássico. Valores Percentuais.

		Posição Gerada θ, ϕ							
		A	B	C	D	E	F	G	H
Percebida θ, ϕ	A	**54,6**	27,2	6,1	30,3	0	0	0	0
	B	18,2	**15,2**	21,2	18,2	0	0	0	0
	C	3,0	36,4	**60,6**	12,1	0	0	0	0
	D	24,2	21,2	12,1	**39,4**	0	0	0	0
	E	0	0	0	0	**51,5**	36,4	27,2	33,3
	F	0	0	0	0	36,3	**42,4**	48,5	6,1
	G	0	0	0	0	6,1	9,1	**9,1**	12,1
	H	0	0	0	0	6,1	12,1	15,2	**48,5**

Tabela 5: Tabela de confusão. Percepção da posição para o método IPTF. Valores Percentuais.

		Posição Gerada θ, ϕ							
		A	B	C	D	E	F	G	H
Percebida θ, ϕ	A	**57,6**	12,2	6,1	57,6	0	0	0	0
	B	0	**24,2**	21,2	9,1	0	0	0	0
	C	0	24,2	**60,6**	9,1	0	0	0	0
	D	42,4	39,4	12,1	**24,2**	0	0	0	0
	E	0	0	0	0	**48,5**	57,6	27,2	27,3
	F	0	0	0	0	45,5	**27,3**	12,1	3,0
	G	0	0	0	0	6,0	9,1	**15,2**	9,1
	H	0	0	0	0	0	6,0	45,5	**60,6**

Tabela 6: Tabela de confusão. Percepção da posição para o método KLT incremental. Valores Percentuais.

		Posição Gerada θ, ϕ							
		A	B	C	D	E	F	G	H
Percebida θ, ϕ	A	**45,5**	21,2	15,2	39,4	0	0	0	0
	B	24,1	**27,3**	15,2	3,0	0	0	0	0
	C	15,2	24,2	**39,4**	9,1	0	0	0	0
	D	15,2	27,3	30,2	**48,5**	0	0	0	0
	E	0	0	0	0	**51,5**	30,3	30,3	18,2
	F	0	0	0	0	30,3	**42,5**	12,1	9,1
	G	0	0	0	0	6,1	24,2	**21,2**	15,1
	H	0	0	0	0	12,1	3,0	36,4	**57,6**

de acertos igual a 77,7%. O segundo melhor é o método KLT incremental (67,5%), praticamente junto com o método clássico (66,7%).

Dessa forma, conclui-se que os métodos testados podem ser considerados bons substitutos para o método clássico, com uma certa vantagem para o KLT incremental, que na comparação direta é o mais eficaz. Ao se desconsiderar a confusão frente/trás, o método IPTF também se mostra um bom substituto.

CONCLUSÕES

Neste trabalho, mostrou-se o conjunto de resultados de uma avaliação subjetiva realizada para dois métodos de interpolação de HRTFs recentemente desenvolvidos pelos autores do presente artigo. Esses resultados mostram a equivalência entre os métodos de interpolação propostos recentemente e o clássico.

Os resultados obtidos com os métodos KLT incremental e IPTF podem ser considerados um pouco melhores que o clássico.

Portanto, chega-se à conclusão de que os métodos IPTF e KLT incremental são fortes candidatos a substituir o método

clássico. O método KLT incremental é especialmente cotado quando se trata do caso com múltiplas fontes, onde sua baixa complexidade o torna bem mais vantajoso [8].

É importante notar que as taxas de acerto aparentemente baixas (em torno de 40%) devem-se ao fato de não ter sido realizado nenhum treinamento dos ouvintes antes dos testes (os avaliadores foram apresentados aos tipos de som no momento da avaliação). Além disso, o teste exigia muito da capacidade de abstração de cada um, já que o ambiente virtual não é completo, só tratando da localização da fonte. Espera-se que o modelamento de outros efeitos como as primeiras reflexões (*early reverberation*) e a compensação do movimento da cabeça [16] possam melhorar muito esses resultados [17].

REFERÊNCIAS BIBLIOGRÁFICAS

[1] F. P. Freeland, "Geração eficiente de som tridimensional," tese de doutorado, Universidade Federal do Rio de Janeiro, Programa de Engenharia Elétrica–COPPE, Dezembro 2005.

[2] D. R. Begault, *3D Sound for Virtual Reality and Multimedia*. Cambridge, MA, USA: Academic Press, 1994.

[3] L. Savioja, *Modeling Techniques for Virtual Acoustics*. Ph.D. thesis, Helsinki University of Technology, Departament of Computer Science and Engineering, Telecomunications Software and Multimedia Laboratory Espoo, Finland, December 1999.

[4] J.-M. Jot, S. Wardle, and V. Larcher, "Approaches to binaural synthesis," in *AES 105th Convention*, (California, USA), AES, September 1998. (preprint 4861).

[5] V. R. Algazi, R. O. Duda, and D. M. Thompson, "Motion-tracked binaural sound," *J. Audio Eng. Soc.*, vol. 52, pp. 1142–1156, November 2004.

[6] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," Technical Report 280, MIT Media Lab., Cambridge, MA, USA, May 1994.

[7] F. P. Freeland, L. W. P. Biscainho, and P. S. R. Diniz, "Interpositional transfer function for 3D-sound generation," *J. of the Audio Eng. Soc.*, vol. 52, pp. 915–930, September 2004.

[8] F. P. Freeland, L. W. P. Biscainho, and P. S. R. Diniz, "Interpolation of head-related transfer functions (HRTFs): A multi-source approach," in *Proceedings of the XII European Signal Processing Conference*, (Vienna, Austria), pp. 1761–1764, EURASIP, September 2004.

[9] J.-M. Jot, V. Larcher, and O. Warusfel, "Digital signal processing issues in the context of binaural and transaural stereophony," in *98th AES Convention*, (Paris, France), AES, February 1995. (preprint 3980).

[10] F. P. Freeland, L. W. P. Biscainho, and P. S. R. Diniz, "Interpolação bilinear generalizada de HRTFs para geração de som tridimensional," in *Anais da VIII Convenção Nacional da AES Brasil*, (São Paulo, SP, Brasil), AES, Junho 2004.

[11] A. Kulkarni, S. K. Isabelle, and H. S. Colburn, "On the minimum-phase approximation of head-related transfer functions," in *IEEE Workshop on Applications of the Signal Processing to Audio and Acoustics*, (New Paltz, New York), IEEE, October 1995.

[12] J. Chen, B. D. V. Veen, and K. E. Hecox, "A spatial feature extraction and regularization model for virtual auditory display," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 129–132, April 1993.

[13] S. Moshier. Internet, November 2003. http://www.moshier.net/pink.html.

[14] E. W. Weisstein, "Anova." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/ANOVA.html.

[15] MATLAB, "Statistics toolbox." Math Works Inc.

[16] D. R. Begault, "Perceptual efects of synthetic reverberation on three-dimensional audio systems," *J. Audio Eng. Soc.*, vol. 40, pp. 895–904, November 1992.

[17] C.-J. Tan and W.-S. Gan, "Direct concha exitation for the introduction of individualized hearing cues," *J. Audio Eng. Society*, vol. 48, pp. 642–653, July/August 2000.

Sessão 2

Processamento Digital de Áudio, Voz e Sistemas Eletrônicos de Áudio

(Digital Audio and Speech Processing, and Audio Electronic Systems)

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Comparison of Speech Enhancement / Recognition Methods Based on Ephraim and Malah Noise Suppression Rule and Noise Masking Threshold

Francisco J. Fraga[1], André Godoi Chiovato[2] e Lidiane K. S. Abranches[2]
[1]Laboratório de Sistemas Integráveis da Escola Politécnica da USP (LSI-EPUSP)
[2]Instituto Nacional de Telecomunicações - Inatel
São Paulo, SP, CEP 05508-900, Brasil
franciscojfraga@gmail.com, agodoi@radial.br, lidiane@inatel.br

ABSTRACT

The proposed speech enhancement system uses a noise-masking threshold in a frame-by-frame basis in order to perform some important modifications in the original Ephraim and Malah (EM) algorithm. These increased the amount of noise reduction and simultaneously provide a more efficient elimination of the musical noise phenomenon. Perceptual evaluation results have shown that the new algorithm outperforms the standard EM algorithm for all types of nearly stationary noise considered in the experiment, in a wide signal–to–noise ratio range of noisy signals from SpEAR database.

INTRODUCTION

The widespread use of mobile communications in a variety of real environments, including those with high ambient noise levels, highlighted the importance of having good single-channel speech enhancement algorithms.

In this class of algorithms there is no reference channel available for noise estimation, which is realized only during speech pauses. Usually, single-channel speech enhancement systems are based on short–time spectral attenuation, which is the working principle of the so called subtractive–type algorithms. These subtractive–type algorithms are often used because they are easy to implement and offer several possibilities of varying the subtraction parameters according to the intended application. However, the major drawback of these methods is the appearing of the "musical residual noise" in the enhanced speech, which presents a very unnatural disturbing quality.

The noise suppression rule proposed by Ephraim and Malah [1] made it possible to obtain a moderate noise reduction while avoiding completely the musical noise phenomenon. On the other hand, at low signal–to–noise ratios (SNR < 10 dB), the Ephraim and Malah noise Suppression Rule (EMSR) did not offer a strong attenuation of the unwanted noise.

Based on this reasons, we proposed a new speech enhancement scheme, which kernel is based on EMSR, but with some modifications added in order to deal with noisy speech presenting low signal–to–noise ratios. It was done by

introducing the concept of noise–masking threshold, which is a well–known property of the human auditory system [2]. The basic gain function proposed by Ephraim and Malah was modified by adapting its parameters based on the calculation of the noise-masking threshold. This allows us to find a good tradeoff between the amount of noise reduction and the speech distortion in a perceptual sense.

MASKING PROPERTIES IN SHORT-TIME SPECTRAL ATTENUATION ALGORITHMS

If we assume that $y(n)$, the discrete–time noisy input signal, is composed by a clean speech signal $s(n)$ and an uncorrelated additive noise signal $d(n)$, then we can represent it as:

$$y(n) = s(n) + d(n) \tag{1}$$

In the class of short–time spectral attenuation algorithms, also known as subtractive-type algorithms, the processing is done on a frame-by-frame basis in the frequency domain:

$$\left|\hat{S}(\omega)\right| = G(\omega)\cdot\left|Y(\omega)\right| \quad \text{with} \quad 0 \le G(\omega) \le 1 \tag{2}$$

The phase of the noisy speech is used in order to re–synthesize the enhanced speech signal. The best result achievable by any kind of *subtractive-type* algorithms is given by the combination of the clean speech spectral magnitude with the noisy spectral phase. Following Virag [2], this situation is called the *theoretical limit.* Berouti et al. [3] proposed a flexible form of subtractive-type algorithm. In their algorithm, the gain function used to estimate the magnitude of the short-time Fast Fourier Transform (FFT) of the clean speech signal is given by:

$$G(\omega) = \begin{cases} \left[1 - \alpha.\left(\dfrac{\left|\hat{D}(\omega)\right|}{\left|Y(\omega)\right|}\right)^{\gamma}\right]^{1/\gamma} & \text{, if } \left(\dfrac{\left|\hat{D}(\omega)\right|}{\left|Y(\omega)\right|}\right)^{\gamma} < \dfrac{1}{\alpha+\beta} \\ \left[\beta\cdot\left(\dfrac{\left|\hat{D}(\omega)\right|}{\left|Y(\omega)\right|}\right)^{\gamma}\right]^{1/\gamma} & \text{, otherwise} \end{cases} \tag{3}$$

where α is the O*versubtraction factor* ($\alpha > 1$), β is the *Spectral Flooring factor* ($0 \le \beta << 1$) and the *Exponent* γ determines the sharpness of the transition from $G(\omega) = 1$ to $G(\omega) = 0$. The choice of these three parameters allows flexibility, but at low SNRs, it is impossible to minimize speech distortion and residual noise, simultaneously.

The idea of exploiting the masking properties of human auditory system was taken from a successful speech enhancement system proposed by Nathalie Virag [2]. In her paper, she adapted the classical subtraction parameters in (3) using a perceptual model. This model, with some adaptations, presents some steps for the calculation of a noise-masking threshold:

The signal critical band analysis

The first step calculates the present energy in each critical band, assuming discrete non-overlapping critical bands.

$$B_i = \sum_{\omega=bl_i}^{bh_i} P(\omega) \tag{4}$$

where bl_i and bh_i are the lower and upper boundaries of the ith critical band and $P(\omega)$ is the power spectrum.

Spreading function

A spreading function S_i is then convolved with the critical band spectrum B_i, generating the critical-band spread spectrum:

$$C_i = S_i * B_i \tag{5}$$

where S_i is given by [4], in dB :

$$S_i = 15{,}81 + 7{,}5(i + 0{,}4) - 17{,}5\sqrt{1 + (i + 0{,}474)^2} \tag{6}$$

The noise–masking threshold calculation

The noise–masking threshold is obtained by subtraction of a relative threshold offset O_i depending on the noise-like or tone-like nature of the masker and the maskee signals.

$$T_i = 10^{\log_{10}(C_i) - (O_i/10)} \tag{7}$$

In Sinha and Tewfik's method [5], O_i is given by a simple estimation, based on the fact that often the speech signal has a tone–like nature in lower critical bands and a noise–like nature in higher bands, as shown in Fig. 1

.

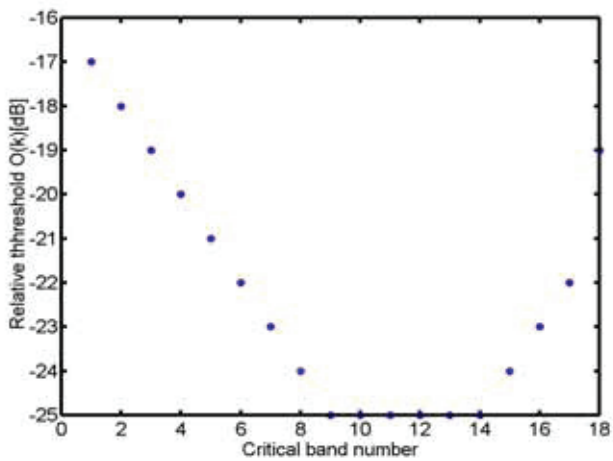


Fig. 1: Relative threshold offset (after Sinha and Tewfik[5])

Renormalization

The renormalization procedure is performed by a simple multiplication of each T_i by the inverse of the energy gain obtained by the convolution.

Accounting for absolute thresholds

In order to consider absolute thresholds, any critical band that has a calculated noise–masking threshold lower than the absolute threshold of hearing is replaced by the absolute threshold in that critical band.

In the method presented to noise masking threshold calculation described above, the noise-masking threshold must be calculated from the power spectrum of the clean speech. However, in practice only the original noisy signal is available. Then a rough estimate of the clean speech signal is computed using a simple power spectral subtraction scheme. Virag used the masking threshold to adjust the spectral subtraction parameters α and β of (3), for each frequency ω of a given speech frame q:

$$\begin{aligned} \alpha(q,\omega) &= F_\alpha[\alpha_{\min}, \alpha_{\max}, T(q,\omega)] \\ \beta(q,\omega) &= F_\beta[\beta_{\min}, \beta_{\max}, T(q,\omega)] \end{aligned} \tag{8}$$

where $\alpha_{\min}$, $\alpha_{\max}$ and $\beta_{\min}$, $\beta_{\max}$ are the minimal and maximal values of the oversubtraction and spectral flooring parameters, respectively, and $T(q,\omega)$ is the calculated noise–masking threshold for each frequency ω of the current speech frame q. The function F_α performs a linear interpolation according to the following boundaries:

$$\begin{aligned} F_\alpha &= \alpha_{\max} \quad \text{if} \quad T(q,\omega) = T(q,\omega)_{\min} \\ F_\alpha &= \alpha_{\min} \quad \text{if} \quad T(q,\omega) = T(q,\omega)_{\max} \end{aligned}$$

where $T(q,\omega)_{\min}$ and $T(q,\omega)_{\max}$ are the minimum and maximum values of $T(q,\omega)$, respectively.

The function F_β operates in a similar way. N.Virag [2] has chosen $\alpha_{\min} = 1$, $\beta_{\min} = 0$, $\alpha_{\max} = 6$, $\beta_{\max} = 0{,}02$ for an acceptable tradeoff between residual noise and speech distortion. The parameter γ was fixed to 2.

But we have found out that with this scheme it was not possible to eliminate completely the musical noise phenomenon. In our work, the information given by the noise masking threshold was used to adapt the Ephraim and Malah noise suppression rule, as explained in next section.

PROPOSED SPEECH ENHANCEMENT SYSTEM

The standard Ephraim and Malah Suppression Rule (EMSR) is a special type of *short–time spectral attenuation* algorithm where the spectral gain $G(q,\omega)$ applied to each short–time spectral component $|Y(q,\omega)|$ of the current speech frame is given by:

$$G = \frac{\sqrt{\pi}}{2}\sqrt{\left(\frac{1}{1+R_{post}}\right)\cdot\left(\frac{R_{prio}}{1+R_{prio}}\right)}.\,M[\theta] \tag{9a}$$

$$\begin{aligned} \theta &= \left(1+R_{post}\right)\cdot\left(\frac{R_{prio}}{1+R_{prio}}\right), \\ \mathrm{M}[\theta] &= \exp\left(-\frac{\theta}{2}\right)\cdot\left[(1+\theta)\cdot I_0\left(\frac{\theta}{2}\right)+\theta\cdot I_1\cdot\left(\frac{\theta}{2}\right)\right] \end{aligned} \tag{9b}$$

where I_0 and I_1 are the modified Bessel functions of zero and first order, respectively [1]. In (9a) and (9b), the frame index q and the frequency index ω have been omitted for compactness reasons. The spectral gain depends on two parameters:

$$R_{post}(q,\omega) = \begin{cases} \dfrac{|Y(q,\omega)|^2}{|\hat{D}(\omega)|^2} - 1 & , \text{ if } \dfrac{|Y(q,\omega)|^2}{|\hat{D}(\omega)|^2} > 1 \\ 0 & , \text{ otherwise} \end{cases} \tag{10}$$

$$\begin{aligned} R_{prio}(q,\omega) &= (1-\mu)\cdot R_{post}(q,\omega) + \\ &\mu\cdot G^2(q-1,\omega)\cdot\frac{|Y(q-1,\omega)|^2}{|\hat{D}(\omega)|^2} \end{aligned} \tag{11}$$

where $G(q-1, \omega)$ stands for the gain function (9) estimated in the previous frame. A detailed explanation about the effect of each parameter of (10) and (11) in the gain function expressed by (9) can be found in [6].

The *a priori* SNR $R_{prio}(q,\omega)$ is evaluated by the nonlinear recursive relation of (11) and is the dominant parameter in (9), as we can see in Fig. 2. Strong attenuations are obtained only if R_{prio} is low and low attenuations are obtained only if R_{prio} is high. When R_{prio} is low and the *a posteriori* SNR R_{post} is high, there is a very strong attenuation (left–hand part of Fig. 2). This behavior is a consequence of the disagreement between *a priori* and *a posteriori* SNRs and it is actually useful in the elimination of the musical noise.

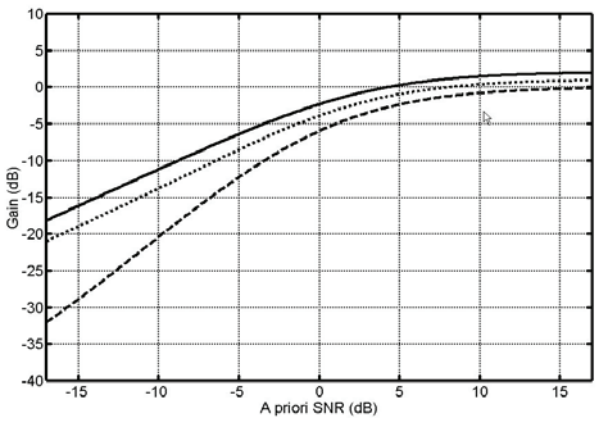


Fig. 2: EMSR gain versus a priori SNR for different values of R_{post}; top curve: $R_{post} = -17$ dB; middle curve: $R_{post} = 0$ dB; bottom curve: $R_{post} = 17$ dB.

In our algorithm, the gain function is also calculated by (9), but *a priori* and *a posteriori* SNRs are derived by means of the following relations:

$$R_{post}(q,\omega) = \frac{|Y(q,\omega)|^2}{\alpha(q,\omega)\cdot|\hat{D}(\omega)|^2} - 1 \tag{12}$$

$$\begin{aligned} R_{prio}(q,\omega) &= (1-\mu)\cdot R_{post}(q,\omega) + \\ &+ \mu\cdot\nu\cdot G^2(q-1,\omega)\cdot\frac{|Y(q-1,\omega)|^2}{\alpha(q,\omega)\cdot|\hat{D}(\omega)|^2} + \\ &+ \mu\cdot(1-\nu)\cdot G^2(q-2,\omega)\cdot\frac{|Y(q-2,\omega)|^2}{\alpha(q,\omega)\cdot|\hat{D}(\omega)|^2} \end{aligned} \tag{13}$$

where μ and ν were experimentally set to 0.96 and 0.75, respectively, and the calculation of $\alpha(q,\omega)$ follows (8). The time-frequency dependant perceptual overattenuation factor $\alpha(q,\omega)$ operates in a way similar to parameter α in (3) and depends on the noise masking threshold $T(q,\omega)$ (now with $\alpha_{min} = 0.75$, $\alpha_{max} = 2.5$), which is calculated for each frame q as explained in the previous section.

Other important difference between our algorithm and the standard EMSR is the presence of a third term in (13), which was empirically proved to be efficient in increasing the smoothness of R_{prio} over successive frames, thus allowing better reduction of the musical noise. It occurs because the main cause of the musical noise is the inaccurate estimation of R_{prio}, which normally lead to great variations of this parameter over successive frames.

RESULTS

In order to compare the performance of our algorithm to the performance of the standard EMSR algorithm, we performed an objective evaluation of the enhanced speech quality using the PESQ–MOS [7] score. The noisy signals and the reference clean signals were obtained from the SpEAR [8] (tables I, II and III) and Aurora 2 [9] (tables IV, V an VI) databases. In the first database (SpEAR), the noisy signals were obtained by acoustically adding the clean signal and the noise in a controlled environment. With several types of noise combined with clean speech at different SNRs, the results were presented in the form of averages (of both SNRs and PESQ scores) from a total of 33 WAVE files.

TABLE I

AVERAGE PESQ–MOS MEASURES AT SNR FROM 0 TO 5 dB

NOISE TYPE → (Average SNR)	WHITE (3,22dB)	PINK (2,78dB)	F16 (2,65dB)	FACTORY (3,49dB)
No processing	1,980	1,917	2,094	2,414
EMSR (μ = 0.96)	2,487	2,386	2,484	2,756
Proposed algorithm	**2,601**	**2,512**	**2,591**	**2,854**
Theoretical limit	3,879	3,728	3,801	3,877

TABLE II

AVERAGE PESQ–MOS MEASURES AT SNR FROM 5 TO 10 dB

NOISE TYPE → (Average SNR)	PINK (6,97dB)	F16 (6,21dB)	CAR (7,89dB)	FACTORY (5,17dB)
No processing	1,878	2,194	3,183	2,213
EMSR (μ = 0.96)	2,489	2,749	3,667	2,622
Proposed algorithm	**2,663**	**2,883**	**3,695**	**2,744**
Theoretical limit	3,620	3,910	4,143	3,747

TABLE III

AVERAGE PESQ–MOS MEASURES AT SNR FROM 10 TO 15 dB

NOISE TYPE → (Average SNR)	PINK (14,85dB)	F16 (12,13dB)
No processing	2,499	2,647
EMSR (μ = 0.96)	3,254	3,257
Proposed algorithm	**3,410**	**3,298**
Theoretical limit	3,957	4,064

In addition, we carried out an experiment using the proposed speech enhancement algorithm as a pre–processing step of a standard HMM connected–word speech recognition system. The AURORA 2 experimental framework (based in a carefully prepared noisy database using the original clean TIDIGITS) was used exactly as described in [9], with the same front–end and back–end, allowing direct comparison of performance with other systems.

TABLE IV

AVERAGE WORD ACCURACY RECOGNITION RATE (%) – TEST A AND B (PARTIAL SNRS AND NOISE TYPES) MULTI–CONDITION TRAINING FROM AURORA 2 DATABASE

SNR (dB)	TEST A			TEST B		
	CAR			TRAIN-STATION		
	ORIG	**OUR**	EMSR	ORIG	**OUR**	EMSR
15	97.61	**98.09**	98.15	95.53	**97.69**	97.50
5	87.80	**92.81**	93.05	83.52	**87.29**	87.63
0	53.44	**80.50**	81.39	56.12	**69.82**	69.81

TABLE V

AVERAGE WORD ACCURACY RECOGNITION RATE (%) – TEST A AND B (PARTIAL SNRS AND NOISE TYPES) CLEAN TRAINING FROM AURORA 2 DATABASE

SNR (dB)	TEST A			TEST B		
	CAR			TRAIN-STATION		
	ORIG	**OUR**	EMSR	ORIG	**OUR**	EMSR
15	90.04	**95.35**	96.69	83.65	**92.38**	93.77
5	34.09	**73.37**	77.66	27.92	**62.79**	67.42
0	14.46	**45.27**	49.93	11.57	**34.59**	38.94

TABLE VI

AVERAGE PESQ-MOS – TEST A AND B (PARTIAL SNRS AND NOISE TYPES) CLEAN TRAINING FROM AURORA 2 DATABASE

SNR (dB)	TEST A			TEST B		
	CAR			TRAIN-STATION		
	ORIG	**OUR**	EMSR	ORIG	**OUR**	EMSR
15	2.493	**2.937**	2.880	2.577	**2.929**	2.886
5	1.878	**2.377**	2.284	1.937	**2.339**	2.272
0	1.618	**2.036**	1.946	1.638	**1.997**	1.935

CONCLUSION

The perceptual results (PESQ-MOS) showed that our speech enhancement system outperforms the standard EMSR algorithm, for all noise types and SNRs considered in both databases. The improvement can be mainly explained by the effect of the introduction of a perceptual–dependent overattenuation factor in the derivation of R_{prio} and R_{post}. Regarding the speech recognition results, we can observe just the opposite: the EMSR showed a slight better performance, probably because it causes less distortion to the speech signal.

REFERENCES

[1] Y. Ephraim and D. Malah, *Speech enhancement using a minimum mean–square error short-time spectral amplitude estimator*, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP–32, no. 6, pp. 1109-1121, 1984

[2] N. Virag, *Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System*, IEEE Trans. Speech Audio Processing, vol. 7, no. 2, pp. 126–137, March 1999.

[3] M. Berouti, R. Schwartz, and J. Makhoul, *Enhancement of speech corrupted by acoustic noise*, in Proc. IEEE ICASSP, Washington, DC, pp. 208–211, Apr. 1979.

[4] M. R. Schroeder, B.S. Atal and J.L. Hall, *Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear*, in Journal of Acoustical Soc. of America, pp. 1647-1652, 1979.

[5] D. Sinha and A.H. Tewfik, *Low bit rate transparent audio compression using adapted wavelets*, Trans. Signal Processing, vol.41, pp. 3463-3479, December 1993.

[6] O. Cappé, *Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor*, IEEE Trans. Speech Audio Processing, vol. 2, no. 2, pp. 345–349, April 1994.

[7] Antony W.Rix et. al., *Perceptual Evaluation of Speech Quality (PESQ). The New ITU Standard for End–to–End Speech Quality Assessment*, Journal of Audio Eng. Soc., vol. 50, no. 10, pp. 755–778, October 2002.

[8] E. Wan, A. Nelson, and Rick Peterson. *Speech Enhancement Assessment Resource (SpEAR) database*. http://cslu.ece.ogi.edu/nsel/data/SpEAR_database.html. Beta Release v1.0. CSLU, Oregon Graduate Institute of Science and Technology.

[9] H.G. Hirsch, D. Pearce, *The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions,* ISCA ITRW ASR2000, Paris, France, September 18–20, 2000.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

A Visual Sound Description for Speech Corpora's Manual Phonemic Segmentation

She Kun[1], Chen Shu-zhen[1]
[1]School of Electronic Information, Wuhan University, Wuhan 430079, China
intel_ghost@sohu.com, szchen@whu.edu.cn

ABSTRACT

A visual sound description, called sound dendrogram is introduced to simplify speech corpora's manual annotation. Sound dendrogram is a lattice structure, constructed by an iterative procedure of mergence from a group of "seed regions". It can present the corresponding speech excerpt's rich structure information ranging from coarse to fine. Tests show that all phonemic boundaries are contained in this lattice structure and easy to identify. If integrated into the existed speech analysis programs, sound dendrogram can provide essential information for speech corpora's manual annotation.

INTRODUCTION

Building speech corpora is a vital task for developing almost all the currently available speech processing systems, including large vocabulary speech recognition systems [1, 2], speaker recognition systems [3] and language identification systems [4] etc.. Segmentation of speech, on phoneme level or word level, is a standard annotation work within speech corpora. In the literature, much effort is put to make this work automatic [5, 6], but however, the scores achieved by machine yet match those by a trained phonetician, and "true value" is still given by manual annotation.

Some speech analysis tools, like Praat[1], can provide some assist to this tedious manual procedure. These tools usually display speech's waveform, along with intensity and pitch contours, and sometimes short-time spectrogram, too. However, the clues on phonemic boundaries, provided by these descriptions are obscure, if not lacking, because in natural speech, there are many cases where intensity or pitch doesn't vary abruptly at the transition from one phoneme to another. Spectrogram is the most used visual description of an acoustic sound, by which an experienced phonetician can even "see" rather than "hear" speech, but spectrogram cannot provide speech's structure information directly. And, because of speech's continuous nature (that is, articulation gesture changes continuously), the boundaries between the realizations of two adjacent phonemes are blurred, so a human annotator will hesitate on where to flag the phonemic boundary. So for the most cases, it is still by repeatedly listening to playback that a boundary can be confirmed. Thereby, speech annotation remains time-consuming, which limits the scale of speech corpora.

In this paper, a kind of visual sound description, called sound dendrogram, is presented as a supplement to those mentioned above. It is a lattice structure automatically constructed from a group of "seed regions" and through an iterative procedure of mergence. Not like the other sound descriptions, sound dendrogram directly presents the structure information of an acoustic sound. The evaluation to sound dendrogram will show that all of a speech excerpt's phonemic boundaries are contained in the lattice structure of its sound dendrogram. With the assist of sound

[1] http://www.fon.hum.uva.nl/praat/

dendrogram, we believe, speech corpora's annotation work could be much easier.

CONSTRUCTION OF SOUND DENDROGRAM

Sound dendrogram is built by a local clustering procedure. First, the audio signal is divided by some means into a sequence of small sections, called "seed regions", whose borders are all potential phonemic boundary (These regions and their borders locate at the bottom level of the dendrogram). Then, distance of every two adjacent regions is computed and every couple of regions with local minimum distance is merged to form a new region. In this way, a new set of regions are born and they locate at the second level in the dendrogram. After, a new turn of mergence of closest regions follows and the dendrogram keeps growing upwards. This process repeats until only a single region remains. The mergence step is illustrated by Figure 1.

Since whether to merge relies only on relative distance, no threshold is needed. If the segmentation of "seed regions" is appropriate, several consecutive "seed regions" together will match a phoneme nicely, and they should merge into a single region at some higher level in the lattice structure, as acoustic characters usually keep well stable through the duration of a phoneme in speech. On the other hand, there is great difference between two regions separated by a phonemic boundary, so these two regions will resist merging and this boundary can spread to a very high level. Figure 2 shows a dendrogram produced in this way and several other sound descriptions such as waveform, spectrogram, etc. All of the phonemic boundaries (known by manual annotation) are contained in the dendrogram and easy to identify, while the other descriptions fail to give any information.

Signal Representation

The segmentation of "seed regions" and the distance metric are both based on a certain signal representation of acoustic sound. This paper adopts the third stage output of an auditory model proposed by Seneff, which is a multi-dimensional representation and can be identified with the average rate of neural discharge [7]. Rather than the strategy of "framing before processing" applied by short-time analysis, such as Mel-frequency cepstrum coefficients, signal representation based on this auditory model is reached by "sampling after processing" [8]. So, the dynamic information in speech has been preserved in this signal representation through much "smoother" transition and thereby, it is capable of locating indistinctive phonemic boundaries.

Segmentation of "Seed Regions"

To ensure that every phonemic boundary is among the borders of seed regions, the acoustic landmarks in speech are taken as seed region's border, since at these points the signal is undergoing significantly more change than in the neighboring environment, which always implies a phoneme's onset or offset. As mentioned above, the audio signal is represented by a multi-dimensional parameter $\boldsymbol{S}(t)$, so in this paper the magnitude of its first order derivative $\dot{\boldsymbol{S}}(t)$ is taken to indicate the rate of the signal's change.

Since most analysis of speech is performed in a discrete manner, the derivative operation has to be approximated by some discrete operator, such as smoothing the discrete signal representation $\boldsymbol{S}[n]$ in some degree and then carrying a difference operation to it. Smoothing and differencing can be done in a single step, by convolving each dimension of $\boldsymbol{S}[n]$ with the samples of the minus of a Gaussian's derivative, that is,

$$d[n] = -\frac{\mathrm{d}}{\mathrm{d}t} g(t)\Big|_{t=nT}, \; g(t) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{t^2}{2\sigma^2}} \quad (1)$$

where T denotes the signal representation's sample period, and σ is the parameter of the Gaussian function $g(t)$. Then a new function for rate of change is given by

$$c_\sigma[n] = \left\| \boldsymbol{S}[n] * d[n] \right\| \quad (2)$$

where the operator $\|\cdot\|$ takes the magnitude of a vector. In order to have a fine level of sensitivity in $c_\sigma[n]$, σ should be set to a small value.

Finally, the local maximum points in $c_\sigma[n]$ are detected and used to form the seed regions. Since the nonlinear modules in the 3rd stage of Seneff's model sharpen acoustic transition in speech [7], all real phonemic boundaries can be surely found. Some spurious borders may be found too, but it does not matter much as these borders will vanish quickly in the process of mergence when constructing sound dendrogram.

Distance Metric

At each level of sound dendrogram, a region is described by the mean of the signal representation vectors of all samples belonging to this region, that is,

$$\bar{\boldsymbol{S}}_{r_x} = \frac{1}{n_1 - n_0 + 1} \sum_{i=n_0}^{n_1} \boldsymbol{S}[i] \quad (3)$$

where the samples indexed by $n_0 \sim n_1$ belong to region r_x. Then, the distance between region r_1 and region r_2 is defined as

$$d(r_1, r_2) = \left\| \bar{\boldsymbol{S}}_{r_1} - \bar{\boldsymbol{S}}_{r_2} \right\| \times (1 - \cos\alpha)$$
$$\cos\alpha = \frac{\bar{\boldsymbol{S}}_{r_1} \bullet \bar{\boldsymbol{S}}_{r_2}}{\left\| \bar{\boldsymbol{S}}_{r_1} \right\| \left\| \bar{\boldsymbol{S}}_{r_2} \right\|} \quad (4)$$

where $\left\| \bar{\boldsymbol{S}}_{r_1} - \bar{\boldsymbol{S}}_{r_2} \right\|$ is the Euler distance between vectors $\bar{\boldsymbol{S}}_{r_1}$ and $\bar{\boldsymbol{S}}_{r_2}$, and $\cos\alpha$ is their normalized dot product. The Euclidean metric over-emphasize the gain difference between two regions, and therefore two regions belonging to the same phoneme may keep from merging as a result of the sound intensity's fluctuation. As shown in Figure 3, if two adjacent regions belong to the same phoneme, the according $\cos\alpha$ approaches 1, and much less than 1 if not. Glass [9] weights the Euler distance with $1/\cos\alpha$ to magnify the distance between two regions separated by a phonemic border. However, the Euler distance between these two regions is significant, too, so the effect of weighting is not obvious (See Figure 3). So, $1-\cos\alpha$ is adopted instead to suppress region distances within a phoneme so that regions belong to the same phoneme

merge much easily.

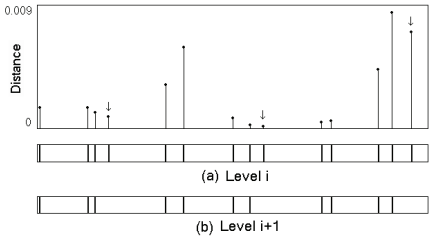


Fig. 1 A turn of region mergence
(a) The set of regions locating at the ith level and the distances between two adjacent regions (all local minimum distances are marked with downward arrows); (b) The set of regions at the $i+1^{th}$ level

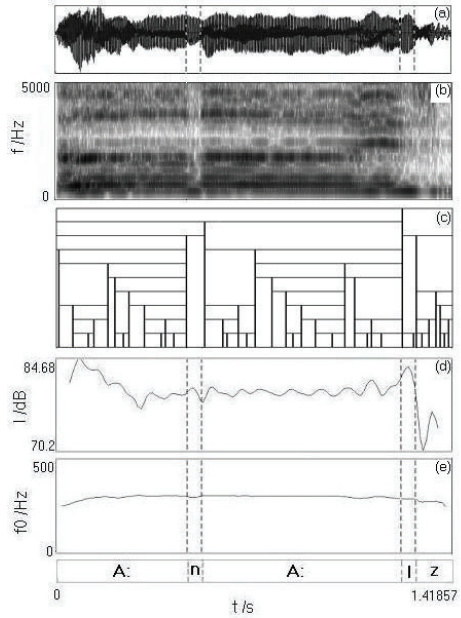


Fig. 2 Some speech sound descriptions
(a) The waveform; (b) The "wide band" spectrogram; (c) The lattice structure of sound dendrogram; (d) The intensity contour; (e) The Pitch contour. The phonemic boundaries are marked on the bottom ("A", "n", etc. are phonetic symbols signed with the SAM Phonetic Alphabet)

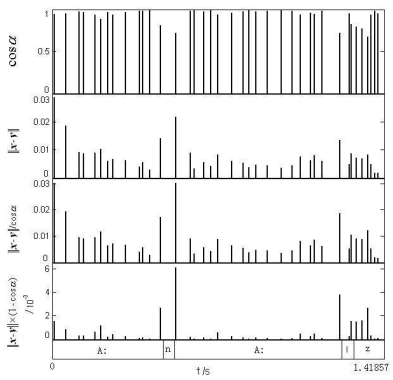


Fig. 3 Several distance metrics
Each stem locates on the borderline between two adjacent regions

EVALUATION AND DISCUSSION

The benefit from sound dendrogram was evaluated in several ways. First, a path through each dendrogram which best matched a time-aligned phonetic transcription was found using an automatic time alignment tool developed by us, and then, the deletion and insertion errors of these paths were tabulated. Next, the time difference between the boundaries found and the actual boundaries as provided by the transcriptions was compared. Finally, the height distributions of the valid/invalid boundaries in these dendrograms were examined. The evaluation was carried out using several sentences spoken by three subjects (two male, one female); these speeches were sampled at 16 kHz in a noisy computer room, and contained 165 units, phoneme or syllable[1].

The best-path alignment procedure gave almost none deletion error and 13% insertion error, respectively. The tradeoff between deletion and insertion error is met by all phonemic segmentation algorithms. Since sound dendrogram is used to provide clue for manual annotation, it is crucial to get the deletion error as little as possible. Relative higher insertion error rate may be due to coarse annotation. In fact, the insertion error was well suppressed by adopting the distance metric illustrated in equation (4). To prove that, the distance metric adopted by Glass [9] was used instead, and the insertion error became 20%. The sound dendrogram of the speech excerpt in Figure 2 was constructed again with the latter distance metric, and is showed in Figure 4. The regions belonging to phoneme /z/ failed to merge together as a result of the reason mentioned above.

The Analysis of the time difference between the boundaries found and the boundaries provided by the transcriptions showed that more than 74% of the boundaries were within 10ms of each other, while 80% of them were within 20ms. This degree of accuracy is comparable with those acquired by normal manual annotation [5, 6]. Finally, the statistics of boundary heights, valid and invalid, are shown in Figure 5. The valid boundaries are typically higher, so they can be distinguished easily from those invalid.

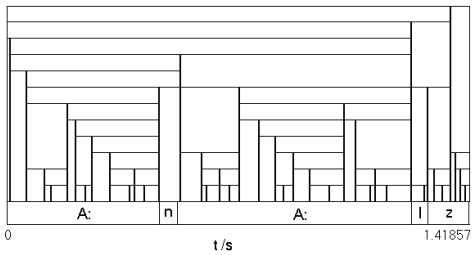


Fig. 4 The dendrogram with a different distance metric

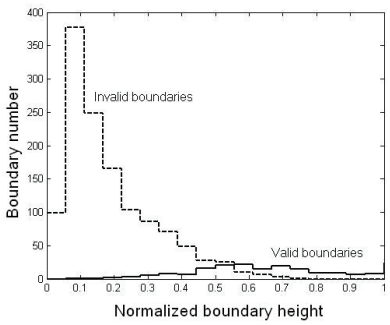


Fig. 5 Histogram of boundary height
Every boundary height is normalized by the total level number of its host sound dendrogram

[1] Some phonemes, especially stop consonants, like /p/, /b/, /t/, /d/ are transient, noncontinuant sound. Their properties are highly influenced by the vowels that follow them and few distinguishing features are shown in their own waveforms [10]. Since separating stop consonant and its following vowel is much difficult, they are not separated in the phonetic transcription.

Defining a metric to measure how much convenience sound dendrogram can bring to manual annotation is hard, if not impossible. Therefore several more typical examples are given, instead (Figure 6-8). With sound dendrogram available, the manual phonemic segmentation work becomes "observing (for example, the spectrogram) and choosing (the phonemic border from the dendrogram)", much easier than deciding where to put phonemic borders without any reference.

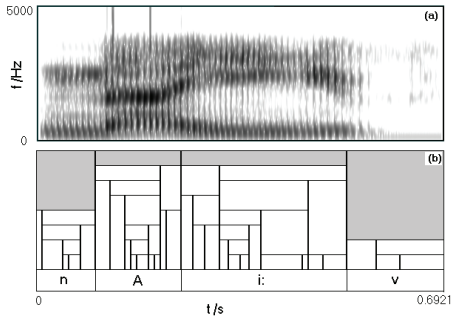


Fig. 6 The sound dendrogram of speech excerpt 0_1
(a) The spectrogram; (b) The sound dendrogram (The shadow lattices are the path best matched with the phonetic transcription and found automatically by the time alignment tool, the same in Figure 7 and Figure 8)

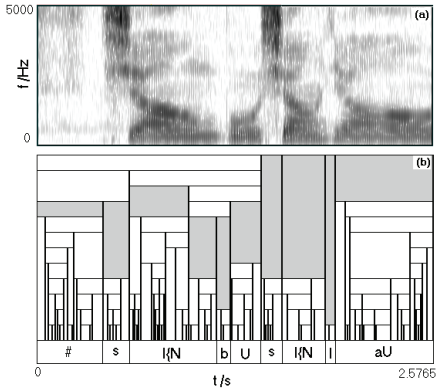


Fig. 7 The sound dendrogram of speech excerpt 5_2

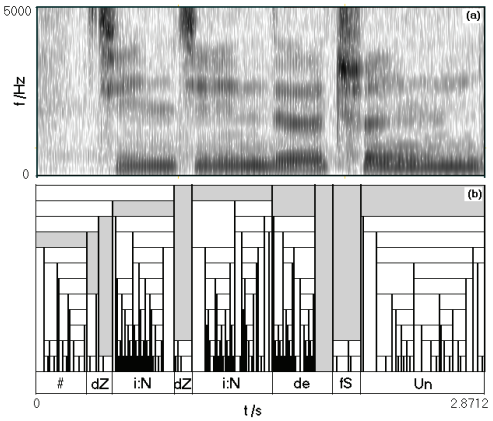


Fig. 8 The sound dendrogram of speech excerpt 4_1

CONCLUSION

The sound dendrogram proposed by this paper can reliably capture all phonemic boundaries in a speech. When it is integrated into the existed sound analysis tools, we believe, the efficiency of annotating speech corpora can be improved significantly. Moreover, some automatic method based on dendrogram for phonemic segmentation can be found in the literature, like Husson [11], which providing an automatic path-finding algorithm. Although there is still large developing space for these methods [12], the automatic found path can provide a useful reference. So, a reliable path-finding method is worthy of further research.

REFERENCES

[1] Tang M. Large Vocabulary Continuous Speech Recognition Using Linguistic Features and Constraints. Ph. D. thesis, the Massachusetts Institute of Technology, 2005.

[2] Campbell J, Reynolds D. Corpora for the Evaluation of Speaker Recognition Systems. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. Phoenix, pp. 829-832, May 1999.

[3] Furui S. 50 Years of Progress in Speech and Speaker Recognition. http://www.furui.cs.titech.ac.jp/publication/2005/SPCOM05.pdf.

[4] Padró M, Padró L. Comparing Methods for Language Identification. http://www.lsi.upc.edu/~nlp/papers/2004/sepln04-pp.pdf.

[5] Laureys T, Demuynck K, Duchateau J, Wambacq P. An Improved Algorithm for the Automatic Segmentation of Speech Corpora. Proceedings of the 3rd International Conference on Language Resources and Evaluation. Las Palmas, pp. 1564-1567, May 2002.

[6] Sharma M, Mammone R. "Blind" Speech Segmentation: Automatic Segmentation of Speech without Linguistic Knowledge. Proceedings of the 4th International Conference on Spoken Language Processing. Philadelphia, pp. 1237-1240, October 1996.

[7] Seneff S. A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing. Journal of Phonetics, Special Issue, Vol. 16, No. 1, pp. 55-76, 1988.

[8] Cosi P. Evidence Against Frame-Based Analysis Techniques. www.pd.istc.cnr.it/Papers/PieroCosi/cp-NATO98.pdf

[9] Glass J R. Finding Acoustic Regularities in Speech: Application to Phonetic Recognition. Ph. D. thesis, the Massachusetts Institute of Technology, 1988.

[10] Rabiner L, Juang B H. Fundamentals of Speech Recognition. Prentice Hall, 1993.

[11] Husson J L, Laprie Y. A New Search Algorithm in Segmentation Lattices of Speech Signals. Proceedings of the 4th International Conference on Spoken Language Processing, Philadelphia, pp. 2099 -2102, October 1996.

[12] Husson J L. Evaluation of A Segmentation System Based on Multi-Level Lattices. Proceedings of the 6th European Conference on Speech Communication and Technology. Budapest, pp. 471-474, September 1999.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Equalizador gráfico digital de alta seletividade em VST

Leonardo de O. Nunes[1], Alan F. Tygel[1], Rafael A. de Jesus[1], e Luiz W. P. Biscainho [1]

[1]LPS – PEE/COPPE & DEL/Poli, UFRJ
Caixa Postal 68504, Rio de Janeiro, RJ, 21941-972, Brasil
lonnes,alan,rjesus,wagner@lps.ufrj.br

RESUMO

Este trabalho apresenta a implementação de um equalizador gráfico digital de 1024 canais lineares agrupados em 10 oitavas, com alta seletividade. A estrutura escolhida foi um *Fast Filter Bank* (FFB), banco de filtros altamente seletivos que preserva a baixa complexidade da FFT, em que se baseia. Os ganhos atribuídos a cada oitava são interpolados suavemente através dos ganhos de cada subcanal. A implementação é realizada na linguagem C++, sendo gerado um *plug-in* no padrão VST.

INTRODUÇÃO

A extraordinária evolução dos processadores digitais no último quarto do século XX abriu as portas para uma verdadeira revolução que aproximou as aplicações de ciência avançada do usuário comum. Especificamente na área de áudio, o processamento digital pode ser encontrado desde nos equipamentos domésticos de som até numa quantidade de aplicativos para manipulação e reprodução de áudio disponíveis em 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. Será apresentado, então, o procedimento de projeto de um equalizador gráfico digital de 10 oitavas baseado em um *Fast Filter Bank* de 1024 canais lineares. Este banco de filtros combina alta seletividade com baixa complexidade. A fim de permitir a fácil utilização e portabilidade do sistema, utilizou-se o padrão de *plug-in* VST[1], amplamente aceito por fabricantes e usuários de aplicativos de áudio profissional.

Após esta Introdução, o artigo é organizado da seguinte forma. Uma breve revisão da estrutura chamada FFB (*Fast Filter Bank*) é seguida do detalhamento de sua implementação proposta no trabalho. Na seção seguinte especifica-se o equalizador gráfico que serve de aplicação ao FFB, fazendo-se a correspondência entre os ganhos definidos pelo usuário e os ganhos reais do banco de filtros. Após uma breve discussão do *plug-in* em VST, apresentam-se as conclusões.

FAST FILTER BANK (FFB)

Definição

Esta seção descreve o *Fast Filter Bank* (FFB), que é a estrutura adotada como base do equalizador descrito neste trabalho.

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

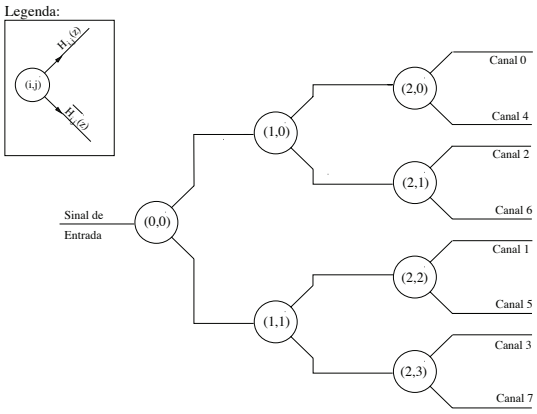


Figura 1: Construção dos canais de um FFB de oito canais a partir das versões modificadas dos filtros-*kernel* dos três níveis da estrutura.

A ferramenta mais popular de análise espectral para sinais discretos no tempo é a *Discrete Fourier Transform* (DFT) [1], definida como

$$X[k] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]\, e^{-j\frac{2\pi kn}{N}},$$

onde $x[n]$ é o sinal no tempo, $X[k]$ é sua representação no domínio da freqüência, na forma de um par (módulo,fase) associado à componente $e^{j\frac{2\pi kn}{N}}$. Ela admite implementações rápidas, genericamente chamadas de *Fast Fourier Transform* (FFT), das quais as mais usuais são as de raiz 2 [2].

É possível representar a FFT na forma de um banco de filtros em árvore [3], conforme se vê na Figura 1. Diferentente da FFT usual, que opera sobre blocos do sinal de entrada, nessa estrutura cada amostra da entrada origina N amostras na saída, uma para cada canal. O j-ésimo filtro de cada nível, i, da árvore é obtido pela modificação de um mesmo filtro-*kernel*

$$H(z) = 1 + z^{-1}, \tag{1}$$

de acordo com a expressão

$$H_{ij}(z) = H(W_N^{-\tilde{j}} z^{2^{L-i-1}}), \tag{2}$$

onde $L = \sqrt{N}$, $W_N = e^{-j\frac{2\pi}{N}}$ e $\tilde{j}$ é j com os *bits* na ordem reversa. Com isso, o filtro-*kernel* é deslocado na freqüência e estreitado por interpolação dos seus coeficientes, de acordo com sua posição na árvore. As réplicas indesejadas na resposta de um dado filtro, decorrentes da interpolação, são estruturalmente eliminadas nos níveis subseqüentes da árvore.

Os filtros dos canais resultantes apresentam fase linear e o mesmo atraso de grupo. Dessa forma, apesar das ordens elevadas dos filtros envolvidos, o único efeito significativo sobre a fase do sinal é um atraso global.

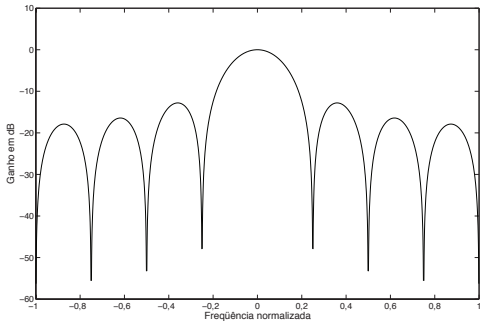


Figura 2: Resposta de módulo na freqüência de um filtro da FFT.

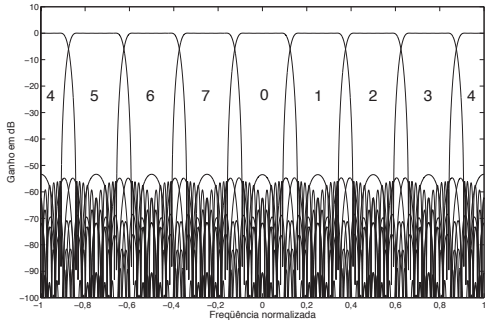


Figura 3: Resposta de módulo na freqüência dos canais do FFB. O índice de cada canal está indicado na figura.

Como se pode observar na Figura 2, a resposta de módulo na freqüência de um canal do banco de filtros correspondente à FFT apresenta baixa atenuação na faixa de rejeição, da ordem de 13 dB. Com o intuito de melhorar essa característica, em [4] propôs-se o *Fast Filter Bank* (FFB), onde o filtro-*kernel* da FFT pode ser substituído por filtros de ordem mais alta, potencialmente mais seletivos. Essa generalização admite filtros-*kernel* diferentes para cada nível, $H_i(z)$.

A título de ilustração, a Figura 3 mostra a resposta de módulo na freqüência para todos os canais de um FFB de ordem 8 com filtros-*kernel* de ordens 23, 19 e 7 na ordem crescente dos níveis i, onde se pode notar a elevada atenuação na banda de rejeição.

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

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

evita operações redundantes. A saída $\overline{y}(n)$ do filtro complementar $\overline{H_{ij}}(z)$ para uma entrada $x(n)$ pode ser

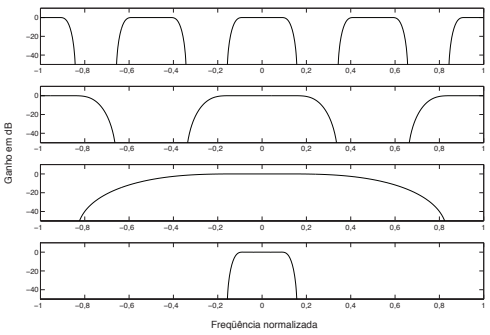


Figura 4: Construção do canal 0 de um FFB de oito canais a partir das versões modificadas dos filtros-*kernel* dos três níveis da estrutura. Os gráficos representam, de cima para baixo, a resposta de módulo na freqüência dos filtros (0,0), (1,0) e (2,0) e o filtro resultante para o Canal 0 da Figura 1.

obtida através de:

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

onde

$$y(n) = h_{ij}(n) * x(n)$$

é própria a saída do filtro $H_{ij}(z)$.

O projeto dos filtros pode ser feito através do método FRM (*Frequency Response Masking*) [5], que permite a obtenção de filtros com banda de atenuação estreita, por interpolação de coeficientes.

Em [6] pode ser encontrada uma discussão detalhada do projeto dos filtros e da complexidade do FFB. Os filtros obtidos em cada estágio da estrutura referente à Figura 3, bem como o filtro resultante para o canal 0, podem ser vistos na Figura 4.

Implementação

Será descrita a seguir a estratégia de implementação do FFB adotada neste trabalho.

O FFB foi implementada em C++ [7], tendo sido criadas duas classes, a `FfbFilter` e a `FfbFilterTree`. A primeira descreve um único filtro dentro da estrutura em árvore, enquanto que a outra descreve a própria árvore. Será feita agora uma descrição detalhada de cada classe.

Os filtros utilizados pelo FFB possuem uma estrutura muito particular que permite um número reduzido de operações. Após as transformações necessárias descritas em (2), os coeficientes dos filtros se apresentam como na Tabela 1. Como pode ser visto, o número de elementos não-nulos e não-unitários para os filtros $H_{ij}(z)$ continua o mesmo do filtro-*kernel* $H_i(z)$.

Os filtros foram implementados na forma direta não-causal, multiplicando-se a saída da memória pelo seu respectivo coeficiente e somando os resultados, apenas para os coeficientes não-nulos e não-unitários.

$W_N^{\tilde{j}M} h_i[-M]$
$2^{L-i} - 1$ zeros
$W_N^{\tilde{j}(M-1)} h_i[1-M]$
$\vdots$
$W_N^{\tilde{j}2} h_i[-2]$
$2^{L-i-1} - 1$ zeros
$W_N^{\tilde{j}} h_i[-1]$
1
$W_N^{-\tilde{j}} h_i[1]$
$2^{L-i-1} - 1$ zeros
$W_N^{\tilde{j}(-2)} h_i[2]$
$\vdots$
$W_N^{\tilde{j}(1-M)} h_i[M-1]$
$2^{L-i} - 1$ zeros
$W_N^{\tilde{j}(-M)} h_i[M]$

Tabela 1: Valores dos coeficientes dos filtros $H_{ij}(z)$, considerando um filtro-*kernel* $H_i(z)$ de ordem $2M+1$.

Para tal foi necessário 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 5, 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 (lembrando que para os coeficientes $h_i[1]$ e $h_i[-1]$ 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 `FfbFilter` 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.

A filtragem é feita levando-se em conta o fato de os coeficientes do filtro serem conjugados-simétricos; para isso foi criada uma função que utiliza essa propriedade, requerendo o armazenamento de apenas metade dos coeficientes, além de reduzir o número de operações aritméticas.

Os dois principais métodos da `FfbFilter` são o `set_param`, no qual são passados a posição do filtro dzentro da árvore (i e j) e os seus coeficientes; e o `filter`, que recebe um valor complexo correspondente à entrada e retorna a amostra filtrada por ele e pelo seu complementar.

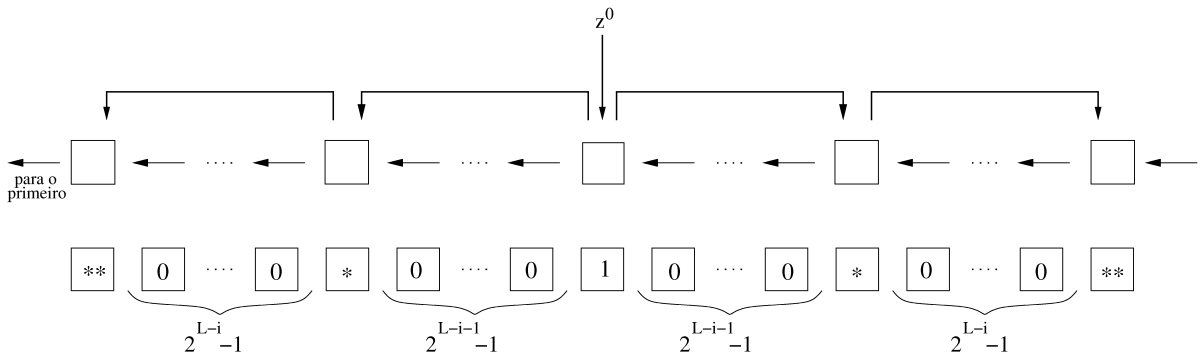


Figura 5: Diagrama da organização da memória de um sub-filtro do nível i, mostrando sua correspondência com os coeficientes do filtro (abaixo). As casas marcadas com asterisco indicam os coeficientes não-nulos. As setas indicam os ponteiros.

A classe `FfbFilterTree` possui um vetor contendo $\sqrt{N} - 1$ objetos do tipo `FfbFilter`, onde os filtros estão ordenados externamente por i e internamente por j, ou seja, o primeiro elemento desse vetor corresponde ao par (i, j) é $(0, 0)$, o segundo é $(1, 0)$, o terceiro é $(1, 1)$, e assim por diante.

O construtor da `FfbFilterTree` lê os coeficientes de cada filtro apartir de um arquivo-texto denominado `coefs.fir`. Cada linha desse arquivo contém o valor de metade dos coeficientes não-nulos e não-unitários de cada filtro, suficientes para o cálculo.

O método que realiza a filtragem nessa classe é denominado `filter`; recebe um valor em ponto flutuante como entrada e retorna um vetor complexo contendo as saídas de todos os canais. A saída de cada filtro é armazenada no próprio vetor de saída (*in place*), da mesma maneira que na FFT [8].

Em [9] é mostrada uma simplificação adicional da estrutura do banco de filtros para o caso de sinais de entrada reais, utilizando sua simetria no domínio da freqüência. Com isso, apenas metade dos filtros é utilizada, reduzindo o número necessário de operações. A ordenação dos canais na saída do filtro, originalmente em *bit-reversal*, é perdida. Mais adiante será proposto um algoritmo para realizar a leitura dos canais, após essa simplificação.

O EQUALIZADOR

Idéia Geral

Em processamento de sinais, um equalizador se destina a corrigir distorções lineares (de módulo e fase) sofridas por um sinal. Equalizadores para sinais de áudio normalmente objetivam corrigir modificações introduzidas no sinal pelo sistema e pelo ambiente de reprodução do som. Os tipos mais comuns de equalizadores de amplitude (módulo) são: o paramétrico, em geral com um número reduzido de filtros com freqüência central, ganho e largura de faixa ajustáveis; e o gráfico, em geral com diversos filtros passa-faixa com ganhos independentes por faixa. Tipicamente, os filtros atuam de 20 Hz a 20 kHz, limites aproximados da audição humana.

Um equalizador gráfico analógico emprega um potenciômetro para controlar o ganho de cada filtro ativo. Sua versão digital segue o mesmo princípio, sendo o ganho definido por constantes multiplicadoras aplicadas à saída de cada filtro digital. O usuário atua sobre uma interface gráfica amigável que freqüentemente simula o painel do equalizador analógico.

Uma configuração típica de equalizador gráfico divide o espectro de áudio em oitavas, partindo do limite superior. Assim, considerando que se vai operar sobre sinais digitais com qualidade de CD, cuja taxa de amostragem é de 44,1 kHz, o espectro útil se estende até 22,05 kHz. A última (décima) oitava vai de 11,025 a 22,05 kHz, a penúltima de 5,5125 a 11,025 kHz e assim sucessivamente, até a faixa restante, de 0 a aproximadamente 43,07 Hz.

Tendo-se decidido implementar o equalizador com base no FFB, cujo espaçamento entre filtros é linear, o número de filtros que permite alcançar a resolução de 43,07 Hz é 1024. Nesse contexto, o filtro 0 fica em torno de DC e o filtro 512, em torno de 22,05 kHz. Em se tratando de sinais reais, cada par de filtros (i, 1024-i), $1 \leq i \leq 1023$, receberá ganhos iguais e responderá pela i-ésima faixa do espectro, entre $21{,}53i$ e $21{,}53(i+2)$ Hz. Por sua vez, os filtros 0 e 512 podem ter seus ganhos zerados sem prejuízo do desempenho, já que isso apenas limitará a faixa útil ao intervalo de 21,53 Hz a 22,03 kHz.

A especificação de cada filtro do FFB determina, naturalmente, a complexidade global do sistema, que, em última análise, viabilizará ou não a sua operação em tempo-real. Os filtros utilizados neste trabalho têm 40 dB de atenuação na faixa de rejeição, resultando em filtros-*kernel* com 15, 11, 7, 3, 3, 3, 3, 3, 3 e 3 coeficientes, em ordem crescente de i, equivalendo a 16 multiplicações complexas por canal. Vale observar que os filtros podem ser alterados pela simples troca de um arquivo-texto, sem a necessidade de alteração

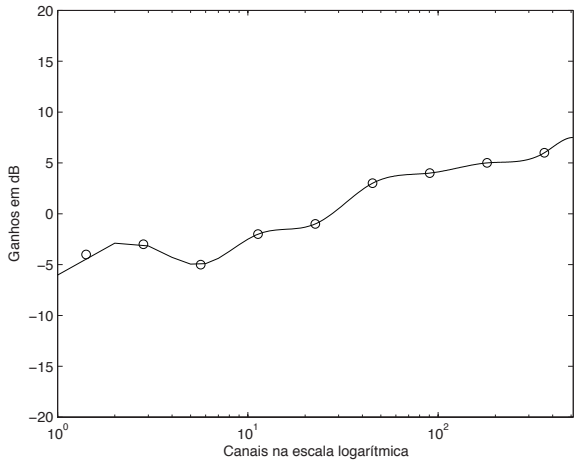


Figura 6: Curva de ganhos do FFB interpolados a partir dos ganhos fornecidos pelo usuário (○).

do código.

No sistema implementado, o usuário determinará 10 ganhos entre -12 e 12 dB, referentes às oitavas de atuação do equalizador, centralizadas aproximadamente em 30,5, 60,9, 122, 244, 487, 974, 1950, 3900, 7800 e 15600 Hz. Para obter os sub-ganhos lineares correspondentes aos filtros do FFB, interpolaram-se os ganhos fornecidos através de uma curva suave, a *cubic spline*. O procedimento é descrito na próxima seção.

Interpolação dos Ganhos

O problema da determinação dos ganhos pode ser resumido no seguinte: dada uma função tabelada $y_i = f(x_i)$, $i = 1 \ldots N$, deseja-se obter o valor da função num ponto localizado no intervalo $[x_j, x_{j+1}]$.

Uma possível solução seria a interpolação linear, que encontra o ponto buscado sobre o segmento de reta que liga os dois pontos conhecidos. Obviamente, essa solução possui a segunda derivada nula no intervalo considerado e infinita ou indefinida nos limites deste. A *cubic spline* [10] é uma função de comportamento suave na primeira derivada e contínuo na segunda, sendo definida pela equação

$$y = Ay_j + By_{j+1} + Cy''_j + Dy''_{j+1}, \tag{3}$$

onde

$$A = \frac{x_{j+1} - x}{x_{j+1} - x_j}$$
$$B = 1 - A$$
$$C = \frac{1}{6}\left(A^3 - A\right)\left(x_{j+1} - x_j\right)^2$$
$$D = \frac{1}{6}\left(B^3 - B\right)\left(x_{j+1} - x_j\right)^2.$$

A interpolação da *cubic spline* envolve duas etapas. Na primeira, recebem-se os pares (x_i, y_i) de entrada disponibilizados e as derivadas nas extremidades do intervalo ($i = 1$ e $i = N$, para as quais foi adotado o valor 0) e calculam-se as derivadas de segunda ordem. Na segunda etapa recebem-se os pares de entrada, as derivadas de segunda ordem e a abcissa x do ponto que se deseja interpolar, e calcula-se o valor de y correspondente.

Para obter uma curva mais suave foi necessário adicionar 2 pontos exteriores aos 10 pontos originalmente disponíveis na entrada, nas freqüências de 22 e 22 kHz, respectivamente. Suas ordenadas foram determinadas por uma simples extrapolação linear.

A rotina implementada recebe os 10 ganhos definidos pelo usuário em dB, e retorna os 511 ganhos para os canais de saída do FFB, também em dB. Um exemplo do resultado da interpolação descrita pode ser visto na Figura 6.

Implementação como VST

Nesta subseção será mostrado como as classes que implementam o banco de filtros FFB e o interpolador descrito na subseção anterior são combinados para formar o equalizador gráfico proposto. Também é mostrado um algoritmo capaz de ordenar a saída do FFB modificada para sinais reais.

Um *plug-in* pode ser definido como um programa que interage com outro de modo a oferecer novas funcionalidades, sendo geralmente distribuído como bibliotecas compartilhadas (*shared libraries*). O VST, do inglês *Virtual Studio Technology*, é um padrão desenvolvido pela empresa *Steinberg*, utilizado em uma variedade de aplicativos para áudio. Maiores informações sobre o padrão VST, bem como as bibliotecas necessárias, podem ser encontradas em [11].

O *plug-in* implementado conta com 10 parâmetros de entrada (os ganhos de cada oitava); esses ganhos são interpolados para se obter os ganhos de cada canal através do método descrito na seção anterior. No padrão VST qualquer parâmetro sempre é fornecido como um valor em ponto flutuante no intervalo entre 0 e 1; conseqüentemente faz-se necessário o mapeamento dos valores recebidos para a faixa de -12 dB a 12 dB. Os valores mapeados são, então, passados para o interpolador e o ganho de cada canal é armazenado num vetor estaticamente alocado.

A função responsável pelo processamento do sinal recebe um bloco de amostras e retorna um bloco de mesmo comprimento. Para cada amostra do bloco, ela utiliza o método `filter` da `FfbFilterTree` para obter a saída de todos os canais. Cada saída é multiplicada pelo ganho correspondente ao seu canal e esses produtos são somados, gerando a amostra de saída atual.

Devido à simplificação realizada sobre a estrutura em árvore do FFB, sua saída não possui uma ordenação simples. Como demonstrado em [9], para entradas reais as saídas dos filtros $\overline{H_{i,2}(z)}$ podem ser descartadas; isso implica o desaparecimento dos

Quadro 1: Algoritmo para localização dos canais de saída na estrutura simplificada.

```
enquanto(contador<(N/2))
   contador2 = 0;
   enquanto(contador2 < contador)
      se(bit_reversal(contador2+(2*contador),LL)>N/2)
         Posição do canal (N-bit_reversal(contador2+(2*contador))) = contador2+(2*contador);
      senão
         Posição do canal bit_reversal(contador2+(2*contador)) = contador2+(2*contador);
      contador2++;
   contador <<= 1;
```

canais numa progressão geométrica de razão 2, pois ao se retirar um filtro do nível i da árvore, 2^{9-i} canais de saída desaparecerão. Por exemplo, ao se eliminar a resposta do filtro $\overline{H_{1,2}}(z)$, os 2^8 últimos canais desaparecem da estrutura em árvore. Para localizar os canais na saída é necessário percorrer o vetor de saída em incrementos crescentes de acordo com a progressão geométrica, lembrando que as saídas para os canais $k > 512$ são equivalentes às saídas para $1024 - k$. O algoritmo no Quadro 1 descreve esse procedimento.

Esse algoritmo é utilizado apenas uma vez dentro do *plug-in*; a posição de cada canal é, então, salva num vetor, de modo a diminuir o número de operações dentro do bloco de processamento do sinal.

Devido à complexidade global do *plug-in*, sua implementação corrente ainda não permite a execução em tempo real, o que requererá otimização adicional do código.

CONCLUSÕES

Este trabalho apresentou uma implementação em C++ do algoritmo FFB aplicado à realização de um equalizador gráfico digital no padrão VST. A motivação inicial foi empregar o FFB numa aplicação típica de áudio que pudesse usufruir de sua alta seletividade e baixa complexidade. O sistema final implementado foi testado com sinais de áudio reais de alta qualidade, tendo sido bem avaliado em testes informais. Outras aplicações para o FFB poderão utilizar a implementação geral aqui apresentada.

Como continuação deste trabalho, pretende-se aumentar a velocidade de execução do processamento pela substituição da estrutura em árvore, mais flexível, pela formulação matricial descrita em [12].

AGRADECIMENTOS

Os autores gostariam de agradecer a Filipe C. da C. B. Diniz, Iúri Kothe e Sergio L. Netto pelas valiosas discussões ligadas ao trabalho; e às agências de fomento CNPq e FAPERJ pelo apoio na forma de bolsas de iniciação científica e de auxílio ao projeto de pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] S. Haykin and B. V. Veen, *Signals and Systems*. John Wiley & Sons, 1996.

[2] J. W. Cooley and J. W. Tukey, "An algorithm for the machine computation of complex fourier series," *Mathematics of Computation*, vol. 19, pp. 297–301, 1965.

[3] Y. C. Lim and B. Farhang-Boroujeny, "A comment on the computational complexity of sliding FFT," *IEEE Transaction on Circuits and Systems - II: Analog and Digital Signal Processing*, vol. 39, no. 12, pp. 875–876, 1992.

[4] Y. C. Lim and B. Farhang-Boroujeny, "Fast filter bank (FFB)," *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 39, pp. 316–318, May 1992.

[5] Y. C. Lim, "Frequency-response masking approach for the synthesis of sharp linear phase digital filters," *IEEE Transactions on Circuits and Systems*, vol. 33, pp. 357 – 364, April 1986.

[6] Y. C. Lim and B. Farhang-Boroujeny, "Analysis and optimum design of the FFB," *IEEE International Symposium on Circuits and Systems*, vol. 2, pp. 509 – 512, June 1994.

[7] B. Stroustrup, *The C++ Programming Language*. Addison-Wesley, 2000.

[8] P. S. R. Diniz, E. A. B. da Silva, and S. L. Netto, *Digital Signal Processing: System Analysis and Design*. United Kingdom: Cambridge, 2002.

[9] J. W. Lee and Y. C. Lim, "Efficient implementation of real filter banks using frequency response masking techniques," *Asia-Pacific Conference on Circuits and Systems*, vol. 1, pp. 69 – 72, 2002.

[10] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge, 1992.

[11] Steinberg, "Steinberg VST plugin." webpage, 2005. `http://www.steinberg.de/Steinberg/Developers8b99.html`.

[12] Y. C. Lim and J. W. Lee, "Matrix formulation: fast filter bank," *IEEE International Conference on Audio, Speech and Signal Processing*, vol. 5, pp. V – 133–6, May 2004.

Sociedade de Engenharia de Áudio
Artigo de Congresso
Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Aplicação em Áudio da Aproximação Mínimo Erro Médio Quadrático

Sidnei Noceti Filho, Calisto Schwedersky e Luiz Fernando Micheli

LINSE - Laboratório de Circuitos e Processamento de Sinais
Depto. Engenharia Elétrica, Universidade Federal de Santa Catarina
Telefone: (48)3331-9504, Fax: (48)3331-9091
88040-900, Florianópolis, SC, Brasil
sidnei@linse.ufsc.br , calisto@linse.ufsc.br , lfmicheli@linse.ufsc.br

RESUMO

Este artigo apresenta considerações sobre uma função pouco conhecida na literatura, aqui chamada função de Mínimo Erro Médio Quadrático (ME). Ela se caracteriza por apresentar, na banda de passagem, a magnitude da resposta em freqüência mais próxima da ideal. É feita uma comparação entre a função ME com outras funções clássicas usadas em divisores de freqüência para caixas acústicas. Além disso, é mostrado como determinar uma função de transferência ME.

INTRODUÇÃO

Em síntese de filtros, a solução (ou soluções) pode(m) ser obtida(s) com o uso de otimização. No entanto, uma solução analítica é possível com a utilização de funções de aproximação clássicas cujas características já foram exaustivamente estudadas. Nesse caso, a determinação da função de transferência (FT) de um filtro passa primeiramente pela determinação da função passa-baixa normalizada. Após isso, faz-se uma simples desnormalização (no caso de um filtro passa-baixa) ou de uma desnormalização acompanhada de uma transformação em freqüência (nos casos de filtros passa-alta, passa-faixa e rejeita-faixa) [1].

Conseguir uma sonoridade agradável em um sistema completo (fonte sonora + amplificação + caixa acústica + ambiente) não é uma tarefa trivial em vista da enorme variedade de parâmetros envolvidos (elétricos, mecânicos e acústicos). Por exemplo, o ouvinte pode conjugar a melhor fonte sonora, o melhor processamento eletrônico e a melhor caixa acústica. Se o ambiente acústico não for adequado, a sua resposta pode produzir efeitos desagradáveis ao ouvido em função, por exemplo, das possíveis reflexões das ondas sonoras.

O objetivo deste trabalho não é discutir estes aspectos de projeto relativos à iteração entre filtros e alto-falantes, com suas complexas impedâncias e variadas *SPL* (*sound pressure level*), o complexo modelo eletro-mecânico-acústico de alta ordem, a influência da disposição espacial dos alto-falantes nas caixas, etc., mesmo porque isto é assunto para um livro completo. O objetivo é discutir a opção de uso da função de aproximação Mínimo Erro Médio Quadrático (ME) e compará-la com as funções mais usadas no projeto de *crossovers*. Esse trabalho mostra a forma de determinação de funções ME de qualquer ordem, baseado nas poderosas ferramentas computacionais hoje disponíveis. Em adição, é mostrada uma tabela com funções características até a ordem 15 e um procedimento de cálculo das constantes de ganho, o que facilita sobremaneira a obtenção das FTs dos filtros ME. É

importante salientar que FTs digitais também podem ser obtidas a partir das correspondentes funções analógicas.

COMENTÁRIOS SOBRE AS FUNÇÕES CLÁSSICAS

As funções clássicas usadas no projeto de *crossovers* para caixas acústicas sempre apresentam características otimizadas em algum aspecto. A seguir, são comentadas as características principais destas funções, considerando-se a mesma ordem n e a mesma atenuação A_p no limite da banda passante.

Funções Butterworth (BT)

A aproximação BT é monotônica e apresenta a magnitude da resposta em freqüência mais plana na banda passante dentre todas as funções de aproximação polinomiais. As aproximações polinomiais são aquelas cujas FTS passa-baixa apresentam todos os zeros no infinito.

Funções Chebyshev (CB)

A aproximação CB se caracteriza por ser *equiripple* na banda passante e por apresentar o corte mais abrupto na banda de rejeição dentre todas as funções de aproximação polinomiais.

Funções Legendre (LG)

A aproximação LG, dentre todas as aproximações polinomiais monotônicas, se caracteriza por apresentar a maior declividade da magnitude na freqüência limite da banda passante (o que a faz mais seletiva do que a BT). No entanto, a sua determinação não é tão trivial quanto a da aproximação BT.

Funções Linkwitz-Riley (LR)

A aproximação LR [2] é uma tentativa de se obter aproximação do tipo passa-tudo em sistemas de duas vias, quando se soma uma função passa-baixa e uma passa-alta. Nesse caso, teoricamente não são introduzidas distorções na magnitude dos sinais. Na prática, utiliza-se apenas aproximações LR de segunda e quarta ordem. A aproximação de segunda ordem é obtida a partir da cascata de dois filtros de primeira ordem. A aproximação de quarta ordem é obtida a partir da cascata de dois filtros BT de segunda ordem. É importante observar que a vantagem dos filtros LR não existe no caso de sistemas de três ou mais vias.

Funções Bessel (BS)

A aproximação BS, dentre todas as aproximações polinomiais clássicas com pólos complexos, se caracteriza por apresentar a fase mais linear dentro da banda passante. Essa característica não é preservada nos filtros BS passa-alta e passa-faixa.

Funções Gauss

A aproximação GS, dentre todas as aproximações polinomiais clássicas com pólos complexos, se caracteriza por apresentar a melhor resposta temporal, isto é, o menor tempo de atraso e o menor *overshoot* na resposta ao degrau.

COMENTÁRIOS SOBRE AS FUNÇÕES ME

As funções ME se caracterizam por melhor aproximar as características reais da magnitude da resposta em freqüência na banda de passagem, em relação às características ideais. A Fig. 1 mostra uma comparação entre as respostas passa-baixa normalizadas ME com as aproximações clássicas utilizadas em *crossovers*. Todas as funções comparadas apresentam ordem $n = 4$ e atenuação de $A_p = 3\,\text{dB}$ no limite da banda de passagem normalizada $\omega_p = 1\,\text{rad/s}$. Uma função LR de ordem quatro apresenta naturalmente uma atenuação de $A_p = 6\,\text{dB}$ em $\omega_p = 1\,\text{rad/s}$. Assim, com o intuito de melhor comparar todas as funções, a aproximação LR foi escalada pelo fator $\omega_N \cong 0{,}80224$, de modo a apresentar também $A_p = 3\,\text{dB}$ no limite da banda.

Quando se compara as características de atenuação (CAA) com as características de fase (CAF) de funções de aproximação passa-baixas clássicas (CB, LG, BT, LR, BS e GS) utilizadas em *crossovers*, observa-se que sempre existe um compromisso entre tais características. Quanto melhores são as CAA, piores são as CAF e *vice-versa*.

Considera-se um filtro com melhores CAA aquele que atenda aos requisitos de seletividade com menor ordem. Considera-se um filtro com melhores CAF aquele que apresenta uma menor dispersão do atraso de grupo na banda de interesse. Nesse contexto, as aproximações CB, LG e BT são as que apresentam melhores CAA, nessa ordem.

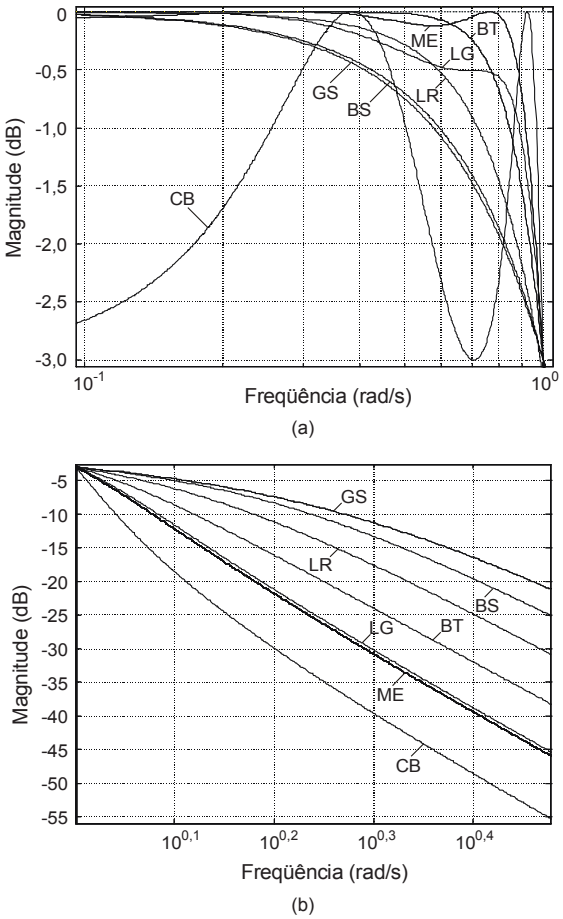


Fig. 1. Comparação da magnitude da resposta em freqüência da função ME com outras funções clássicas. (a) Detalhe na banda de passagem. (b) Detalhe na banda de rejeição.

É interessante discutir, neste ponto, primeiramente a razão da utilização das funções LG. Elas são monotônicas

e apresentam características intermediárias de magnitude e fase (ou atraso de grupo) entre as funções CB e BT. Suas CAA são melhores que as de um BT e piores que as de um CB. Por outro lado, suas CAF são melhores que as de um CB e piores que as de um BT. Sendo assim, as funções LG têm sido uma opção de uso entre as funções CB e BT.

Considere agora a comparação entre as funções LG e ME. A Fig. 2 mostra a comparação, para $n = 5$, entre as respostas passa-baixa normalizadas ME e LG, com a característica ideal (*brick wall filter*). Como pode se observar na Fig. 1 (b), as funções LG e ME apresentam características de atenuação semelhantes a partir de $\omega_p = 1$ rad/s. A vantagem principal da função ME é que esta apresenta um menor erro na banda de passagem em relação à resposta do *brick wall filter* do que a função LG (e também em relação a todas as outras funções de aproximação). Então qual a razão da pouca popularidade da função ME? Em primeiro lugar, para sua determinação são necessárias ferramentas computacionais que não eram facilmente disponíveis no passado. Em segundo lugar, porque a referência [3] faz apenas uma menção a este tipo de aproximação e a referência [4] apresenta as funções características básicas até a ordem nove e não apresenta uma forma sistemática de cálculo da constante de ganho. Assim, se o projetista procura uma função alternativa à função CB (que apresenta o corte mais abrupto dentre todas as funções polinomiais, porém com um *ripple* igual à atenuação em $\omega_p = 1$ rad/s) e à função BT (que apresenta magnitude da resposta em freqüência plana e melhores características de fase), a melhor opção é sem dúvida a função ME, ao invés da função LG.

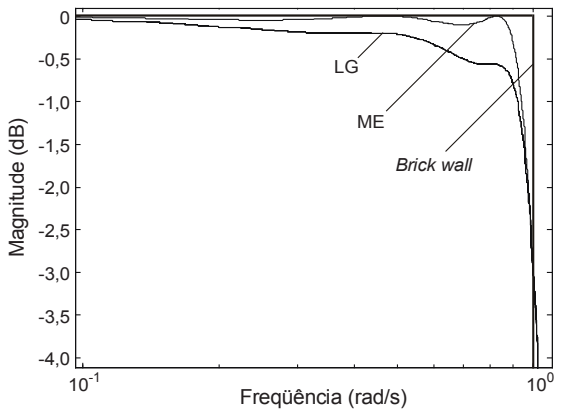


Fig. 2. Comparação da magnitude da resposta em freqüência das funções ME, LG e *brick wall filter*.

DETERMINAÇÃO DAS FUNÇÕES ME

A função atenuação $H(\omega)$ de um filtro é encontrada a partir de sua função característica $K(\omega)$ usando (1).

$$\left|H(\overline{\omega})\right|^2 = 1 + \left|K(\overline{\omega})\right|^2 \tag{1}$$

Usando continuação analítica (da teoria de variáveis complexas), substituindo $\overline{\omega}^2$ por $-\overline{s}^2$, é obtida (2), a chamada equação de Feldtkeller. Após encontrar as raízes de $H(\overline{s})H(-\overline{s})$, para que se obtenha uma rede estável, escolhe-se aquelas localizadas no semiplano lateral esquerdo (são os pólos do filtro).

$$H(\overline{s})H(-\overline{s}) = 1 + K(\overline{s})K(-\overline{s}) . \tag{2}$$

A partir de (1), obtém-se a atenuação em dB $A(\overline{\omega}) = \left|H(\overline{\omega})\right|_{dB}$:

$$A(\overline{\omega}) = 10\ \log\left(1 + \left|K(\overline{\omega})\right|^2\right) . \tag{3}$$

Definindo ε como a máxima distorção na banda passante normalizada $\overline{\omega}_p = 1$ rad/s (em alguns casos ε é o *ripple*) da função característica $K(\overline{\omega})$, tem-se que:

$$K(1) = \varepsilon .$$

Então $A(1) = A_p = 10\log(1+\varepsilon^2)$ dB

e
$$\varepsilon = \left[10^{A_p/10} - 1\right]^{1/2} . \tag{4}$$

A função característica de um filtro ME é dada por (5), onde $M_n(\overline{\omega})$ é um polinômio de grau n em ω .

$$K(\overline{\omega}) = \varepsilon\, M_n(\overline{\omega}) . \tag{5}$$

Conseqüentemente, $\left|H(\overline{\omega})\right|^2 = 1 + \varepsilon^2 M_n^2(\overline{\omega})$. (6)

Usando (2), obtém-se:

$$H(\overline{s})H(-\overline{s}) = 1 + \varepsilon^2 M_n^2(\overline{\omega})\Big|_{\overline{\omega}^2 = -\overline{s}^2} . \tag{7}$$

A partir de (7), obtêm-se numericamente as raízes $\overline{s}_k$ do semiplano lateral esquerdo. A função $H(\overline{s})$ é dada por:

$$H(\overline{s}) = \prod_{k=1}^{n} (\overline{s} - \overline{s}_k) = \overline{s}^{\,n} + b_{n-1}\,\overline{s}^{\,n-1} + \cdots + b_1\overline{s} + b_0 . \tag{8}$$

A função ganho $T(\overline{s})$ é:

$$T(\overline{s}) = \left(\frac{1}{H(0)}\right)\frac{b_0}{\overline{s}^n + b_{n-1}\overline{s}^{n-1} + \cdots + b_1\overline{s} + b_0} , \tag{9}$$

onde por (6), $H(0) = (1 + \varepsilon^2 M_n^2(0))^{1/2}$. (10)

Note na Tabela 1 que $M_n^2(0) = 0$ para n ímpar e, neste caso, $H(0) = 1$. Porém, para n par $M_n^2(0) \neq 0$. A informação sobre a constante $H(0)$ foi inserida em (9) porque ela é perdida no cálculo das raízes de $H(\overline{s})H(-\overline{s})$.

A magnitude da resposta em freqüência da função ME é obtida de forma que $M_n(\omega)$ seja o mais próximo de zero na banda de passagem normalizada, usando o critério do mínimo erro médio quadrático. Em adição, é estabelecida a condição $M_n(1) = 1$ de tal forma que $K(1) = \varepsilon\, M_n(1) = \varepsilon$.

O polinômio $M_n(\omega)$ tem a forma apresentada em (11), no caso de funções pares e a forma apresentada em (12), no caso de funções ímpares. Essa diferença é necessária para que a função ao quadrado tenha apenas coeficientes em $\overline{\omega}^2$. Assim, após a substituição de $\overline{\omega}^2$ por $-\overline{s}^2$, os coeficientes resultantes são reais.

$$M_n(\omega) = a_0 + a_2\omega^2 + ... + a_n\omega^n \quad \text{para } n \text{ par} \quad (11)$$

e $$M_n(\omega) = a_1\omega + a_3\omega^3 + ... + a_n\omega^n \quad \text{para } n \text{ ímpar} \quad (12)$$

Os coeficientes são escolhidos de forma que a seguinte integral (erro médio quadrático) seja minimizada:

$$E = \int_0^1 (M_n(\omega) - 0)^2 d\omega = \int_0^1 M_n^2(\omega) d\omega . \quad (13)$$

Por simplicidade, mas sem perda de generalidade, é vista a seguir a determinação dos polinômios $M_4(\omega)$ e $M_4^2(\omega)$. Para $n = 4$ tem-se:

$$M_4(\omega) = a_0 + a_2\omega^2 + a_4\omega^4.$$

Para que a condição $M_4(1) = 1$ seja satisfeita, então $a_0 + a_2 + a_4 = 1$. Isolando a_0 obtém-se $a_0 = 1 - a_2 - a_4$. Assim, pode-se escrever (13) como:

$$E = \int_0^1 M_n^2(\omega) d\omega = \int_0^1 (1 - a_2 - a_4 + a_2\omega^2 + a_4\omega^4)^2 d\omega.$$

Um sistema de equações lineares é formado, em função dos coeficientes a_2 e a_4, baseando-se na condição de minimização do erro médio quadrático, ou seja, $\frac{\partial E}{\partial a_k} = 0$. Assim:

$$\frac{\partial E}{\partial a_2} = \int_0^1 [(2a_4 + 2a_2 - 2) + \omega^2(2 - 4a_2 - 2a_4) + \\ + \omega^4(2a_2 - 2a_4) + 2a_4\omega^6] d\omega = 0$$

e

$$\frac{\partial E}{\partial a_4} = \int_0^1 [(2a_2 + 2a_4 - 2) - 2a_2\omega^2 + \\ + (2 - 2a_2 - 4a_4)\omega^4 + 2a_2\omega^6 + 2a_4\omega^8] d\omega = 0.$$

Resolvendo as integrais, obtém-se as duas equações que compõem o sistema linear:

$$(16/15)a_2 + (128/105)a_4 - 4/3 = 0$$

$$(128/105)a_2 + (64/45)a_4 - 8/5 = 0 .$$

A solução do sistema é $a_2 = -7/4$ e $a_4 = 21/8$. Sabendo-se que o somatório dos coeficientes é igual a 1, encontra-se $a_0 = 1/8$. Assim, o polinômio $M_4(\omega)$ é dado por:

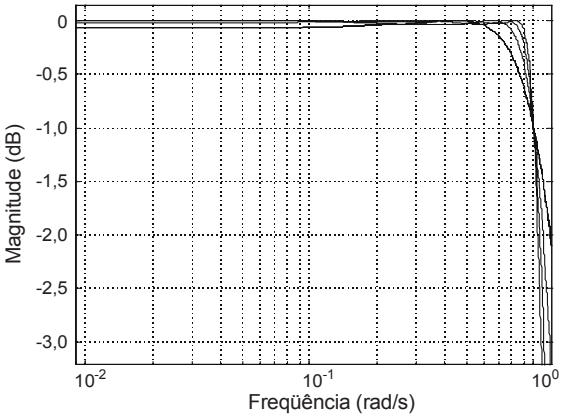
$$M_4(\overline{\omega}) = (21/8)\overline{\omega}^4 - (7/4)\overline{\omega}^2 + 1/8 .$$

Conseqüentemente, $M_4^2(\overline{\omega}) = M_4(\overline{\omega}) \times M_4(\overline{\omega})$ é:

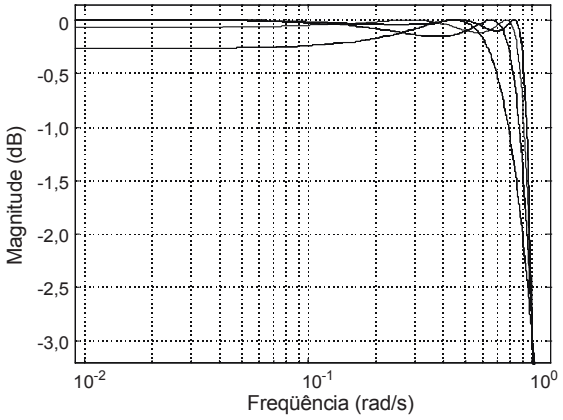
$$M_4^2(\overline{\omega}) = \frac{441}{64}\overline{\omega}^8 - \frac{147}{16}\overline{\omega}^6 + \frac{119}{32}\overline{\omega}^4 - \frac{7}{16}\overline{\omega}^2 + \frac{1}{64} .$$

O processo descrito para ordem quatro pode ser estendido para outras ordens. Quando a ordem aumenta, é conveniente utilizar recursos computacionais para resolver as integrais e o sistema de equações lineares. Isso foi feito para ordens *n* de 1 a 15 e os polinômios $M_n(\overline{\omega})$ e $M_n^2(\overline{\omega})$ encontrados são apresentados na Tabela 1. É importante observar que os coeficientes de $M_n^2(\overline{\omega})$ estão aproximados para $n \geq 6$. Assim, se for necessário operar com maior exatidão, pode-se optar em trabalhar com o produto $M_n(\overline{\omega}) \times M_n(\overline{\omega})$.

A Fig. 3 (b) apresenta a magnitude da resposta em freqüência das funções ME passa-baixa normalizadas de ordem dois a cinco, com máxima atenuação na banda de passagem A_p de 1 dB e 3 dB, respectivamente. Quanto menor é o valor de A_p, mais a resposta da função ME se aproxima da resposta do *brick wall filter* na banda de passagem. No entanto, o preço que se paga é que as funções com menores A_p são menos seletivas na banda de rejeição.



(a)



(b)

Fig. 3. Magnitude da resposta em freqüência das funções ME com $n = 2$ a 5 com (a) $A_p = 1$dB e (b) $A_p = 3$ dB.

EXEMPLO DE DETERMINAÇÃO DE UMA FUNÇÃO DE TRANSFERÊNCIA ME

Como exemplo, é mostrada a determinação da FT de um filtro passa-faixa de ordem 4 para um *crossover* de três vias, apresentando máxima atenuação na banda passante $A_p = 1\,\text{dB}$, freqüência de corte inferior $f_i = 300\,\text{Hz}$ e freqüência de corte superior $f_s = 2500\,\text{Hz}$.

O primeiro passo é determinar a FT de um filtro passa-baixa normalizado de ordem $n = 2$ e $A_p = 1\,\text{dB}$ no limite da banda de passagem normalizada $\overline{\omega}_p = 1\,\text{rad/s}$. Uma conveniente desnormalização e uma transformação em freqüência serão efetuadas.

Tabela 1 – Polinômios $M_n(\overline{\omega})$ e $M_n^2(\overline{\omega})$ para ordens de 2 a 15

n	Polinômios $M_n(\overline{\omega})$
2	$(5/4)\overline{\omega}-1/4$
3	$(7/4)\overline{\omega}^3-(3/4)\overline{\omega}$
4	$(21/8)\overline{\omega}^4-(7/4)\overline{\omega}^2+1/8$
5	$(31/8)\overline{\omega}^5-(15/4)\overline{\omega}^3+(5/8)\overline{\omega}$
6	$(429/64)\overline{\omega}^6-(495/64)\overline{\omega}^4+(135/64)\overline{\omega}^2-5/64$
7	$(715/64)\overline{\omega}^7-(1001/64)\overline{\omega}^5+(385/64)\overline{\omega}^3-(35/64)\overline{\omega}$
8	$(2431/128)\overline{\omega}^8-(1001/32)\overline{\omega}^6+(1001/64)\overline{\omega}^4-(77/32)\overline{\omega}^2+7/128$
9	$(4199/128)\overline{\omega}^9-(1989/32)\overline{\omega}^7+(2457/64)\overline{\omega}^5-(273/32)\overline{\omega}^3+(63/128)\overline{\omega}$
10	$(14697/256)\overline{\omega}^{10}-(31492/256)\overline{\omega}^8+(23205/256)\overline{\omega}^6-(6825/256)\overline{\omega}^4+(682/256)\overline{\omega}^2-11/256$
11	$(26001/256)\overline{\omega}^{11}-(31089/128)\overline{\omega}^9+(53295/256)\overline{\omega}^7-(19635/256)\overline{\omega}^5+(361/32)\overline{\omega}^3-(115/256)\overline{\omega}$
12	$(92863/512)\overline{\omega}^{12}-(245157/512)\overline{\omega}^{10}+(239827/512)\overline{\omega}^8-(53295/256)\overline{\omega}^6+(10519/256)\overline{\omega}^4-(1485/512)\overline{\omega}^2+17/512$
13	$(167153/512)\overline{\omega}^{13}-(482885/512)\overline{\omega}^{11}+(265587/256)\overline{\omega}^9-(138567/256)\overline{\omega}^7+(17321/128)\overline{\omega}^5-(7293/512)\overline{\omega}^3+(215/512)\overline{\omega}$
14	$(605927/1024)\overline{\omega}^{14}-(1901357/1024)\overline{\omega}^{12}+(2323883/1024)\overline{\omega}^{10}-(697165/512)\overline{\omega}^8+$ $(424361/1024)\overline{\omega}^6-(60623/1024)\overline{\omega}^4+(3191/1024)\overline{\omega}^2-27/1024$
15	$(1104927/1024)\overline{\omega}^{15}-(1871247/512)\overline{\omega}^{13}+(5033009/1024)\overline{\omega}^{11}-(3417475/1024)\overline{\omega}^9+$ $(1230291/1024)\overline{\omega}^7-(112331/512)\overline{\omega}^5+(8915/512)\overline{\omega}^3-(403/1024)\overline{\omega}$
n	Polinômios $M_n^2(\overline{\omega})$
2	$(25/16)\overline{\omega}^4-(5/8)\overline{\omega}^2+1/16$
3	$(49/16)\overline{\omega}^6-(21/8)\overline{\omega}^4+(9/16)\overline{\omega}^2$
4	$(441/64)\overline{\omega}^8-(147/16)\overline{\omega}^6+(119/32)\overline{\omega}^4-(7/16)\overline{\omega}^2+(1/64)$
5	$(1089/64)\overline{\omega}^{10}-(495/16)\overline{\omega}^8+(615/32)\overline{\omega}^6-(75/16)\overline{\omega}^4+(25/64)\overline{\omega}^2$
6	$(14513/323)\overline{\omega}^{12}-(4666/45)\overline{\omega}^{10}+(14184/161)\overline{\omega}^8-(14184/161)\overline{\omega}^6+(2382/421)\overline{\omega}^4-(675/2048)\overline{\omega}^2+(25/4096)$
7	$(4618/37)\overline{\omega}^{14}-(52770/151)\overline{\omega}^{12}+(18573/49)\overline{\omega}^{10}-(8617/43)\overline{\omega}^8+(5969/112)\overline{\omega}^6-(1612/245)\overline{\omega}^4+(419/1401)\overline{\omega}^2$
8	$(23085/64)\overline{\omega}^{16}-(185359/156)\overline{\omega}^{14}+(20444/13)\overline{\omega}^{12}-(12839/12)\overline{\omega}^{10}+(38533/97)\overline{\omega}^8-$ $(14558/185)\overline{\omega}^6+(5123/683)\overline{\omega}^4-(539/2048)\overline{\omega}^2+49/16384$
9	$(36589/34)\overline{\omega}^{18}-(126419/31)\overline{\omega}^{16}+(70204/11)\overline{\omega}^{14}-(31993/6)\overline{\omega}^{12}+(133467/52)\overline{\omega}^{10}-$ $(28649/40)\overline{\omega}^8+(9841/89)\overline{\omega}^6-(8188/975)\overline{\omega}^4+(961/3967)\overline{\omega}^2$
10	$(32957/10)\overline{\omega}^{20}-(98871/7)\overline{\omega}^{18}+(127704/5)\overline{\omega}^{16}-(101451/4)\overline{\omega}^{14}+(120655/8)\overline{\omega}^{12}-$ $(32963/6)\overline{\omega}^{10}+(27696/23)\overline{\omega}^8-(36350/243)\overline{\omega}^6+(2240/241)\overline{\omega}^4-(255/1166)\overline{\omega}^2+65/38638$
11	$(82529/8)\overline{\omega}^{22}-(49338)\overline{\omega}^{20}+(607685/6)\overline{\omega}^{18}-(116708)\overline{\omega}^{16}+(414446/5)\overline{\omega}^{14}-$ $(150023/4)\overline{\omega}^{12}+(248360/23)\overline{\omega}^{10}-(47952/25)\overline{\omega}^8+(18661/95)\overline{\omega}^6-(8987/883)\overline{\omega}^4+(332/1631)\overline{\omega}^2$
12	$(197375/6)\overline{\omega}^{24}-173690\overline{\omega}^{22}+399185\overline{\omega}^{20}-524091\overline{\omega}^{18}+433681\overline{\omega}^{16}-(470865/2)\overline{\omega}^{14}+$ $(338491/4)\overline{\omega}^{12}-(258129/13)\overline{\omega}^{10}+(26335/9)\overline{\omega}^8-(16113/64)\overline{\omega}^6+(9313/842)\overline{\omega}^4-(375/2006)\overline{\omega}^2+25/24072$
13	$106582{,}4778\overline{\omega}^{26}-615809{,}870049\overline{\omega}^{24}+1566893{,}99802\overline{\omega}^{22}-2310328{,}22266\overline{\omega}^{20}+$ $2185649{,}02877\overline{\omega}^{18}-1387643{,}67687\overline{\omega}^{16}+6{,}00896{,}573196\overline{\omega}^{14}-176835{,}9303\overline{\omega}^{12}+$ $34600{,}64944588\overline{\omega}^{10}-4308{,}54616\overline{\omega}^8+316{,}278316867\overline{\omega}^6-11{,}935030717\overline{\omega}^4+0{,}17551515583\overline{\omega}^2$
14	$350139{,}380306\overline{\omega}^{28}-2197426{,}20618\overline{\omega}^{26}+6133428{,}33053\overline{\omega}^{24}-10039120{,}059\overline{\omega}^{22}+$ $10697288{,}69914554\overline{\omega}^{20}-7789321{,}825211772\overline{\omega}^{18}+3958585{,}655396469\overline{\omega}^{16}-1408885{,}23005206\overline{\omega}^{14}+$ $347204{,}3143101698\overline{\omega}^{12}-57672{,}81216922631\overline{\omega}^{10}+6158{,}744061257684\overline{\omega}^8-390{,}6375583054145\overline{\omega}^6+$ $12{,}80912169951406\overline{\omega}^4-0{,}1631734536301306\overline{\omega}^2+6{,}85602425722\times10^{-4}$
15	$1164306{,}090639831\overline{\omega}^{30}-7887234{,}559682801\overline{\omega}^{28}+23964383{,}15839925\overline{\omega}^{26}-43129096{,}84118479\overline{\omega}^{24}+$ $51145274{,}41351520\overline{\omega}^{22}-42062332{,}44050433\overline{\omega}^{20}+24589788{,}81250985\overline{\omega}^{18}-10304237{,}83456609\overline{\omega}^{16}+$ $3081950{,}074008407\overline{\omega}^{14}-647274{,}4767984994\overline{\omega}^{12}+92596{,}79890341107\overline{\omega}^{10}-8584{,}204065794445\overline{\omega}^8+$ $475{,}5323885284167\overline{\omega}^6-13{,}67782331091915\overline{\omega}^4+0{,}15426114153\overline{\omega}^2$

Segundo a Tabela 1, para $n = 2$, a função $M_n^2(\overline{\omega})$ é:

$$M_4^2(\overline{\omega}) = (25/16)\overline{\omega}^4 - (5/8)\overline{\omega}^2 + (1/16)\,. \qquad (14)$$

Assim, a função atenuação ao quadrado é:

$$\left|H(\overline{\omega})\right|^2 = 1 + \varepsilon^2\left[(25/16)\overline{\omega}^4 - (5/8)\overline{\omega}^2 + (1/16)\right]. \qquad (15)$$

Calculando-se ε usando (4) e substituindo $\overline{\omega}^2$ por $-\overline{s}^2$ em (15), obtém-se $H(\overline{s})H(-\overline{s})$ dada por:

$$H(\overline{s})H(-\overline{s}) = 1 + 0{,}2589254118\left[\frac{25}{16}\overline{s}^4 + \frac{5}{8}\overline{s}^2 + \frac{1}{16}\right]. \qquad (16)$$

As raízes de $H(\overline{s})H(-\overline{s})$ são

$$\overline{s}_{1,2,3,4} = \pm a \pm b\,j = \pm 0{,}832121237 \pm 0{,}944682885\,j\,.$$

Escolhendo as raízes localizadas no semiplano lateral esquerdo, forma-se o polinômio $H(\overline{s})$:

$$H(\overline{s}) = \overline{s}^2 + b_1\overline{s} + b_0\,,$$

onde $\qquad b_1 = 2a = 1{,}664242474$

e $\qquad b_0 = a^2 + b^2 = 1{,}58485150628\,.$

Usando (10) calcula-se $H(0)$ como:

$$H(0) = \left[1 + 0{,}2589254\left(\frac{1}{16}\right)\right]^{1/2} = 1{,}0080589458\,.$$

A função de transferência do filtro ME é

$$T(\overline{s}) = \frac{1}{H(0)}\frac{b_0}{H(\overline{s})} \qquad (17)$$

ou $\quad T(\overline{s}) = \dfrac{1{,}5721814}{\overline{s}^2 + 1{,}664242474\overline{s} + 1{,}58485150628}\,.$

A equação (18) permite transformar uma FT passa-baixa normalizada em uma passa-faixa com simetria geométrica [1]. Assim, a freqüência central do filtro é $\omega_0 = (\omega_s\omega_i)^{1/2}$. Em (18), B representa a banda passante dada por $B = \omega_s - \omega_i = 2\pi(f_s - f_i)$.

$$\overline{s} = \frac{s^2 + \omega_0^2}{Bs}\,. \qquad (18)$$

Substituindo (18) em (17), obtém-se

$$T(s) = \frac{b_0 / H(0)}{\left(\dfrac{s^2 + \omega_0^2}{Bs}\right)^2 + b_1\left(\dfrac{s^2 + \omega_0^2}{Bs}\right) + b_0}\,,$$

$$T(s) = \frac{(Bs)^2\left[b_0 / H(0)\right]}{(s^2 + \omega_0^2)^2 + b_1 Bs(s^2 + \omega_0^2) + b_0(Bs)^2}\,. \qquad (19)$$

Colocando (19) em uma forma conveniente e substituindo as variáveis literais por valores numéricos, obtém-se

$$T(s) = \frac{3{,}00405411877 \times 10^8 \quad s^2}{s^4 + 23004{,}8365s^3 + 362043986s^2 + \\ + 681145907000s + 8.76681819 \times 10^{14}}\,.$$

A Fig. 4 mostra a magnitude da resposta em freqüência do filtro passa-faixa projetado.

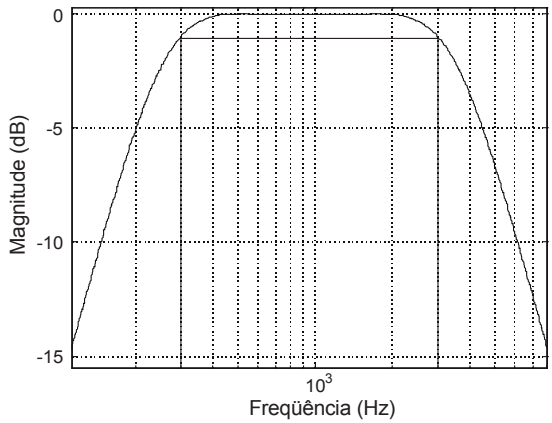


Fig. 4 Magnitude da resposta em freqüência do filtro ME passa-faixa.

CONCLUSÕES

Neste trabalho, foi discutida a função de aproximação Mínimo Erro Médio Quadrático que apresenta o menor erro da magnitude da resposta em freqüência na banda de passagem em relação à resposta ideal do *brick wall filter*, dentre todos os outros tipos de funções de aproximação clássicas conhecidas. Essa função apresenta características intermediárias de seletividade e de fase entre as aproximações Butterworth e Chebyshev e, portanto, é uma interessante opção de uso em lugar da aproximação Legendre. Foi mostrada a forma de obter essas funções e determinada uma simples equação para o cálculo do ganho.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] S. Noceti Filho, *Filtros Seletores de Sinais*, 2ª ed. Florianópolis: Edufsc, 2003.

[2] V. Dickason, *Caixas Acústicas e Alto-falantes*, 5ª ed. Rio de Janeiro: H. Sheldon, 1997.

[3] H. J. Blinchikoff and A. I. Zverev, *Filtering in the Time and Frequency Domain*, New York: Jonh Wiley and Sons, 1976.

[4] D. S. Humpherys, *The Analysis, Design, and Synthesis of Electrical Filters*, N.J.: Prentice-Hall, Englewood Cliffs, 1970.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

O Método FCC de Correção para Amplificadores Chaveados Operando no Esquema Sigma Delta – Resultados Fundamentais

Marcelo H. M. Barros
Grupo de Materiais e Dispositivos, Departamento de Física e Engenharia Física, Universidade Federal de São Carlos, 13565-905, São Carlos, São Paulo.
HotSound. Ind. Com. de Equipamentos Eletrônicos Ltda, 13.270-294, Valinhos, São Paulo.
marcelo@hotsound.com.br

RESUMO

Este artigo irá expor as bases e os resultados fundamentais do método FCC de correção para amplificadores chaveados. Centrado no tratamento matemático, via técnica variacional, este procedimento introduziu melhoras muito significativas no sistema amplificador chaveado, chegando a ter desempenho completamente similar a um amplificador linear de alto padrão, em termos da distorção, da resposta em frequências, do módulo da impedância de saída e do ruído residual de fundo, mas preservando a alta eficiência energética típica de um amplificador chaveado.

1. DESCRIÇÃO GERAL

O método FCC consiste em um procedimento sistemático para implementação de um conformador de ondas (wave-shaping) [5,6,7] em estrutura recorrente na malha de realimentação de amplificadores chaveados, a fim de se obter a modulação 1-bit sigma-delta [6,7] com o máximo de fidelidade ao sinal original. O procedimento consiste em postular um grupo de operadores, ALPHA, BETA e GAMMA-*i*, onde cada um deles representa uma etapa deste conformador de ondas, mas com vários parâmetros livres. Nestes operadores aplicaram-se técnicas variacionais [8,9] a fim de encontrar os melhores valores para os parâmetros livres que minimizam os erros introduzidos nas diversas partes do amplificador chaveado. O resultado surge na forma de equações de vínculos, que inter-relacionam os parâmetros livres e diminuem os graus de liberdade para apenas alguns poucos dados, que foram posteriormente identificados como dados de sistema. Partindo destes poucos dados de sistema, inerentes a um dado conversor acoplado a um módulo de potência chaveado classes AD ou BD (que daqui a diante chamaremos simplesmente *plataforma*) e por meio das equações de vínculo obtidas, puderam-se determinar os parâmetros livres de forma fechada e assim, estes operadores, inicialmente genéricos, se tornaram específicos para uma dada plataforma e puderam ser finalmente convertidos em circuito eletrônico, por meio dos métodos usuais. Esse conformador de ondas, assim obtido, foi inserido em uma plataforma classe BD [4]. A adição desse conformador caracterizou o sistema como um grande modulador sigma-delta [6], com o estágio de saída fazendo parte desse loop [7]. Este procedimento de otimização introduziu melhoras muito significativas no amplificador chaveado sigma-delta, chegando a ter um desempenho muito próximo, e até melhor em alguns aspectos, aos amplificadores lineares de potência compatível, mas preservando sua principal virtude – a alta eficiência energética, algo em redor de 95%, independente da potência de saída, contra os típicos 50-60% dos amplificadores lineares (mas somente na máxima potência).

2. DESCRIÇÕES FUNCIONAIS DE ALPHA, BETA E GAMMA-*i*, AS EQUAÇÕES DE VÍNCULO E O MÉTODO VARIACIONAL

Um amplificador chaveado (classe-AD) típico é descrito por Attwood [2,3], Vanderkooy [4] e citado por Duncan [1] e consiste na seguinte estrutura básica:

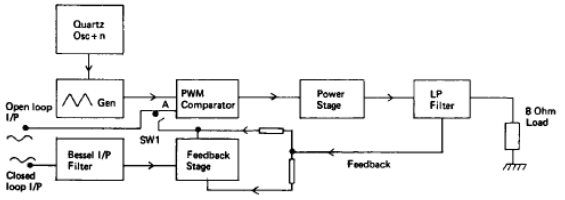


Fig.1 – Amplificador chaveado básico (após 1983), como proposto por Attwood

Esse modelo pode ser considerado padrão. Em [4], Vanderkooy cita a classe BD como uma variante da classe AD original. A alta eficiência energética destas plataformas é largamente discutida na literatura e não será considerada aqui. O sinal aplicado (da banda de áudio, 20-20kHz) é convertido no bloco *PWM Comparator*, onde emerge como um sinal binário, de apenas 2 estados e de frequência constante; no caso de Attwood e Vanderkooy seguindo o esquema PWM (*pulse width modulation*) [2,3,4]:

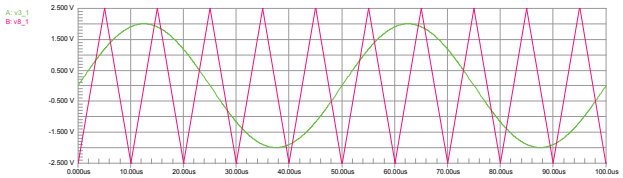


Fig. 2 – Amostragem PWM, com Fb=100kHz, Fs=20kHz e M=0,8, segundo [4]

Na figura 2, o sinal aplicado *Fs* é senoidal puro com F=20kHz e o relógio (clock), chamado *bias* em [4], tem frequência de 100kHz e é do tipo rampa (triangular). *M* é o índice de modulação [4]. O sinal PWM aparece na figura 3:

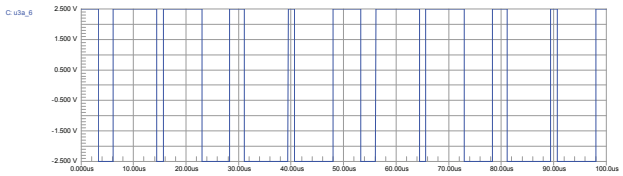


Fig. 3 – Sinal PWM, para a amostragem da figura 2, segundo [4]

A conversão 1-bit sigma-delta é descrita (neste artigo) por Klugbauer-Heilmeier e por Esslinger [6,7]. Para o mesmo sinal aplicado (Fs = 20kHz) tem a forma como segue:

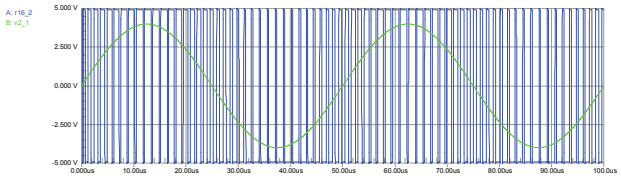


Fig. 4 – Conversão 1-bit sigma-delta para Fs = 20kHz e M = 0,8 segundo [6,7]

Onde se percebe claramente a diferença fundamental entre o esquema PWM e o SDM (*sigma-delta modulation*): a densidade de amostras não é mais uma constante do sistema.

O sinal portador (*carrier*, ou *bias* em [4]) pode ser removido por um filtro passa-baixas (bloco *LP Filter*, na fig. 1) a fim de recuperar o sinal original. Procedimentos deste tipo já foram extensivamente tratados na literatura.

Attwood [2,3], Vanderkooy [4] e vários outros pesquisadores dedicaram grande parte de seu trabalho à procura de um método de correção que elevasse o padrão de qualidade sonoro dos amplificadores chaveados ao mesmo nível dos amplificadores lineares. Uma leitura destes trabalhos e outros, pode esclarecer o quão custoso é esse objetivo, especialmente em situações de grande potência, como é o caso do áudio profissional.

O método NDFL, proposto por Cherry para amplificadores lineares [5], forneceu a inspiração básica para a procura de uma técnica recorrente de realimentação negativa aplicada a amplificadores chaveados e essa é a proposta do método FCC, consistindo em um conformador de ondas de estrutura recorrente, otimizado matematicamente para realimentar um conversor SDM operando em classe AD ou BD. Sua estrutura guarda certa semelhança com a proposta por Cherry [5] para amplificadores lineares.

O sistema inicia com um bloco alimentador ALPHA (α), que distribui o sinal para $n+1$ conformadores diferenciadores, iniciando com BETA (β) e se estendem através dos GAMMA-*i* (Γ_i), (i = 1,2,...,n). O índice i pode assumir qualquer valor inteiro positivo e será mostrado que os erros introduzidos pelas não-linearidades do conversor SDM + estágio de potência decrescem quando i aumenta, ocorrendo uma melhora muito significativa de todos os parâmetros do amplificador chaveado.

ALPHA, BETA e os GAMMA-*i* são funcionalmente descritos pelas seguintes funções transferência (não são funções realimentação), no domínio da variável complexa s (transf. de Laplace), como está definido em [9]:

$$\alpha(s) = \frac{1}{A}\frac{1}{Bs+1} \qquad (1)$$

$$\beta(s) = C\frac{Ds+1}{Es+1} \qquad (2)$$

$$\Gamma i(s) = F_i\frac{G_i s+1}{H_i s+1} \qquad (3)$$

onde i = 1,2,...n e os parâmetros A, B, C, D, E, F, G e H são inicialmente desconhecidos. A, C e F_i são adimensionais e B, D, E, G_i e H_i têm dimensões de inverso da frequência angular.

Através de um procedimento sistemático de otimização, realizado por métodos variacionais [8,9], foi possível determinar os melhores parâmetros A, B, C, D, E, F, G e H, de modo que os graus de liberdade, inicialmente 8, foram drasticamente reduzidos. O processo variacional forneceu algumas *equações de vínculo*, que permitiram diminuir os graus de liberdade. Os parâmetros restantes, três no total, foram identificados como *dados de sistema*, chamados A, A_0 e $\Delta\omega_L$ e são oriundos da plataforma adotada, seguindo a prescrição:

A = ganho desejado em malha fechada;
A_0 = ganho de malha aberta;
$\Delta\omega_L$ = largura de banda, definida como sendo o inverso da resolução máxima da plataforma adotada que por sua vez é definida como sendo o pulso de duração mais curta que a plataforma é capaz de produzir.

O índice i, que a princípio poderia assumir qualquer valor inteiro positivo, na prática depende da largura de banda da plataforma utilizada e da largura de banda do conformador de ondas. Como i afeta diretamente a função sensibilidade, existirá um valor para o qual o sistema se tornará instável. Também está diretamente relacionado com a complexidade do sistema. Assim, optou-se por escolher um i que fornecesse ao amplificador chaveado o melhor desempenho, a partir do qual um incremento não traz nenhuma melhora significativa (convergência). Com a plataforma classe BD utilizada no protótipo a convergência foi rapidamente atingida, com $i = 3$.

2.1 O Método Variacional

Um dos métodos mais interessantes e antigos usados na física-matemática é o do *cálculo das variações* [8,9]. A idéia central está em minimizar (ou maximizar) uma certa função estática, chamada *funcional*, por meio de pequenas variações em alguns de seus parâmetros. Daremos aqui apenas um exemplo de como essa técnica pode ser usada para tratar problemas em muitas dimensões, tal como foi realizado com o método FCC, mas apenas para ilustrar o seu uso. O caso da *corda distendida*, como uma corda de piano ou de violão é típico e de grande interesse.
Uma corda distendida pode ser considerada como um sistema com infinitos graus de liberdade, cada elemento dx sendo tratado como uma partícula de massa ρdx. Portanto a energia cinética de um sistema de partículas como esse

$E_{cin} = \frac{1}{2}\sum_{i=1}^{N} m_i v_i^2$ torna-se uma integral $E_{cin} = \frac{1}{2}\int_0^L (\rho dx)(\frac{\partial u}{\partial t})^2$

A energia potencial da corda deformada é mais facilmente calculada como sendo o trabalho efetuado contra a força de tensão T. O comprimento da corda deformada é um pouco maior do que o comprimento original L e é dado por

$$L' = \int ds = \int_0^L \sqrt{1+\left(\frac{\partial u}{\partial x}\right)^2}\; dx$$

Para deformações pequenas, temos que

$\sqrt{1+\left(\frac{\partial u}{\partial x}\right)^2} \cong 1+\frac{1}{2}\left(\frac{\partial u}{\partial x}\right)^2$, por conseguinte, a extensão ΔL da corda é aproximadamente $\Delta L = L' - L \cong \frac{1}{2}\int_0^L \left(\frac{\partial u}{\partial x}\right)^2 dx$, e a energia potencial (trabalho realizado contra T) é dada por $E_{pot} \cong T\Delta L \cong \frac{T}{2}\int_0^L \left(\frac{\partial u}{\partial x}\right)^2 dx$, esta análise nos permite escrever o Lagrangeano do sistema como sendo

$$\pounds = E_{cin} - E_{pot} = \int_0^L \left[\frac{\rho}{2}\left(\frac{\partial u}{\partial t}\right)^2 - \frac{T}{2}\left(\frac{\partial u}{\partial x}\right)^2\right] dx$$

Segundo o princípio de Hamilton, o movimento da corda deve ser tal que a integral

$J = \int_{t_0}^{t_1}\int_0^L \left[\frac{\rho}{2}\left(\frac{\partial u}{\partial t}\right)^2 - \frac{T}{2}\left(\frac{\partial u}{\partial x}\right)^2\right] dxdt$, onde t_0 e t_1 são dois instantes arbitrários no tempo, seja *estacionária*. A equação de Euler-Lagrange para J toma então a forma

$\frac{D}{Dt}\frac{\partial \pounds}{\partial u_t} + \frac{D}{Dx}\frac{\partial \pounds}{\partial u_x} = 0$, Em que a quantidade

$\pounds = \frac{\rho}{2}\left(\frac{\partial u}{\partial t}\right)^2 - \frac{T}{2}\left(\frac{\partial u}{\partial x}\right)^2$, é geralmente chamada de *densidade Lagrangeana*. Procedendo com as operações necessárias, reduz-se a equação de Euler-Lagrange à forma familiar:

$$-T\frac{\partial^2 u}{\partial t^2} + \rho\frac{\partial^2 u}{\partial x^2} = 0$$

que pode agora ser resolvida pelos métodos usuais de EDP. Demonstrou-se assim, como o método variacional pode reduzir um problema de muitas dimensões para formas mais brandas. Estes métodos são fartamente descritos na literatura usual de física-matemática.

3. RESULTADOS

Um protótipo com potência na faixa de 2kWavg (@ 2 ohms) realizado seguindo as definições encontradas em [4] para a classe BD foi implementado para análise, inicialmente em malha aberta. Numa segunda etapa, para comparação, foi aplicado o método de realimentação proposto em [4] e finalmente, em uma terceira etapa foi aplicado o método FCC. O conformador de ondas FCC foi construído usando-se os métodos usuais de análise, após a obtenção dos parâmetros a partir das equações de vínculo e para os seguintes dados de sistema:

A = 24dB
A_0 = 26dB
$\Delta\omega_L = \frac{2\pi}{8\times10^{-7}\,s}$

Nas medições foi empregado o analisador Audio Precision System One + DSP com software APWin 2.24, interfaceado pelo filtro auxiliar Audio Precision AUX-0025, conforme prescrito em [11]. As cargas são puramente resistivas. Todas as medições foram executadas em conformidade com as referências encontradas em [1], [10] e [11].
A análise para malha aberta, com Fb=192kHz forneceu THD+N=0,6%, em regime permanente senoidal de 1kHz, carga fixa resistiva de 8 ohms e potência média na carga, conforme definida em [1,10], de aprox. 625 Wavg.
Em [4] está definida uma técnica de realimentação para amplificadores chaveados. Ela foi implementada no mesmo protótipo classe BD, a fim de se fazer uma comparação direta com o método FCC. A medição forneceu THD+N = 0,9%,

nas mesmas condições. E como citado por Vanderkooy [4], a realimentação produzida por um integrador introduz distorção, apesar de alguma melhora em outras figuras de mérito.

Na próxima etapa, foi introduzido o conformador FCC, mantendo-se a mesma plataforma utilizada para as medidas anteriores. Foram executadas medidas sucessivas nas mesmas condições e a cada uma incrementava-se o índice *i* de uma unidade, a fim de atestar a diminuição da THD+N com o aumento de *i*, como foi antecipado em 2.

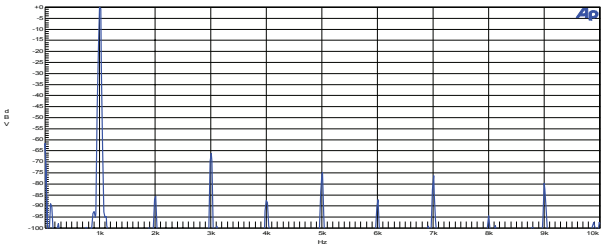


Fig. 5 – Análise espectral da tensão (normalizada) na carga, para i = 1. THD+N=0,057%

Com *i* = 1 já foi possível obter um valor bem superior aos registrados anteriormente. Aumentando *i* ainda mais, vem:

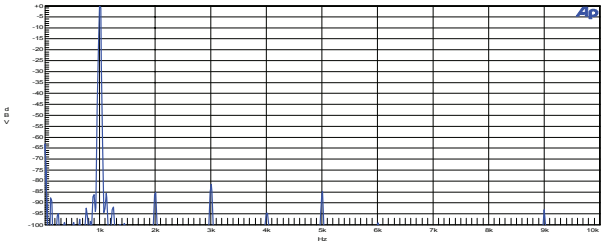


Fig. 6 – Análise espectral da tensão na carga, para i = 2. THD+N=0,013%

Já próximo da região de convergência. Incrementando *i* de mais uma unidade:

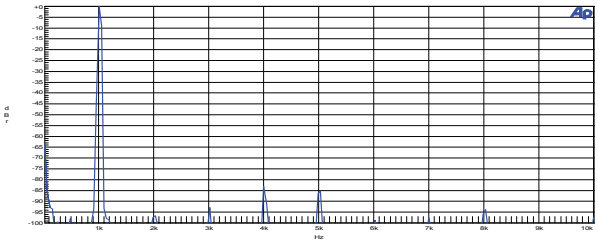


Fig. 7 – Análise espectral da tensão na carga, para i = 3. THD+N=0,009%

A convergência foi visivelmente alcançada para *i* = 3 e a THD+N alcançou um valor mais que dez vezes melhor que o nível de referência de 0,1%.

O método FCC provê o melhoramento de todas as principais figuras de mérito do amplificador chaveado, tais como: resposta em frequências, ruído residual de fundo, módulo da impedância de saída e a já (parcialmente) analisada, distorção harmônica+ruído. Na próxima sub-seção serão apresentados os resultados para todas essas figuras de mérito, mantendo fixo *i* = 3 e fazendo-se imediata referência aos valores obtidos com o método de [4] e os obtidos com um amplificador linear de alto padrão (de potência compatível, que chamaremos *amplificador linear de referência*) e ter-se-á uma exata idéia da posição em que o método FCC colocou a plataforma chaveada classe BD.

3.1 Resposta em Frequências

Inicialmente a magnitude da resposta em frequências, para o método FCC:

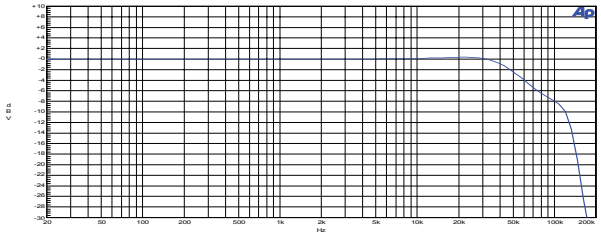


Fig. 8 – Magnitude normalizada da resposta em frequências para uma carga resistiva de 2 ohms, método FCC, exibindo a atuação do filtro AUX-0025

Percebe-se que, na banda de áudio, a resposta é perfeitamente plana, pois o que se vê é quase que totalmente a "marca" do filtro AUX-0025 [11]. Pode-se fazer a mesma medida para o método proposto em [4] e nas mesmas condições.

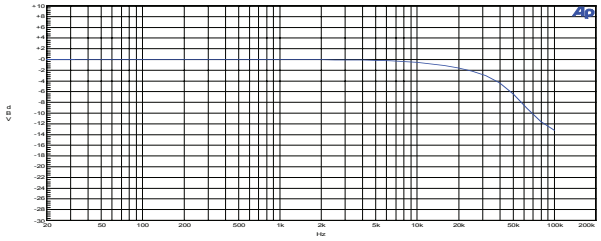


Fig. 9 – A mesma resposta em magnitude, obtida para o método proposto em [4]

Onde fica evidente a superioridade do método FCC em altas frequências.

O próximo passo será examinar a fase da resposta, em relação à entrada, conforme definido em [1,10]. Somente para o método FCC encontra-se:

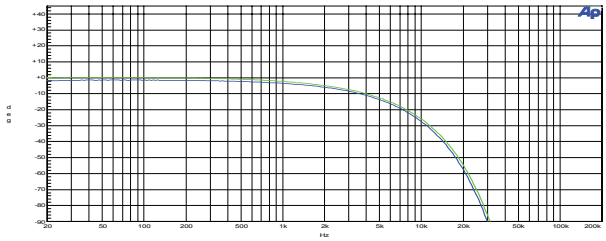


Fig. 10 – Fase da resposta em frequências para o método FCC (2Ω). Acima a resposta do filtro AUX-0025 e abaixo a resposta do protótipo interfaceado pelo filtro AUX-0025

Este gráfico mostra o pouco atraso introduzido pelo amplificador chaveado assistido pelo método FCC, com carga resistiva de 2 ohms. Para o método proposto em [4] o atraso introduzido chegou a -90deg em 30kHz, mostrando a sua inabilidade de reproduzir as frequências mais altas do espectro de áudio. No caso do amplificador linear usado como referência, os resultados são bastante compatíveis com os obtidos pelo protótipo assistido pelo método FCC.

3.2 Ruído Residual de Fundo

Agora a análise espectral por FFT do ruído residual de fundo presente na saída do amplificador quando sua entrada é desconectada [1,10]. Inicialmente para o método FCC:

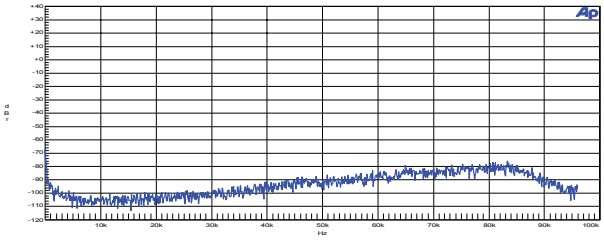


Fig. 11 – Análise espectral do ruído residual de fundo para o método FCC. dBr=dBu

Nota-se que dentro da banda de áudio o range dinâmico, conforme definido em [1,10] é extremamente grande, com SNR(22-22kHz) = 109,8dBr. O amplificador linear de referência possui SNR(22-22kHz) = 100dBr.

Para o protótipo assistido pelo método proposto em [4];

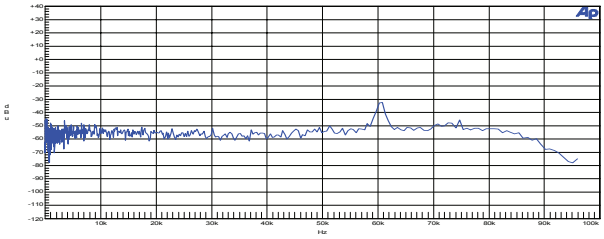


Fig. 12 – Análise espectral do ruído residual de fundo para o método proposto em [4]

o quadro é bastante inferior, com SNR(22-22kHz) = 82dBr. Verifica-se que, no parâmetro ruído residual de fundo, o amplificador chaveado assistido pelo método FCC obteve um resultado que supera o amplificador linear de referência.

3.3 Módulo da Impedância de Saída e Fator de Amortecimento

O fator de amortecimento (damping factor), como está definido em [1,10] pode ser facilmente obtido relacionando a resposta em frequências para uma carga conhecida com a resposta em frequências para uma carga infinitamente grande (amplificador com a saída em aberto) e calculando-o de acordo com as definições encontradas em [1,10]. A partir do fator de amortecimento pode-se calcular o módulo da impedância de saída, ainda conforme [1,10]. Um bom e suficiente valor para o fator de amortecimento se situam entre algumas centenas (200-600). Calculando-se o fator de amortecimento na frequência de 50Hz, obtém-se, para o método FCC, D ≈ 400 @ 8 ohms. A partir desse valor a impedância de saída (módulo) é obtida, |Z| = 0,02 ohms. Para o amplificador linear de referência, nas mesmas condições, é obtido D ≈ 570 com um respectivo |Z| = 0,014 ohms.

O mesmo procedimento para o protótipo assistido pelo método proposto em [4] obteve D ≈ 28 @ 8 ohms, com um respectivo |Z| = 0,286 ohms. O valor original (obtido com a plataforma sem nenhuma realimentação) foi D ≈ 10 @ 8 ohms, com |Z| = 0,8 ohms. Mais uma vez, os resultados para o amplificador chaveado + FCC concordam muito bem com os obtidos para o amplificador linear de referência.

3.4 Distorção Harmônica Total + Ruído

Agora serão feitas análises detalhadas da THD+N. O primeiro procedimento consiste em se fixar a frequência do sinal senoidal e variar sua amplitude [1,10]; para cada incremento na amplitude é feita uma medida da THD+N. Primeiro, para o método FCC:

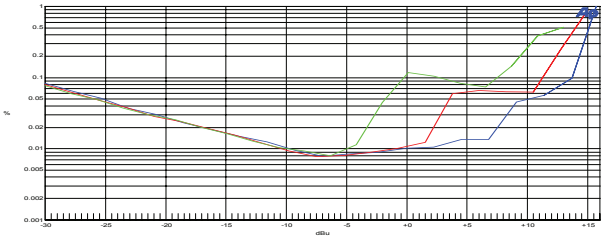


Fig. 13 – THD+N versus amplitude de entrada com sinal de teste de 1kHz para o método FCC. A amplitude de saída é 24dB maior. Abaixo carga de 8 ohms, ao centro carga de 4 ohms e acima carga de 2 ohms

Onde se fez a mesma medida para três cargas diferentes. Pode-se ver que a THD+N fica restrita a valores inferiores ao valor de referência na maior parte do intervalo, elevando-se somente nos limites de sua potência máxima. Por outro lado, atinge valores excepcionais (0,008%) em potências medianas. Para comparação, na próxima figura, a mesma medida, para o protótipo assistido pelo método de [4], onde confirmamos o resultado fornecido no início dessa seção.

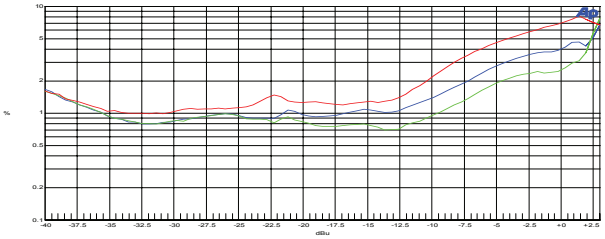


Fig. 14 – THD+N versus amplitude de entrada com sinal de teste de 1kHz, para o método proposto em [4]. A amplitude de saída é 32dB maior. Abaixo carga de 8 ohms, ao centro carga de 4 ohms e acima carga de 2 ohms.

A próxima figura exibe a THD+N versus amplitude para o amplificador linear de referência, que possui potência ligeiramente inferior ao do protótipo FCC apresentado, mas pôde fornecer uma comparação útil.

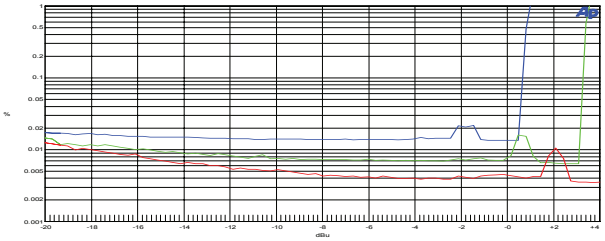


Fig. 15 – THD+N versus amplitude de entrada com sinal de teste de 1kHz para o amplificador de referência. A amplitude de saída é 32dB maior. Abaixo carga de 8 ohms, ao centro carga de 4 ohms e acima carga de 2 ohms.

Este excelente amplificador de tecnologia linear fornece uma

base segura do ponto onde o método FCC colocou a plataforma chaveada classe BD utilizada no protótipo.
Agora, em lugar de se fixar a frequência e varrer a amplitude será feito o contrário. Se Fixa a amplitude e varre-se a frequência [1,10]. O gráfico assim obtido é o de THD+N *versus* frequência.

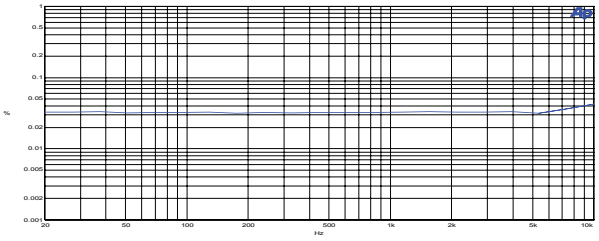


Fig. 16 – THD+N versus frequência @ -6dB do máximo sinal admissível, para carga de 2 ohms. Amplificador chaveado + FCC

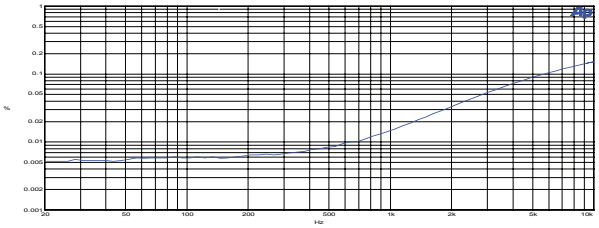


Fig. 17 – THD+N versus frequência @ -6dB do máximo sinal admissível, para carga de 2 ohms. Amplificador linear de referência

Onde se verifica, por comparação, a excepcional linearidade proporcionada pelo método FCC em relação às diferentes frequências do espectro de áudio. O amplificador linear de referência apresenta níveis excepcionalmente baixos de THD+N nas frequências mais baixas, contudo, nas mais altas o comportamento não é tão bom. Em um sistema de alta qualidade este amplificador provavelmente seria indicado para as frequências mais baixas (sistema de graves), já o amplificador chaveado FCC poderia ser utilizado em qualquer faixa de frequências.
Em toda a seção 3.4 a banda passante considerada pelo analisador foi de 22-22kHz.

4. FOTO DO PROTÓTIPO

O protótipo utilizado nas análises media cerca de 27x15cm e pesava cerca de 500g, com potência na faixa de 2kWavg.

Fig. 18 – Protótipo classe BD + FCC

5. CONCLUSÕES

O amplificador de áudio foi criado logo após a invenção da válvula eletrônica, na década de 1910. A sua enorme importância econômica logo se tornou óbvia e atualmente o número de pessoas, cujas atividades dependem, direta e indiretamente desse objeto é continuamente crescente. A importância econômica de se gerar tecnologias de alto rendimento energético vai desde uma simples redução do volume e peso transportado (menor custo com transporte) até uma redução no consumo de energia elétrica.
Atualmente, universidades e empresas do mundo todo buscam desenvolver seus próprios métodos em amplificação chaveada e sempre com os mesmos objetivos: alta eficiência energética e grande fidelidade sonora.
O método FCC visa a implementar a modulação sigma-delta de maneira otimizada para grande qualidade sonora, mas mantendo a alta eficiência energética. No futuro, com o aperfeiçoamento destes métodos de alta eficiência, os amplificadores lineares poderão estar no mais completo desuso.
Torna-se, portanto imperativo, que o meio acadêmico do Brasil, bem como às suas indústrias do setor de áudio profissional, dominem métodos próprios e competitivos de amplificação chaveada. Lembrando que os vários métodos recentemente desenvolvidos no mundo todo são proprietários e mantidos sob proteção.

7. REFERÊNCIAS BIBLIOGRÁFICAS

[1] Duncan, Ben; “High Performance Audio Power Amplifiers”, Butterworth-Heinemann, 1996;
[2] Attwood, Brian E.; “Very High Fidelity Quartz Controlled PWM (class D) Stereo Amplifiers for Consumer and Professional Use”, An Audio Engineering Society PrePrint, 1978;
[3] Attwood, Brian E.; “Design Parameters Important for the Optimization of Very-Fidelity PWM Audio Amplifiers”, An Audio Engineering Society PrePrint, 1982;
[4] Vanderkooy, J.; “New Concepts in Pulse-Width Modulation”, An Audio Engineering Society PrePrint, 1994;
[5] Cherry, Edward M; ”Nested Differentiating Feedback Loops in Simple Audio Power Amplifiers”, J. Audio Eng. Soc., Vol. 30, No. 5, 1982 May;
[6] Klugbauer-Heilmeier, Josef; “A Sigma Delta Modulated Switching Power Amp”, An Audio Engineering Society Preprint, preprint 3227,1992;
[7] R. Esslinger, G. Gruhler and R.W. Stewart; “Digital Audio Power Amplifiers Using Sigma Delta Modulation – Linearity Problems in the Class-D Power Stage”, Audio Engineering Society Convention Paper, 2001;
[8] Arfken, G.B. & Weber, H.J., “Mathematical Methods for Physicists”, Academic Press, 1995;
[9] Butkov, E., “Mathematical Physics”, Addison-Wesley Publishing Company, Inc., 1968;
[10] Metzler, B. “Audio Measurement Handbook”, Audio Precision, Inc., 1993;
[11] Hofer, B., “Measuring Switch-Mode Power Amplifiers”, Write paper, Audio Precision, Inc., 2003.

Sessão 3

Sonorização Espacial, Som 3D, Acústica de Salas e Ambientes II

(Spatial sound systems, 3D Sound, Environmental and Room Acoustics II)

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Parâmetros Acústicos em Salas de Música: análise de resultados e novas interpretações

Fábio Leão Figueiredo, Fernando Iazzetta
Departamento de Música - Universidade de São Paulo
São Paulo – SP - Brasil
fabioflf@hotmail.com, iazzetta@usp.br

RESUMO

Este artigo apresenta análises e conclusões sobre resultados de medições de parâmetros acústicos estabelecidos como critérios para avaliação da qualidade acústica de salas de música. As medições foram realizadas em seis importantes salas de concerto de São Paulo, durante o ano de 2005, dentro do projeto Acmus desenvolvido na Universidade de São Paulo. Primeiramente exibimos um quadro geral dos resultados para cada parâmetro. Em seguida, buscamos nas características arquitetônicas das salas as causas ou explicações para os resultados observados. Verificamos as limitações de alguns parâmetros, e sugerimos novas interpretações que podem enriquecer a compreensão sobre a avaliação da qualidade acústica das salas de música.

INTRODUÇÃO

Em 2003 iniciamos na Universidade de São Paulo, Brasil, um projeto de pesquisa em acústica de salas voltado para questões musicais. O núcleo de trabalho, intitulado AcMus [1], concentra-se no desenvolvimento de ferramentas computacionais para projeto, medição e simulação do comportamento acústico de salas destinadas à música.

O presente trabalho focaliza os resultados obtidos nas pesquisas de medições acústicas efetuadas com base na norma ISO 3382 [2]. Os resultados das medições foram processados de modo a levantarmos os parâmetros acústicos reconhecidos como critérios para avaliação da acústica de salas.

Os parâmetros acústicos subjetivos são critérios que definem a qualidade acústica de uma sala de música. A apreciação musical dentro da sala é afetada por diversas impressões acústicas que ocorrem ao mesmo tempo. Cada uma dessas impressões é associada a um parâmetro acústico de natureza subjetiva que está correlacionado a uma grandeza física mensurável, constituindo um conjunto de parâmetros acústicos objetivos que formam uma base científica para a análise acústica das salas de música.

Determinamos a metodologia experimental mais adequada [3] e efetuamos medições em seis importantes salas de concerto em São Paulo, comparando os resultados. Realizamos uma análise crítica a respeito dos parâmetros acústicos obtidos e aprofundamos a compreensão sobre seus significados e suas utilidades. Por fim, fizemos uma análise subjetiva de júri correlacionando os parâmetros acústicos medidos às respectivas impressões acústicas sobre amostras musicais gravadas nas salas, que está detalhada na referência [3].

Os parâmetros analisados aqui são: RT60 (tempo de reverberação), BR e TR (razão de graves e razão de agudos), RDR (razão entre som direto e som reverberante), EDT (*early decay time*), e C80 (clareza).

Realizamos as medições nas salas do Teatro Municipal de São Paulo, Teatro Sérgio Cardoso, Anfiteatro Camargo Guarnieri (USP), Teatro Municipal de Diadema, Teatro São Pedro e Teatro do Memorial da América Latina.

RESULTADOS

Tempo de reverberação (RT60):

Os resultados de reverberação foram, em geral, condizentes com a fórmula de Sabine, ou seja: maiores valores de reverberação para salas com maior razão entre volume e capacidade de absorção. As salas menores (Camargo Guarnieri, São Pedro e Diadema) apresentaram menores tempos de reverberação em comparação com as maiores (Municipal, Memorial e Sérgio Cardoso). Porém, algumas sutilezas do comportamento do tempo de reverberação em função da freqüência podem ser melhor entendidas quando observamos as particularidades do tratamento acústico de cada teatro.

O Memorial, que tem praticamente todas as paredes cobertas por carpetes, e o Municipal, que também é bastante acarpetado, são as salas que mais dispõem de material absorvedor. O Camargo Guarnieri e principalmente o São Pedro têm relativamente pouca quantidade de material de absorção. Isso explica porque esses teatros apresentam tempos de reverberação mais estáveis nas altas freqüências quando em comparação com teatros maiores, porém mais absorvedores.

O Sérgio Cardoso, que também é um teatro usado para arte dramática, possui um palco com 13.676 metros cúbicos, que é por si só um volume maior do que o de alguns teatros. Isso resulta numa câmara reverberante cujos efeitos podem ser comprometedores, quando não bem controlados. A presença de alguns painéis em torno do espaço da orquestra não se mostrou suficiente para diminuir os efeitos do excesso de reverberação causado pela câmara reverberante e o resultado pode ser observado no gráfico 1.

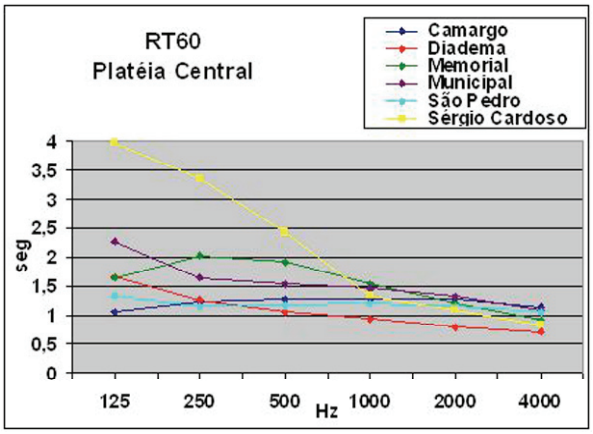


Fig. 1: RT60 nas platéias centrais dos teatros

Com exceção do Teatro Municipal, que apresenta maior variedade de locais para escuta, os tempos de reverberação se mostraram, em geral, uniformes para cada teatro, ou seja, não detectamos grandes variações de RT60, para cada faixa de freqüência, dentro de uma mesma sala. Entretanto, existem grandes diferenças na percepção auditiva conforme mudamos de lugar num mesmo teatro, como pode ser verificado através das amostras musicais gravadas para a análise do parâmetro RDR. Isso mostra como o parâmetro RT60 é absolutamente insuficiente para caracterizar a acústica de uma sala.

Verificamos que mesmo impressões como vivacidade e reverberação, usualmente atribuídas ao RT60, mudavam bastante de acordo com os diversos locais de escuta dentro de uma mesma sala, ainda que o parâmetro RT60 não apresentasse variações na mesma proporção. Certamente, outros parâmetros exercem, juntamente com o RT60, uma forte influência sobre a impressão de reverberação, conforme veremos mais adiante.

Conforme as indicações de Beranek [4] os resultados de RT60 para o Teatro Municipal o colocam essencialmente como um teatro bom para ópera, os teatros São Pedro e Camargo Guarnieri propícios para música de câmara ou reduzidas formações orquestrais.

Equilíbrio entre graves e agudos (BR e TR):

O parâmetro BR é usualmente relacionado ao calor acústico, ou à presença de graves. O parâmetro TR é normalmente relacionado ao brilho acústico. Os valores de BR e TR apresentam relativamente pouca variação entre as diversas posições de captação numa mesma sala.

As grandes dimensões da câmara reverberante no palco do Sérgio Cardoso, e suas laterais de alvenaria, fazem com que as ondas de baixas freqüências tenham longos tempos de reverberação, gerando valores de BR demasiadamente altos.

O Teatro de Diadema apresenta aberturas incomuns nas laterais do palco, ocasionando um aumento considerável na largura desse setor. Essa região torna-se propícia para o confinamento de ondas de baixas freqüências, ocasionando valores de BR relativamente altos.

Além de ser o teatro mais estreito, o Camargo Guarnieri é o único que apresenta em toda a extensão lateral grande quantidade de superfície de madeira funcionando como membranas dissipadoras de energia das ondas de baixa freqüência, o que resultou nos menores valores de BR.

As paredes descobertas e lisas dos teatros São Pedro e Camargo Guarnieri resultaram nos maiores índices de TR e o excesso de material absorvedor no Memorial causou os menores valores desse parâmetro.

Seguindo as orientações bibliográficas, analisamos as amostras musicais gravadas nos teatros de maior BR esperando perceber maior presença de graves nesses teatros. Isso não aconteceu. A presença dos graves percebida nas amostras não acompanhava a indicação dos valores de BR, isto é, teatros que apresentaram grande diferença nos valores de BR não apresentaram a mesma diferença na percepção auditiva da presença dos graves, o que pode ser verificado fazendo-se uma comparação entre as amostras gravadas e o gráfico geral de BR e TR.

Basta uma observação mais atenta na definição do parâmetro BR para concluirmos que de fato não faz muito sentido esperarmos que ele seja bem correlacionado com a presença de graves. O parâmetro BR engloba variáveis de RT60, que informam a *rapidez* do decaimento da energia acústica. A presença de graves deve estar mais relacionada à *intensidade* com que as

ondas de baixas freqüências atingem um determinado ponto de captação. Devemos, portanto, esperar melhor correlação entre tal impressão e o parâmetro G (*strength*), tomado para baixas freqüências.

Embora ainda referências modernas apresentem o referido equívoco, a conclusão anterior é confirmada por referências mais específicas e atualizadas. Em seu mais recente trabalho, Beranek [4] associa a impressão de presença dos graves ao novo parâmetro G_{low} que é a média dos valores do parâmetro G entre 125 Hz e 250 Hz.

A análise do parâmetro TR revelou fato semelhante. Encontramos amostras que eram muito mais "opacas" do que outras, apresentando, entretanto, praticamente os mesmos valores de TR. Seguindo o mesmo raciocínio usado para o parâmetro BR, podemos esperar que a impressão de brilho acústico esteja relacionada não a uma razão entre valores de RT60, mas à quantidade de energia de ondas de alta freqüência captadas. Embora não apareça em nenhuma referência estudada, torna-se natural propor o emprego de outro novo parâmetro, o G_{high}, média dos valores do parâmetro G entre 2 KHz e 4 KHz, o qual, espera-se, esteja melhor relacionado ao brilho acústico.

A audição das amostras gravadas revelou uma outra utilidade, bastante importante do ponto de vista musical, para os parâmetros BR e TR. Ao contrário do que foi constatado anteriormente, essa nova utilidade está em perfeito acordo com a definição dos parâmetros.

O parâmetro BR é a razão entre os RT60 de graves e médios e o TR é razão entre os RT60 de agudos e médios. Observamos que as salas que apresentavam valores de BR próximos aos de TR soavam mais equilibradas com respeito à reverberação entre graves e agudos, enquanto que nas salas que apresentavam maiores discrepâncias entre esses parâmetros ouvia-se um desequilíbrio indesejável na reverberação entre graves e agudos.

Nas salas onde BR é maior do que TR (Teatro de Diadema, Municipal e Memorial) há uma perceptível "sobra" de graves quando em comparação com teatros em que os valores de BR e TR são mais próximos (Camargo Guarnieri e São Pedro) nos quais o decaimento sonoro entre graves e agudos é mais uniforme e agradável.

As conseqüências musicais desse desequilíbrio vão desde uma execução aparentemente infiel do texto musical (notas de mesma duração soando com diferentes durações) até a sensação de que os naipes estão tocando de forma desencontrada.

Concluímos, portanto, que a utilidade dos parâmetros BR e TR se restringe à importância que eles apresentaram como critérios eficientes para a avaliação do equilíbrio entre freqüências dentro de uma sala.

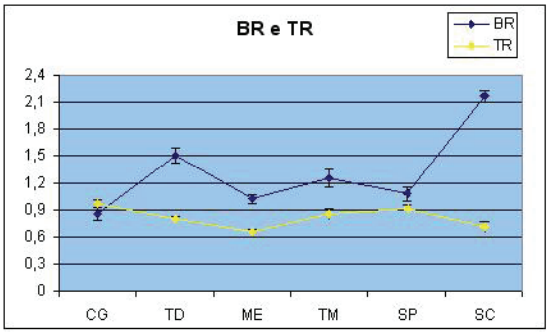


Fig. 2: Valores médios de BR eTR

As legendas no eixo horizontal são definidas por:

Sigla	Teatro
CG	CamargoGuarnieri
SP	São Pedro
TD	Diadema
SC	Sérgio Cardoso
ME	Memorial
TM	Municipal

Tabela 1: Legenda dos teatros

Clareza (C80):

O parâmetro C80 mede a razão entre a energia acústica que chega em um ponto de captação nos primeiros 80 ms e a energia remanescente. Essa distribuição de energia ao longo do tempo é determinada por características peculiares de cada teatro. Dada a diversidade de peculiaridades observadas nas salas que analisamos, é de se esperar também uma variedade no comportamento de C80.

Por exemplo, a platéia do Teatro São Pedro têm forma de concha e há pouco material absorvedor nas superfícies. Isso faz com que as ondas de alta freqüência transitem mais pelo teatro, causando os menores valores de C80 para essa faixa de freqüência. No Memorial há um excesso de material absorvedor e a distância entre as paredes laterais é muito grande. Além disso, o teto parabólico transforma os fundos da platéia num calabouço para ondas de alta freqüência. Como resultado, os valores de C80 para essa faixa de freqüência na posição central do Memorial foram os maiores.

Os valores de C80 no palco foram maiores que os da platéia em todos os teatros. Isso é desejável por facilitar o trabalho do maestro e tornar a audição mais agradável para o público.

A partir de um ponto de vista conceitual, somos induzidos a esperar que quanto maior a reverberação numa sala, menor será a clareza. De fato, os aglomerados de curvas de RT60 em função da freqüência são descendentes, enquanto que os de C80 em função das mesmas freqüências são ascendentes.

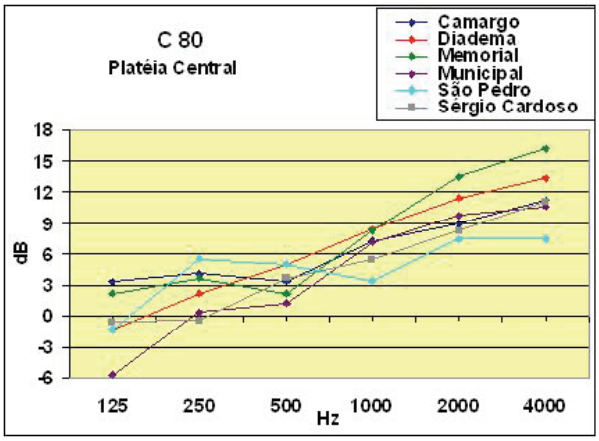


Fig. 3: C80 nas platéias centrais dos teatros

Esse resultado deve-se ao fato de que ondas de baixas freqüências são mais difusas e transpõem melhor os obstáculos, enquanto que as de altas freqüências são mais direcionais e mais suscetíveis de serem absorvidas em cada incidência sobre uma superfície. Assim, as ondas de baixas freqüências serão captadas por mais tempo e sofrerão um decaimento menos acentuado do que as de altas freqüências, ou seja, maior RT60 e menor C80. O mesmo raciocínio se aplica às ondas de alta freqüência, levando a um RT60 menor e C80 maior.

Porém, constatamos que essa regra geral vale para tendências estatísticas com respeito à freqüência, mas nem sempre para comparação direta entre valores isolados; isto é, dada uma determinada freqüência, não podemos olhar no gráfico de RT60, tomar o valor de um teatro que esteja abaixo de todos os outros e afirmar que ele estará acima de todos os outros no gráfico de C80. Por exemplo, o Municipal apresenta os maiores valores de C80 no palco, entretanto seus valores de RT60 estão numa região intermediária com relação aos outros teatros. O teatro de Diadema é o que apresenta menores valores de RT60, porém, é o que tem menores valores de C80 na região dos graves, e na região dos agudos está numa zona intermediária.

Podemos compreender tais possibilidades se observarmos os conceitos mais atentamente. O RT60 informa *quanto tempo* dura o decaimento, mas o C80 informa *como* esse decaimento se dá. Para um mesmo tempo de decaimento podemos ter várias possibilidades de distribuição de energia ao longo do tempo, ou seja, para um mesmo valor de RT60 há diversos valores possíveis de C80.

Esse fato pode ser facilmente verificado quando observamos os parâmetros medidos em alguns teatros. Por exemplo, as três diferentes posições de captação na platéia central do São Pedro apresentaram praticamente o mesmo RT60, porém seus valores de C80 são bastante diferentes; o mesmo vale para as posições do balcão daquele mesmo teatro. Um caso ainda mais acentuado é o dos pisos superiores (balcões e galeria) do Teatro Municipal, que também apresentam valores de RT60 semelhantes entre si, mas os valores de C80 divergem fortemente. Certamente outros parâmetros devem estar influenciando a Clareza.

Como já mencionamos, a referência temporal para o cálculo de C80 é 80 ms. Considerando um decaimento linear em dB, já observado nos resultados da salas, e utilizando uma regra de três simples, podemos calcular que em 80 ms o decaimento de energia é de 3,2 dB, para o caso de um RT60 de 1,5 s. Seria grosseiro demais tentar estimar o que acontece nos primeiros 3,2 dB a partir de um resultado válido para o decaimento de 60 dB. É mais razoável esperar uma correlação melhor entre C80 e um valor referente ao intervalo de tempo relacionado aos primeiros instantes de decaimento da energia.

Já conhecemos um parâmetro relacionado ao decaimento de energia nos primeiros instantes da reverberação. Este parâmetro é o EDT (*early decay time*), que é calculado tomando-se a inclinação de decaimento apenas para os primeiros 10 dB. Na análise dos gráficos de decaimento notamos como é possível que haja valores de EDT muito diferentes para valores de RT60 bastante semelhantes.

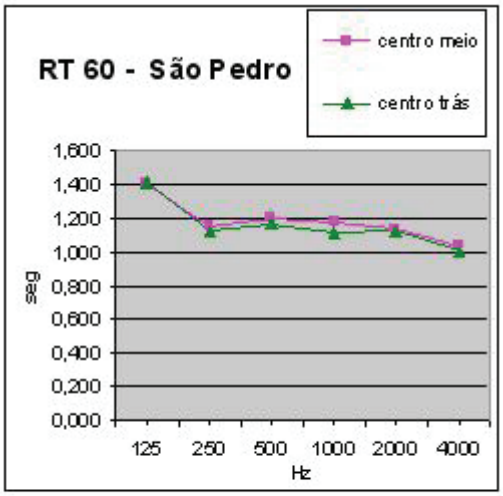
Seguindo o raciocínio anterior, podemos esperar que seja mais provável uma relação entre EDT e C80, de tal forma que olhando para o gráfico de um poderíamos estimar o comportamento do outro, algo que como já vimos é mais difícil entre C80 e RT60. O próximo passo é comparar os três parâmetros (RT60, C80 e EDT) nas mesmas posições de captação.

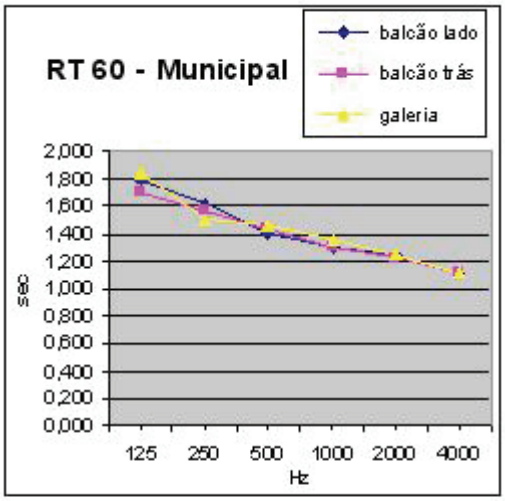
Dentro de cada setor os valores de RT60 são praticamente os mesmos. Os valores de C80 assumem valores diferentes entre as posições de cada setor. Os valores de C80 na posição centro-meio do São Pedro são menores que os da posição centro-trás. Para o mesmo setor o comportamento de EDT é inverso: a posição centro-meio apresenta maiores valores de EDT do que a posição centro-trás. No setor dos balcões e galeria do Municipal esse fato se repete: a posição que estava em cima no gráfico de C80 está em baixo no gráfico de EDT.

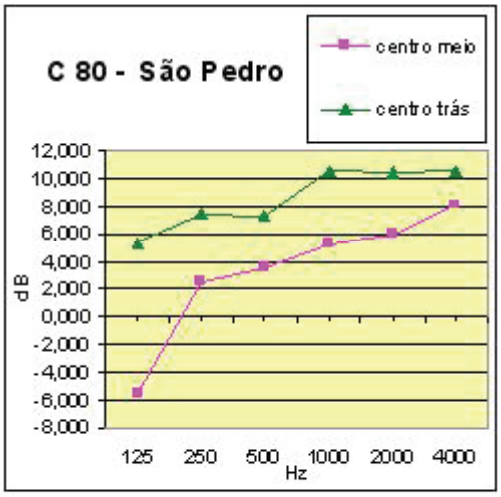
O que esses gráficos querem expressar vai de encontro à hipótese inicial segundo a qual quanto maior a clareza menor a reverberação e vice-versa. O detalhe importante é que essa relação diz respeito aos primeiros instantes da reverberação (EDT) e não à reverberação total (RT60). No caso em que o EDT sofre pouca variação entre as posições de captação observamos que o C80 também apresenta variações menores.

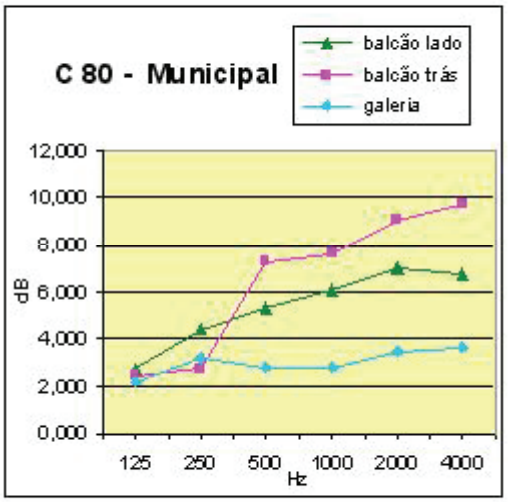
Em alguns casos, a relação de proporcionalidade inversa entre EDT e C80 não se verificou para todas as freqüências. Apesar desse fato, a conclusão mais importante a ser tomada, e que permanece válida para todos os casos observados, é que Clareza musical é muito mais sensível ao decaimento nos primeiros instantes de reverberação do que na reverberação total.

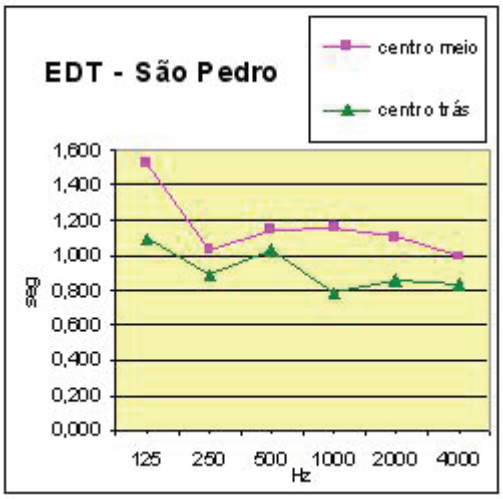
Essa conclusão aparece em trabalhos mais recentes Beranek [4] e é importante para compreendermos alguns resultados acústicos observados. Por exemplo, o Teatro Sérgio Cardoso apresenta excessivos valores de RT60. Antes da conclusão a que chegamos, poderíamos ficar temerosos quanto à Clareza percebida naquele teatro. Entretanto, observamos que seus valores de EDT são bem menores que os de RT60, principalmente nas baixas freqüências. Os valores de EDT no Sérgio Cardoso estão dentro da média com relação aos outros teatros, isso explica a posição intermediária ocupada pelo Teatro Sérgio Cardoso no gráfico geral de C80, que também pode ser verificada nas amostras gravadas.











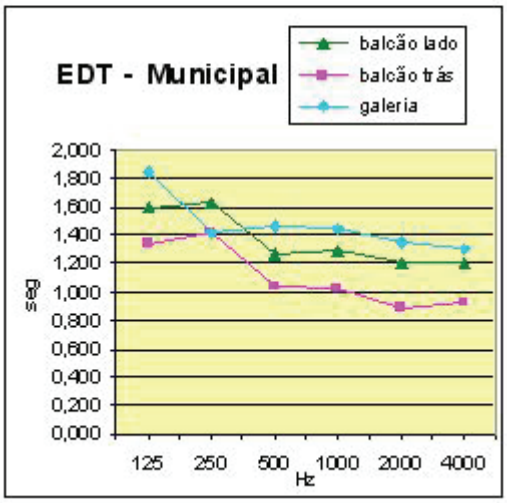


Fig. 4: Comparações entre RT60, C80 e EDT

Razão Direto / Reverberante (RDR):

O parâmetro RDR é a razão entre a energia direta e a energia reverberante captadas em determinado ponto. O valor do parâmetro RDR é obtido tomando-se como referência o instante de chegada da primeira reflexão. A energia compreendida entre a captação do som direto e da primeira reflexão é a energia direta, e após o instante da primeira reflexão é a energia reverberante.

O gráfico a seguir mostra os valores de RDR calculados em três setores diferentes para cada teatro, conforme o que foi obtido na seção de resultados :

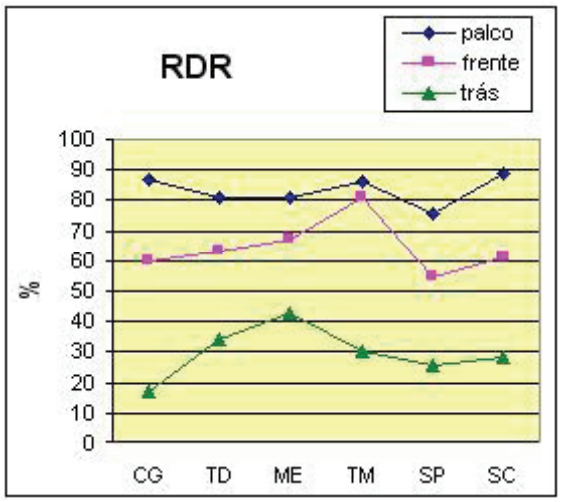


Fig. 5: Valores de RDR

Observamos um interessante padrão no qual as posições de palco apresentam altos valores de RDR, as posições do fundo da platéia apresentam RDR baixo e as posições centrais, valores intermediários. Isso mostra que o RDR é um bom parâmetro para indicar a distância entre fonte sonora e local de captação, grandezas referentes à impressão de intimismo.

Através da análise auditiva das amostras gravadas, percebemos que a sensação de intimismo e mesmo a de reverberação muda bastante conforme a posição de captação, embora o parâmetro RT60 se mantenha constante. Isso indica que ao lado do parâmetro RT60, o parâmetro RDR também é determinante para a impressão subjetiva de reverberação.

Quanto às suas aplicações, o parâmetro RDR pode ser útil como ferramenta auxiliar em simulações acústicas ou como monitoração do ponto de mixagem nos estúdios de gravação.

RESUMO DAS CONCLUSÕES

- RT60 se mantém razoavelmente constante para as várias posições de captação dentro de uma sala.
- A impressão de reverberação muda conforme a posição de audição dentro de uma sala, embora os valores de RT60 muitas vezes não acompanhem tal mudança.
- Além do RT60, o parâmetro razão de som direto / reverberante tem forte influência sobre a impressão de reverberação.
- O parâmetro BR não se mostrou bem correlacionado com a presença dos graves. Tal impressão é melhor correlacionada ao parâmetro G (strength) tomado nas baixas freqüências.
- O parâmetro TR nem sempre foi um bom indicador de brilho.
- A utilidade dos parâmetros BR e TR se restringe à importância que eles apresentaram como critérios eficientes para a avaliação do equilíbrio entre freqüências dentro de uma sala.
- O parâmetro razão direto / reverberante se mostrou mais estável e coerente do que o ITDG, no que diz respeito à impressão de intimismo.
- Ao contrário do RT60, o parâmetro C80 sofre forte variação conforme o local de captação na sala.
- O parâmetro C80 é muito melhor correlacionado ao EDT (early decay time) do que ao RT60.

REFERÊNCIAS

[1] Iazzetta, F., Kon, F. and Silva, F. S. C. AcMus: *Design and Simulation of Music Listening Enviroments*, Anais do XXI Congresso da Sociedade Brasileira de Computação, Fortaleza, Brazil, 2001.

[2] ISO 3382 *Acoustics – Measurement of the reverberation time of rooms with reference to other acoustical parameters*, 1997.

[3] Figueiredo, F. L. *Parâmetros Acústicos Subjetivos: Critérios para Avaliação da Qualidade Acústica de Salas de Música*. 2005. 258p. Dissertação de Mestrado. Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2005.

[4] Beranek, L. *Concert halls and opera houses: music, acoustics, and architecture*, Springer-Verlag, New York, 2004

AGRADECIMENTOS

Esta pesquisa é financiada pela FAPESP (processo n.º 02/02678-0) e apoiada pela Roland Brasil.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Experimentações de espacialização orquestral sobre a arquitetura AUDIENCE

Leandro Ferrari Thomaz[1], Regis Rossi A. Faria[1], Marcelo K. Zuffo[1] e João Antônio Zuffo[1]
[1]LSI – Escola Politécnica da USP
São Paulo, SP, 05508-900, Brasil
{lfthomaz, regis, mkzuffo, jazuffo}@lsi.usp.br

RESUMO

Descrevemos neste artigo a implementação de uma aplicação de espacialização orquestral desenvolvida sobre o sistema AUDIENCE. O objetivo principal do AUDIENCE é prover soluções flexíveis e escaláveis para imersão sonora multicanal. Abordamos um dos problemas típicos em orquestração: a configuração espacial do corpo orquestral, erudito ou popular, com impacto direto sobre a apreciação da peça musical ou multimídia. A aplicação proposta tem a finalidade de ampliar as possibilidades em orquestração explorando aspectos espaciais relevantes, e dando suporte para montagens usuais ou incomuns. Concebemos para tal uma cena musical virtual com três instrumentos, apresentamos o sistema construído e resultados.

INTRODUÇÃO

Uma música ou trilha sonora ao ser concebida carrega com ela alguns atributos que devem ser reproduzidos da forma mais fidedigna possível à idéia do compositor ou produtor, para que seja recebida em sua plenitude expressiva pelos ouvintes. Um desses atributos é a distribuição do som no espaço.

A capacidade de posicionar ou redistribuir as fontes sonoras no espaço ao redor do ouvinte é uma característica muito solicitada na exibição de peças musicais, nas trilhas sonoras e em jogos eletrônicos interativos. Ela é importante tanto para garantir a expressão da idéia original do compositor, como para o regente, produtor ou arranjador, bem como para calibrar um ótimo resultado final da apresentação considerando a acústica do local. Entretanto, nem sempre é possível realizar experimentações de espacialização complexas ou sofisticadas em um ensaio orquestral ou em apresentações reais.

O sistema proposto contribui para a evolução da engenharia de áudio na área de espacialização sonora, tornando possíveis diversas experimentações de espacialização orquestral por parte do compositor, regente ou produtor musical, através da facilidade de testar livremente o posicionamento de fontes sonoras virtuais no espaço 2D/3D.

Neste artigo descrevemos o problema musical escolhido para a aplicação do sistema, no caso uma orquestração composta por três instrumentos dentro de uma sala, que podem ser deslocados livremente no espaço 3D, assim como a posição do ouvinte.

PROBLEMA MUSICAL ABORDADO

A configuração espacial do corpo orquestral é um problema que vem sendo explorado sistematicamente por compositores e regentes por mais de meio século. Peças que utilizam a espacialização foram compostas por compositores como I. Xenakis (*Terretektorh*, 1965-66), para 88 instrumentistas espalhados pela platéia; R. Murray Schaffer (*Apocalypsis*, 1976-77), para 12 coros dispostos em um círculo; e K. Stockhausen (*Gruppen*, 1955-57 e *Spiral*, 1970), para três orquestras envolvendo a audiência

e para alto-falantes espalhados em forma esférica em torno da audiência. Umas destas montagens pode ser vista na figura 1 [1].

No Brasil, experimentos com a espacialização foram feitos principalmente por Flô Menezes, em peças como *Parcours de l'Entité* de 1994, para duas flautas, percussão e sons eletroacústicos, e *Harmonia das Esferas*, de 2000, para sons eletroacústicos octofônicos [2]. Na primeira peça, os flautistas se deslocam pelo espaço cênico durante toda a apresentação.

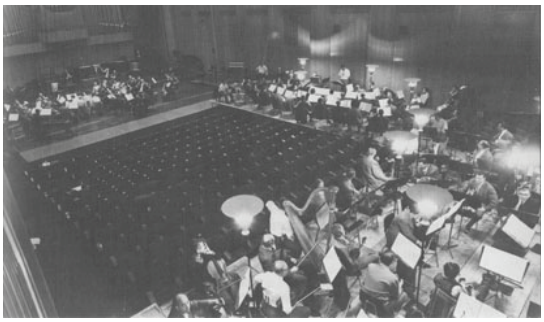


Figura 1 Ensaio da peça *Gruppen*, de Stockhausen, para 3 orquestras.

Com essa evolução, o compositor tem grandes possibilidades para aumentar o interesse por sua composição, mas torna-se muito difícil para ele conseguir prever os resultados de suas idéias espaciais sem que a peça seja realmente executada, muitas vezes sem a possibilidade de avaliar previamente por meio de um ensaio real.

Idealmente, ele poderia ter uma orquestra com a formação escolhida para a peça a sua disposição, fazendo tantas experiências com a posição de cada instrumento quanto necessárias. É claro que esta situação é praticamente impossível atualmente, devido ao custo de mobilizar uma orquestra para este fim experimental, deixando para o compositor apenas a alternativa da imagem mental da formação orquestral e seu resultado musical final.

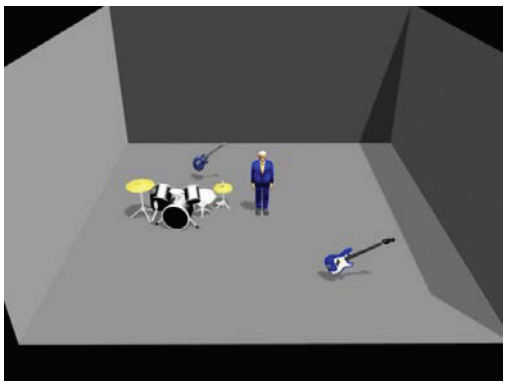


Figura 2 Cena tri-dimensional do problema musical abordado, com o posicionamento do ouvinte e das fontes sonoras.

Seria muito interessante que ele dispusesse de uma ferramenta que o auxiliasse nessa espacialização da obra, sem que fosse necessária a presença dos músicos. O sistema descrito neste artigo pode ser utilizado para a resolução deste problema em música, auxiliando o compositor e o regente na espacialização interativa orquestral.

O problema musical abordado neste experimento refere-se à espacialização de uma pequena orquestra, composta de três instrumentos contemporâneos: contrabaixo elétrico, guitarra elétrica e bateria tocando dentro de uma sala cúbica, conforme mostrado na figura 2. A posição do ouvinte e dos instrumentos nesta cena pode ser alterada livremente, permitindo uma apreciação imediata e o impacto sonoro da disposição desejada.

Esta formação é útil também quando o ouvinte também é um instrumentista que deseja simular uma sessão (ensaio) tocando junto com os instrumentos virtuais, e assim avaliar a melhor disposição relativa entre todos, segundo seus propósitos. A formação atual pode ser expandida explorando a escalabilidade do sistema, chegando mesmo a poder considerar problemas musicais de grande porte, efetivamente auxiliando o trabalho do compositor e/ou regente.

ARQUITETURA AUDIENCE

O projeto AUDIENCE – *Audio Immersion Experience by Computer Emulation* – está sendo conduzido na CAVERNA Digital da Universidade de São Paulo [5], um ambiente de realidade virtual imersiva completa. O objetivo principal é o de investigar e prover soluções flexíveis e escaláveis para imersão sonora multicanal, integradas ou não a ambientes de realidade virtual, conforme descrito em [3] e [4].

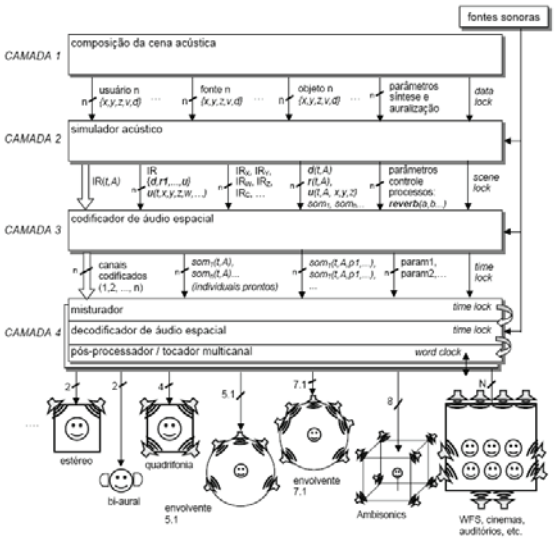


Figura 3 Arquitetura genérica de camadas do AUDIENCE.

A arquitetura de produção de som espacial do AUDIENCE, proposta por Faria em [3], está baseada em uma abordagem modular de quatro camadas funcionais, ilustradas na figura 3, permitindo a utilização de técnicas diferentes na implementação das funções executadas em cada camada e mantendo a comunicação entre elas via uma interface pré-definida e conhecida.

A camada de composição da cena acústica faz a interface com o compositor ou regente, que define a configuração da sala, a posição dos instrumentos e sua localização virtual dentro da sala de concerto.

Em seguida, a camada do simulador acústico calcula a propagação acústica da fonte sonora até o ouvinte, posicionando-a no espaço, e criando a ambiência da sala.

No codificador de áudio espacial, o sinal anecóico é convolucionado com as repostas impulsivas geradas na camada anterior, codificando os sinais de áudio espacial no formato da técnica de auralização escolhida.

A última camada é responsável pela mixagem das fontes sonoras já codificadas, decodificando o sinal de áudio e reproduzindo o campo sonoro através de uma matriz de alto-falantes.

TECNOLOGIA E INFRA-ESTRUTURA UTILIZADA

Ambiente Virtual

No presente experimento o ambiente virtual sonoro é produzido por oito alto-falantes dispostos em uma forma octogonal em torno do ouvinte, como mostra a figura 4.

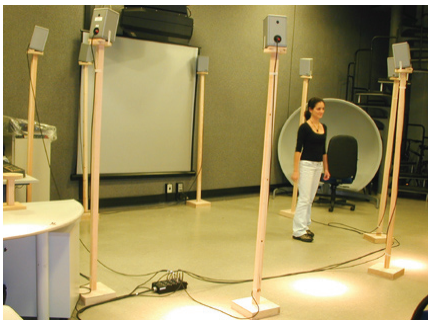


Figura 4 Configuração octogonal (2D) de decodificação Ambisonics utilizada no experimento

Estes alto-falantes são alimentados por dois amplificadores de potência de quatro canais cada, que por sua vez recebem o sinal de áudio de uma placa multicanal.

Técnica de espacialização

A técnica de espacialização utilizada é o Ambisonics, definida por Gerzon em diversos artigos como [6] e [7]. Ela permite a gravação, manipulação e reprodução de espaços sonoros tri-dimensionais, naturais ou artificiais.

O Ambisonics é uma solução tecnológica de duas partes, pois a codificação e reprodução funcionam separadamente, de forma que não é necessário preocupar-se com o sistema de reprodução no momento da gravação ou da síntese (artificial) do espaço sonoro. O formato de transmissão é conhecido por *B-Format*, e consiste em um feixe multicanal de no mínimo quatro canais individuais (Ambisonics de 1ª ordem).

Parâmetros psico-acústicos podem ser levados em consideração na decodificação, incrementando as indicações necessárias ao sistema auditivo no reconhecimento da posição da fonte sonora. Um filtro é utilizado de forma a tratar separadamente o sinal de áudio, acima e abaixo de aproximadamente 700 Hz, uma vez que nosso sistema auditivo discerne a localização dos sons graves principalmente pela diferença de fase, enquanto que dos agudos pela diferença de intensidade ou amplitude [8].

De acordo com Gerzon [6], quanto maior a ordem do sistema, maior o grau de realidade na reprodução do espaço sonoro e do espaço de audição estável (*sweet spot)*. A ordem do sistema determina o número de canais a ser utilizado. A técnica é escalável e ordens superiores são obtidas adicionando canais aos já existentes. O limite é o processamento computacional do sistema e a banda utilizada para transmissão destes canais.

Uma das grandes vantagens do Ambisonics é utilizar um número fixo de canais (de acordo com a ordem do sistema), independente do número de alto-falantes utilizados na reprodução. Desta forma, pode-se montar um arranjo de oito alto-falantes em cubo para uma reprodução tri-dimensional utilizando-se apenas quatros canais. Isto não ocorre nos sistemas de espacialização (ou *surround*) usuais, como o Dolby® Digital 5.1[1], que necessita de um canal para cada alto-falante.

Embora o número e a disposição de alto-falantes possam ser variados, melhores resultados são obtidos com um número maior e dispostos de forma regular em torno do ouvinte [7].

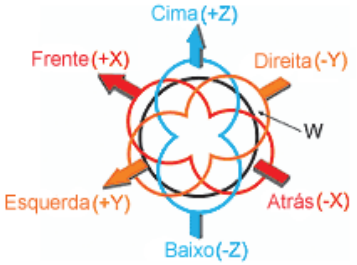


Figura 5 Representação em coordenadas cartesianas da cobertura dos sinais do Ambisonics de 1ª ordem

Neste trabalho estamos utilizando um sistema de primeira ordem, onde quatro canais são necessários (W, X, Y, Z). Esta configuração impõe requisitos mínimos para um eventual sistema de transmissão multicanal deste formato por radiodifusão. A cobertura espacial destes canais pode ser vista na figura 5.

Plataforma de programação

Estamos utilizando o PureData (PD) como plataforma para a construção dos blocos de *software* do sistema e suas conexões. O PD, desenvolvido por Miller Pucket [9], é um ambiente de programação gráfico para aplicações musicais e de áudio, amplamente utilizado nas comunidades afins.

A escolha desta plataforma foi feita por ser uma ferramenta aberta, flexível e com um tempo de reposta com baixa latência para o processamento de áudio, além de permitir a lógica de ligação entre o subsistema de áudio e o de visualização.

O PD é utilizado no projeto AUDIENCE como a ferramenta que liga os diferentes módulos, operando em cada uma das camadas apresentadas, e renderiza o áudio para reprodução final. As funções de cada camada são implementadas em blocos no PD. O *software* também oferece recursos para que esses módulos possam se comunicar com o navegador de realidade virtual e o sistema operacional, tornando possível a passagem de parâmetros da navegação para o sistema que trata o áudio.

IMPLEMENTAÇÃO

A seguir apresentamos os quatro blocos implementados, correspondentes a cada camada do AUDIENCE, bem

[1] Dolby® Digital 5.1 é marca registrada de Dolby Laboratories, Inc

como o *patch* final, que faz a ligação entre os blocos e renderiza o áudio. Eles são ilustrados na figura 8 adiante.

Sceneparser

Este bloco faz a comunicação com o navegador do sistema de realidade virtual auditiva+visual ou somente auditiva. Sua função principal está na extração (*parsing*) das propriedades e atributos da cena acústica. Este envia as posições atuais das fontes e do ouvinte, que são recebidas pelo `sceneparser` e repassadas para a próxima camada (`acousticsim`). Para otimizar o funcionamento, as posições só são passadas quando de sua mudança, evitando cálculos desnecessários pelo `acousticsim`.

O `sceneparser` é mostrado no bloco 4 na figura 8.

Acousticsim

A função principal deste módulo (bloco 5 na figura 8) é executar a simulação acústica da sala. Para este experimento, foi considerada uma sala de geometria simples (retangular), sem obstruções, e uma técnica que calcula as reflexões sonoras, obtendo uma resposta impulsiva artificial.

Utilizamos um simulador acústico baseado no traçado de raios, utilizando uma adaptação do método de fonte-imagem descrito por Allen em [10]. Uma reflexão nesta técnica vem de uma fonte-imagem virtual, localizada atrás da parede, baseada nas leis da geometria óptica, como pode ser visto na figura 6. Desta forma é possível calcular todas as reflexões[2] de uma onda sonora e o caminho destas até o ouvinte.

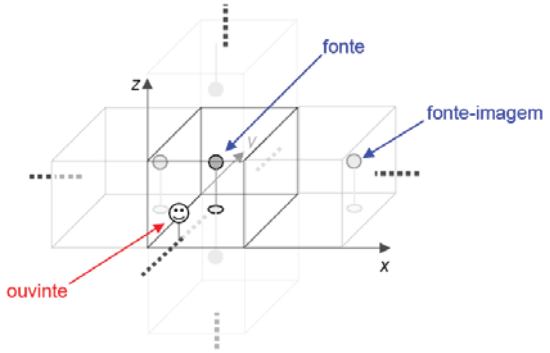


Figura 6 Técnica de traçado de raios baseado no método de fonte-imagem.

Os parâmetros necessários para o cálculo das respostas impulsivas são as dimensões da sala, o coeficiente de absorção das paredes, a posição da fonte e do ouvinte, bem como o tamanho (em amostras) da resposta impulsiva.

A saída gerada pelo `acousticsim` consiste em quatro respostas impulsivas (IRW, IRX, IRY, IRZ), correspondentes aos quatro canais do padrão B-Format do Ambisonics de 1ª ordem (W, X, Y, Z). Neste ponto, temos somente uma codificação da resposta do ambiente aos impulsos no espaço tri-dimensional.

[2] número limitado apenas pela capacidade de processamento em tempo real para um dado comprimento da resposta impulsiva.

Spatialcoder

Este módulo (bloco 6 na figura 8) tem o papel de codificar o sinal de áudio anecóico da fonte, utilizando as respostas impulsivas geradas pelo `acousticsim`.

Para isso, foi implementado um algoritmo de convolução de sinais. O método utilizado é o da convolução *overlap-add* usado em sinais de grande comprimento, caso do sinal de áudio. Desta forma, temos uma convolução contínua com baixa latência.

A biblioteca FFTW [11] foi utilizada para efetuar as transformadas rápidas de Fourier, devido sua rapidez e fácil integração ao código, tanto no sistema operacional Linux como no Windows. Otimizações foram feitas no código original para possibilitar a execução de várias convoluções ao mesmo tempo, visto que para cada fonte sonora temos um bloco `spatialcoder`. Estas otimizações, basicamente de acesso a memória, diminuem consideravelmente o uso de CPU.

Ao final do processamento pelo `spatialcoder`, temos os quatros canais codificados em *B-Format* para uma fonte posicionada em algum ponto do espaço sonoro tri-dimensional.

Spatialdecoder

O decodificador espacial desenvolvido no atual sistema é basicamente um decodificador Ambisonics de primeira ordem, com seu diagrama de blocos mostrado na figura 7.

Este módulo (bloco 7 na figura 8) recebe o sinal de áudio em *B-Format* (quatro canais) e o decodifica para o número de alto-falantes presentes, reproduzindo o espaço sonoro codificado na fase anterior. Uma mixagem é feita antes, através do bloco misturador, para que os sinais das diversas fontes sonoras sejam misturados em apenas um vetor *B-Format*, que alimentará o `spatialdecoder`.

A matriz de ganhos para a decodificação de diversas configurações de alto-falantes foi previamente calculada por R. Furse, e estão disponíveis em [12]. Aos sinais de entrada são aplicados ganhos (α, β, γ, δ), com um peso específico para cada alto-falante de saída n, e somadas.

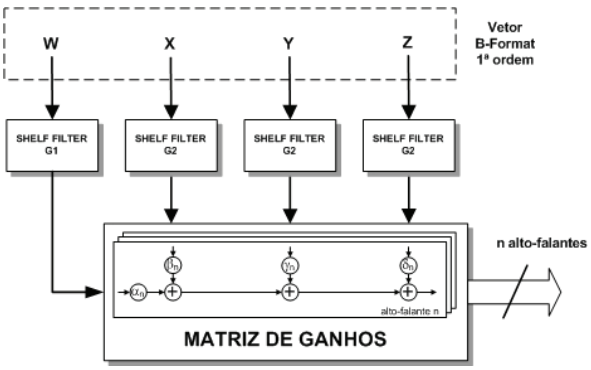


Figura 7 Diagrama de Blocos de um Decodificador Ambisonics de 1ª Ordem

Os filtros psico-acústicos utilizam dois ganhos, G1 para o sinal W e G2 para os outros, sendo que para cada ganho, temos dois valores para contemplar a divisão de freqüência em 700 Hz.

Nesta versão do sistema, usamos ganhos unitários para os filtros psico-acústicos descritos por Gerzon em [7].

Filtros equalizadores, para de-reverberação acústica local, não foram utilizados neste experimento.

Patch do experimento

A montagem do *patch* do experimento, mostrado na figura 8, utiliza os blocos descritos anteriormente, além de blocos internos do PD. Como estamos utilizando no experimento três fontes sonoras, são necessários três pares de blocos `acousticsim-spatialcoder` para gerar o áudio espacializado a partir dos sinais anecóicos (secos), dos parâmetros da sala, e da posição das fontes e do ouvinte.

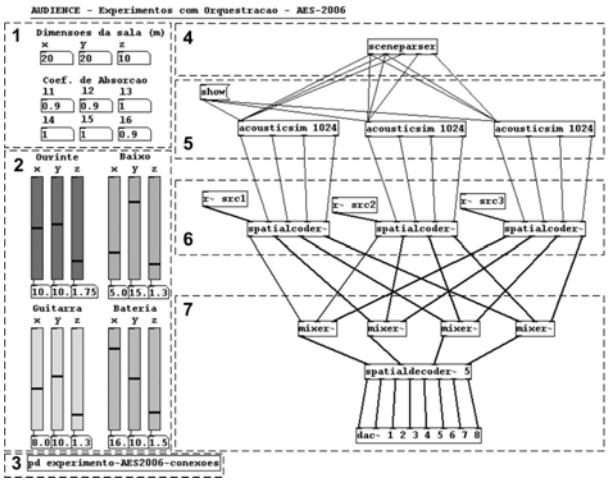


Figura 8 *Patch* feito em Pure Data do experimento

O controle dos parâmetros é feito através de campos onde pode ser modificada a configuração da sala (1). As posições das fontes e do ouvinte são controladas através de *sliders*, determinando as coordenadas *xyz* dentro da sala (2). Os blocos principais e suas conexões podem ser vistos no lado direito do *patch*. Um *patch* interno (3) faz as outras conexões, de forma que o principal não fique poluído visualmente. As referências de (4) a (7) correspondem às quatro camadas do AUDIENCE.

EXPERIMENTO

Montamos um cenário flexível que considera três instrumentos (baixo, bateria e guitarra) posicionáveis no ambiente sonoro virtual através da interface gráfica do *patch*, manipulada pelo usuário através de *sliders*, podendo também alterar o tamanho da sala e os coeficientes de absorção das paredes. A partir disto, testes foram feitos alterando a posição das fontes e do ouvinte.

A figura 9 mostra uma posição fixa deste experimento, onde um hipotético compositor quer ver os resultados de se colocar a bateria próxima ao ouvinte, ao seu lado direito, a guitarra à sua frente, distante e à esquerda, e o baixo atrás. Uma visão em perspectiva da cena é mostrada na figura 2.

Para comparar a espacialização gerada pelo simulador acústico com uma gerada por um espacializador sem ambiência, outro *patch* foi montado que não utiliza o bloco `acousticsim`. A espacialização e codificação do sinal é feita utilizando as equações de codificação do Ambisonics, apresentadas em [13], As equações mostradas a seguir indicam como calcular o sinal de cada canal em B-Format, baseado no sinal anecóico (S) e nos ângulos de rotação (θ) e elevação (ϕ) da fonte sonora com relação ao ouvinte.

$$W = 0{,}707 * S \quad (1)$$
$$X = \cos(\theta) * \cos(\phi) * S \quad (2)$$
$$Y = \mathrm{sen}(\theta) * \cos(\phi) * S \quad (3)$$
$$Z = \mathrm{sen}(\phi) * S \quad (4)$$

Neste caso, posicionam-se as fontes sonoras sobre a superfície de uma esfera de referência ao redor do ouvinte.

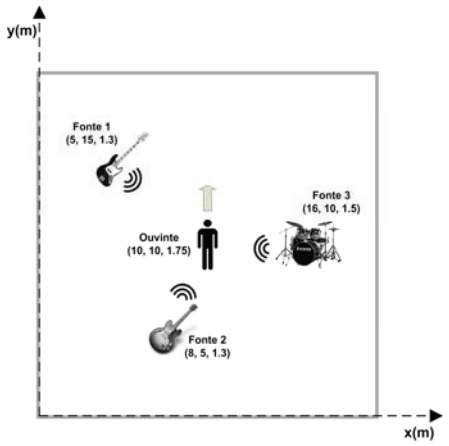


Figura 9 Planta do experimento, mostrando as posições dos instrumentos e ouvinte.

Este *patch* não permite o controle individual do distanciamento da fonte sonora, além de não acrescentar reverberação ao sinal anecóico. A comparação está relacionada somente com a percepção da direção da fonte.

Algumas simplificações forem consideradas no experimento, como a reprodução bi-dimensional através de do anel com oito alto-falantes (figura 4) e a utilização de ganhos unitários nos filtros psico-acústicos.

RESULTADOS PRELIMINARES

O simulador acústico utilizado atualmente apresenta uma boa reprodução da reverberação da sala, possibilitando uma percepção da profundidade do ambiente e das distâncias das fontes sonoras. No presente experimento, contudo, ele não apresentou uma resposta estável para a direcionalidade das fontes.

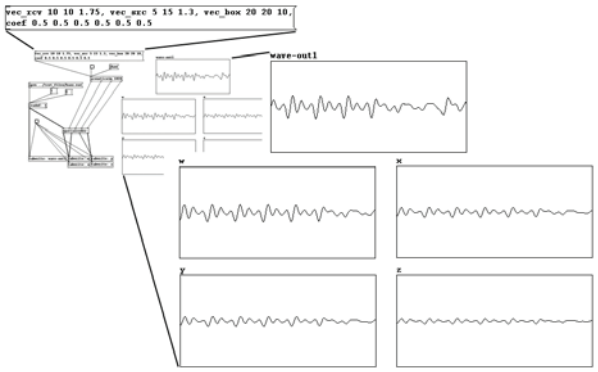


Figura 10 Formas de onda para o baixo: sinal anecóico original (em cima) e sinais B-Format do baixo posicionado (em abaixo)

Outro problema que se manifestou durante o experimento foi o pequeno *sweet-spot* conseguido dentro da montagem dos alto-falantes. Assim, para uma percepção

estável da cena, era necessário que o usuário do sistema buscasse se posicionar no centro da montagem e reduzir seus movimentos no espaço de audição.

A figura 10 mostra a espacialização específica do baixo elétrico, na posição indicada anteriormente. Acima da figura, pode-se ver os parâmetros de entrada do módulo `acousticsim`. À direita, gráficos representam as formas de onda do sinal original seco (acima) e dos quatro sinais codificados em *B-Format* (W, X, Y e Z, abaixo na figura).

Os testes feitos sem o simulador acústico mostraram uma sensação de posição das fontes muito mais nítida, embora a ambiência fosse perdida, tornando a experiência menos próxima da realidade.

CONCLUSÃO E TRABALHOS FUTUROS

Nas condições em que foram realizados os experimentos, a estabilidade da imagem espacial nas imediações do centro mostrou-se crítica com relação ao posicionamento do usuário, e a percepção da direcionalidade mostrou-se sensível quando gerada somente através do algoritmo de simulação acústica. Com os resultados globais obtidos, podemos concluir que o sistema mostra-se bastante promissor para a finalidade proposta, e que possibilita uma forma inédita para compositores e regentes executarem seus experimentos orquestrais com um baixo custo.

Uma das grandes vantagens do sistema é sua fácil utilização, acessível aos usuários não técnicos, que é o público alvo desta aplicação. Além disso, o sistema pode ser implantado domesticamente, devido ao seu relativo baixo custo, popularizando a ferramenta entre músicos.

O sistema encontra-se em estágio de desenvolvimento, e muitas melhorias serão ainda incorporadas. Primeiramente, melhorias no simulador acústico devem ser consideradas para aprimorar a percepção da direcionalidade das fontes.

A adição de novas fontes sonoras tornará o sistema mais útil para que músicos possam fazer seus experimentos, bem como uma melhora na interface para o compositor/regente programar a espacialização orquestral como, por exemplo, o uso de um *joystick* para controlar as posições.

Prevê-se num futuro breve a cooperação com compositores e regentes para que os testes possam ser também balizados por especialistas da área musical.

A montagem do sistema de alto-falantes deve ser ajustada para que o *sweet-spot* seja maior. Também se prevê o aumento da ordem do Ambisonics, para segunda e terceira ordens, e a adição de mais alto-falantes ao sistema, o que acarretaria em um aumento significativo esperado na qualidade e estabilidade do campo sonoro reproduzido.

Embora o sistema tenha sido projetado para o uso de três dimensões, para simplificar a experiência apenas simulamos um campo bi-dimensional. Uma configuração tri-dimensional acrescentando a noção de elevação à fonte sonora é prevista em experimentos próximos. Esta mudança apenas necessita da montagem de uma nova configuração de alto-falantes, sendo que o *software* já permite esse tipo de reprodução.

Finalmente, o decodificador Ambisonics poderá prever filtros de equalização e ganhos não unitários para o filtro psico-acústico descrito, quesito importante para salas de reprodução pequenas de acordo com Malham [13], para que a espacialização torne-se mais fiel ao ouvido humano.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Griffiths, P. *Modern Music.* World of Art, 1994.

[2] Menezes, F. *Atualidade Estética da Música Eletroacústica*. Editora Unesp, 1999.

[3] Faria, R. R. A. *Auralização em ambientes audiovisuais imersivos*. Tese de Doutorado em Engenharia Eletrônica, Escola Politécnica da Universidade de São Paulo, 2005.

[4] Faria, R. R. A., Thomaz, L., Soares, L., Santos, B., Zuffo, M., Zuffo, J. *AUDIENCE – Audio Immersion Experiences in the CAVERNA Digital*. Anais do 10º Simpósio Brasileiro de Computação Musical, pg. 106-117, Outubro, 2005.

[5] Zuffo, J. A et al. "CAVERNA Digital – Sistema de Multiprojeção Estereoscópico Baseado em Aglomerados de PCs para Aplicações Imersivas em Realidade Virtual. In: 4th Symposium of Virtual Reality, Florianópolis, 2001. Proceedings.

[6] Gerzon, M. *Periphony: With-Height Sound Reproduction*. J. Audio Eng. Soc., Vol. 21, No. 1, pg. 2-10, January/February, 1973.

[7] Gerzon, M. *Practical Periphony: The Reproduction of Full-Sphere* Sound. Preprinted at the 65th Audio Engineering Society Convention, London, 1980.

[8] Gerzon, M. *Surround-sound psychoacoustics.* Wireless World, pg. 483-485, December, 1974.

[9] Puckette, M. *Pd Documentation*. http://crca.ucsd.edu/~msp/Pd_documentation/. Acessado em: 14 de fevereiro de 2006.

[10] Allen, J. B., Berkley, D. A. *Image method for efficiently simulating small-room acoustics*. Journal of the Acoustical Society of America, v.65, n.4, pg. 943-950, Abril, 1979.

[11] *FFTW*. www.fftw.org. Acessado em: 14 de fevereiro de 2006.

[12] Furse, R. *First and Second Order Ambisonic Decoding Equations*. www.muse.demon.co.uk/ref/speakers.html. Acessado em: 14 de fevereiro de 2006.

[13] Malham, D., Myatt, A. *3-D Sound Spatialization using Ambisonic Techiniques*. Computer Music Journal, 19:4, pg. 58-70, Winter 1995.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Impactos na Qualidade Acústica das Salas de Aula e Atelier de uma Faculdade de Arquitetura e Urbanismo por seus Alunos e Professores

José Geraldo Querido[1], Cesar Augusto Alonso Capasso[2]
[1]Universidade de Taubaté - Taubaté, São Paulo, 12020 270, Brasil
[2]Universidade Santa Cecília - Santos, São Paulo, 11702 160, Brasil
jgquerido@bighost.com.br - cesarcapasso@unisanta.br

RESUMO

A acústica ambiental e a arquitetônica são dos principais parâmetros dos projetos dos espaços escolares urbanos. A pesquisa apresentada trata do desempenho do espaço interno de uma faculdade relatado pelos seus usuários: professores e estudantes de arquitetura e urbanismo. Eles descrevem impactos acústicos, suas reações usuais a eles e como interviriam para a sua mitigação. Este artigo propõe discutir ensino da acústica ambiental e arquitetônica numa abordagem didática, educativa e gestora, sensibilizando o arquiteto a partir de suas experiências pessoais.

INTRODUÇÃO

O desenvolvimento da Arquitetura e Urbanismo no Brasil, enquanto área do conhecimento e profissão, vem se desenhando nos moldes contemporâneos nos últimos 50 anos. Tanto no âmbito acadêmico da graduação quanto na atuação dos profissionais, um dos seus principais objetivos é a tentativa de aliar conhecimentos da arte e da técnica, trabalhando principalmente com questões relacionadas ao binômio "forma e função".[1]

O Conforto Ambiental é classificado como Matéria Profissional pela Portaria Nº. 1.770 – Ministério da Educação e Cultura (MEC), de 21 de Dezembro de 1994, e desmembrado em quatro segmentos básicos: o estudo das condições acústicas, térmicas, lumínicas e energéticas.[2]

O papel do segmento acústico da disciplina de conforto ambiental pode ser compreendido pelo estudo da defesa contra o ruído e pelo condicionamento sonoro no recinto. Encontra-se ao longo de seu desenvolvimento nas faculdades de arquitetura autores que, dedicando-se a desenvolver bibliografias específicas para o acompanhamento de cursos de graduação convergem nesta linha, são, por exemplo: CARVALHO[3], DE MARCO[4] e SILVA[5].

No atual momento histórico, a discussão sobre a incorporação das ciências ambientais na arquitetura está tratando das formas com as quais o aluno da graduação em arquitetura e urbanismo deve receber os conhecimentos necessários para a sua incorporação no projeto do edifício e dos espaços urbanos. Discutem-se como estes conhecimentos, cujo desenvolvimento científico mais sistemático é recente, serão incorporados nos currículos das escolas de arquitetura e urbanismo. Neste processo não se tem esquivado de discussões, tais como, a escassa bibliografia nacional e a necessidade do estudo da física aplicada, coisa para a qual se supõe um conhecimento prévio que o arquiteto não adquiriu e que depende do fortalecimento da informação e formação técnica na área e a aplicação de novas metodologias e instrumentos de ensino.[6]

Discute-se a implantação de atividades laboratoriais em complemento às aulas em sala, e a necessidade de que o aluno experimente a expressão prática e teórica dos conhecimentos que suas competências e habilidades requerem.[7]

A Portaria MEC Nº 1.770/94 preconiza uma formação de profissional generalista ao arquiteto. Afirma que deve ser apto a compreender e traduzir as necessidades de indivíduos, grupos sociais e comunidades, com relação à concepção, organização e construção do espaço interior e exterior, abrangendo o urbanismo, a edificação, o paisagismo, bem como a conservação e a valorização do patrimônio construído, a proteção do equilíbrio do ambiente natural e a utilização racional dos recursos disponíveis.[8]

Pressupõem-se, portanto que as decisões projetuais especificamente relacionadas à acústica arquitetônica são estudadas num nível no qual em determinados projetos o arquiteto consultará acústicos. Porém, a utilização de especialistas não se justifica na maior parte dos casos, aos projetos, atualmente, se exige a garantia da satisfação do usuário e da eficiência energética, coisa para a qual o arquiteto deve estar preparado, já que é um dos maiores responsáveis pela qualidade ambiental final do espaço arquitetônico e urbano.[9]

O trabalho apresentado propõe a realização de exercícios utilizando-se instrumentos subjetivos, cujas bases são impressões dos alunos e professores em relação ao seu desempenho pessoal durante o decorrer das aulas, atividade na qual a acústica é fundamental.

A escolha da sala de aula como principal objeto de estudo visa demonstrar a importância da qualidade acústica, associada ao projeto de um edifício para o qual não cabe a presença de especialistas. O conforto acústico é fundamental para o bom desenvolvimento das atividades didáticas e preservação da qualidade da saúde de seus usuários, principalmente a dos professores, "profissionais da voz", por vezes, inconscientes do fato.[10]

Outro fator importante na escolha do ambiente escolar é a atual necessidade da avaliação institucional continuada, preconizada pelo MEC e que inclui a avaliação das instalações, na qual se aborda questões do conforto ambiental das salas de aula.[11]

O exercício não está relacionado a qualquer disciplina, trata-se de atividade livre desenvolvida por ocasião de uma pesquisa de mestrado, porém, poderá ser incorporado às atividades do laboratório de conforto ambiental e repetido com freqüência torna-se instrumento didático, de educação ambiental e contribui para a gestão acústica do espaço pela comunidade acadêmica e pela mantenedora.

METODOLOGIA

A pesquisa buscou um universo onde houvesse indícios de problemas relativos ao conforto ambiental e identificou num trabalho do Núcleo de Avaliação Institucional (NAI) de uma Universidade, dados que relatam a insatisfação do corpo discente de uma das suas faculdades em relação às instalações das salas de aula. Criaram-se dois instrumentos que abordam aspectos subjetivos na forma de questionários: o primeiro direcionado ao corpo discente e outro ao docente. Desenvolveram-se levantamentos espaciais de diversas tipologias além de testes e cálculos. Todos os instrumentos são voltados à caracterização do conforto acústico dos usuários durante o desenvolvimento das atividades didáticas.

O questionário do aluno se dirige a sala de aula que ele utiliza durante as atividades do ano letivo e o do professor aborda a sua experiência nas salas de aula em que atua. As questões são elaboradas de forma que leigos possam respondê-las, bastando a vivência do espaço a ser pesquisado. São utilizadas perguntas optativas e dissertativas.

O questionário foi encaminhado a todos os 43 professores da faculdade por arquivo de texto na forma de anexo em mensagem eletrônica via Internet. As respostas foram enviadas à caixa de mensagens e impressas sem que se identificasse o respondente, a amostragem composta pelas respostas enviadas pelos professores atingiu a 30,23%.

O questionário foi aplicado aos alunos no interior da sala de aula, pelo pesquisador, que inicialmente esclarece que o procedimento é autorizado pelo NAI e que os respondentes não seriam identificados.

Cada turma do primeiro ao quinto ano do curso teve aplicado o questionário em horário de aula normal, sendo pesquisada uma turma por dia no período de cinco dias consecutivos. A amostragem é composta pela totalidade de alunos que compareceram a aula no dia e horário da sua aplicação e atingiu 60,07% do total de 273 alunos do curso.

Neste artigo apresenta-se tabulação geral, porém, pode-se realizar tabulação por cada uma das cinco salas.

Como há perguntas que possibilitam ao respondente fornecer mais de uma resposta a tabulação considerou a porcentagem da recorrência da pergunta no total de respondentes, portanto a somatória dos valores porcentuais pode exceder aos 100% em alguns casos.

Questionário aos professores

Por favor, responda a partir de agora, especificamente quanto ao Conforto Acústico das salas de aula (quinto andar) desta Faculdade:

1- Você sente dificuldade em ouvir e/ou entender as frases formuladas pelos alunos em sala de aula?

Em curta distância: entre a primeira fila e o meio da sala.

( )sim ( )não

Em média distância: Entre o meio e o fundo da sala.

( )sim ( )não

1.1- Em caso de resposta positiva, você procura superar o problema? ( )sim ( )não

1.2- Em caso de resposta positiva, consegue resolver a questão? ( )sim ( )não ( ) parcialmente

1.3- Em caso de resposta positiva, você consegue identificar a origem do problema?

( )sim ( )não ( ) não tem certeza

1.4- Qual é?

2- Você percebe alguma dificuldade por parte dos alunos em ouvir suas palavras e/ ou compreendê-las?

Em curta distância: entre a primeira fila e o meio da sala.

( )sim ( )não

Em média distância: Entre o meio e o fundo da sala.

( )sim ( )não

2.1- Em caso de resposta positiva, como você procura superar o problema?

2.2- Em caso de resposta positiva, consegue resolver a questão? ( )sim ( )não ( ) parcialmente

2.3- Em caso de resposta positiva, você consegue identificar a origem do problema?

( )sim ( )não ( ) não tem certeza

2.4- Qual é?

3- Há ruídos externos à sala de aula que são percebidos por você durante as atividades didáticas? ( )sim ( )não

3.1- Quais são?

3.2- Por favor, classifique o grau de incômodo pelos ruídos externos:

( )não incomoda ( )incomoda pouco

( )incomoda medianamente ( )incomoda muito

3.3- Com qual freqüência ele (ruído externo) ocorre?

( )nunca ( )eventualmente ( )freqüentemente ( )sempre

3.4- Em caso de provocar incômodo você procura superar o problema? ( )sim ( )não

3.5- Em caso de resposta positiva, descreva como?

3.6- Em caso de resposta positiva, consegue resolver a questão? ( )sim ( )não ( ) parcialmente

4- Há ruídos internos na sala de aula que são percebidos por você durante as atividades didáticas? ( )sim ( )não

4.1- Quais são?

4.2- Por favor, classifique o grau de incômodo pelos ruídos internos:

( )não incomoda ( )incomoda pouco

( )incomoda medianamente ( )incomoda muito

4.3- Com qual freqüência ele (ruído interno) ocorre?

( )nunca ( )eventualmente ( )freqüentemente ( )sempre

4.4- Em caso de provocar incômodo você procura superar o problema? ( )sim ( )não

4.5- Em caso de resposta positiva, descreva como?

4.6- Em caso de resposta positiva, você consegue resolver a questão? ( )sim ( )não ( ) parcialmente

5- Você classificaria o desempenho da acústica das salas de aula como:

( ) péssimo ( ) sofrível ( ) regular ( ) bom ( )excelente

6- Você identifica problemas relacionados diretamente a acústica arquitetônica no atelier, que de alguma forma comprometem o desempenho das suas atividades e/ou lhe incomodam? ( )sim ( )não

6.1- Quais são?

Questionário aos alunos

Por favor, responda a partir de agora, especificamente quanto ao Conforto Acústico desta sala de aula:

1- Você tem dificuldade em ouvir e/ou compreender as palavras dos professores?

Em curta distância: até quatro metros.

( )sim ( )não

Em média distância: acima de quatro metros.

( )sim ( )não

1.1- Em caso de resposta positiva, você procura superar o problema? ( )sim ( )não

1.2- Em caso de resposta positiva, consegue resolver a questão? ( )sim ( )não ( ) parcialmente

1.3- Em caso de resposta positiva, você consegue identificar a origem do problema?

( )sim ( )não ( ) não tem certeza

1.4- Qual é?

2- Você sente dificuldade em ouvir e/ou entender as frases formuladas pelos outros alunos durante as aulas?

Em curta distância: num raio de no máximo quatro carteiras.

( )sim ( )não

Em média distância: num raio acima de quatro carteiras.

( )sim ( )não

2.1- Em caso de resposta positiva, como você procura superar o problema?

2.2- Em caso de resposta positiva, consegue resolver a questão? ( )sim ( )não ( ) parcialmente

2.3- Em caso de resposta positiva, você consegue identificar a origem do problema?

( )sim ( )não ( ) não tem certeza

2.4- Qual é?

3- Você percebe alguma dificuldade por parte dos professores em ouvir suas palavras e/ ou compreendê-las?

Em curta distância: até quatro metros.

( )sim ( )não

Em média distância: acima de quatro metros.

( )sim ( )não

3.1- Em caso de resposta positiva, como você procura superar o problema?

3.2- Em caso de resposta positiva, consegue resolver a questão? ( )sim ( )não ( ) parcialmente

3.3- Em caso de resposta positiva, você consegue identificar a origem do problema?

( )sim ( )não ( ) não tem certeza

3.4- Qual é?

4. Há ruídos externos à sala de aula que são percebidos por você durante as atividades didáticas? ( )sim ( )não

4.1- Quais são?

42- Por favor, classifique o grau de incômodo pelos ruídos externos:

( )não incomoda ( )incomoda pouco

( )incomoda medianamente ( )incomoda muito

4.3- Com qual freqüência ele (ruído externo) ocorre?

( )nunca ( )eventualmente ( )freqüentemente ( )sempre

4.4- Em caso de provocar incômodo você procura superar o problema? ( )sim ( )não

4.5- Em caso de resposta positiva, descreva como?

4.6- Em caso de resposta positiva, consegue resolver a questão? ( )sim ( )não ( ) parcialmente

5- Há ruídos internos na sala de aula que são percebidos por você durante as atividades didáticas? ( )sim ( )não

5.1- Quais são?

5.2- Por favor, classifique o grau de incômodo pelos ruídos internos:

( )não incomoda ( )incomoda pouco

( )incomoda medianamente ( )incomoda muito

5.3- Com qual freqüência ele (ruído interno) ocorre?

( )nunca ( )eventualmente ( )freqüentemente ( )sempre

5.4- Em caso de provocar incômodo você procura superar o problema? ( )sim ( )não

5.5- Em caso de resposta positiva, descreva como?

5.6- Em caso de resposta positiva, você consegue resolver a questão? ( )sim ( )não ( ) parcialmente

6. Você classificaria o desempenho da acústica das salas de aula como:

( ) péssimo ( ) sofrível ( ) regular

( ) bom ( )excelente

7- Você identifica problemas relacionados diretamente a acústica arquitetônica no atelier, que de alguma forma comprometem o desempenho das suas atividades e/ou lhe incomodam? ( )sim ( )não

7.1- Quais são?

ANÁLISES E CONCLUSÕES

A primeira conclusão que se apresenta é a ocorrência de impactos acústicos que interferem no desempenho de alunos e professores, eles são levantados nos instrumentos objetivos e subjetivos. Portanto, conclui-se que a avaliação do NAI é referendada pela atual pesquisa, e que o conforto acústico representa, junto aos usuários do espaço da faculdade, uma preocupação, influenciando diretamente sobre as atividades didáticas desenvolvidas. (Tab. 1 e 2)

Em curta distância: até quatro metros.	
SIM	61,54%
NÃO	38,46%
Em média distância: acima de quatro metros.	
SIM	92,31%
NÃO	7,69%

Tabela 1 - Ocorrência de dificuldade na audição ou entendimento das frases formuladas pelos alunos - pelo professor

Em curta distância: até quatro metros.	
SIM	28,66%
NÃO	70,73%
Em média distância: acima de quatro metros.	
SIM	53,66%
NÃO	45,12%

Tabela 2 - Ocorrência de dificuldade na audição ou entendimento das frases formuladas pelos professores - pelo aluno

Conclui-se também que os instrumentos de pesquisa objetivos e subjetivos são complementares e importantes para as conclusões que levem ao entendimento global da avaliação de um espaço, isso é reforçado neste caso em se tratando de uma avaliação pós-ocupação e onde os instrumentos subjetivos representam a vivência dos seus usuários, alunos e professores.

Algumas das questões levantadas com os instrumentos objetivos têm o seu impacto destacado pelas citações recorrentes entre os respondentes. Pode-se citar como exemplo disso os ruídos provenientes do corredor interno de acesso às salas, que se destaca em relação ao ruído do "buffet" localizado em edifício vizinho. A quantidade de citações que o corredor recebe de professores e de alunos é significativa e supera em freqüência a de ruídos externos.

A ocorrência de problemas na garganta e nas pregas vocais de professores e a reação de elevar a voz para sobrepor-se aos ruídos de fundo, indicam que a intensidade e a freqüência do impacto acústico são grandes. (Tab.3 e 4)

Sudorese	15,38%
Mal estar	7,69%
Desidratação	7,69%
Cansaço	30,77%
Desgaste	7,69%
Problemas na garganta	30,77%
Comprometimento das cordas vocais	15,38%
Dispersão	10,00%
Incômodo	30,00%
Desconforto provocado pela temperatura	20,00%

Tabela 3 - Sintomas físicos relacionados pelos professores ao mau desempenho do conforto ambiental do espaço

Quanto às questões relativas à audibilidade os dados obtidos nos instrumentos objetivos apresentam pontos antagônicos em relação aos levantados pelos questionários. Por exemplo: em todas as salas os testes de articulação indicaram resultados muito bons, porém, contestados pelas respostas de professores e alunos que atestam problemas de audibilidade.

Ao aprofundarmos a análise com outros instrumentos objetivos verificamos que o tempo de reverberação calculado para cada uma das salas apresenta níveis muito superiores ao tempo ótimo de reverberação determinado pela norma, isso, aliado a ocorrência de ruídos de fundo pode piorar muito a articulação da sala, solicitando melhoria da relação sinal/ ruído – o que pode explicar as dificuldades relatadas nos questionários.

Neste mesmo sentido há outros resultados dos questionários dos professores que corroboram com a hipótese de que o tempo de reverberação superior ao tempo ótimo de reverberação aliado aos ruídos de fundo causam problemas na relação sinal/ruído nas salas. Reações tais como "falar vagarosamente", "pedir silêncio" e "falar mais alto", obtidas dos professores quando inquiridos sobre dificuldades na audibilidade de suas palavras pelos alunos podem relacionar-se a este tipo de impacto. (Tab. 4)

Neste caso os instrumentos subjetivos foram significativamente importantes para a valorização da dúvida em relação aos resultados do teste de articulação, já que, em todas as salas, quando questionados sobre a inteligibilidade da comunicação, os alunos e professores atestam dificuldades, o que intensifica a necessidade de abordagens que levem a diagnosticar os causadores dos impactos quanto à audibilidade e articulação da sala.

Pedindo silêncio	7,69%
Falando mais alto	84,62%
Deslocando-me pela sala	7,69%
Falando vagarosamente	7,69%
Resolve a questão	30,77%
Não resolve a questão	0,00%
Resolve parcialmente a questão	61,54%

Tabela 4 - Reações individuais dos professores contra a má audição ou compreensão das suas palavras pelos alunos e eficácia das reações

Quando se trata de aspectos relacionados à voz de professores, não se pode desconsiderar que o curso em questão é noturno, e quase a totalidade de professores, tem outras atividades profissionais diurnas nos seus dias de aula, não sendo possível, portanto, relacionar ao ambiente as possíveis patologias. Para isso os resultados obtidos nos instrumentos objetivos e subjetivos não se mostram conclusivos.

Outro aspecto importante verificado é que os dados levantados pelos questionários junto ao corpo discente são respaldados pelos levantados junto ao corpo docente. Na pesquisa confirma-se que a identificação dos impactos é coincidente e reforça a similaridade entre os levantamentos com instrumentos objetivos e aqueles obtidos segundo a percepção e vivência dos espaços pelo corpo discente e docente.

A grande questão que fica em aberto refere-se a identificação dos graus de incômodo, sua freqüência e a real interferência dos impactos no conforto acústico, nas atividades didáticas e até na saúde dos professores. Para isso será necessária a inclusão de medições acústicas, porém, pode-se determinar a tipologia de ensaios a partir das informações obtidas.

Verificaram-se pontos de divergência nos dados levantados junto ao corpo discente no que se refere aos graus de incômodo e a sua freqüência. Nota-se uma tendência de crescimento da intensidade das classificações do impacto coincidente com o tempo de curso do aluno.

No que se refere à utilização de dados recolhidos nos instrumentos subjetivos que podem gerar diretrizes e influenciar nas ações de gestão ambiental do espaço pode-se concluir que além da simples identificação dos impactos acústicos há a reação de boa parte dos usuários quanto a sua mitigação, seja por parte de professores (Tab. 5 e 6) ou alunos. (Tab. 7 e 8)

Reagem	81,82%
Não reagem	18,18%

Tabela 5 - Ocorrência de reação individual ao incômodo por ruídos de fundo internos à sala de aula – professores

Reagem	69,23%
Não reagem	30,77%

Tabela 6 - Ocorrência de reação individual ao incômodo por ruídos de fundo externo à sala de aula - professores

Reagem	58,43%
Não reagem	41,57%

Tabela 7 - Ocorrência de reação individual ao incômodo por ruídos de fundo internos à sala de aula – alunos

Reagem	51,83%
Não reagem	48,17%

Tabela 8 - Ocorrência de reação individual ao incômodo por ruídos de fundo externo à sala de aula – alunos

Isso indica um potencial de utilização de mecanismos que dependam da participação ativa dos usuários. Pode-se prever que a participação ativa nos processos de implementação de melhorias garanta a conservação e preservação de materiais e equipamentos a serem instalados, assim como, um potencial reconhecimento das melhorias. As hipóteses iniciais levantadas junto aos instrumentos subjetivos que podem ser diretrizes iniciais para a correção ou mitigação dos impactos acústicos são:

- Necessidade de diminuição da influência de ruídos internos e externos na sala de aula, que atualmente mascararam a comunicação verbal e contribuem para a dispersão e desconforto dos usuários do espaço, além de impactos na saúde do corpo docente;

- Tratamento das salas quanto aos problemas de inteligibilidade durante as atividades didáticas, que atualmente comprometem o entendimento da fala e contribuem para a dispersão e desconforto dos usuários do espaço, além de impactos na saúde do corpo docente.

Conclui-se que ações diretas no sentido da gestão dos espaços da universidade e da faculdade são identificadas nos instrumentos subjetivos e podem gerar ações no sentido da educação ambiental:

- Os alunos poderiam ser orientados no sentido de não se reunirem nos corredores durante os períodos de aula, ocupando para conversas e reuniões ocasionais o espaço do hall da escadaria ou o espaço do atelier, onde as atividades corriqueiras não seriam comprometidas pela influência destas ações.

- A diminuição da velocidade dos ventiladores em 20% muitas vezes diminui em 90% o ruído gerado por eles e também poderia ser alvo de discussão entre os usuários.

- A criação de políticas ambientais que provoquem a discussão da conduta acústica ética, junto ao corpo discente e docente da faculdade. Pode iniciar-se pela discussão sobre o uso dos aparelhos celulares durante as aulas, assim como abordar as conversas paralelas. Estas ações poderiam ser ampliadas para todo o Campus, por exemplo, abordando o ruído por uso de carros com som ligado acima dos limites necessários para a audição pelos seus passageiros nas ruas do entorno do Campus.

- Programas que sensibilizem os professores para notarem-se como "profissionais da voz" e da necessidade de ações no sentido do uso correto do aparelho fonador e para os procedimentos básicos da higiene vocal são importantíssimos.

- Gerenciar as atividades do Campus de forma integrada e considerando a interferência entre os edifícios do ginásio de esportes, da piscina e da sala de musculação são medidas que mitigariam os impactos externos sem qualquer custo inicial. Pode iniciar-se imediatamente pela adequação de calendários e horários de aulas e competições realizadas na piscina e ginásio.

Conclui-se também que ações diretas no sentido da interferência física dos espaços da universidade e da faculdade são identificadas nos instrumentos subjetivos:

- Intervenções relacionadas ao corredor interno no sentido de diminuir a interferência dos ruídos gerados neste espaço em relação ao interior das salas de aula.

- Adequações dos pisos das salas de aula e dos seus mobiliários são identificadas nos instrumentos subjetivos, pois barulhos provenientes de ruídos das carteiras são mencionados por alunos e professores.

- O nível de ruído dos ventiladores pode ser analisado, daí, tomadas atitudes no sentido de programação de manutenções temporárias com o objetivo de evitar a

emissão de ruídos por vibrações oriundas de problemas mecânicos. Diretriz que indiquem futuras aquisições de aparelhos de baixo nível de ruído, em médio prazo, é boa alternativa para minorar o impacto dos ventiladores.

- A melhoria da capacidade de absorção dos revestimentos da sala é medida que auxiliaria no sentido da diminuição do potencial de impactos de inteligibilidade e em menor escala dos ruídos internos. Depende do aprofundamento da prospecção acústica e de uma avaliação mais profunda, já que os instrumentos subjetivos não esgotam a questão. O cálculo do tempo de reverberação das salas, medida inicial desenvolvida junto aos instrumentos objetivos, também indica a necessidade da ampliação do potencial de absorção dos revestimentos.

Quando a pesquisa aborda as questões de qualidade acústica do atelier, conclui-se que os impactos identificados pelos alunos estão de acordo com as hipóteses levantadas pelos levantamentos "in loco". Os instrumentos subjetivos demonstram-se eficientes, porém, quando se analisa a importância dada ao impacto, verifica-se que as turmas mais antigas de alunos tendem a valorizar mais as interferências em relação àquelas que estão iniciando o curso.

Vale ressaltar que a surpresa em relação aos instrumentos objetivos foi a citação, por parte dos alunos, da interferência de ruídos externos no atelier, o que havia sido desconsiderado. Essa identificação leva a necessidade de aprofundamento da verificação da interferência de ruídos externos incluindo-se o atelier em futuras medições.

Nas respostas dos professores e nas demais considerações dos alunos, quanto aos impactos no atelier, os pontos levantados pelos instrumentos objetivos são ratificados: necessidade de elementos que contribuam na absorção da energia sonora, isolamento entre cobertura metálica e ambiente interno e sua compartimentação acústica, possibilitando eventos simultâneos.

Sobre o exercício realizado conclui-se que as contribuições dos instrumentos subjetivos utilizados, no sentido propositivo são válidas.

Conclui-se que como primeiro passo no sentido da delimitação dos problemas acústicos do espaço pela instituição, os dados obtidos junto aos usuários, abordando o seu desempenho nas atividades didáticas é confiável e os questionários junto ao corpo docente e discente são complementares.

As perguntas de caráter classificatório da sala de aula (item 6 do questionário aos alunos e item 5 do questionário aos professores) não se mostraram significativas para as conclusões quanto ao impacto acústico vivido pelos respondentes. Considera-se que elas poderiam ser retiradas do questionário sem trazer prejuízos para a pesquisa.

Notadamente os aspectos dos ruídos de fundo foram mais bem delimitados do que os demais, relacionados ao condicionamento sonoro no recinto, isso se considerando a contribuição propositiva, porém, sob a ótica da análise do impacto acústico no desempenho pessoal, a identificação de problemas de audibilidade e compreensão das palavras entre os usuários do espaço da sala de aula é significativa.

Conclui-se que a delimitação do real impacto do espaço em relação à audibilidade só será possível com medições acústicas, assim como, o nível das ações em relação à interferência dos ruídos de fundo serão eficazes na medida em que se estabelecerem comparações dos dados quantitativos "in loco" previstos pelas normas técnicas.

O próximo passo no sentido do desenvolvimento de instrumentos de sensibilização da comunidade acadêmica, para a importância da utilização de elementos do conforto acústico no exercício do projeto de arquitetura, é a avaliação da real contribuição da experiência descrita neste artigo junto aos estudantes e professores. A sua implantação como instrumento didático regular, de discussão ambiental e de gestão acústica do espaço também deve ser .

REFERÊNCIAS BIBLIOGRÁFICAS

[1] ARTIGAS, J. B. V. **A Função Social do Arquiteto**. São Paulo: Nobel, 1989.

[2] BRASIL. Ministério da Educação e Cultura. **Portaria 1.770/94**. Trata das diretrizes curriculares para cursos de arquitetura e urbanismo. Brasília: DF, 1994.

[3] CARVALHO, B. A. **Acústica aplicada à Arquitetura.** Rio de Janeiro: Livraria Freitas Bastos, 1967.

[4] DE MARCO, C. S. **Elementos de Acústica Arquitetônica.** São Paulo: Nobel, 1982.

[5] SILVA, P. **Acústica Arquitetônica & Condicionamento de Ar.** Belo Horizonte: Termo Acústica Ltda., 1997.

[6] PEREIRA, F. O. R.; BITTENCOURT, L. **Configuração de Laboratórios de Conforto Ambiental e Preservação de Energia**. In: IX Congresso Nacional da Associação Brasileira de Escolas de Arquitetura - XVI Encontro Nacional Sobre Ensino de Arquitetura e Urbanismo – UEL. Londrina, PR. Novembro, 1.999.

[7] MEIRA, M. E. **Laboratórios, LABINF / LABCON / LABTEC: Configurações Preconizadas**. In: IX Congresso Nacional da Associação Brasileira de Escolas de Arquitetura - XVI Encontro Nacional Sobre Ensino de Arquitetura e Urbanismo – UEL. Londrina, PR. Novembro, 1.999.

[8] BRASIL. Ministério da Educação e Cultura. **Portaria 1.770/94**. Trata das diretrizes curriculares para cursos de arquitetura e urbanismo. Brasília: DF, 1994.

[9] PEREIRA, F. O. R.; BITTENCOURT, L. **Configuração de Laboratórios de Conforto Ambiental e Preservação de Energia**. In: IX Congresso Nacional da Associação Brasileira de Escolas de Arquitetura - XVI Encontro Nacional Sobre Ensino de Arquitetura e Urbanismo – UEL. Londrina, PR. Novembro, 1.999.

[10] BEHLAU, M., DRAGONE M. L. S. e NAGANO L. **A Voz que Ensina**. Rio de Janeiro: Revinter, 2004.

[11] BRASIL. Ministério da Educação e Cultura. **Lei de Diretrizes e Bases da Educação Nacional (LDB), Lei 9394/96.** Brasília: DF, 1996.

Sessão 4

Síntese, Modelagem de Instrumentos e Computação Musical

(Synthesis, Instrument modelling and Computer Music)

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Population-Based Generative Synthesis: A Real-Time Texture Synthesizer based on Real-World Sound Streams

César Costa1,2, Jonatas Manzolli1, Fernando Von Zuben2
1Interdisciplinary Nucleus for Sound Studies (NICS)
2Laboratory of Bioinformatics and Bio-inspired Computing (LBiC/FEEC)
University of Campinas (Unicamp)
PO Box 6101, 13083-970, Campinas, SP, Brazil
{cesar;jonatas}@nics.unicamp.br, vonzuben@dca.fee.unicamp.br

ABSTRACT

The Population-Based Generative Synthesis (PBGS) is a real-time texture synthesizer - based on granular synthesis - with a novel grain generation methodology. Real-world sound streams are used as a systemic control source, bringing more versatility to the task of representing the final sonic objective. Therefore, PBGS is a perceptual-friendly alternative to parametric methods of synthesis. Bio-inspired algorithms are conceived to self-organize a population of sound grains in response to sonority and dynamical compositional stimuli. Based on a variety of experiments, the outcome of the PBGS device resembles complex textures with a colorful timbre palette, and inherits sonic attributes from the provided control references.

INTRODUCTION

Xenakis' Screens [15] and subsequent Granular Synthesis [14] surged on the 70's as a new sound generative paradigm bringing more complexity and colorfulness to digitally generated audio. It is based on Gabor's discoveries on the limitations of human's fast frequency variation perception (acoustic quanta theory) [9]. An analogy to the acoustic quanta theory is shown in Figure 1. Human visual space resolution has equivalent limitations. On the left, a low-resolution quarter of circle is shown and quantization could be easily perceived. On the right, a high-resolution image is presented. Although quantized, it invokes a continuum perception. The way sound is perceived is equivalently limited, being in frequency or in time.

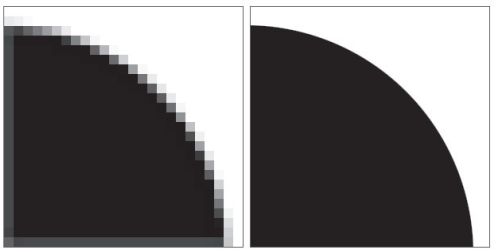


Figure 1. Effect of resolution on perception.

Xenakis wrote that complex sounds could be reproduced by playing a book of screens with a regular rate (just like a movie with frames, see Figure 2). He defines a screen as a low-duration sound with well defined spectrum distribution. In his work, a stochastic generative

methodology oriented by deterministic events is applied to the screen generation process.

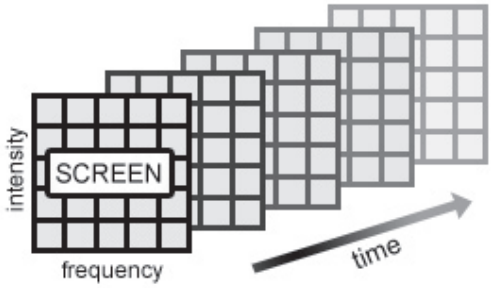


Figure 2. Book of Screens: sound seen as a movie.

Xenakis' method allows the user to compose sound material with rich spectral and dynamical complexity. However, due to its parametrical nature, it is quite limited concerning intuitiveness of user interaction. Other traditional granular synthesizers suffer from the same limitation. The so-called Ecologically-based GS [11] has arised as an alternative paradigm correlating synthesis methology with dynamic and perception of natural sounds, but the sound organization is still assigned to the user. Hence, to synthesize a desired sonority it is necessary to know how to properly organize the sonic material. Nonetheless, the use of natural sounds makes Ecologically-based GS the approach more akin to the one to be presented here.

As will be explained in the section devoted to the bio-inspired model, bio-inspired computation allows the integration between sonic features and compositional strategies, controlling various aspects in the evolution of a population of sound material. We developed a sonic control model based on a population-based search where we envisaged that a composer, helped by a bio-inspired algorithm, will be able to find a stimulating diversity of sounds. Given inherent self-organization on sound populations, we hope to generate variety and complexity in the sound domain such as biological systems produce [8].

The paper is organized as follows. The next section presents an overview of the Population-Based Generative Synthesis (PBGS), followed by the presentation of relevant aspects surrounding bio-inspired models. Next, a description of the implementation is outlined, followed by the experiments and the analysis of the obtained results. Some concluding remarks are then presented in the last section.

OVERVIEW OF THE PBGS METHOD

On PBGS we take advantage of Xenakis model synthesis capabilities, explored in the context of a new interface paradigm. We defined sonic scenario (SS) as the group of sounds featured with a certain set of sonic qualities. The composer expects the output material to be included in a desired sonic scenario. Instead of controlling numerical attributes in a parametric interface, we adopt bio-inspired models as strategies to create distinct sonic control layers. The essence of our approach has already been explored in other contexts by the same research group [3,4,5]. We have replaced Xenakis' original stochastic frame generation process by a bio-inspired algorithm, with unusual and strongly desired attributes like diversity maintenance and advanced search capabilities in feature spaces.

Our proposal is to use real-world sound streams as a way of representing a desired sonority and defining the objective sonic scenario. We apply bio-inspired techniques to adapt the synthesizer behavior in order to make it capable of producing sonic material associated with a specified sonic scenario.

Going deeper on the application of real-world sound streams, they are also used as dynamical control of the synthesizer. The goal is not only to promote the achievement of complex behavioral sound, but also to control the synthesis with desired complexity.

BIO-INSPIRED MODEL

To provide the functioning reported above, it is necessary to find a methodology to automatically extract sonic features from a screen sequence and store them in a computer based structure. This extraction procedure is a hard task due to its high-dimensionality and to the fuzzy notion of what should be a relevant sonic feature for human perception. It is also necessary to develop a screen sequence generation technique guided by these sonic features. These demands are not fulfilled by exact mathematical procedures.

Bio-inspired computation is a set of techniques based on natural processes such as evolution, self-organization and social behavior. The purpose is to bring, by means of computer simulation, attributes like self-adaptation. Our aim is to exploit transforming environments and self-regulation to develop new operational conditions [8]. Some common applications that have some relation to our needs are self-organization (in the self-organizing process of the Representative Structure) and pattern recognition (when automatically obtaining the relevant features).

A population-based approach has been adopted. The idea is to obtain the most representative population of screens which could identify different details of the representative set. This way, the sonic features can be stored in the form of reference prototypes. The Representative Structure would be composed of a population of screens. In this task, self-organization has an important role on the process of identifying, organizing and separating screens with different features. These are well-known attributes of Self-Organizing Maps (SOM) [12]. However, we have tried some alternative population-based self-organizing algorithms, based on Artificial Immune Systems (AIS) [6] and evolutionary computation (EC) [10]. Under the existence of reference prototypes, the self-organizing process in denoted in the literature as Learning Vector Quantization (LVQ) [13]. Figure 3 depicts the outcome of a two-dimensional LVQ process. The gray circles are the input samples that will be represented by the black circles. Of course, the two-dimensional scenario should be interpreted solely as a pictorial view of what would happen

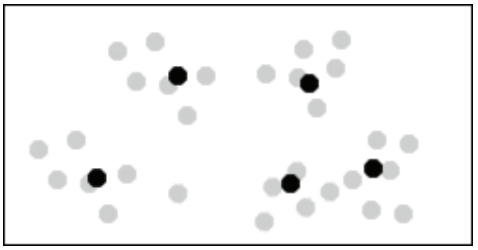


Figure 3. Learning Vector Quantization bi-dimensional graphical sample. Gray: input samples. Black: representative population.

in practice, with the gray and black circles residing in spaces of a much higher dimension.

The black circles correspond to the population of prototypes that will pass through a self-organizing process responsible for the final spatial configuration presented in Figure 3. Notice that the black circles are organized to capture the most relevant aspects of the input samples. They are called representative prototypes because they can be interpreted as concise representations of the input samples, generally expressing a consensual explanation of the local variability in the neighboring input samples.

Self-Organized Map (SOM)

Results in Figure 3 can be obtained by means of a self-organizing map (SOM). A Kohonen's SOM associates high-dimensional data with a population of output nodes arranged in a low-dimensional grid. Output nodes are extensively interconnected with many local connections. Based on neuron's organization principles, topologically close nodes are sensitive to physically similar stimulus. Thus, the output nodes are ordered in a natural manner without external interference in a process called unsupervised learning. After a repeated presentation of the input dataset, output node positions will specify clusters or vector centers that sample the input space such that the density function of the vector centers tends to approximate the probability density function of the input vectors [12]. A deeper explanation can be found in [5] where SOM has been applied in a timbre design methodology.

Artificial Immune Systems (AIS)

Artificial Immune Algorithms are adaptive procedures inspired by the biological immune system and devoted to the solution of challenging computational problems [6]. Biological Immune Systems are capable of recognizing a wide range of antigens with a reduced number of antibodies, applying two mechanisms: clonal selection and affinity maturation. Once these principles are applied in the realm of computer systems, it is possible to create a limited population of digital antibodies to represent a wide rage of digital antigens (or input data). AIS has already been used in sonic applications as reported in [3]. Antibody networks for self-organization are similar to self-organizing maps, except for the absence of a local neighborhood to guide the interaction of the antibodies. Besides, the size of the population is self-regulated [7].

Evolutionary Computation (EC)

The Genetic Algorithm (GA) is an Evolutionary Computation paradigm that consists of a set of computational techniques based on Darwin's Evolutionary Theory and the survival of the fittest principle. Given a population of individuals whose physical features are coded in a digital DNA, simple genetic operators like mutation, crossover and selection are repeatedly applied to produce the next generations. The fitness of each individual in the population is provided by an objective function. The genetic operators promote a parallel exploration of the search space with a concentration of the individuals in the most promising regions, i.e. regions whose samples are given high fitness values. It happens because individual with high fitness values are favored in the reproduction phase, having a higher probability of spreading his genetic material to the future generations. On PBGS, the fitness of an individual is proportional to its similarity to those on the sonority reference input screen sequence. In [4], GA is applied in a sound synthesis method and the paper supplies important considerations about its use in sonic applications.

Contrary to traditional applications of GA, PBGS is interested in the whole population and not solely in the best individual of the population. Notice that, given the fitness function, the population at a given generation operates as an LVQ device.

THE ARCHITECTURE

The task of PBGS is to produce sound material guided by a reference dynamic and that could be included in a given sonic scenario. The architecture is presented in Figure 4. On PBGS, we propose that the composer expresses his desired sonic scenario into a set of sound samples arranged in a sequential sound stream, named Sonority Reference (SR).

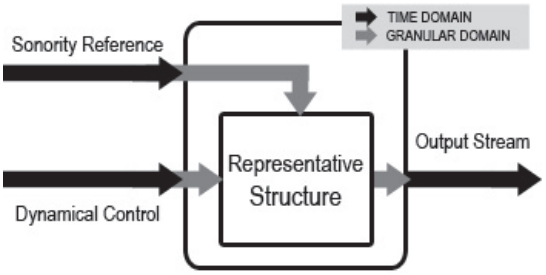


Figure 4. The PBGS Architecture

As screens, we have used low-duration sample frames extracted from a source stream and windowed by a Gaussian-like envelope. A sound stream converted into a screen sequence is said to be on a Granular Domain (GD).

The first action of the system is to convert the sonority reference into a screen sequence in the granular domain. At this point, bio-inspired algorithms are applied to the sequence with the purpose of extracting prototypes with noticeable features and storing them in a computational structure denoted Representative Structure (RS). To accomplish this task, self-organizing maps or artificial immune systems could be considered in isolation or integrated in a hybrid framework. In the experiments to be presented, self-organizing maps have been considered in isolation.

PBGS provides a second control level associated with the synthesis process. The Dynamical Control (DC) input receives a sound stream that works as a guideline for the output generation. Just as the sonority reference, the dynamical control is converted into a screen sequence in the granular domain.

Further, the synthesizer applies the dynamical control sequence to generate prototypes in the Representative Structure in order to obtain a screen sequence that once converted to a sound stream must be included in the sonic scenario expressed in the sonority reference, and having its dynamic related to the dynamical control. At this stage, an evolutionary algorithm is implemented, so that the output stream is composed of individuals with better fitness extracted from a population of prototypes at a given generation of the evolutionary algorithm. The dynamical control screen sequence acts as a setpoint. It is expected that the output screen carries sonic features provided by the

population at the Representative Structure and follows the dynamics specified by the dynamical control.

An interesting feature of PBGS architecture is that both main procedures, sonority reference LVQ and output generation, could flow independently. So, it is possible to vary system sonority during output generation. Thus, the synthesizer allows real-time operation in both of its inputs. It is possible to vary output sonority maintaining the learning process during presentation. In the other input, working with the dynamic guidance gives the opportunity to the composer to operate the synthesizer as a musical instrument.

Screen Comparison

All mechanisms presented for the self-organizing procedure of the Representative Structure needs a specific metric to compare its individuals. Our approach is to calculate similarity on spectral domain applying the traditional FFT algorithm. Thus, for optimal performance grain sizes are chosen to be power of two.

Screen Context

We could not see an individual screen isolated in time since time evolution is one of the most remarkable features of sound for our perception. Thus, we define a Screen Context as the temporal circumstances that trigged the appearance of a certain spectral event. Again, determining what relevant features must be considered is a fuzzy task.

In our method, the individuals used in the population were composed of the screens itself and their respective context. The context is implementation-specific and its completeness may vary according to the computational resources available.

IMPLEMENTATION

The PBGS was implemented on two different architectures. At first, a non real-time prototype on the MATLAB environment was conceived, intended to work as a base for PBGS' architecture development. Afterwards, a C++ version under LINUX OS was programmed to yield real-time performance.

MATLAB version

In the first attempt, the MATLAB environment has been chosen due to its easiness of reusing already available bio-inspired algorithms (developed by the research group) and signal processing tools. It has been focused on the development of the architecture and in the set up of algorithm details, having no real-time performance requisites. The resultant software has two modulates: one for the RS training and another for the synthesis process itself. On this implementation, the sound streams were coded in 16-BIT PCM and encapsulated on WAVE audio format.

On the first module, a SOM algorithm from Helsinki University of Technology CIS SOM Toolbox[1] [1] was used. It receives as input the sonority reference stream and functional parameters of the learning algorithm: grain size (in samples), population size (number of SOM's neurons) and training epochs (number of times that a grains is presented to the SOM). As output, it returns a population of grains that works as the RS.

The second module receives as input the RS and the dynamical control stream. It chops the input stream in a grain sequence which is submitted to the SOM algorithm. A sequence of best match grains (SOM's best matching units) is obtained as a result and the output stream is then reconstructed by an overlap technique.

This implementation is sample rate independent. However, the frequency rate must be equal on both sonority reference and dynamical control streams.

C++ version

Focusing on the real-time performance, a second implementation was developed on C++ to work on Linux OS with PortAudio Sound API[2] [2]. The main difference from the MATLAB version is that the training and the synthesis modules could work in parallel as different threads, allowing real-time sonority variation. Also, it has to be optimized to avoid unnecessary latency to output. At this time, an evolutionary algorithm was adopted to perform LVQ.

The real-time implementation uses PCM 16bit coded audio originated by a live microphone input or a RAW file for both sonority reference and dynamical control. The output could be directed to soundcard output, to a RAW file or both.

EXPERIMENTS AND RESULTS

Four experiments have been considered and are listed in Table 1. Table 2 presents experiments' parametric space, considering: grain size (GS), population size (PS) and the sonic population variety (SPV).

Exp.	*Objective*
1	Verify sonority and dynamic transference to output
2	Verify if real-time performance can be achieved
3	Verify the influence of system parameters on behavior
4	Verify spectral and dynamical tracking behavior

Table 1. Experiments and Objectives

Exp.	*Parametric Space*		
	GS (ms)	PS	SPV
1	22	128	High
2	11-92	128/256/512	Low
3	11-92	32/64/128/256/512	Low
4	11-92	128	Fixed

Table 2. Parametric Space: GS (grain size); PS (population size, in power of two); SPV (sonic population variety, i.e., number of sounds in the population from different sources).

Experiment 1

Using MATLAB simulation, we verified if there were traces of the sonority reference at the output stream and also if the dynamical control was operating correctly. We used three different sonic scenarios: a male voice, a guitar solo and a synthetic harmonically well-defined sound. They were cross-presented to both inputs and the output was further analyzed.

[1] http://www.cis.hut.fi/projects/somtoolbox

[2] http://www.portaudio.com/

The results indicate that the obtained output presents relevant features derived from the sonority reference and the dynamics inherent to the control stimuli guided the generation of the sound material. This effect can be verified even visually using a sonogram (see Figure 5). Please refer to the online reference[3] for the sound files and all the results. On Figure 5, the synthetic sound was used as a sonority reference and the voice was used as a dynamical guideline. On the left, the voice signal is in gray and the output is in black. On the right, the output sonogram shows that the high-energy peaks, generally associated with voice sounds, are present, but mixed with harmonically well-defined lines, characteristic of the sonority reference.

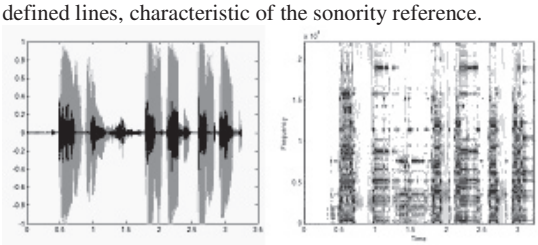


Figure 5. Experiment 1 Output: voice as dynamical control and harmonic sound as sonority reference. Left, output dynamic in black and control dynamic in gray. Right, output sonogram.

On some runs, with a voice used as a dynamical control, it was possible to discern the phrase spelt at the same time that the sonority variance could be recognized.

As a perceptual comparison, three listeners confirmed the existence of traces of reference's sonority on resultant sound material.

Experiment 2

On real-time implementation we have estimated the computational demand of the method, with a clear indication that real-time performance can be achieved without much effort. Larger sound streams were used for RS training and a microphone as dynamical control. Processor usage and latency times were verified.

Concerning the performance, in a mid-range personal computer, it had no problems on running in real-time. It has used a maximum rate of 5% on an INTEL PENTIUM IV 2.2GHz with an overall rate of 3%. On the learning task, it trained a 4 minutes file in an overall rate of 1 minute with the worst result lasting 1'06". During execution there weren't experienced sound faults. The latency observed was caused by the accumulation of the grains and that was expected (always < 100ms). The latency could be calculated since the dynamical of the input and the output were very similar.

Regarding the influence of parametric variation over system performance, it was observed that the increase in population size implies a higher computational demand. Also, a smaller grain configuration implies a higher grain density over time. Thus, it demands a more intense populational search and consequently it becomes more computationally expensive. It is important to notice that the grain size has little effect on a single populational search due to the FFT computation complexity nature.

Experiment 3

We verified how sensitive the synthesis behavior and the human perception are with respect to variation in the system's parameters. Regarding the influence of grain size on perception, the experiments have shown that smaller grains implies in a poorer frequency definition (perceived in both listening and visual media). Figure 6 shows the result of the execution with a small grain – 11ms (left) and with a large grain – 185ms (right).

Concerning to population size, output sound complexness decreased dramatically when using few individuals. With the increase of population size it had low effect over sound complexity and caused a noticeable depreciation on system performance.

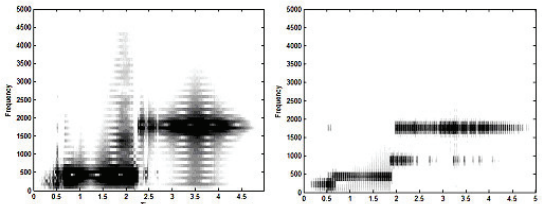


Figure 6. Influence of Grain Size: sonogram with small grains (left), and with large grains (right)

Experiment 4

This experiment was conceived in order to better comprehend how the synthesizer acts on some specific circumstances. We intended to test the dynamic and spectral tracking capability (i.e. we presented sound with well defined spectral distribution and dynamical behavior and analyzed the output). For this execution (see Figure 7), sine samples have been considered as a sonority reference (left) and a sine-based linear spectral evolution sound as a dynamical guidance (right).

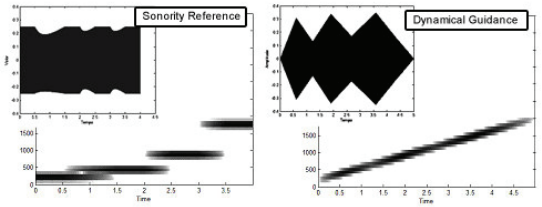


Figure 7. Tracking Experiment: Left, Sonority Reference in time (top-left) and in spectrum (back). Right, Dynamical Control.

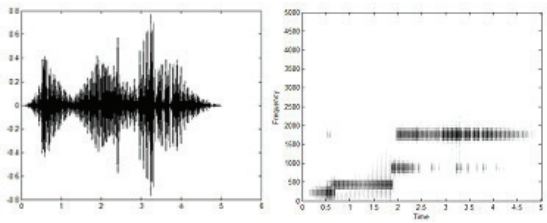


Figure 8. Experiment 4: output's dynamic (left); and spectral behavior (right).

The tracking experiment (Figure 7) produced the results depicted in Figure 8. The resulting sound successfully followed the reference dynamics, being more accurate with smaller grains. The sonogram shows that the spectral evolution had a positive slope, just as the control stimuli, and was composed of well defined sinusoidal samples.

DISCUSSION

A summary of experimental results is shown in Table 3. Experiment 1 indicated that PBGS successfully preserves reference's sonority features on the output. This is the main

[3] http://www.nics.unicamp.br/~cesar/granular

functionality of the proposed method together with the control features.

On Experiment 3, grain sizes have affected quality of the sonority transference. Small grain setup implies poor output-reference correlation. However, as seen on Experiment 4, it produced sound material with high dynamical fidelity related to control stream. Also, small grains made screen context representation overweigh perceptual inexpressive details, i.e., depreciating more structured - and relevant to sonority perception – dynamical behavior. On the other side, excessive large grains may force the screen context to ignore fast dynamical nuances. It was also verified on Experiment 2 that small grains are more computational expensive.

Exp.	*Results*
1	Sonograms and perceptual inspection have been achieved.
2	Real-Time was made feasible. ↑Grain Size → ↓Computational Effort; ↑Population Size → ↑Computational Effort.
3	↑Grain Size → ↑Frequency Definition; ↑Population Size → ↑Spectral Variations
4	Output successfully tracks dynamic control guidance. ↑Grain Size → ↑Dynamic Fidelity

Table 3. Summary of experimental results. Legend: ↑ increase, ↓ decrease, → implies.

Experiment 3 indicated that small populations produce poor spectral variation at the output. Complex sounds emphasize such property better than pure sine waves streams. Enriched sound scenarios required an increment in the size of the population in order to be correctly represented. When using larger-size populations with simpler sounds, much of the representative power is wasted and many individuals stored redundant data. This is actually a common problem with LVQ procedures [13].

CONCLUSION

The PBGS method has been proposed, implemented and the obtained results have been analyzed. The obtained sound at the output of the synthesizer inherits sonic qualities from the reference provided. We have developed a bio-inspired synthesis procedure that is not dependent on composer capability of translating his sonic expectations into a parametric and well-structured mathematical domain.

Electroacustic composition and real time performance are straightforward applications of the PBGS. Also, by working on a real-time basis it opens new possibilities for computational-based synthesizers. Based on a sound stream control paradigm, the control stimuli can be provided by other sound interfaces on live presentations or improvisation. For example, it could be controlled by a guitar in a rock solo, working like a varying sonority effect processor.

Regarding future perspectives, they include the development of more elaborate bio-inspired algorithms, more comprehensive screen contexts and the conception of alternative synthesis techniques. We also intend to further release a user-friendly software package based on the C++ implementation.

ACKNOWLEDGMENTS

This work has been supported by grants from Fapesp and CNPq.

REFERENCES

[1] Alhoniemi, E., Himberg, J., Parhankangas J. and Vesanto J. *SOM Toolbox for MATLAB 5* Report A57 Libella Oy, Espoo 2000, April 2000.

[2] Bencina, Ross and Burk, Phil, PortAudio - an Open Source Cross Platform Audio API, *In Proceedings of the 2001 International Computer Music Conference*, Havana Cuba, September 2001. pp. 263-266.

[3] Caetano, M., Manzolli, J., Von Zuben, F. J. Application of an Artificial Immune System in a Compositional Timbre Design Technique, in *Proceedings of ICARIS 2005*, Alberta, Canada. In Press. 2005.

[4] Caetano, M., Manzolli, J., Von Zuben, F. J. Interactive Control of Evolution Applied to Sound Synthesis. In *Proceedings of the 18th International Florida Artificial Intelligence Research Society* (FLAIRS), Clearwater Beach, EUA. 2005.

[5] Caetano, M., Costa, R.C., Manzolli, J., and Von Zuben, F. J. Self-Organizing Topological Timbral Design Methodology Using a Kohonen Neural Network. In *Proceedings of the 10th Brazilian Symposium on Computer Music (SBCM), Belo Horizonte, Brazil, 2005, 94-105.*

[6] de Castro, L. N., Timmis, J. *Artificial Immune Systems: A New Computational Intelligence Approach*, Springer-Verlag, 2002.

[7] de Castro, L.N., Von Zuben, F.J. aiNet: An Artificial Immune Network for Data Analysis. *in* Abbass, H.A., Sarker, R.A. & Newton, C.S. (eds.) *Data Mining: A Heuristic Approach*, Idea Group Publishing, pp. 231-259, 2002

[8] de Castro, L.N., Von Zuben, F.J. (eds.) *Recent Developments in Biologically Inspired Computing*. Idea Group Inc., 2004.

[9] Gabor, D. Acoustical Quanta and the Theory of Hearing. *Nature*, vol. 159, 1946, 591-594.

[10] Goldberg, D.E. *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley. 1989.

[11] Keller, D., Truax, B. Ecologically-based Granular Synthesis. *ICMC 1998*, Ann Arbor, Michigan. 1998.

[12] Kohonen, T. *Self-organizing maps*. Springer-Verlag. 2000.

[13] Kohonen, T. *Learning Vector Quantization for Pattern Recognition*. Technical Report TKK-F-A601. Helsinki University of Technology, 1986

[14] Roads, C. Introduction to Granular Synthesis, *Computer Music Journal* vol.12 n.2, 1988.

[15] Xenakis, I. *Formalized Music*. Indiana University Press. 1971.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Síntese por Modelagem Física de Instrumentos de Sopro

Luís Carlos de Oliveira[1], Ricardo Goldemberg[2], Jônatas Manzolli[2]
[1]FEEC-NICS-Bolsita do CNPq, [2]IA-NICS; UNICAMP
CEP: 13083-970, Campinas, SP, Brasil
{luis,rgoldem,jonatas}@nics.unicamp.br

RESUMO

Este artigo está centrado na revisão bibliográfica de métodos de síntese de som de instrumentos musicais de sopro, especificamente o naipe das madeiras; clarinetas, os saxofones entre outros. A síntese por modelagem física é uma técnica que vem ampliando seu grau de importância, pois oferece maior interação entre o músico e o modelo computacional que representa o instrumento simulado. Este artigo é dividido em três secções: Síntese Musical por Modelagem Física, Modelagem Física de Instrumentos de Sopro e Métodos Experimentais necessários para determinação e análise de parâmetros envolvidos no processo de geração sonora destes instrumentos.

INTRODUÇÃO

Desde 2003, temos investigado a natureza das sonoridades de instrumentos de sopro, principalmente da clarineta, utilizando um mecanismo de simulação experimental que tem comportamento físico análogo ao envolvido na performance de instrumentos de sopro. Durante a nossa pesquisa, percebemos que existem muitos fatores que são determinantes na construção e no entendimento de um modelo que possa elucidar todas as relações e variáveis intrínsicas ao processo de geração sonora dos instrumentos musicais.

Frente a esta complexidade, optamos por utilizar um método de pesquisa denominado Projeto Fatorial, que possibilitou comparar e verificar a importância relativa dos diversos fatores envolvidos na produção sonora dos instrumentos analisados. Todavia, ficou claro que o controle de todas as variáveis de um sistema experimental é um problema complexo o que nos estimulou a ampliar o escopo da nossa pesquisa no sentido de utilizar, também, simulação computacional para depois, com a mesma, podermos interpretar as medidas do nosso modelo experimental. Neste momento, nos pareceu relevante, iniciar um estudo da síntese sonora que mais se aproxima do nosso modelo experimental. Iniciamos a pesquisa com um levantamento de referências sobre a utilização de modelagem física para a construção de modelos de síntese sonora de instrumentos musicais.

Apesar da preocupação central deste trabalho residir na discussão e revisão da bibliografia sobre a síntese de instrumentos musicais de sopro, em particular, do naipe das madeiras (incluem a clarineta, o saxofone, o oboé, etc.), também são mencionados artigos voltados à síntese por modelagem física de outros instrumentos musicais. A trajetória adotada neste artigo parte de uma pequena exposição histórica sobre pesquisadores que se preocuparam em estudar os instrumentos mencionados do ponto de vista científico. Os comentários partem de trabalhos feitos a partir da metade do século XIX até meados do século XX.

Em seguida, serão apresentadas referências mais recentes divididas em três seções. A primeira, dividida em três sub-seções, refere-se às etapas envolvidas na modelagem física propriamente dita. Nela são detalhados os procedimentos envolvidos na elaboração de modelos físicos que têm aplicação para a síntese de instrumentos

musicais de sopro. São apresentadas algumas equações e estratégias adotadas na modelagem.

A segunda seção é dedicada aos artigos que têm por preocupação a determinação e análise dos parâmetros oriundos dos modelos físicos. Este tratamento é obtido especificamente através de experimentos. Nela são apresentados modelos empíricos para a amplitude e frequência de notas em três regiões distintas da clarineta. Estes resultados são componentes de nossa pesquisa.

A última seção trata genericamente dos algorítmos que abordam os modelos voltados para a síntese de instrumentos musicais. O foco está direcionado para os resultados na performance artística.

Finalmente, este artigo se encerra com um levantamento dos problemas apontados pelos diversos autores e são apresentadas algumas propostas para a continuidade deste trabalho.

PANORAMA HISTÓRICO

No tutorial dedicado ao estudo da modelagem física de instrumentos de sopro, Keefe [1] faz um pequeno apanhado histórico. Ele cita Helmholtz [2] onde, na primeira edição de "On the Sensations of Tone" de 1862, estabeleceu os princípios para classificar os instrumentos de sopro em duas classes: instrumentos com palheta ("reed pipe") e instrumentos sem palheta ("flue pipe"). Para os instrumentos com palheta ele fez uma divisão em três sub-classes: 1) palheta fixa, como o órgão de tubo com palheta e a gaita; 2) palheta construída de bambu (arundo donax), incluindo os de palheta simples como a clarineta e o saxofone e os de palheta dupla como o oboé e o fagote; 3) vibração labial, incluindo os instrumentos onde os lábios atuam com ação valvular como é o caso do trompete, trombone, trompa, etc. A segunda classe, a de instrumentos sem palheta, inclui as flautas e os órgãos de tubo sem palheta.

Quinze anos mais tarde, na edição de 1877, Helmholtz formulou teorias quantitativas sobre o mecanismo pelo qual oscilações são mantidas em tubos com palhetas. Este trabalho estabeleceu a base para toda pesquisa posterior sobre este assunto. Sua teoria, ao contrário de formular um modelo detalhado da dinâmica envolvida, incorporou restrições que precisavam ser satisfeitas para se criar oscilações em estado de regime permanente.

Pouco tempo depois, em 1894, Rayleigh [3] apontou sobre a importância do estudo de sistemas dinâmicos não–lineares para o desenvolvimento de teorias de instrumentos musicais. Ele elaborou a primeira descrição quantitativa de oscilações auto-sustentadas que serviram de pano de fundo para o estudo de processos mecânicos não-lineares e modelos de acústica musical a partir da década de 1960. Utilizando uma nova terminologia, analisou sistemas que possuem estreita ligação com o oscilador de Van der Pol. Ele mostrou que existe resistência negativa no processo de geração de oscilações auto-sustentadas, indicou a existência das bifurcações de Hopf, bem como desenvolveu a teoria das instabilidades transversas em jatos de ar.

Tanto Helmholtz quanto Rayleigh entenderam que a característica essencial para a sustentação de uma nota em um instrumento de sopro é a existência de dissipação – parte da energia é transmitida sob a forma de radiação acústica, mas a maior parte é perdida na forma de atrito e dissipação térmica. Desta forma, um instrumento de sopro necessita de uma fonte externa de suprimento de energia, pois o próprio processo de produção sonora consome a energia intrínsica do sistema. Quanto mais energia é suprida, mais é dissipada mas ainda assim a amplitude de oscilação cresce. Vale a recíproca quando a energia suprida diminui.

O trabalho de Bouasse [4] marca a transição entre os trabalhos desenvolvidos por Helmholtz e a era moderna.

Mais recentemente, Benade [5, 6, 7, 8, 9] também desenvolveu uma série de trabalhos teóricos e experimentais sobre instrumentos de sopro de madeira bem como um conjunto de modelos de tais instrumentos. Além dele, Fletcher e Rossing [10] detalharam modelos de vários outros instrumentos em um minucioso trabalho.

MODELAGEM FÍSICA

Segundo Smith [11], existem basicamente dois tipos de modelos físicos utilizados para a síntese de som de instrumentos musicais: os modelos globais ("lumped model") e os modelos distribuídos ("distributed model").

O modelo global consiste em equações que não descrevem microscopicamente os fenômenos envolvidos em um sistema. Ele é uma aproximação física global do sistema como por exemplo, o conjunto formado pela boquilha, lábios e palheta. Por outro lado, os modelos distribuídos têm por preocupação a descrição do fenômeno a nível microscópico e divide o sistema em blocos funcionais. Estas duas categorias de modelos podem tanto representar um sistema dinâmico, onde as propriedades variam com o tempo, quanto um sistema estático, onde não há variação de propriedades com o tempo.

Etapas de Modelagem

Keefe [1] enumera sete etapas na elaboração de um modelo no domínio do tempo. A primeira etapa (I) corresponde à formulação propriamente dita do sistema dinâmico que, para ele, é a mais crucial de todas. Várias simplificações devem ser consideradas no modelo com vistas a tornar o tratamento computacional factível.

O modelo dinâmico proposto por Keefe, válido tanto para clarinetas e saxofones (a diferença está na geometria) como para metais (a diferença está nos valores dos parâmetros), consiste de um sistema de três equações diferenciais ordinárias de primeira ordem acoplados por um hiato de tempo ("time delay").

As três variáveis consideradas fundamentais são: deslocamento da palheta (x), velocidade da palheta (ur) e vazão volumétrica através da abertura da palheta (u), todas representadas na Eq. (1). As demais variáveis são obtidas em função destas e o modelo dinâmico com as correspondentes equações está representado na Eq. (1). A nomenclatura das demais variáveis e parâmetros encontram-se na Tab. (1), com valores no S.I..

Ainda segundo Keefe [1], desconsiderando-se as propriedades do acoplamento temporal associadas com a resposta linear da coluna de ar (assumida por hipótese), o espaço de fase correspondente a este sistema dinâmico é tri-dimensional, pois há três variáveis fundamentais. A presença do hiato de tempo ("time delay") proporciona ao espaço de fase uma dimensão muito maior, porém, as notas musicais estão restritas a um subespaço (do espaço de fase) de dimensão menor.

Para os instrumentos de palheta (arundo donax) assume-se que esta fecha com o aumento da pressão de ar. Esta característica destes instrumentos é representada escolhendo δ=1 na Equação (1). Nos instrumentos de vibração labial (metais) assume-se que o processo é inverso e impõe-se a abertura com o aumento da pressão escolhendo-se δ=–1.

$$\dot{x}(t) = \frac{1}{S_r} u_r(t)$$

$$\dot{u}_r(t) = S_r \left\{ \begin{array}{l} -\left(g_r S_r^{-1} + \frac{\delta}{u_r} Z_c \right) u_r(t) - \omega_r^2 [x(t) - H] - \\ -\frac{\delta}{u_r} [P_0 - p_h(t)] - Z_c u(t) \end{array} \right\}$$

$$\dot{u}(t) = \frac{1}{I_e(x)} \left\{ \begin{array}{l} P_0 - p_h(t) - Z_c [u(t) - u_r(t)] - \\ -C|u|^{\alpha} x^{-\beta} Sgn(u) \end{array} \right\}$$

Eq. 1: Equações que configuram o modelo utilizado para simular clarineta. Ver Keefe [1].

Parâmetro	Denominação do Parâmetro
c	Velocidade do som
ρ	Densidade do ar
S	Área da coluna de ar da entrada
Z_c	$\rho c/S$, Impedância característica na entrada
S_r	Área dinâmica da palheta
ω_r	Frequência de ressonância da palheta (rad/s)
f_r	ω_r, freq. de ressonância da palheta em Hz
μ_r	Massa dinâmica por unidade de área da palheta
g_r	ω_r/Q_r Onde Q_r tem valor 3 para madeiras e é variável para metais
H	Abertura de equilíbrio da ponta da palheta
l	Comprimento da abertura da ponta da palheta
w	Largura da abertura da ponta da palheta
I_e	$pl/(wH)$, Inertância da ponta da palheta
C	Constante de controle de fluxo (44,4 para madeiras)
α	1,5 (palheta simples), 2(palheta dupla e metais)
β	2 (palhetas simples, dupla e metais)
P_0	Pressão do ar
$R(t)$	Função de reflexão da coluna de ar na sua entrada
$p_h(t)$	Pressão da boquilha convoluída com a função de reflexão da coluna de ar
δ	Para madeiras (1) para metais (-1)

Tab.1: Parâmetros e símbolos da Eq. 1.

A vazão volumétrica de ar que efetivamente passa pelo tubo é a diferença entre a vazão que chega até a abertura da palheta e a vazão que é "varrida" pela palheta, isto é:

$$u_d(t) = u(t) - u_r(t) \qquad (2)$$

A pressão na boquilha, $p(t)$, é calculada a partir de:

$$p(t) = p_h(t) + Z_c u_d(t) \qquad (3)$$

Onde a variável $p_h(t)$ guarda os valores passados da pressão através da convolução da função de reflexão da coluna de ar, $r(t)$, com a pressão da boquilha e vazão efetiva através da coluna, isto é:

$$p_h(t) = r(t) * [p(t) + Z_c u_d(t)] \qquad (4)$$

A teoria subjacente para a solução das equações (2), (3) e (4) está apresentada em McIntyre et alii [12].

A segunda etapa (II), bastante importante, diz respeito à escolha dos parâmetros envolvidos no modelo. Além das variáveis que o descrevem, existe um conjunto de parâmetros no sistema dinâmico. Por exemplo, a massa, dureza e umidade da palheta, geometria da coluna de ar, da palheta, etc. De acordo com Keefe [1], o parâmetro central é a pressão de ar que entra no tubo. Esta pressão representa a fonte externa de energia que contrabalança a perda por dissipação térmica e viscosa.

Do ponto de vista experimental, para compreender o processo de produção sonora faz-se necessário determinar a faixa de valores plausíveis, do ponto de vista físico, dos parâmetros. A resposta a este questionamento virá das duas etapas seguintes.

Análise Paramétrica

Um conjunto de valores plausíveis dos parâmetros é escolhido. Em seguida, estabelece-se uma condição inicial fisicamente viável para as três variáveis. Na seqüência, o sistema de equações diferenciais é integrado numericamente no tempo. A terceira etapa (III) corresponde à simulação no domínio do tempo (dinâmica).

Em qualquer instante o sistema dinâmico está em um ponto do espaço de fase e a evolução do sistema no tempo corresponde às trajetórias no espaço de fase (TEF). Após um período inicial de transientes, as TEF tendem a se aproximar de um *conjunto limite* que é dependente das condições iniciais e dos valores dos parâmetros escolhidos.

Dada a terminologia de sistemas dinâmicos não-lineares, um conjunto limite que pode ser observado experimentalmente é chamado de *atrator*. Um atrator periódico é o atrator cuja trajetória no espaço de fase descreve uma curva fechada. Benade e Kouzoupis [5] estabeleceram que "*um regime de oscilação é uma oscilação multicomponente, estável e não-linear, na qual vários picos de ressonância descrevem um controlador de fluxo para manter uma oscilação cujos componentes espectrais são membros de uma série harmônica exata*". Este é o conceito de um atrator periódico quando aplicado a instrumentos de sopro, afirma Keefe [13].

Uma vez estabelecida a simulação dinâmica, a etapa (IV) corresponde ao estudo da sensibilidade paramétrica. Isto é, deve-se estudar o quanto o sistema dinâmico é sensível a variações nos valores dos parâmetros. Provavelmente, alguns valores de parâmetros deverão ser obtidos através de dados experimentais e este é o foco da secção seguinte.

Com os valores dos parâmetros estabelecidos, pode-se seguir às três últimas etapas, onde novas questões podem ser levantadas: (V)simulação em tempo real (que depende da tecnologia de hardwares e softwares disponíveis); (VI)percepção e cognição musical onde é discutido o "quão" próximo de um instrumento real o sistema dinâmico está; ,e finalmente, (VII)aplicação no desenvolvimento de sonoridades de instrumentos musicais e performance. Este último será o assunto da penúltima seção deste artigo.

No seu artigo, Keefe [1] utilizou um oscilador harmônico simples como modelo para a palheta. Para resolver este conjunto de Equações Diferenciais Ordinárias foi utilizado um método numérico implícito de segunda ordem. Isto resultou numa única equação não linear que foi resolvida pela regra de Newton. O artigo apresenta os resultados da simulação para uma clarineta e analisa o

efeito de diversos fatores, como pressão na boquilha, deslocamento da ponta da palheta, etc sobre a sonoridade.

Modelagem por Waveguides

No entanto, antes de passarmos às secções seguintes, vale a pena fazermos um corte nesta exposição para analisarmos uma modelagem distinta. Ela tem particular interesse para o propósito de síntese musical. Trata-se da modelagem através de "waveguides".

Borin et al. [14] apresentaram que os modelos físicos de instrumentos musicais podem ter suas partes decompostas, geralmente, em dois blocos: de *ressonância* ("resonator") e de *excitação* ("excitation").

Os modelos utilizados para os blocos de excitação, como para a palheta da boquilha de uma clarineta ou para o arco em contato com a corda de um violino são geralmente não lineares. A descrição de um bloco de ressonância, sem perda de generalidades, é redutível a um sistema dinâmico linear cujas características formam a base para aplicação em análise de instrumentos e síntese musical. Um dos modelos mais eficientes para este bloco é o modelo "waveguide". Ele modela a propagação de onda em um meio distribuído como cordas, tubos e instrumentos de sopro, ,segundo Smith [11].

Os modelos globais são implementados, para síntese sonora, comumente por filtros digitais de segunda ordem. Por outro lado, os modelos distribuídos são implementados por linhas de atraso (*delay lines*), que são denominadas por "digital waveguides" quando usadas em modelagem física.

Os modelos distribuídos podem ser combinados livremente com os modelos globais, sempre segundo Smith. Por exemplo, a modelagem de um saxofone pode consistir de um modelo global para o conjunto palheta-boquilha e um modelo distribuído para o tubo.

Desta forma, estruturas complexas podem ser construídas através da montagem e acoplamento destes elementos. Aí reside a sua importância para a síntese musical.

Outro exemplo de modelagem do elemento de ressonância por "waveguide" pode ser encontrada no trabalho de Ducasse [15, 16]. Ele afirma que a simulação no domínio do tempo (dinâmica) da operação física de instrumentos musicais permite criar transitórios e fenômenos perceptivos que são difíceis de se obter por outro método de processamento de dados.

No seu artigo ele sugere fazer a modelagem através de uma estrutura modular. O instrumento com palheta simples é constituído por módulos cujos elementos se interconectam. Ele inova quando inclui no seu modelo de boquilha com palheta simples a ação do instrumentista, representado pela ação da língua, dos lábios e da respiração.

A Fig. 1 mostra os quatros módulos elementares do modelo: 1)Boquilha, 2)Tubo, 3)Furo com Cobertura e 4)Campana. Esta representação por módulos também é bastante útil na programação orientada-objeto devido a sua flexibilidade. Cada elemento possui uma ou duas entradas de comunicação com os outros elementos do sistema, cada uma sendo caracterizada por uma entrada, p_{in} e uma saída, p_{out}.

Fazendo a aproximação por uma onda plana em cada entrada, a pressão média p em cada secção transversal S e a vazão volumétrica de ar u atravessando esta secção são dadas pela Eq. (5). A densidade do ar e a velocidade do som estão representadas pelas letras ρ_o e c, respectivamente.

$$\begin{cases} p(t) = p_{in}(t) + p_{out}(t) \\ u(t) = \dfrac{S}{\rho_o c}\left[p_{in}(t) - p_{out}(t)\right] \end{cases} \qquad (5)$$

Eq. 5: Equações que representam a interconexão entre os módulos.

Esta modelagem conduzirá também ao modelo "digital waveguide" para a porção cilíndrica do tubo e a um modelo de entrada dupla para um tubo de secção transversal variável.

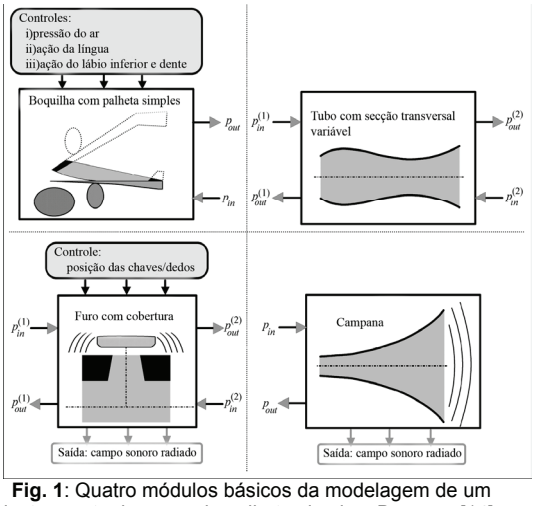


Fig. 1: Quatro módulos básicos da modelagem de um instrumento de sopro de palheta simples. Ducasse [14].

MÉTODOS EXPERIMENTAIS

Modelo Empírico

Numa trajetória oposta à apresentada até aqui, as *relações* entre as variáveis que caracterizam o sistema formado pelo ar escoando através do instrumento musical podem ser totalmente obtidas através de experimentos.

Em trabalhos realizados pelos autores [17, 18] tratamos o sistema como uma caixa preta. Isto é, procuramos obter uma relação entre um conjunto de variáveis de entrada (independentes) com a amplitude e freqüência (variáveis de saída ou dependentes) de algumas notas de uma clarineta. Obtivemos um modelo linear para cada uma das componentes espectrais de três notas em regiões distintas, tanto para a amplitude como para a freqüência. Os modelos obtidos têm potencial de servir como referência para uma posterior elaboração de síntese sonora, empregando softwares como o MATLAB e PD através de síntese aditiva para cada componente espectral e usando a envoltória de Blackman-Harris, como na obtenção dos dados experimentais.

O aparato experimental está apresentado na Fig. 3 e os detalhes podem ser examinados nos trabalhos indicados. Nos experimentos as variáveis independentes consideradas foram: 1)Volume vazio do tanque pulmão ($\mathbf{x_1}$), 2)Dureza da palheta ($\mathbf{x_2}$), 3)Posição de contato na palheta ($\mathbf{x_3}$), 4)Abertura da boquilha ($\mathbf{x_4}$), 5)Área de contato com a palheta ($\mathbf{x_5}$) e 6)Quantidade de material absorvente sonoro ($\mathbf{x_6}$).

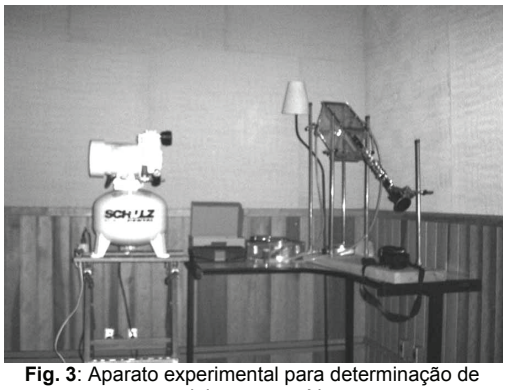


Fig. 3: Aparato experimental para determinação de modelagem empírica.

Tanto para a frequência (Hz) como para a amplitude (dB) o modelo linear obtido é da forma:

$$Y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6$$

A título de exemplo, .para a região "chalumeau" (grave), estudamos o efeito destas variáveis sobre a nota $\mathbf{E_3}$ da clarineta ($\mathbf{D_3}$ do piano) utilizando um projeto de experimentos. As tabelas 2 e 3 indicam os valores dos coeficientes obtidos (a_i, i=0,...,6) para as variáveis (dependentes) intensidade (Y_i) e frequência (Y_f) dos modelos da fundamental e das componentes espectrais (até a 12ª).

Y_i	$-a_o$	a_1	a_2	a_3	a_4	a_5	a_6
D3	21	-1,6	1,4	0,2	-0,9	-3,2	-0,2
2	59	0,4	-0,2	1,6	-2,6	-6,4	-0,4
3	23	-0,3	1,0	-0,8	-0,3	-2,0	-0,3
4	47	0,5	1,8	-0,3	-0,3	-3,2	-1,0
5	32	0,8	-1,0	-1,5	-1,6	2,8	1,0
6	34	0,6	1,9	-1,9	-1,2	-0,4	0,9
7	35	0,6	-0,9	-0,2	-1,2	0,6	0,2
8	28	-1,6	0,6	-0,4	1,6	-1,4	-0,2
9	42	1,0	-1,8	-0,3	-2,0	1,0	1,8
10	36	1,5	-1,0	-1,8	0,3	2,5	0,5
11	41	2,5	-2,2	-0,2	-0,5	-3,0	1,2
12	36	2,2	0,6	-2,2	0,4	4,2	0,6

Tab.2: Coeficientes do modelo empírico da amplitude da nota $\mathbf{D_3}$ (piano).

Y_f	$-a_o$	a_1	a_2	a_3	a_4	a_5	a_6
D3	146	-0,2	-1,0	-1,0	0,2	0,2	1,0
2	295	2,4	0,4	-1,9	-0,9	-1,6	0,4
3	440	-1,5	1,5	2,2	0,8	-1,5	0,0
4	588	0,9	1,6	0,4	-1,6	-1,9	0,4
5	735	-1,6	0,4	0,9	0,4	-1,6	0,4
6	885	0,0	0,5	1,5	0,8	-2,8	-0,8
7	1031	-0,2	2,5	0,5	-0,2	-3,8	-1,0
8	1178	-1,5	2,8	2,2	0,0	-4,0	-1,2
9	1325	-1,0	2,5	1,8	0,2	-3,0	-1,0
10	1473	-0,6	2,9	3,4	0,9	-4,2	-2,2
11	1620	-1,2	2,2	1,6	0,9	-3,6	-0,4
12	1770	-0,4	2,4	3,2	1,2	-5,2	-0,9

Tab.3: Coeficientes do modelo empírico da frequência da nota D_3 (piano).

Análise Paramétrica

A modelagem física, como apresentada pela Eq. (1), envolve o emprego de vários parâmetros. Estes podem ser determinados através da simulação, porém, os resultados devem sempre ser confrontados com valores obtidos empiricamente.

O tratamento experimental para a modelagem física foi inicialmente considerado por Keefe [13] tratando o instrumento de sopro de madeira como sendo uma coleção de orifícios em várias posições e comprimento ao longo do tubo principal. Ou seja, com esta hipótese, ele considera que os parâmetros de impedância associados com cada furo são independentes dos demais furos. Os parâmetros de impedância mencionados correspondem à indutância ("inertance") e capacitância ("compliance") acústicas e serão determinados em função da freqüência.

Ele utilizou o modelo de um circuito de secção em T de uma linha de transmissão para representar os furos tonais de um tubo de instrumento de sopro, conforme representado na Fig. (2). As impedâncias em série possuem o índice *a* para indicar o caso antissimétrico e a impedância cruzada possui o índice *s* para indicar o caso simétrico. As impedâncias em série *Za* e a impedância *Zs* estarão associadas tanto com o furo aberto quanto com o furo fechado.

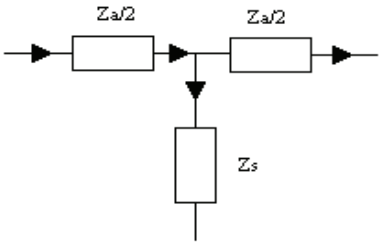


Fig. 2: Circuito elétrico para analogia com o sistema físico.

MANIPULAÇÃO MUSICAL DE SÍNTESE POR MODELAGEM FÍSICA

Waveguides

Conforme as palavras de Smith [11], um instrumento musical precisa ter "vida" nas mãos do instrumentista. A característica principal reside na interatividade entre o músico e seu instrumento. A principal fonte de "vida" na maior parte dos instrumentos acústicos (deixando de lado a performance do artista) reside nas suas formas de *ressonância*.

Smith [11] exemplifica através do violoncelo. As cordas ressoam para fornecer a altura da nota ("pitch") e ainda todo o corpo do instrumento ressoa proporcionando pequenas variações da nota tocada. A ressonância, ele continua, "*fornece memória e caráter variável ao som. O músico interage com a ressonância corporal de maneira imprevisível, algumas vezes reforçando outras cancelando parcialmente o estado de ressonância acumulado*".

O autor perfaz um apanhado geral sobre o estado da arte da modelagem física de instrumentos musicais. Ele não se restringe apenas à família dos instrumentos de sopro. Aborda ainda os instrumentos de corda, metais, voz, instrumentos de percussão e ambientes acústicos.

Sua abordagem, no entanto, restringiu-se ao uso das "digital waveguides", sua especialidade. Uma das razões pelas quais o método de Smith teve grande repercussão no contexto da computação musical foi a facilidade com que as waveguides podem ser implementadas computacionalmente.

Modelos de Síntese e Manipulação Musical

Atualmente, sintetizadores musicais que têm por base modelos que procuram descrever o mecanismo de produção sonora possibilitam ao músico ferramentas mais eficientes para o controle e produção tanto de sonoridades novas como tradicionais.

Smith [11] apresenta uma análise de vários algoritmos de síntese a partir do ponto de vista estrutural. Para o caso de algoritmos que utilizam as estruturas contidas na síntese aditiva ou granular faz-se necessário especificar vários parâmetros e o resultado dependerá da coerência com que estes parâmetros foram escolhidos. Esta coerência não é intrínseca à estrutura e precisa ser garantida durante a especificação dos parâmetros.

Uma segunda categoria de algoritmos diz respeito à estrutura de multi-blocos "feed-forward", na qual alguns blocos geram um sinal que será alimentado a outros blocos para posterior processamento. Esta estrutura inclui técnicas lineares e não-lineares tais como a síntese subtrativa, síntese FM, síntese AM e algumas remotas sínteses por modelagem física. A principal característica desta classe de algoritmos é o surgimento de uma complexidade sonora intrínseca à estrutura. Isto é, escolhendo a síntese através desta técnica damos à estrutura a tarefa de produzir nuances que caracterizam a complexidade do som sintetizado.

A última classe de algoritmos é caracterizada por uma estrutura de multi-blocos interativos. A síntese por modelagem física é um caso especial desta classe de algoritmos que possui ainda uma interpretação física precisa. Esta interpretação é útil para a identificação dos parâmetros de controle do modelo.

Finalmente, a síntese por amostragem (*sampling synthesis*) oferece, para o caso de uma única nota tocada, uma grande possibilidade de interação entre músico e instrumento. A técnica baseada em modelos físicos, no entanto, oferece uma maior expressividade musical além de exigir menor capacidade de memória, ainda que implique na necessidade de uma máquina com maior poder de cálculo.

CONCLUSÕES E PROPOSTAS

Este trabalho teve como espinha dorsal os artigos do Keefe [1, 13], Smith [11] e Ducasse [15]. Entretanto, estes trabalhos não forneceram detalhes dos procedimentos adotados. Portanto, existe um conjunto de conhecimentos essenciais que deveremos adquirir para, realmente, testarmos os modelos apresentados nestes artigos.

Tal postura, pode dar condições de entender os processos computacionais, a modelagem matemática e, principalmente, verificar a natureza e a qualidade sonora de simulações. Para nós, o objetivo final é desenvolver um modelo que possa estabelecer uma ponte entre o mundo real dos instrumentos musicais e as simulações que estudamos.

Propomos, inicialmente, reproduzir os resultados apresentados naqueles artigos através de simulações. Isto permitirá produzir uma análise quantitativa teórica com apoio em resultados experimentais, gerando resultados mais precisos que os disponíveis no momento.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] KEEFE, D.H. (1992). *Physical Modeling of Wind Instruments*. Computer Music Journal **16**(4): 57-73.

[2] HELMHOLTZ, H.L.F. (1954). *On the Sensations of Tone*. Tradução em inglês da versão alemã de 1877 por A.J. Ellis. New York: Dover Publitions.

[3] RAYLEIGH, Lord (1894). *The Theory of Sound*. Macmillan, New York: reeditado pela Dover, 1945.

[4] BOUASSE, H. (1929-30). *Instruments à Vent*. Paris: Librairie Delagrave.

[5] BENADE, A.H. e KOUZOUPIS, S.N. (1988). *The clarinet spectrum: Theory and experiment.* J. Acoust. Soc. Am. **83**, 292-304.

[6] BENADE, A.H. e LARSON, C.O. (1985). *Requirements and Techniques for measuring the musical spectrum of the clarinet.* J. Acoust. Soc. Am. **78**, 1475-1498.

[7] BENADE, A.H. (1976). *Fundamentals of Musical Acoustics.* Oxford University Press, New York.

[8] BENADE, A.H. e GANS, D.J. (1968). *Sound Production in wind instruments.* Ann. N.Y. Acad. Sci. **155**, 247-263.

[9] BENADE, A.H. (1966). *Relation of air-column resonances to sound spectra produced by wind instruments.* J. Acoust. Soc. Am. **40**, 247-249.

[10] FLETCHER, N.H. E ROSSING, T.H. (1991). *The Physics of Musical Instruments*. 2nd ed, New York: Springer-Verlag

[11] SMITH, J.O. (1996). *Physical Modeling Synthesis Update*. Computer Music Journal **20**(2): 44-56.

[12] McINTYRE, M.E., SCHUMACHER, R.T. e WOODHOUSE, J. (1983) *On the Oscillations of Musical Instruments.* J. Acoust. Soc. Am. **74**, 1325-1345.

[13] KEEFE, D.H. (1983). *Theory of the Single Woodwind Tone Hole* e *Experiments on the Single Woodwind Tone Hole*. Journal of the Acoustical Society of America **72**(3): 676-699.

[14] BORIN, G., De POLI, G., SARTI, A. (1992). *Algorithms and Structures for Synthesis Using Physical Models*. Computer Music Journal. **16**(4): 30-42.

[15] DUCASSE, E. (2003). *A Physical Model of Single-Reed Wind Instrument, Including Actions of the Player*. Computer Music Journal. **27**(1): 59-70.

[16] DUCASSE, E. (2002). An Alternative to the Traveling-Wave Approach for Use in Two-Port Descriptions of Acoustic Bores. Journal of the Acoustical Society of America **112**(6): 3031-3041.

[17] OLIVEIRA, L.C, GOLDEMBERG, R., MANZOLLI, J. (2005). *Estudo Experimental da Sonoridade Chalumeau da Clarineta através de Projeto Fatorial (I)*, Anais da IX Convenção Nacional da AES, SP.

[18] OLIVEIRA, L.C, GOLDEMBERG, R., MANZOLLI, J. (2005). *Estudo Experimental da Sonoridade Chalumeau da Clarineta através de Projeto Fatorial (II)*, Anais do XV Congresso da ANPPOM, RJ

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Sintetizador Evolutivo de Segmentos Sonoros

José Fornari 1, Jônatas Manzolli 2, Adolfo Maia Jr. 3
Núcleo Interdisciplinar de Comunicação Sonora – NICS - UNICAMP
Rua da Reitoria, 165 - Cidade Universitária "Zeferino Vaz"
CEP: 13 091 - 970 - Caixa Postal: 6166.
Campinas, São Paulo, Brasil
[fornari, jonatas, adolfo]@nics.unicamp.br

RESUMO

Apresentamos nesse trabalho a implementação em software do método da síntese evolutiva de segmentos sonoros, (SESS), conforme descrita em [1]. A síntese evolutiva é inspirada nos processos biológicos de reprodução e seleção de indivíduos em uma população em função do meio. Na SESS segmentos sonoros (*waveforms*) são tratados como indivíduos pertencentes a uma população onde o som sintetizado é o caminho evolutivo dos melhores indivíduos de cada geração da população de sons. A implementação foi feita utilizando a linguagem de programação PD (*Pure Data*).

INTRODUÇÃO

Desde o surgimento dos primeiros processos elétricos e eletrônicos com objetivos musicais, vem-se desenvolvendo uma grande quantidade e variedade de métodos de síntese sonora. Estes métodos podem ser organizados em três categorias: 1) métodos lineares, tais como a síntese aditiva [2], métodos não-lineares, como a síntese FM [3] e métodos de edição, como é o caso da síntese *wavetable* [4]. Todas estas categorias de métodos de síntese sonora apresentam algo em comum: são métodos determinísticos, pois apresentam um único tipo ou padrão fixo de saída (o som sintetizado) para uma condição fixa dos parâmetros de controle do processo de síntese. A síntese evolutiva é, ao que sabemos, o primeiro método não-determinístico de síntese sonora uma vez que o som sintetizado evolui ao longo do tempo no sentido de se adaptar a determinadas características ou regras, mesmo que os parâmetros de controle da síntese permaneçam inalterados.

A Síntese Evolutiva de Segmentos Sonoros (SESS) é um método computacional de síntese sonora baseado na Computação Evolutiva [5], que por sua vez, inspira-se na teoria Darwiniana da evolução das espécies biológicas, através dos processos de reprodução e seleção.

Existem diversos outros métodos musicais baseados na computação evolutiva, tais como o GenJam [6], um algoritmo genético para simular improvisos de Jazz; um processo evolutivo de geração automática de processos de síntese sonora [7]; um processo de geração evolutiva de padrões rítmicos [8]; e o VoxPopuli [9], um software de composição musical interativa que utiliza algoritmos genéticos e funções de adequação para criação de seqüências musicais. A SESS é, ao que sabemos, o primeiro método evolutivo de síntese sonora pois utiliza algoritmos genéticos e função de adequação não para a manipulação do controle de um método determinístico de síntese mas para a síntese sonora em si, agindo intrinsecamente no segmento sonoro.

O MÉTODO DA SESS

Na SESS os indivíduos são amostras discretas (digitais) de segmentos sonoros com uma dada taxa de amostragem (amostras/s) e resolução (bits). O conjunto de todos os indivíduos compõe a **população**, onde ocorre a evolução. O caminho da evolução da população é condicionado através de uma medida de distância dada por uma função de adequação, ***fitness***, que mede a distância entre as características perceptuais sonoras dos indivíduos da população com os de outro conjunto de indivíduos, o **conjunto alvo**. A evolução da população ocorre em estágios, chamados de **geração**.

A evolução da população é feita por dois processos: a reprodução e a seleção. Em cada geração a reprodução gera novos indivíduos e a seleção escolhe o melhor indivíduo da população, ou seja, o mais adaptado aos critérios dados pelo conjunto alvo.

No processo de reprodução agem dois operadores genéticos: ***crossover*** e **mutação**. O *crossover* permuta características sonoras dos indivíduos em reprodução (os progenitores). A mutação insere modificações aleatórias nessas características, aumentando assim a diversidade da população. Chamamos de genótipo do indivíduo o conjunto de características perceptuais sonoras que o compõem, ou seja, suas grandezas psicoacústicas. O processo de evolução atua sobre os genótipos dos indivíduos.

Na reprodução, o genótipo é modificado pelo crossover e pela mutação. Na seleção, pela escolha do indivíduo mais adequado, ou seja, o melhor indivíduo. O grau de adequação de cada indivíduo é medido pela **distância** entre o seu genótipo e um conjunto de genótipos dos indivíduos do conjunto alvo, que condicionam a evolução da síntese evolutiva.

O resultado sonoro deste método de síntese é o segmento sonoro escolhido como **melhor indivíduo**. A cada geração da população o processo de seleção busca pelo melhor indivíduo da população, ou seja, aquele com menor distância em relação ao alvo. Ao longo das gerações tem-se uma sucessão de melhores indivíduos que, como segmentos sonoros, tendem a convergir para indivíduos cada vez mais similares, isso considerando que o conjunto alvo permaneça inalterado ao longo das gerações.

O método da síntese evolutiva é extensivamente explicado em [10], serviu de inspiração para dois pedidos de patente nacionais [11] e [12] e vem sendo desenvolvido no NICS (www.nics.unicamp.br/~fornari) patrocinado pela FAPESP, sob a forma de projeto de PosDoc no Brasil, processo: 04/00499-6R

IMPLEMENTAÇÃO DO SESS EM PD

Pure Data (PD) é uma linguagem de programação visual desenvolvido inicialmente por Miller Puckette [13]. Trata-se de uma ferramenta gráfica de programação em tempo real, para áudio, video, e processamento gráfico. Ele é a terceira maior ramificação da família de linguagem de programação modular, conhecida como Max (Max/FTS, ISPW Max, Max/MSP, jMax, etc.) originalmente desenvolvida por Miller Puckette (IRCAM). O núcleo do Pd é escrito e mantido por Miller Puckette, com a contribuição de muitos outros desenvolvedores.

Pd é um software livre e pode ser baixado em um pacote para um sistema operacional específico, um pacote com fontes, ou direto do CVS. O Pd é desenvolvido em multi-plataformas, portanto completamente portável; existem versões para Win32, IRIX, GNU/Linux, BSD, MacOS X e rodando em qualquer coisa desde um PocketPC, um Mac antigo ou um novo PC. Usando softwares como "Flext" e "Cyclone" pode-se escrever "*externals*" e "*patches*" que rodam no Max/MSP e no Pd. (www.puredata.org).

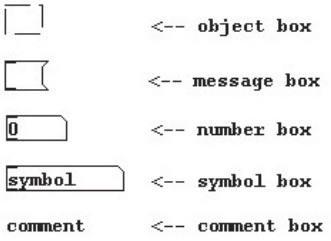


Fig. 1. Alguns módulos básicos do PD.

A figura acima mostra alguns módulos básicos do PD. Estes podem ser conectados entre si para compor os algoritmos de processamento de áudio.

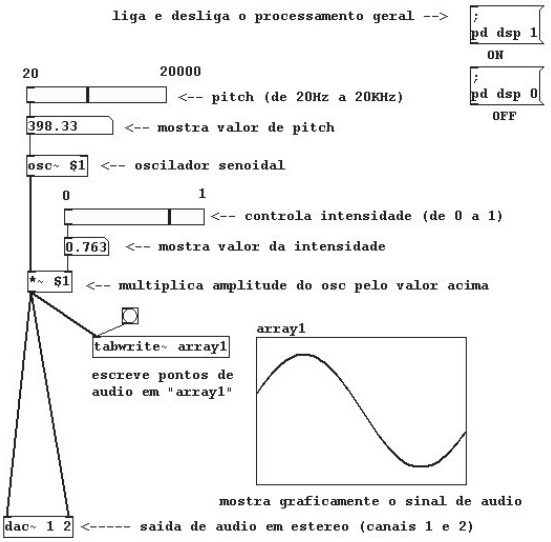


Fig. 2. Exemplo de síntese sonora básica, com processamento e controle em tempo-real.

Os *object boxes* irão conter métodos de processamento (terminados por "~") ou controle de áudio. Cada versão de PD acompanha uma ampla biblioteca de métodos, mas é também possível criar novos métodos, escrito em linguagem C ou C++.

O SESS foi desenvolvido em PD utilizando *subpatches*. Existem duas maneiras de cria-los em PD, o primeiro, que são salvos como parte do código são representados em PD por um *object box* contendo as letras "pd" seguidas pelo nome do subpatch. A segunda maneira, é a utilização de um *subpatch* escrito como código separado, que deve ser previamente salvo como um arquivo do tipo *.pd e acessado através de um *object box*

contendo o esse nome, sem a extensão .pd. Estes são chamados de *abstractions*.

Em PD *subpatches* podem conter um número ilimitado de *subpatches* dentro de outros *subpatches*, A implementação do SESS utiliza extensivamente essa característica. Abaixo tem-se a implementação do SESS em PD através de quatro *subpatches* principais.

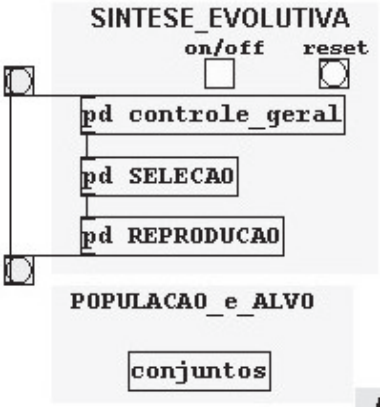


Fig. 3. Implementação da SESS

Observe que o *subpatch* "conjunto" é um *abstraction* e, portanto é um arquivo *.pd distinto, salvo separadamente do arquivo da SESS. A razão disso é que "conjuntos" contêm uma grande quantidade de dados (todos os *arrays* que contêm os segmentos sonoros da população e alvo) o que torna conveniente mantê-los separadamente.

Cada *subpatch* possui uma grande quantidade de *boxes* e outros *subpatch*s, em particular, um chamado "contador" que serve para a contagem dos pontos de cada *array*. O controle da SESS é feito por MIDI (note e velocity) e pelos controles descritos no método da síntese evolutiva, a saber, as taxas dos operadores genéticos: crossover e mutação e a velocidade de proliferação, que determina a velocidade de execução do ciclo de cada geração (em ms).

Fig. 4. Controle da SESS

Os processos de seleção e reprodução estão contidos dento dos *subpatches* "pd SELECAO" e "pd REPRODUCAO" mostrados na figura 3.

Quando expandidos, esses subpatches mostram os algoritmos das figuras 5 e 6. Estes também contêm outros *subpatches*.

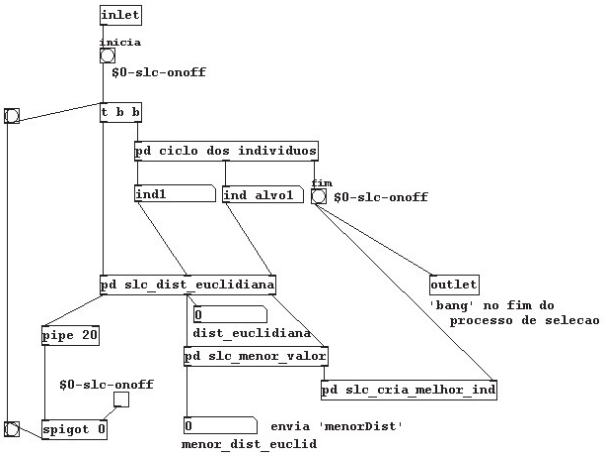


Fig. 5. Implementação do processo de seleção da SESS.

No processo de seleção, tem-se um *subpatch* que calcula a distancia euclidiana, utilizada para o cálculo da métrica L2, usada aqui como a função de adequação entre indivíduos.

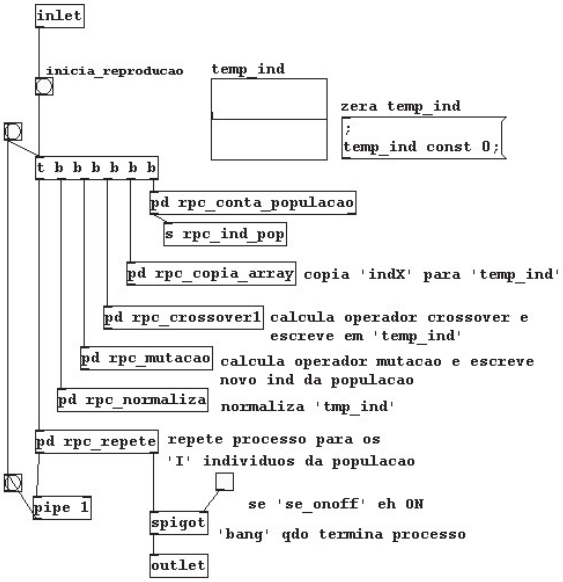


Fig. 6. Implementação do processo de reprodução da SESS

No processo de reprodução tem-se um *array* temporário utilizado para armazenar os cálculos dos operadores genéticos: crossover e mutação. O crossover escolhe uma secção em posição e tamanho aleatório que é misturado entre cada indivíduo da população a a respectiva secção do melhor indivíduo. A mistura ocorre de acordo com a taxa de crossover dada na figura 4. A mutação mistura o array com um segmento de números aleatórios normalizados entre [-1,1], de acordo com a taxa de mutação, também dada na figura 4.

RESULTADOS EXPERIMENTAIS

Foi aqui implementada a versão mais simples do SESS, conforme descrita em [1] onde os indivíduos são segmentos sonoros de áudio digital (16bits, 44.1KHz) contidos em *arrays* de 1024 pontos (equivalente a 23,21ms de áudio). O conjunto população contém 12 indivíduos que são senoides normalizadas ([-1,1]), em diferentes freqüências. O conjunto alvo é formado por 3 indivíduos que são ruídos-brancos normalizados.

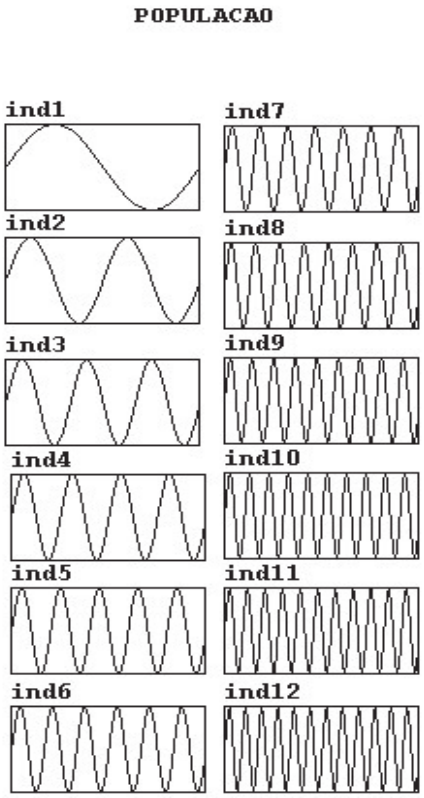


Fig. 7. População de indivíduos do SESS

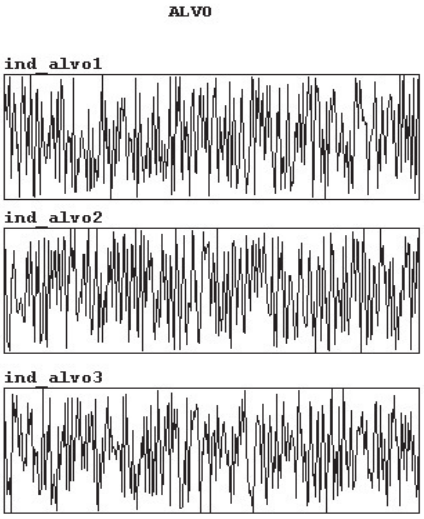
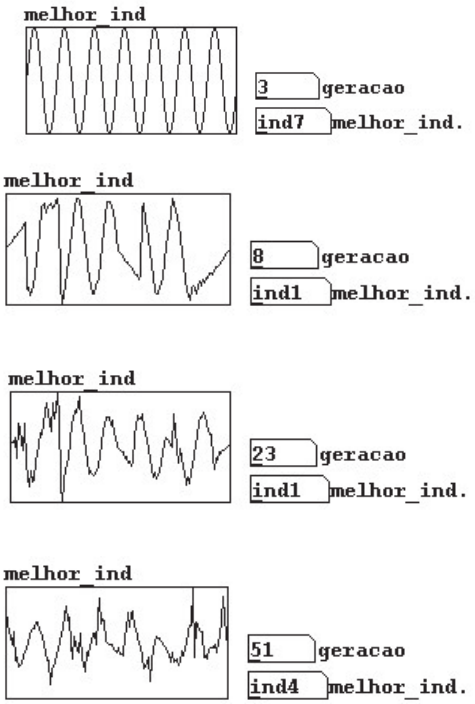


Fig. 8. Conjunto alvo do SESS

Os operadores genéticos crossover e mutação agem diretamente sobre o segmento sonoro (correspondente ao modelo da *waveform* como genótipo do indivíduo) de acordo com as taxas de crossover e mutação.

O SESS vai inicialmente selecionar o individuo da população que é mais próximo dos indivíduos do conjunto alvo. Este será o primeiro melhor indivíduo. Logo em seguida o processo de reprodução modifica todos os elementos do conjunto população através das operações genéticas entre cada indivíduo da população e o melhor indivíduo escolhido anteriormente. Finalmente o processo seleção escolhe um novo indivíduo mais próximo do alvo. Este equivale a uma geração da SESS. O som sintetizado corresponde a seqüência de melhores indivíduos de cada geração. Abaixo tem-se a ilustração da *waveform* de alguns desses melhores indivíduos ao longo do processo de síntese.



Observa-se que na medida em que o tempo passa, sob a forma do avanço dos ciclos de processamento do SESS representado pelas gerações da população, o segmento sonoro do melhor indivíduo vai se tornando mais semelhante aos segmentos do alvo. Isso ocorre porque o processo de seleção sempre busca o individuo na população mais semelhante aos indivíduos do alvo e o coloca como melhor individuo, enquanto que o processo de reprodução está sempre criando novos indivíduos descendentes dos indivíduos da geração anterior e o melhor indivíduo. Assim cria-se uma variabilidade fenotípica entre os indivíduos, porém sempre na direção de torná-los mais parecidos ao melhor indivíduo, que é o indivíduo mais bem adaptado da sua geração. Em uma escala muito mais simplificada, isso emula à adaptação biológica que os indivíduos de uma população sofrem pela condição do meio. Na SESS, o "meio ambiente" é simplificadamente representado pelo alvo, e a condição de semelhança com este equivale à pressão condicionante deste meio.

CONCLUSÕES E COMENTÁRIOS

É interessante observar que a implementação da SESS de fato simula o processo adaptativo da evolução das espécies. Note que não há qualquer troca de dados entre os *arrays* do alvo e os da população. Os processo de seleção e reprodução são capazes de criar melhores indivíduos cada vez mais semelhantes aos indivíduos do alvo. Uma vez que utilizamos o segmento como genótipo e fenótipo podemos visualizar a semelhança entre as *waveforms*, no entanto a percepção sonora se baseia em grandezas psicoacústicas para estabelecer a semelhança entre sons. Um próximo modelo de síntese evolutiva levará em conta essas características para medir a distância entre indivíduos. Um extrator de curvas psicoacústica já foi desenvolvido para tal [12] e está em fase de implementação.

Outra característica a ser melhorada é o tamanho dos segmentos sonoros. A utilização de segmentos de 1024 pontos resulta em sons muito curtos (~23ms na taxa de amostragem de 44,1KHz). Isto impede a percepção auditiva das diferenças entre melhores indivíduos. A utilização de *arrays* maiores (acima de 44100 pontos) não é trivial e necessita uma ampla remodelagem da implementação do SESS no PD, que já está em andamento.

A utilização da linguagem de programação PD para a implementação do método de síntese sonora evolutiva, sob a forma de um sintetizador evolutivo em software, foi bastante eficiente e satisfatória. As principais razões para sua utilização são: 1) PD é uma linguagem rápida (considerada tão rápida quanto executáveis em linguagem C), desenvolvida especialmente para o processamento e controle de algoritmos de multimídia operando em tempo-real. 2) PD é gratuita, de código aberto e distribuição livre (nos termos da "Standard Improved BSD License"), 3) PD possui uma grande comunidade de programadores e desenvolvedores na internet, (ver site: www.puredata.org) , 4) PD é multi-plataforma (roda em Windows, Linux, e MacOS, entre outros). 5) PD é expansível (permite criar novos módulos de processamento através de sub-rotinas, na própria linguagem PD, chamadas de abstractions, ou criar externals, criadas em outras linguagens de programação, tais como: C, C++ ou Fortran). 6) PD pode se comunicar em rede e interconectar com outros programas de processamento de áudio em tempo-real, tais como JACK e Ardour.

Este é um trabalho em andamento, onde iremos implementar novas versões do SESS incorporando características mais próximas da realidade biológica que inicialmente motivou e inspirou o desenvolvimento deste método. Entre outras, podemos citar: 1) implementação de população de tamanho variável (conceito de extinção e superpopulação), 2) indivíduos com gênero e tempo de vida (conceitos de sexo e morte). 3) diferentes formas de representação do genótipo do indivíduo (ex: *waveform*, curvas psicoacústicas e espectrograma, entre outros), 4) novas formas operações genéticas (ex: operadores crossover e mutação no domínio da freqüência), 5) novas funções de adequação (utilizando outras métricas além da distancia Euclidiana). 6) Diferentes formas de controle da síntese (ex: MIDI *aftertouch* controlando taxa de operadores genéticos. MIDI *modulation* controlando taxa de proliferação).

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Fornari, José, Jônatas Manzolli, Adolfo Maia, Furio Damiani. “The Evolutionary Sound Synthesis Method”. *Short-paper do ACM multimedia*, E.U.A. 2001.

[2] Kleczkowski, P., “Group additive synthesis”. COMP. MUSIC J. Vol. 13, no. 1, pp. 12-20. 1989.

[3] J. Chowning, "The synthesis of complex audio spectra by means of frequency modulation," Journal of the Audio Engineering Society, vol. 21, pp. 526-534, 1973.

[4] Horner, Andrew; Beauchamp, James; Haken, Lippold. “Methods for multiple wavetable synthesis of musical instrument tones”. J AUDIO ENG SOC. Vol. 41, no. 5, pp. 336-356. 1993.

[5] Fogel, D. B., “Evolutionary Computation: Toward a New Philosophy of Machine Intelligence”, IEEE Press, 46 – 47, 1995.

[6] Biles, J. A., “Gen Jam: A Genetic Algorithm for Generating Jazz Solos”, Proceedings of the 1994 International Computer Music Conference, (ICMC’94), 131—137, 1994.

[7] R Garcia. "Growing Sound Synthesizers using Evolutionary Methods". Proceedings of ALMMA 2002 Workshop on Artificial Models, 2001

[8] N Tokui, H Iba. "Music composition with interactive evolutionary computation.". Proceedings of the third International Conference GA2000, 2000.

[9] Moroni, A., Manzolli, J., Von Zuben, F., Gudwin, R., “Vox Populi: An Interactive Evolutionary System for Algorithmic Music Composition”, Leonardo Music Journal, San Francisco, USA, MIT Press, Vol. 10, 2000.

[10] Fornari, José Eduardo. “Síntese Evolutiva de Segmentos Sonoros”. Dissertação de Doutoramento. DSIF/FEEC/UNICAMP. 2003

[11] Fornari, José, Jônatas Manzolli, Adolfo Maia. “Métodos e Dispositivos Evolutivos para a análise, Processamento e Síntese de sinais digitais unis e multidimensionais, Pedido de Patente. Protocolado no INPI em 23 de Março de 2005, Protocolo: PI0500958-8.

[12] Fornari, José, Jônatas Manzolli. “Método Extrator de Curvas Psicoacústicas de Intensidade Sonora e Freqüência Fundamental”, Pedido de Patente, Protocolado no INPI em 15 de Dezembro de 2005. Protocolo: 01850064017.

[13] M Puckette. "Pure Data: another integrated computer music environment ". Proceedings, Second Intercollege Computer Music Concerts, 1996.

Sessão 5

Psicoacústica, Percepção Auditiva, Análise e Audição Automática

(Psychoacoustics, Auditory Perception, Analysis and Automatic Listening)

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Dead regions and speech perception in subjects with auditory dyssynchrony

Vinay S.N [1] and Vanaja C.S [1]
[1] Department of Audiology, All India Institute of Speech and Hearing
Manasagangothri, Mysore – 570006, India
shrivinyasa@gmail.com

ABSTRACT

Auditory Dyssynchrony (AD) is a hearing disorder in which sound enters the inner ear normally but the transmission of signals from the inner ear to the auditory cortex is impaired (Starr et al. 1996). Studies have shown that speech recognition scores (SRS) are affected in AD subjects (Sininger et al. 1995; Starr et al. 1996). However earlier studies have not identified the presence of dead regions in subjects with AD. The aim of the present study was to identify the presence of dead regions in subjects with AD using TEN (HL) test and to compare the SRS in AD subjects with and without dead regions. The SRS was correlated with the edge frequencies. Audiometric thresholds at different frequencies were compared for AD subjects with and without dead regions. Results of TEN (HL) test on subjects with AD indicated 21 out of 29 ears had a dead region. Results indicated poor SRS in AD subjects with dead region compared to those without dead regions. SRS also reduced as the edge frequency increased. AD subjects with dead region had higher audiometric thresholds than those without dead regions.

INTRODUCTION

Auditory dysynchrony (AD) is a hearing disorder in which sound enters the inner ear normally but the transmission of signals from the inner ear to the auditory cortex is impaired. The term was coined by Starr, Picton, Sininger, Hood & Berlin (1996). It has been showed that patients with AD demonstrate primarily a timing deficit that is consistent with a lack of neural synchrony (Zeng, Oba, Garde, Sininger & Starr, 1999). Although AD is not yet fully understood, researchers believe the condition probably has more than one etiology (Sininger & Starr, 2001). In some cases, it may involve damage to the inner hair cells (IHCs). Other causes may include faulty connections between the inner hair cells and the nerve leading from the inner ear to the auditory cortex, or damage to the nerve itself. A combination of these problems may also occur.

Diagnosis of AD is based upon the results of auditory brainstem response (ABR) and otoacoustic emissions (OAE). The hallmark of AD is a negligible or very abnormal ABR reading together with a normal OAE reading (Sininger & Starr, 2001). The audiometric pattern reveals a rising pattern. Often, speech perception is worse than would be predicted by the degree of hearing loss (Sininger, Hood, Starr, Berlin, & Picton, 1995; Starr, Picton, Sininger, Hood & Berlin, 1996). Subjects with AD show normal frequency resolution and varying degrees of temporal disruption (Sininger, Hood, Starr, Berlin, & Picton, 1995). The severity of this temporal abnormality is strongly correlated to speech perception ability (Rance, Beer & Cone-Wesson, 1999; Wunderlich & Dowell, 2002). Another factor that can affect speech identification scores is the presence of dead regions in the cochlea and/or neurons. It has been reported that speech recognition scores is poor in subjects with dead regions (Vickers, Moore &

Baer et al. 2001; Nagaraj & Moore, 2002). However, there is a dearth in the studies to investigate the presence of dead regions in subjects with AD.

Studies carried out to investigate the potential benefits of hearing aids, cochlear implants, and other technologies for individuals with AD have revealed inconclusive results (Sininger & Starr, 2001). Some investigators have reported that hearing aid is useful in 50% of the subjects, whereas in others, there is deterioration in performance when a hearing aid is prescribed (Rance, Beer & Cone-Wesson, 1999; Starr, Picton, Sininger, Hood, and Berlin (1996). It is possible that subjects who did not benefit from hearing aid had dead regions whereas others did not have dead regions.

Thus the following were the aims of the present study:

i) Identifying the presence of dead regions in subjects with AD.

ii) Comparison of audiometric thresholds in auditory neuropathy subjects with and without dead regions.

iii) To compare speech recognition scores in subjects with AD with and without dead regions.

iv) To investigate the correlation between speech recognition scores and edge frequency of the dead region in subjects with AD.

Studies carried out to investigate the potential benefits of hearing aids, cochlear implants, and other technologies for individuals with AD have revealed inconclusive results (Sininger & Starr, 2001). Some investigators have reported that hearing aid is useful in 50% of the subjects, whereas in others, there is deterioration in performance when a hearing aid is prescribed (Rance, Beer & Cone-Wesson, 1999; Starr, Picton, Sininger, Hood, and Berlin (1996). It is possible that subjects who did not benefit from hearing aid had dead regions whereas others did not have dead regions.

Thus the following were the aims of the present study:

i) Identifying the presence of dead regions in subjects with AD.

ii) Comparison of audiometric thresholds in auditory neuropathy subjects with and without dead regions.

iii) To compare speech recognition scores in subjects with AD with and without dead regions.

iv) To investigate the correlation between speech recognition scores and edge frequency of the dead region in subjects with AD.

METHOD

Subjects

Study consisted of two groups of subjects- Auditory neuropathy subjects with dead regions (21 ears; age ranging from 14 to 45 years; mean age: 23.71 years) and auditory neuropathy subjects without dead regions (8 ears; age ranging from 18 to 37 years; mean age: 25.16 years). The diagnosis of auditory neuropathy was based on the following test results:

i) Normal outer hair cell functioning evident by the presence of TEOAEs amplitude and/or presence of cochlear microphonics (CM)

ii) Abnormal or absent auditory brainstem responses (ABRs)

Instrumentation

The following instruments were used for the present study:

i) A two channel clinical audiometer consisting of supra-aural headphones with earcushions. The audiometer was calibrated to conform to ANSI standards.

ii) A middle ear analyzer to assess the functioning of the middle ear.

iii) A computer connected to the audiometer to present the TEN stimuli.

Materials

i) TEN (HL) compact disc (Moore, Glasberg and Stone, 2004).

ii) Monosyllables word list (Mayadevi, 1974).

Procedure

i) Pure tone audiometry: Air conduction thresholds were determined at the octave/mid-octave frequencies, 250, 500, 750, 1000, 1500, 2000, 3000, 4000, 6000 and 8000 Hz. Bone conduction thresholds were determined at 250, 500, 1000, 2000 and 4000 Hz. The thresholds were measured using the modified Hughson-Westlake procedure proposed by Carhart and Jerger (1959).

ii) Speech audiometry: Speech recognition scores were determined using the monosyllabic word list (Mayadevi, 1974). 20 monosyllables were chosen based upon the frequency of occurrence in Kannada language. Stimuli were presented at 40 dB SL of the pure tone average thresholds. The subjects were asked to repeat the monosyllables that the tester presented. The percentage of correct scores was determined.

iii) Threshold Equalizing Noise (TEN HL) test: The TEN (HL) test was used to check for the presence of dead regions in subjects with AD. The absolute thresholds and masked thresholds in the presence of TEN were measured using the two-channel clinical audiometer with the modified Hughson-Westlake procedure proposed by Carhart and Jerger (1959). The presentation of the TEN level was 10 dB SL of the highest audiometric thresholds. For audiometric thresholds above 80 dB HL, TEN test was carried out for frequencies in which the thresholds are below 90 dB HL as thresholds above 90 dB HL are a definite indication of a dead region (Moore, 2001). The TEN and signal levels was controlled by the use of attenuators on the audiometer. The potentiometers controlling the tape inputs was set to give a reading of 0 dB on the VU meters of the audiometer, while playing the calibration signal. This ensured that the signal and the noise level per ERB were equal to the level indicated on the audiometer.

RESULTS AND DISCUSSION

Table 1 indicates pure tone audiometric thresholds (dB HL), TEN (HL) and speech recognition scores (SRS) values for auditory neuropathy subjects with 'possible' dead regions.

Subjects	TEN	SRS (%)	ABR	OAE	ERB (number)
S1 RE	+	25	Absent	Present	4.85
S1 LE	-	30	Absent	Present	
S2 RE	+	0	Absent	Present	10.37
S2 LE	+	0	Absent	Present	4.85
S3 RE	-	65	Absent	Present	
S3 LE	-	65	Absent	Present	
S4 RE	-	70	Absent	Present	
S4 LE	+	25	Absent	Present	0
S5 RE	+	0	Absent	Present	2.74
S5 LE	+	0	Absent	Present	4.85
S6 RE	+	0	Absent	Present	2.74
S6 LE	-	0	Absent	Present	
S7 RE	+	55	Absent	Present	16.31
S8 RE	+	65	Absent	Present	16.31
S8 LE	+	55	Absent	Present	16.31
S9 RE	+	0	Absent	Present	2.74
S9 LE	+	0	Absent	Present	0
S10 RE	+	50	Absent	Present	16.31
S10 LE	+	70	Absent	Present	16.31
S11 LE	+	0	Absent	Present	2.74
S12 RE	-	0	Absent	Present	
S12 LE	-	0	Absent	Present	
S13 RE	-	0	Absent	Present	
S13 LE	+	0	Absent	Present	2.74
S14 RE	+	60	Absent	Present	2.74
S14 LE	+	65	Absent	Present	0
S15 LE	+	80	Absent	Present	2.74
S16 RE	+	0	Absent	Present	16.31
S16 LE	+	0	Absent	Present	16.31

Subject; RE-Right ear; LE-Left ear; + indicates TEN test result positive – indicates negative

Table 1 TEN (HL) test, ABR and OAE results, speech recognition scores (SRS) and ERB number in Auditory neuropathy subjects

TEN (HL) results revealed two types of patterns were observed in subjects with auditory neuropathy. One type showed abnormally high TEN (HL) thresholds at all frequencies in which TEN (HL) was measured. Results revealed abnormally high TEN (HL) thresholds in subjects S7 (RE), S8 (RE, LE), S10 (RE, LE) & S16 (RE, LE) at all frequencies in the TEN (HL) test. High TEN (HL) thresholds were obtained in spite of 'good' audiometric thresholds at these frequencies in these subjects. This may indicate more of a central problem or other problems related to coding of sounds such as loss of synchrony rather than due to the complete damage to the IHCs and/or auditory neurones. Subjects with auditory neuropathy experience conduction block in the sound transmission pathway at the level of auditory neurones (Starr et al. 1998). This conduction block and the loss of neural synchrony may also lead to high thresholds in TEN (HL). Also, results indicate that subjects with auditory neuropathy have poor speech recognition scores due to a more severe degree of temporal processing problems in these subjects than that are found in subjects having cochlear hearing loss (Moore & Glasberg, 1986; Moore, 1998; Florentine & Buus, 1984). The temporal processing disorder in subjects with auditory neuropathy is associated with impairment in detection of short duration acoustic signals (Sininger & Starr, 2001). The second group of subjects showed high TEN (HL) thresholds at only certain frequencies, in which, high TEN (HL) thresholds were present more at the lower frequencies than at the higher frequencies. It is interesting to note that high thresholds in TEN (HL) in this group of subjects may indicate loss of sound transmission due to neural dysynchrony than due to complete loss of IHCs and/or auditory neurones. These subjects may also have complete damage of the IHCs resulting in loss of transduction.

TEN (HL) results in subjects with Auditory neuropathy

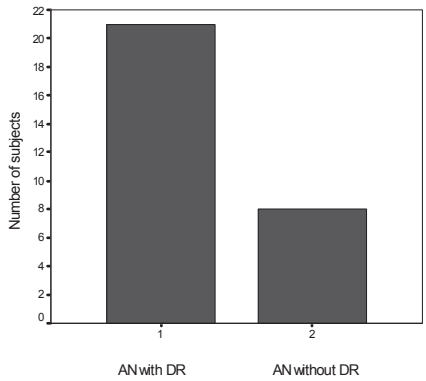


Figura 1 Auditory neuropathy subjects with and without dead regions

TEN (HL) test was administered on subjects with auditory neuropathy (29 ears). 21 ears showed abnormal TEN (HL) results in which the masked thresholds were 10 dB or above than the absolute thresholds. 8 ears obtained masked thresholds within 10 dB of the absolute thresholds. Subjects with auditory neuropathy have a dysynchrony in the auditory neurones.

Comparison of audiometric thresholds in auditory neuropathy subjects with and without dead regions

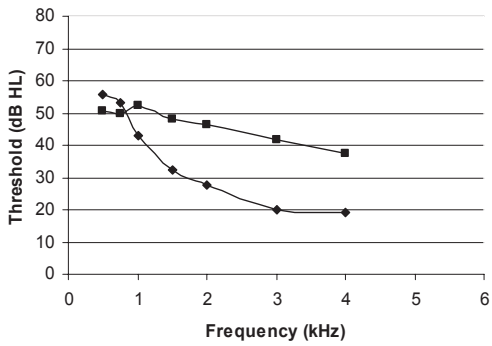


Figura 1 Mean audiometric thresholds for auditory neuropathy subjects with (Diamond filled line) and without (square filled line) dead regions

Audiometric thresholds were compared across auditory neuropathy subjects with and without dead regions. Independent sample 't' test was carried out for statistical significance by comparison of the audiometric thresholds in auditory neuropathy subjects with and without dead regions. Statistical analyses revealed significant difference in audiometric thresholds for auditory neuropathy subjects with and without dead regions at 1500, 2000, 3000 & 4000 Hz. There was no significant difference at 500 (t = 0.754), 750 (t = 0.443), 1000 (t = 1.317), 1500 (t = 2.093), 2000 (t = 2.10), and 4000 Hz (t = 2.363). However, there was a significant difference observed only at 3000 Hz (t = 3.221, $p<0.01$) Results show that high audiometric thresholds at the low frequencies is associated with the presence of asynchrony, that is in tune with the low frequency loss/rising audiogram configuration. The 'audiometric hearing loss' is more due to the dyssynchrony of the auditory neurones rather than due to the damage to the IHCs. For auditory neuropathy subjects with and without dead regions, audiometric thresholds at high frequencies did not show statistically significant results which conclude that the difference in the thresholds is due to the loss of asynchrony in the auditory neurones. The differences in the results obtained at 3000 Hz may be a result of temporal disorder resulting in asynchronous firing.

Comparison of speech recognition scores for auditory neuropathy subjects with and without dead regions

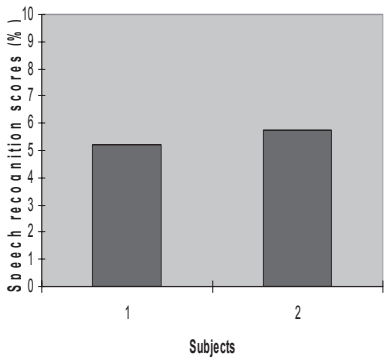


Figura 3 Mean speech recognition scores for auditory neuropathy subjects with (1) and without (2) dead regions

Speech recognition scores for 20 monosyllables were compared in auditory neuropathy subjects with (21 ears) and without (8 ears) dead regions. The scores were calculated in terms of percentage. However, for statistically test analyses, the raw scores were considered. Independent samples 't' test results revealed no statistically significant difference in speech recognition scores in auditory neuropathy subjects with and without dead regions. Speech perception problems in subjects with auditory neuropathy can be related to severe temporal processing disorders (Starr et al. 1996). Also, the speech recognition scores in the subjects do not correlate with the pure tone audiometric thresholds (Yellin et al. 1989). Also, poor speech recognition abilities are reported in subjects with dead regions (Moore, 2001; Vickers et al. 2001; Baer et al. 2002). However, high thresholds in TEN (HL) in these subjects may not indicate the presence of dead regions, but may be due to the loss of synchrony in these subjects.

Comparison of speech recognition scores and extent of dead regions in subjects with auditory neuropathy

The presence of dead regions in subjects with auditory neuropathy was estimated using the TEN (HL) test and the extent of dead regions was expressed in terms of the ERB number. Each frequency represents a corresponding ERB number and the difference in the two ERB numbers indicated the extent of dead regions in these subjects. The ERB number can be calculated using the formula

$$E = 21.4 \log10(4.37F + 1) \quad (1)$$

E = ERB number; F is in kHz (Moore, 2003). The results are demonstrated in the form of a scatter plot (fig. 4).

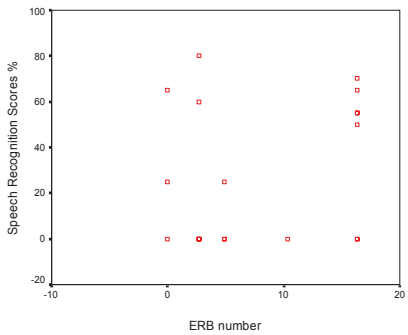


Figura 4 Scatter plot indicating the speech recognition scores (%) in terms of the extent of dead regions (ERB number)

Pearson's correlation was carried out to know the correlation in the speech recognition scores to the extent of dead regions in auditory neuropathy subjects which revealed a positive result in the TEN test. Results revealed a low correlation (ρ=0.285) indicating that the speech recognition scores did not depend on the extent of dead regions in subjects with auditory neuropathy. This may be due to the fact that speech recognition scores are adversely affected in subjects with auditory neuropathy due to loss of synchrony. The role of the presence of dead regions in these subjects may be a compounding factor for poor speech recognition in these subjects. Studies on speech recognition abilities and presence of dead regions reveal statistically significant difference in the scores in high frequency sensorineural hearing impaired subjects with and without dead regions (Moore, 2001; Vickers et al. 2001; Baer et al. 2002). Subjects with dead region do not have any surviving inner hair cells in that regions and hence the transduction of sound stimulus is not possible in those frequencies (Moore et al. 2000). Hence, speech recognition abilities are poor in these subjects. Also, results comparing audiometric thresholds in auditory neuropathy subjects with and without dead regions give a divided opinion. The difference in the audiometric thresholds may be just be a result of loss of synchrony in those frequencies. Speech perception is also affected in the frequency regions where there is asynchrony resulting in the loss of transduction.

CONCLUSION

From the present study, it may be concluded that, dead regions are seen in subjects with AD. Speech perception abilities will be poorer in AD subjects with dead regions than without dead regions. The speech perception scores also depend upon the edge frequency of the dead region.

Speech recognition scores deteriorate, as the edge frequency is higher in terms of frequency.

REFERENCES

[1] Bacon, S.P., & Gleitman, R.M. (1992). Modulation detection in subjects with relatively flat hearing losses. Journal of Speech and Hearing Research, 35, 642-653.

[2] Carhart, R., and Jerger, J. F. (1959). "Preferred method for clinical determination of pure-tone thresholds," Journal of Speech and Hearing Disorders, 24, 330-345.

[3] Formby, C., & Muir, K. (1988). Modulation and gap detection for broadband and filtered noise signals. Journal of the Acoustical Society of America, 84, 545-550.

[4] Mayadevi, N. (1974). The development and standardization of a common speech discrimination test for Indians. An unpublished Master's dissertation submitted to University of Mysore.

[5] Moore, B.C.J. (2001). "Dead regions in the cochlea: Diagnosis, perceptual consequences and implications for the fitting of hearing aids." Trends in Amplification, 5, 1-34.

[6] Moore, B.C.J., Glasberg, B.R., and Stone, M.A. (2004). New version of the TEN test with calibrations in dB HL, Ear and Hearing, 25(5), 478-487.

[7] Moore, B. C. J., Huss, M., Vickers, D. A., Glasberg, B. R., and Alcántara, J. I. (2000). "A test for the diagnosis of dead regions in the cochlea," British Journal of Audiology, 34, 205-224.

[8] Moore, B.C.J., Shailer, M.J., & Schooneveldt, G.P. (1992). Temporal modulation transfer functions for band-limited noise in subjects with cochlear hearing loss. British Journal of Audiology, 26, 229-237.

[9] Rance G., Beer D., Cone-Wesson, B. (1999). Clinical findings for a group of infants and

[10] young children with auditory neuropathy. Ear & Hearing; 20: 238-252.

[11] Sininger, Y., & Starr, A. (2001). Auditory neuropathy: A new perspective on hearing disorders. Singular Publishers.

[12] Sininger, Y., Hood, L.J., Starr, A., Berlin, C.I., & Picton, T.W. (1995). Auditory loss due to auditory neuropathy. Audiology Today, 7, 10-13.

[13] Starr, A., McPherson, D., Patterson, J., Luxford, W., Shannon, R., Sininger, Y., Tonokawa, L., & Waring, M. (1991). Absence of both auditory evoked potentials and auditory percepts dependent on time cues. Brian, 114, 1157-1180.

[14] Starr, A., Picton, T.W., Sininger, Y., Hood, L.J., & Berlin, C.I. (1996). Auditory neuropathy. Brain, 119, 741-753.

[15] Vickers, D. A., Moore, B. C. J., and Baer, T. (2001). Effects of low pass filtering on the intelligibility of speech in quiet for people with and without dead regions at high frequencies, Journal of the Acoustical Society of America, 110, 1164-1175.

[16] Vinay, & Moore, B.C.J. (2002). Effects of high pass filtering on speech intelligibility in subjects with normal hearing and subjects with and without dead regions at low frequencies. Unpublished Master of Philosophy thesis submitted at University of Cambridge, United Kingdom.

[17] Zeng, F.G., Oba, S., Garde, S., Sininger, Y., & Starr, A. (1999). Temporal and speech processing deficits in auditory neuropathy. Neuro Report, 10, 3429-3435..

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Identificação de Notas Musicais de Violão Utilizando Redes Neurais

Alexandre L. Szczupak[1], Luiz W. P. Biscainho[1], e Luiz P. Calôba [1]

[1]LPS – PEE/COPPE & DEL/Poli, UFRJ
Caixa Postal 68504, Rio de Janeiro, RJ, 21941-972, Brasil
aleizor,wagner,caloba@lps.ufrj.br

RESUMO

A identificação de notas musicais em um sinal polifônico pela simples análise de seu espectro de freqüências é dificultada por possíveis superposições dos harmônicos de diferentes notas. Neste trabalho, que aborda especificamente sons de violão, buscamos superar esse problema utilizando redes neurais na análise do espectro freqüencial. Para aproveitar as características particulares dos sinais de música, substituímos como instrumento de representação espectral para sinais discretos a DFT pela *Constant-Q Transform*, que distribui geometricamente as linhas espectrais.

INTRODUÇÃO

Realizar a transcrição de uma peça musical para a partitura exige extenso conhecimento de teoria musical e percepção auditiva aprimorada. Um sistema de transcrição automático, que identifique as notas de um sinal de música, pode se tornar uma ferramente útil na popularização do uso de partituras e no ensino de teoria musical.

Comumente, sinais discretos são representados no domínio da freqüência através da DFT (*Discrete Fourier Transform*), com resultados dispostos sobre uma escala linear de freqüências. Porém, nas escalas musicais de igual temperamento, utilizadas na música ocidental desde o século XVIII [1], as freqüências fundamentais das notas são dispostas em progressão geométrica com razão $2^{\frac{1}{12}}$. Em uma representação através da DFT, o número de linhas espectrais por oitava varia em função da freqüência: oitavas mais altas são descritas com maior densidade de linhas que oitavas mais baixas.

Para otimizar a análise, pode-se utilizar a CQT (*Constant-Q Transform*) [2], uma transformada espectral com seletividade constante e freqüências centrais espaçadas em progressão geométrica, assim como nas escalas de igual temperamento.

Neste estudo utilizamos a CQT para representar o espectro freqüencial de sinais de violão. Essas representações são utilizadas no treinamento e teste de um conjunto de redes neurais projetadas para identificar as notas presentes em gravações do instrumento. A fim de se aferir o grau de dificuldade do reconhecimento de acordo com o número de notas simultâneas, adotou-se a seguinte estratégia: criar 6 redes, cada uma delas especializada na identificação de um número diferente de notas simultâneas.

Essas redes podem ser projetadas para identificar notas de outros instrumentos musicais, desde

que estes também possuam afinação temperada.

A identificação de notas musicais em sinais polifônicos através de redes neurais também foi abordada por Matija Marolt [3, 4]. Em seus estudos, dedicados à identificação de sons de piano, redes especializadas são utilizadas na reconhecimento de cada nota. Seu sistema SONIC apresenta, para diferentes polifonias, erros entre 1,9 e 14% na análise de sinais sintetizados e 11,5 e 14,1% na análise de sinais reais.

Diversas alternativas têm sido propostas para resolução do problema de identificação de notas simultâneas. Uma extensa bibliografia sobre transcrição musical automática pode ser encontrada em [5]. Pode-se destacar um método desenvolvido por Anssi Klapuri para a estimação das notas presentes em sinais polifônicos [6]. Este método, que não utiliza redes neurais, baseia-se em modelos perceptivos da audição humana e também adota a estratégia aqui empregada de aferir separadamente o desempenho do sistema para números diferentes de notas simultâneas.

O VIOLÃO

De um violão de 6 cordas podem ser extraídas 44 notas diferentes, de E2 (82,41Hz) até B5 (987,77Hz). As notas podem soar individualmente ou em combinações de duas até seis notas simultâneas. Dessas 44 notas, 34 podem ser produzidas por um músico utilizando pelo menos duas posições distintas sobre o braço do instrumento. As notas restantes - cada uma das quais só pode ser gerada a partir de uma única posição sobre o braço - são as cinco mais graves e as cinco mais agudas do instrumento.

Para realizar este estudo, gravamos individualmente as 44 notas de 5 violões diferentes. A Figura 1 contém uma representação de um braço de violão. As cordas do instrumento estão desenhadas somente sobre as 78 posições utilizadas durante as gravações. Com essa escolha, todas as 34 notas que podem ser produzidas em posições distintas foram gravadas duas vezes. As 10 notas restantes foram gravadas apenas uma vez. Suas posições estão destacadas na figura.

Para cada posição escolhida, foram realizadas duas gravações: em uma, a corda do violão foi tocada diretamente com os dedos; na outra, com uma palheta.

REPRESENTAÇÃO ESPECTRAL

Transformações espectrais utilizando a CQT resultam em vetores complexos, assim como na DFT, porém com valores dispostos sobre uma escala logarítmica de freqüências. Por conveniência, na análise de sinais de música, essa escala pode ser definida sobre as freqüências das notas de uma escala musical de temperamento igual, ou mesmo

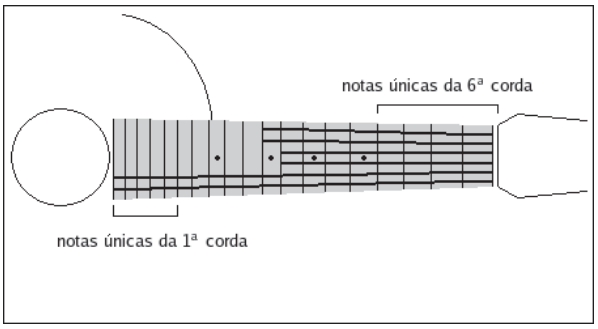


Figura 1: Representação do braço de um violão. As cordas mais agudas estão nas posições inferiores.

incluindo valores intermediários para maior resolução.

A fórmula da CQT pode ser obtida através de algumas alterações na fórmula da DFT direta de uma seqüência x[n] janelada, dada por:

$$X[k] = \frac{1}{N}\sum_{n=0}^{N-1} w[n]x[n]e^{-j\frac{2\pi}{N}kn}, \quad k \in [0, N-1], \tag{1}$$

onde:
N = número de amostras do sinal;
w = função de janelamento.

Para obter seletividade constante e espaçamento logarítmico, o número de amostras analisadas deve variar em função da freqüência desejada, e o índice freqüencial k presente na exponencial deve ser substituído pela seletividade desejada (Q) [2].

$$X_{cq}[k_{cq}] = \frac{1}{N[k_{cq}]}\sum_{n=0}^{N[k_{cq}]-1} w[n, k_{cq}]x[n]e^{-j\frac{2\pi}{N[k_{cq}]}Qn}, \tag{2}$$

onde:
$N[k_{cq}] = \frac{f_s Q}{f_{k_{cq}}}$;
f_s = freqüência de amostragem;
$f_{k_{cq}} = q^{k_{cq}} f_{min}$ = freqüência sob análise;
f_{min} = freqüência mínima escolhida para a análise.

A razão q entre as freqüências adjacentes da CQT deve ser escolhida de acordo com a precisão freqüencial desejada.

Neste estudo utilizamos um algoritmo rápido para cálculo da CQT [7, 8], baseado no algoritmo FFT:

$$X_{cq}[k_{cq}] = \frac{1}{N}\sum_{k=0}^{N-1} X[k]K^*[k, k_{cq}], \tag{3}$$

onde:

$$K[k,k_{cq}] = \sum_{n=0}^{N-1} w[n-(\frac{N}{2}-\frac{N(k_{cq})}{2}),k_{cq}]e^{j2\pi\frac{f_{k_{cq}}}{f_s}(n-\frac{N}{2})}e^{-j\frac{2\pi kn}{N}}. \quad (4)$$

É possível mostrar que, para um dado k_{cq}, a seqüência $K[k,k_{cq}]$ apresenta valores significativos apenas em uma faixa concentrada de valores de k. Considerando nulos os valores muito pequenos da seqüência, podemos reduzir drasticamente o número de multiplicações realizadas, obtendo um algoritmo rápido.

PRÉ-PROCESSAMENTO

A etapa inicial consiste na organização de um banco de dados formado pelas gravações citadas anteriormente.

As gravações foram realizadas em ambiente doméstico, com resolução de 16 bits e taxa de amostragem de 44100 Hz. Para registrar com fidelidade a sonoridade de cada violão, utilizamos um microfone com resposta na freqüência de +/− 1,5 dB sobre a faixa de interesse (de 77,78 Hz até 5274,04 Hz). Cada um dos sinais foi registrado com razão sinal/ruído máxima de aproximadamente 50dB.

Em cada gravação, apenas a corda utilizada na geração da nota podia vibrar. As 5 demais cordas permaneciam abafadas. Todos os sinais foram segmentados em blocos com 1 segundo de duração iniciados no ataque de cada nota, mesmo quando as durações das notas se estendiam além desse limite.

Foram criadas seis rotinas para organização dos sinais, cada uma referente a uma quantidade diferente de notas musicais. Em todas as rotinas os sinais eram divididos em grupos discriminados pelo violão utilizado e pela forma de execução, com ou sem palheta. Na rotina referente a apenas uma nota, os sinais segmentados formavam diretamente os grupos. Nas outras rotinas, foram realizadas combinações dos sinais através da soma de suas amplitudes e posterior divisão do resultado pelo número de notas combinadas[1].

Os sinais formados desse modo simulam combinações de notas tocadas simultaneamente por um músico. Desconsideramos os efeitos de interação entre cordas diferentes tocadas ao mesmo tempo. Em todos as rotinas, após a geração das combinações, os sinais foram multiplicados por uma janela de Hamming.

As notas utilizadas em cada combinação foram escolhidas aleatoriamente dentre as disponíveis, sob a condição de nenhuma nota aparecer mais que uma vez por combinação. Na prática essa situação é possível, ocorrendo quando o músico toca uma mesma nota simultaneamente em cordas diferentes. Como as 5 notas mais graves e as 5 mais agudas são representadas apenas uma vez por grupo de gravações - diferentemente das demais notas, que podem ser tocadas sobre posições diferentes do braço do instrumento - uma cópia adicional de cada é inserida no grupo. Cada grupo passa, assim, a ter 88 gravações diferentes. Dessa forma, todas as notas, em vez de todas as posições, têm a mesma probabilidade de aparecer em uma combinação.

Em cada rotina, 8 grupos de sinais sempre são reservados para a criação do conjunto de treinamento das redes. São formados pelos sinais de 4 violões, produzidos com e sem palheta. Outros dois grupos são reservados para a criação dos conjuntos de teste e validação. São formados pelos sinais do violão restante, produzido com e sem palheta.

Conforme será visto a seguir, existem algumas diferenças nas metodologias aplicadas na primeira e na segunda rotina em relação às demais. As diferenças foram determinadas em função da quantidade de dados disponíveis.

Descrições das Rotinas

Para a primeira rotina, referente a **uma nota** apenas, os conjuntos de treinamento, teste e validação foram criados da seguinte forma:

- Treinamento

  Todos os sinais dos 4 violões reservados para o treinamento das redes foram utilizados na criação do conjunto. As CQTs de cada um deles foram calculadas sobre a faixa que abrange desde 77,78 Hz (um semitom abaixo da nota mais grave do violão) até 5274,04 Hz (suficiente para cobrir até o quinto harmônico de C6, a nota seguinte à nota mais aguda de um violão comum). A precisão freqüencial escolhida foi de 1/8 de semitom. A análise se estende até 5274,04 Hz para evitar perda de informações sobre os harmônicos mais energéticos das notas mais agudas.

  Foram criados vetores com os valores absolutos de cada transformada calculada. Cada vetor foi normalizado de forma a tornar o somatório de seus elementos igual a 1. Em seguida, de cada vetor foi subtraída sua própria média.

  Essas representações espectrais foram armazenadas em uma matriz de representações. Uma matriz de objetivos também foi criada e associada à matriz de representações. Cada uma das colunas da matriz de objetivos é um

[1] Neste trabalho não foi considerada a variação de dinâmica na execução das notas. Tentou-se, no entanto, manter as amplitudes aproximadamente equalizadas.

vetor-objetivo que contém 44 elementos com valores 1 ou -1. Cada elemento pode ser associado a uma das 44 notas encontradas num violão comum, da seguinte forma: se, por exemplo, a representação espectral contida numa coluna x da matriz de representações for da nota G2 (quarta nota a partir de E2), então o quarto elemento da coluna x da matriz de objetivos é igual a 1 e todos os outros elementos na mesma coluna são iguais a -1.

Em seguida, as colunas da matriz de representações são permutadas em ordem aleatória. A mesma ordem é utilizada na permutação das colunas da matriz de objetivos.

- Teste e Validação

  Um procedimento similar foi realizado com os conjuntos de teste e validação, porém desta vez foram utilizados os sinais extraídos do violão restante. Neste caso os sinais são divididos em dois conjuntos com o mesmo número de elementos. A determinação de quais sinais formam os grupos também é aleatória. Não há restrição sobre quantas representações de sinais gravados com ou sem palheta formam cada grupo.

  Através da mesma metodologia aplicada na criação do conjunto de treinamento são criadas matrizes de representações e de objetivos para os grupos de teste e validação.

Para a segunda rotina, para **duas notas simultâneas**, foram avaliadas as possíveis combinações de 2 sinais por grupo que tenham notas diferentes, totalizando 3784 arranjos diferentes. Com essas combinações, o procedimento segue igual ao da primeira rotina, apenas com uma alteração: desta vez os vetores-objetivo são formados com dois elementos iguais a 1, em vez de apenas um.

As 4 demais rotinas, para criação dos conjuntos para análise de **3, 4, 5 ou 6 notas simultâneas**, respectivamente, são similares entre si. Elas se diferenciam da segunda rotina em 3 aspectos: pelo número de notas simultâneas analisadas; porque não foram geradas todas as combinações possíveis devido ao elevado número de possibilidades; e porque em cada rotina são criados 2 conjuntos de 4000 combinações de notas diferentes. Um dos conjuntos determina quais combinações são utilizadas no grupo de treinamento e o outro, quais são utilizadas no grupo de teste. Os dois conjuntos são criados independentemente.

Como as combinações do conjunto de treinamento são geradas de forma aleatória e não abrangem todas as possibilidades, criar os conjuntos de testes e validação a partir de combinações geradas independentemente possibilita a avaliação da robustez das redes.

AS REDES NEURAIS

Foram desenvolvidas 6 redes do tipo *feed-forward / backpropagation* totalmente conectadas, cada uma direcionada para a análise de quantidades diferentes de notas simultâneas.

Todas as redes foram criadas com a mesma topologia:

- duas camadas
- 176 neurônios na primeira camada e 44 neurônios na segunda
- todos os neurônios com função de ativação do tipo tangente hiperbólica

O treinamento buscava minimizar o erro quadrático médio através do método do gradiente descendente.

Critério de Parada

O treinamento das redes era paralisado quando o erro quadrático médio do conjunto de validação tendia a aumentar.

PÓS-PROCESSAMENTO

Para cada rede referente a n notas simultâneas, assumiu-se que as posições dos n maiores valores encontrados nos vetores de saída indicariam as notas que devem ser classificadas como presentes na combinação analisada. Assim, para a rede de uma nota, apenas o maior valor entre os elementos do vetor é considerado. Para a rede de duas notas, os dois maiores valores são considerados, e assim por diante.

RESULTADOS

A seguir, listam-se os resultados finais das simulações descritas acima.

- Para uma nota:
  - MSE = 0,000258.
  - Percentual de erros = 0.
- Para duas notas:
  - MSE = 0,002402.
  - Percentual de erros = 1,64%, sempre com uma só nota errada.
- Para três notas:
  - MSE = 0,003959.
  - Percentual de erros = 6,22%, sempre com uma só nota errada.

- Para quatro notas:
  - MSE = 0,009764.
  - Percentual de erros = 14,37%, sendo 14,22% com uma nota errada e 0,15% com duas notas erradas.
- Para cinco notas:
  - MSE = 0,015622.
  - Percentual de erros = 22,10%, sendo 21,83% com uma nota errada e 0,27% com duas notas erradas.
- Para seis notas:
  - MSE = 0,023388.
  - Percentual de erros: 32,72%, sendo 31,45% com uma nota errada e 1,27% com duas notas erradas.

CONCLUSÕES

Foi apresentada uma topologia baseada em redes neurais para identificação de notas de violão tocadas simultaneamente. Foi definida e executada uma estratégia de simulações para quantificar a dificuldade da tarefa em relação ao número de notas executadas. Como se esperava, o desempenho das redes projetadas associado ao método de análise de seus vetores de saída mostrou depender fortemente do número de notas simultâneas analisadas. Embora erros de mais de 10% possam não ser toleráveis conforme a aplicação em vista, os resultados preliminares para a topologia proposta pareceram promissores, embora tendo sido dissociados os tratamentos de diferentes números de notas executadas. Basta observar que para combinações de até 3 notas, só houve erros de 1 nota, e para combinações de 4 a 6 notas, só houve erros de 1 ou 2 notas. A busca de uma tendenciosidade nesses erros (para que notas ocorriam, e quais as notas erroneamente acusadas?) deve indicar possíveis formas de reduzi-los.

Deve-se observar que a comparação dos resultados do presente artigo com os dos trabalhos referenciados (como também entre estes) não pode ser feita diretamente. Os diferentes critérios de avaliação dos erros e o uso de bancos de dados distintos impedem a comparação coerente entre os métodos.

Próximas metas possíveis neste trabalho, além da investigação minuciosa das ocorrências dos erros: criar um sistema unificado para tratamento de qualquer número de notas, o que pode envolver estratégias heurísticas para determinação de sua arquitetura; estudar diferentes métodos de tratamento das amplitudes, para por fim avaliar o efeito de variações de dinâmica. Também pode ser desenvolvido um sistema que determine apenas os intervalos entre as notas presentes numa combinação, sem se ocupar de suas alturas absolutas.

AGRADECIMENTOS

Os autores gostariam de agradecer ao eng. Gustavo Luis Almeida de Carvalho por sua contribuição na etapa inicial deste trabalho e às agências de fomento CAPES, FAPERJ e CNPq pelo apoio na forma de bolsa de mestrado e de auxílio a projetos de pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] A. Isaacs and E. Martin, *Dicionário de Música*. Rio de Janeiro, RJ: Zahar, 1985.

[2] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, January 1991.

[3] M. Marolt, "A comparison of feed forward neural network architectures for piano music transcriptions," *Proceedings of the 1999 International Computer Music Conference*, Beijing, China, 1999.

[4] M. Marolt, "Sonic: Transcription of Polyphonic Piano Music with Neural Networks," in *Proceedings of the Workshop on Current Research Directions in Computer Music*, (Barcelona, Spain), 2001.

[5] A. Klapuri, *Signal Processing Methods for the Automatic Transcription of Music*. Ph.D. dissertation, Tampere University of Technology, Tampere, Finland, March 2004.

[6] A. Klapuri, "A Perceptually Motivated Multiple-F0 Estimation Method," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Paltz, NY, USA), October 2005.

[7] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant-Q Transform," *J. Acoust. Soc. Amer.*, vol. 92, no. 5, pp. 2698–2701, November 1992.

[8] B. Blankertz, "The constant Q Transform." URL: http://ida.first.fhg.de/publications/drafts/Bla_constQ.pdf.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Accurate and Efficient Fundamental Frequency Determination from Precise Partial Estimates

Adriano Mitre[1], Marcelo Queiroz[1], Regis R. A. Faria[2]

[1]Department of Computer Science,
Institute of Mathematics and Statistics, University of São Paulo

[2]Laboratory of Integrated Systems,
Polytechnic School, University of São Paulo

adriano@mitre.com.br, regis@lsi.usp.br, mqz@ime.usp.br

ABSTRACT

An algorithm is presented for the estimation of the fundamental frequency (F_0) of monophonic sounds. The method relies upon accurate partial estimates, obtained on a frame basis by means of enhanced Fourier analysis. The use of state-of-the-art sinusoidal estimators allows the proposed algorithm to work with frames of minimum length (i.e., about two fundamental periods). The accuracy of the proposed method does not degrade for high pitched sounds, making it suitable for musical sounds.

INTRODUCTION

Extracting the fundamental frequency (F_0) contour of a monophonic sound recording has a number of applications, such as audio coding, prosodic analysis, melodic transcription and onset detection.

Pitch determination in speech signals is a extensively studied topic, mostly motivated by immediate applications in telecommunications. Musical pitch estimation, however, has received considerably less attention.

Speech and musical pitch estimation pose different challenges for pitch determination algorithms (PDA). Fundamental frequency estimation in music signals is in many ways more challenging than that in speech signals. In music, the pitch range can be wide, comprising more than seven octaves, and the sounds produced by different musical instruments vary a lot in their spectral content. The inharmonicity phenomenon has to be taken into account.

On the other hand, the dynamic (time-varying) properties of speech signals are more complex than those of an average music signal. The F_0 values in music are temporally more stable than in speech.

Despite the aforementioned differences, it is occasionally possible to employ speech-tailored PDAs to monophonic musical recordings, with variable degree of success.

The human voice and most pitched musical instruments used in Western music produce quasi-harmonic sounds[1]. The reason for this is encountered in the physics of vibrating strings and tubes. As the pitch of a quasi-harmonic sounds is closely related to its fundamental frequency, both terms were used indistinctly in the present work.

PROPOSED METHOD

A number of techniques have been proposed for pitch estimation, mostly aiming at measuring periodicity in the time or frequency domain. Most funda-

[1]The mallet percussion family is a notable exception.

mental frequency estimation methods may be classified according to the domain on which they operate. The ones which operate directly on the signal waveform are termed time-domain methods. Methods which transform the waveform to a spectral representation are called frequency-domain methods. This transformation is usually carried out by means of constant Q or short-time Fourier transforms (STFT).

Although the proposed method employs the Fourier transform, it does not operate on the complete spectrum signal, but rather on a small set of partials. It requires frequency analysis, followed by extraction and estimation of partials. The list of partials in each frame is the input to the proposed algorithm.

The main steps of the proposed method are shown in Figure 1.

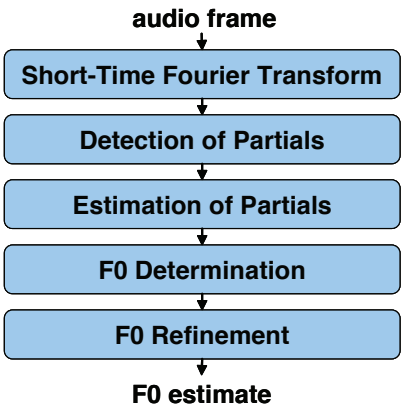


Figure 1: Flowchart of the proposed method.

Detection of Partials

The spectral analysis module produces, for each audio frame, its corresponding complex spectrum. Notwithstanding, we note that only prominent partials are relevant for fundamental frequency estimation.

Under reasonable assumptions, each partial in the input signal produces a local maximum in the magnitude spectrum; the converse is not true due to smearing effects and noise intrinsic to discrete analysis. Therefore several heuristics were proposed to discriminate local maxima induced by partials from those induced by noise. A popular strategy in analysis/resynthesis systems is partial tracking [1, 2], which does not operate on isolated frames and thus suggests an offline partial filtering strategy.

In the present study, the discrimination between genuine and spurious peaks is postponed to the subsequent module. In this approach every peak is estimated "as if it were" a partial. Then, the ones whose frequency estimate depart more than half bin from its original value are discarded as noise.

Estimation of Partials

In order to correctly estimate a 12-tone-equal-tempered pitch from a given fundamental frequency, an accuracy[2] of at least $\mathrm{F0}_{\min}\left(\sqrt[24]{2}-1\right)$ Hz is needed, where $\mathrm{F0}_{\min}$ denotes the lowest expected fundamental frequency in the input signal. In order to accurately follow expressive subtleties such as *vibrati* and *glissandi* a higher accuracy is needed.

Frequency accuracy of conventional STFT is half the inverse of frame length, represented by $\{2\tau\}^{-1}$ Hz. STFT's frequency resolution[3], although constrained by the frame length, depends also on the window shape. More precisely, it is determined by the 6 dB bandwidth of the window power spectrum main lobe and is given by $L_w \cdot \tau^{-1}$ Hz, where L_w depends on the window. For classic windows, such as Hann and Blackman, L_w lies between 1.2 and 3.1 [3, 4].

For instance, in order to discriminate between pitches of a 6-stringed guitar whose lowest pitch is an E corresponding to 82.4 Hz, one needs a frame of duration at least $\left[2 \times 82.4 \times \left(\sqrt[24]{2}-1\right)\right]^{-1} \simeq 207$ ms. Musical signals seldom exhibit quasi-periodic behaviour for so long. Large frames tend to lower temporal precision because of contamination from two or more succesive notes occurring in a single analysis frame. In addition, a temporal accuracy of 20 ms asks for an overlap factor of 90% and therefore raises the computational workload by a factor of ten.

In monophonic quasi-harmonic signals any two partials are at least $\mathrm{F0}_{\min}$ Hz apart and thus a frame length of $L_w \cdot \mathrm{F0}_{\min}{}^{-1}$ s is enough for them to be resolved (i.e., separated). This new bound is much tighter than the previous one. For the guitar example, a Hamming-windowed frame of $1.81 \times 82.4^{-1} \simeq 22$ ms is enough.

Fortunately, several techniques exist for improving the estimates of resolved partials. These generally fall into two categories, phase-based and interpolation-based.

Interpolation-based Techniques

One of the techniques for improving the estimates of sinusoidal components is spectral oversampling. It is usually attained by means of zero-padding, which consists in adding a sequence of zeros to the windowed frame before computing the STFT. The disadvantage of spectral oversampling is that the increase in the computational workload is proportional to the improvement in accuracy.

Another technique is quadratic (or parabolic) interpolation, whose estimates are computed using each local maximum of the spectrum and its adjacent bins. It benefits from the fact that the main lobe of the logarithmic power spectrum of several windows are

[2] In the present work, the term accuracy is used in the sense of exactness. An estimator is thus said to have accuracy ϵ if every estimate is within ϵ of its true value, i.e., $|\hat{f}_i - f_i| < \epsilon$ for all i.

[3] Throughout the text, frequency resolution will refer to how close two sinusoids may get while still being separable in the spectrum. A resolution of Δ means that two sinusoids with same amplitude and frequencies f_1 and f_2 may separated if and only if $|f_1 - f_2| \geq \Delta$ and $\min\{f_1, f_2\} \geq \Delta$. The second inequality is due odd-simmetry of the spectrum of real signals.

very close to a quadratic function. Purposefully designed windows are sometimes employed, which are obtained by taking the inverse transform of a perfect quadratic function. The parabolic interpolation technique is often combined with spectral oversampling.

For the special case of the Hann window, Grandke designed an interpolation technique which considers each peak and its greatest neighbour [5].

A number of interpolation techniques exist for the rectangular-windowed STFT[4], however spectral leakage problems prevent the use of rectangular window for musical signal analysis.

Phase-based Techniques

More sophisticated partial estimation techniques use the phase spectrum in addition to magnitude information. The Derivative Method [6] uses the spectra of the original signal and its derivative (aproximated by a low-pass filter) and the Spectral Reassignment Method [7, 8] associates energy content to the cells of a time-frequency representation in order to improve accuracy of the estimates. Thanks to a trigonometric interpretation of the Derivative Method, an improved estimator was derived in [9]. The new estimator is as precise for close-to-Nyquist frequencies as the Derivative is for low frequencies.

These techniques give better estimates at the expense of additional STFT computations. Comparative studies of these techniques with respect to mean error, variance and bias can be found in [10] and [11].

Amplitude Estimation

Except by quadratic interpolation and spectral oversampling, the aforementioned techniques only estimate the frequency of partials. Nevertheless, one can obtain precise amplitude estimates of partials by applying analytical knowledge about the window used.

Denoting by $\hat{f}_k$ the frequency estimate of the partial at the k-th bin, whose center frequency is f_k, and by W the frequency response of the window, the precise amplitude estimate for the partial is given by the formula

$$\hat{a}_k = \frac{a_k}{W\left(\left|\hat{f}_k - f_k\right|\right)} \tag{1}$$

Prior to fundamental frequency determination, described in the "Fundamental Frequency Determination" section, the magnitude of the partials must be normalized to absolute decibels. This is accomplished by the following formula.

$$\hat{a}_k^{\mathrm{dB-norm}} = \alpha + 20 \cdot \log_{10} \hat{a}_k \tag{2}$$

The term α is set to map the maximum possible amplitude to 70 dB. It is determined by the window size (in samples), the windowing function and the recording bit-depth.

[4]Rectangular-windowed STFT is often misleadingly referred to as unwindowed, instead of *unsmoothed*, STFT.

Finally, non relevant partials are filtered prior to fundamental frequency determination. A partial is considered relevant if its frequency is within human hearing range (20−20,000 Hz) and its magnitude is strictly positive.

Fundamental Frequency Determination

The proposed method assumes that the strongest partial belongs to the main harmonic series, thus its frequency is expected to be multiple of F_0. Letting $f_\star$ denote the frequency corresponding to the strongest partial, the set of candidates for F_0 is composed by submultiples of $f_\star$. Formally,

$$C = \left\{ c_n \stackrel{\text{def}}{=} \frac{f_\star}{n} : 1 \le n \le \left\lfloor \frac{f_\star}{\mathrm{F0}_{\min}} \right\rfloor \right\} \tag{3}$$

The next step consists in collecting the harmonic series corresponding to each F_0 candidate. This is carried out by the following algorithm: firstly, partials are sorted in decreasing order of magnitude; then, each partial is sequentially assigned to the nearest (in a quarter tone vicinity) "empty slot" of the candidate's harmonic series.

As a result of the previous algorithm, the i-th harmonic of the n-th candidate is given by

$$H[n][i] = \arg\max_{p \in \Lambda_i^n} \left\{ p_{\mathrm{mag}} \right\} \tag{4}$$

where p denotes a partial with frequency p_{freq} and magnitude p_{mag}. In words, $H[n][i]$ is the partial with greatest magnitude among the set of potential i-th harmonic of the n-th candidate, given by

$$\Lambda_i^n = \left\{ p : l_i < \frac{p_{\mathrm{freq}}}{i c_n} < h_i \right\} \tag{5}$$

where l_i and h_i ensure smaller than quarter-tone deviation and, in the case of higher order harmonics, prevent single partials from being assigned to multiple adjacent harmonics "slots". Formally,

$$l_i = \max\left\{ \sqrt[24]{2^{-1}}, \sqrt{\frac{i-1}{i}} \right\} \tag{6}$$

$$h_i = \min\left\{ \sqrt[24]{2}, \sqrt{\frac{i+1}{i}} \right\} \tag{7}$$

In short, if the i-th harmonic of the n-th candidate belongs to the spectrum, it will be assigned to $H[n][i]$. Otherwise, it is agreed that $H[n][i]_{\mathrm{mag}} = 0$.

It is further necessary to quantify the prominence of each candidate according to its harmonic series. This takes into account psychoacoustic factors, particularly the critical band [12, §2.4 and §3.4]. The functions Φ and Ψ defined below are based on the harmonic sum model [13, §6.3.3]. The psychoacoustic motivation for these formulas can be found in the same reference.

Formally stating, the prominence of the n-th candidate is given by

$$\Phi(n) = \sum_{i=1}^{I(n)} H[n][i]_{\text{mag}} \cdot \Psi(i) \tag{8}$$

$$I(n) = \max\left\{ j : H[n][j]_{\text{mag}} > 0 \right\} \tag{9}$$

and $\Psi(i)$ denotes the fraction of the critical band which corresponds to the i-th harmonic, given by

$$\Psi(i) = \begin{cases} 1, & \text{if } i \leq 4 \\ \Gamma(i) - \Gamma(i-1), & \text{otherwise} \end{cases} \tag{10}$$

$$\Gamma(n) = \log_{2^{1/3}}\left(n \cdot \sqrt{\frac{n+1}{n}} \right) \tag{11}$$

The fundamental frequency estimation is performed in three steps, given the prominence of the candidates as defined above. The first step selects those candidates with relative prominence of at least $\beta \in [0, 1]$ with respect to the maximal prominence:

$$C^{\Phi} = \left\{ c_n \in C : \Phi(n) \geq \beta \cdot \max_{m | c_m \in C} \{\Phi(m)\} \right\} \tag{12}$$

For each of these candidates the weighted average harmonic magnitude is computed as:

$$\chi(n) = \frac{\sum_{i=1}^{I(n)} H[n][i]_{\text{mag}} \cdot \Psi(i)}{\sum_{i=1}^{I(n)} \Psi(i)} \tag{13}$$

Then the one with the highest value of χ is selected as F_0, whose index is

$$\varphi = \arg\max_{n : c_n \in C^{\Phi}} \{\chi(n)\} \tag{14}$$

Fundamental Frequency Refinement

The exact value of the estimated F_0 was based on the frequency estimate of a single partial: the strongest one. However, the F_0 estimate may be improved by considering frequency estimates of all partials in the harmonic series of the winner candidate. Since partial estimates are expected to be non-biased, individual errors should cancel each other out by averaging.

The realiability of a partial estimate is affected by its signal-to-noise ratio (SNR) and the stability of its absolute frequency. Therefore strong and small indexed harmonics should be privileged, since they have the higher SNR and smallest absolute frequency modulations.

Taking these facts into account, we propose the following formula for further refining the initial fundamental frequency:

$$\hat{F}_0 = \frac{\sum_{i=1}^{I(n)} H[i]_{\text{freq}}/i \cdot H[i]_{\text{mag}} \cdot \Psi(i)}{\sum_{i=1}^{I(n)} H[i]_{\text{mag}} \cdot \Psi(i)} \tag{15}$$

where $H[i]$ denotes the i-th partial of the harmonic series of c_{φ}, which is, $H[i] \stackrel{\text{def}}{=} H[\varphi][i]$.

The F_0 refinement might be thought as an weighted average of local F_0 estimates. Local estimates should be understood regarding the harmonic indice, i.e., the local F_0 estimate for the i-th harmonic is $H[i]_{\text{freq}}/n$.

ADVANTAGES AND DRAWBACKS

It is well known that spectral and temporal resolutions are reciprocals and thus detecting F_0 as low as f Hz requires a window whose length is at least $K \cdot f^{-1}$ s, where K is independent of f. In the case of Fourier spectrum based methods, K is mainly determined by the window [3].

On the one hand, all short-time F_0 estimators suffer from this limitation. On the other hand, while waveform-based PDAs have their precision determined (i.e., fixed) by the signal's sample rate, the precision of F_0 estimates produced by spectrum-based PDAs might be increased by employing longer windows. Notwithstanding, the use of interpolation may be helpful for methods on either domain.

The precision of the proposed method has the same order of magnitude as that of the sinusoid estimator employed, occasionally surpassing it due to the refinement procedure. It must be noted, however, that if spurious peaks in the magnitude spectrum are incorrectly classified as partials and collected to the harmonic series of the winner F_0 candidate, the refinement stage may degrade, instead of enhance, the initial F_0 estimate.

The method is timbre-independent, being robust to the following phenomena:

- weak or absent fundamental
- incomplete series (e.g., only odd harmonics)
- sinusoidal-like sounds
- moderate levels of inharmonicity (as found in acoustic instruments)

It must be noted that although inharmonicity is not explicitly modelled, the tolerance of the harmonic series collector allows for moderately inharmonic low order partials.

Experiments conducted with severely bandlimited (e.g. telephone-like bandpass filtered) versions of musical recordings have shown that the method is robust against bandlimiting. In some sense this is expected, since the method is partially derived from a bandwise multiple-F_0 estimator [14].

IMPLEMENTATION ISSUES

Profiling revealed that the most processing-intense step of the proposed method is the calculation of the STFT, which can be carried out by the Fast Fourier Transform algorithm.

The memory required by the method, excluding the STFT, is proportional to $|C|$, the number of candidates. It can be seen from Equation 3 that $|C|$ is indirectly dependant on the window length, as $\mathrm{F0}_{\min}$ should never be lower than $L_w \cdot \tau^{-1}$. Notwithstanding, the number of candidates can be safely assumed to be smaller than 200, as in musical sounds it is usually the case that $f_\star < 5$ kHz and $\mathrm{F0}_{\min} > 27.5$ Hz.

Thus, not only the processing, but also the memory requirements of the proposed method are dominated by the STFT.

EXPERIMENTS AND RESULTS

By the writing of this article, only informal (although extensive) evaluation was conducted. The results were, in general, very encouraging. Figures 2 and 3 show F_0 contours produced by the proposed method with expressive recordings of acoustic instruments.

There were two main reasons that retarded formal evaluation. The first reason is that there is no standardized *musical* database available for the task of PDA evaluation, i.e., one which provides reference F_0 tracks along with the audio recordings. The second reason is that, to the best of authors knowledge, there is no tool available for automatic generating statistics from reference and estimated F_0 tracks.

In an effort to remedy the situation, an automatic PDA evaluation tool was developed and musical monophonic recordings were collected, comprising most acoustic, electric and electronic instruments. In spite of this, manually obtaining reference F_0 tracks for the recordings is a laborious process which could not be concluded until the article's submission deadline.

It must be stressed that formal evaluation will be carried out. As soon as the work is done, the recordings, reference F_0 tracks, evaluation tool and results will be made available at http://www.mitre.com.br/pda.

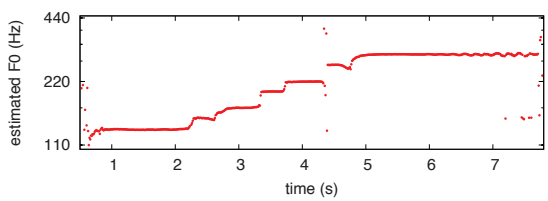


Figure 2: Expressive saxophone performance of the initial notes of a jazz standard.

CONCLUSION

A new algorithm was proposed for monophonic F_0 estimation. The method benefits from state-of-the-art partial estimators to reduce the required analysis frame length to a minimum (i.e., about two fundamental periods). This accounts for increased time resolution and reduced computational workload. The reduced number of configuration parameters makes it easier to fine-tune the method. Furthermore, informal evaluation suggests that the method is very robust for musical sounds.

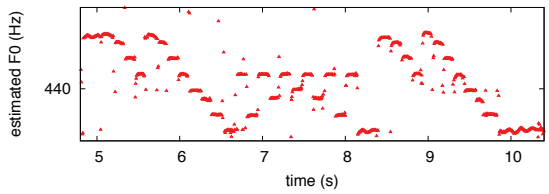


Figure 3: Expressive violin performance of an excerpt from a classical piece.

REFERENCES

[1] Robert J. McAulay and Thomas F. Quatieri. Speech Analysis/Synthesis Based on a Sinusoidal Representation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 34(4):744–754, 1986.

[2] Mathieu Lagrange, Sylvain Marchand, Martin Raspaud, and Jean-Bernard Rault. Enhanced Partial Tracking Using Linear Prediction. In *Proceedings of the 6th International Conference on Digital Audio Effects(DAFx-03)*, Londres, Reino Unido, 2003.

[3] Fredric J. Harris. On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform. *Proceedings of the IEEE*, 66(1), January 1978.

[4] Albert H. Nuttall. Some Windows with Very Good Sidelobe Behavior. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(1):84–91, February 1981.

[5] Thomas Grandke. Interpolation algorithms for discrete Fourier transforms of weighted signals. *IEEE Transactions on Instrumentation and Measurments*, 32(2):350–355, June 1983. 1983.

[6] Myriam Desainte-Catherine and Sylvain Marchand. High Precision Fourier Analysis of Sounds Using Signal Derivatives. *Journal of the Audio Engineering Society*, 48(7/8):654–667, July/August 2000.

[7] Kunihiko Kodera, Roger Gendrin, and Claude de Villedary. Analysis of time-varying signals with small BT values. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):64–76, February 1978.

[8] Francçois Auger and Patrick Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5):1068–1089, May 1995.

[9] Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault. Improving sinusoidal frequency estimation using a trigonometric approach. In *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx-05)*, Madrid, Spain, September 20-22 2005.

[10] Florian Keiler and Sylvain Marchand. Survey On Extraction of Sinusoids in Stationary Sounds. In *Proceedings of the 5th International Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, September 2002.

[11] Stephen Hainsworth and Malcolm Macleod. On Sinusoidal Parameter Estimation. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, London, United Kingdom, September 2003.

[12] Juan G. Roederer. *The Physics and Psychophysics of Music: An Introduction*. Springer-Verlag Telos, 3rd edition, 1995.

[13] Anssi Klapuri. *Signal Processing Methods for the Automatic Transcription of Music*. PhD thesis, Tampere University of Technology, March 2004.

[14] Anssi P. Klapuri. Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6):804–816, November 2003.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Automatic Genre Classification of Musical Signals

Jayme Garcia Arnal Barbedo[1], Amauri Lopes[1]
[1]Department of Communications – FEEC – Unicamp
Campinas, São Paulo, C.P. 6101, Brazil
jgab@decom.fee.unicamp.br , amauri@decom.fee.unicamp.br

ABSTRACT

This paper presents a strategy to perform automatic genre classification of musical signals. The technique divides the signals into 21.3 ms frames, from which 7 features are extracted. The frames are grouped into 1 s analysis segments. Some statistical results of the features along each analysis segment are used to calculate a vector of parameters. An extensive comparison is carried out between such segment vectors and some reference vectors. The procedure points out the genre that best fits the characteristics of each segment. The final classification of the signal is given by the genre that appears more times along all signal segments.

1. INTRODUCTION

The advances in information, communication and media technologies experienced in the last decades have made available a large amount of all kinds of data. This is particularly true for music, whose databases have grown exponentially since the advent of the first perceptual coders early in the 90's. This situation demands tools able to ease searching, retrieving and handling such huge amount of data. Among such tools, automatic musical genre classifiers (AGC) can have a particularly important role, since they could be able to automatically index and retrieve audio data in a human-independent way. This is very useful because a large portion of the metadata used to describe music content is inconsistent or incomplete.

Audio search and retrieval is the most important application of AGC, but is not the only one. There are several other technologies that can benefit from AGC. For example, it would be possible to create an automatic equalizer able to choose which frequency bands should be attenuated or reinforced according to the label assigned to the signal being considered. AGC could also be used to automatically select radio stations playing a particular genre of music.

There are not many previous works that specifically deal with musical genre classification in the literature. The most significant proposal to specifically deal with this task was [1], and some other works followed its paths [2, 3]. Several strategies dealing with related problems have been proposed in research areas such as speech/music discriminators [4-7] and classification of a variety of sounds [8, 9].

The strategy presented here divides the audio signals into 21.3 ms frames from which the following 7 features are extracted: zero-crossing rate (ZCR), spectral centroid, bandwidth, spectral roll-off, spectral flux, loudness and fundamental frequency. The frames are grouped into 1 s analysis segments, and the results of each feature along each analysis segment are used to calculate three parameters: mean, variance, and a third parameter called "prevalence of the main peak". Therefore, a 21-element vector, from now on called "test vector", will be associated to each segment. In the next step, the test vectors are compared to a set of reference vectors that characterize each one of the 13 musical genres here considered. The comparison procedure consists in calculating the Euclidean distance between test and reference vectors, and is carried out in a pair-of-genres basis, meaning that each test vector

is always tested against the reference vectors of only two musical genres at a time. For each pair of genres, the label of the reference vector that is closer to the test vector is taken as winner genre for that specific segment and pair of genres. After all possible combinations of pairs have been considered, the genre that has won more times is taken as the preliminary label for that segment. The procedure is repeated for all segments. The final classification of the signal is given by the genre that has been taken as preliminary label for the greatest number of segments.

2. DISCUSSIONS ON GENRE LABELING

Besides the inherent complexity involved in differentiating and classifying musical signals, the AGC have to face other difficulties that make this a very tricky area of research. In order to work properly, an AGC technique must be trained to classify the signals according to a predefined set of genres. However, there are two major problems involved in such predefinition, which will be discussed next.

Firstly, the definition of most musical genres is very subjective, meaning that the boundaries of each genre are mostly based on individual points-of-view. As a result, each musical genre can have its boundaries shifted from person to person. The degree of arbitrariness and inconsistency of music classification into genres can be found in [10], where the authors compared three different Internet genre taxonomies: allmusic.com, amazon.com and mp3.com. The authors drawn three major conclusions:

- there is no agreement concerning the name of the genres – only 70 words are common to all three taxonomies;
- among the common words, not even largely used names, as “Rock” and “Pop”, denote the same set of songs.
- the three taxonomies have different hierarchical structures.

As pointed out in [11], if even major taxonomic structures present so many inconsistencies among them, it is not possible to expect any degree of semantic interoperability among different genre taxonomies. Despite such difficulties, there have been efforts to develop carefully designed taxonomies [10, 11]. However, no unified framework has been adopted yet.

To deal with such difficulty, the taxonomy adopted in this work was designed using genres and nomenclatures that are largely used by most reference taxonomies (like the three ones cited before), and therefore are most likely to be readily identified by most users. This procedure reduces the inconsistencies and tends to improve the precision of the method, as will be seen in Section 5. However, it is important to emphasize that some degree of inconsistency will always exist due to the subjectiveness involved in classifying music, situation that limits the reachable accuracy.

The second major problem is the fact that a large part of modern songs have elements from more than one musical genre. For example, there are some jazz styles that incorporate elements of other genres, as Fusion (jazz + rock); there are also recent reggae songs that have strong elements of rap; as a last example, there are several rock songs that incorporate electronic elements generated by synthesizers. To deal with this problem, the strategy used in this work is to divide basic genres into a number of subgenres able to embrace such intermediate classes, as will be described in the next Section.

3. TAXONOMY

Figure 1 shows the structure of the taxonomy adopted in the present work.

As can be seen in Figure 1, there is a maximum of 4 hierarchical layers and a total of 13 musical genres in the lowest layer. The description of each box is presented next. Such taxonomy was created aiming to include as many genres as possible, improving the generality of the method, but keeping at the same time the consistency of the taxonomy, as commented in Section 2. It is also important to highlight that as many genres are considered, the more difficult is to perform a correct classification. Therefore, under this point-of-view the strategy proposed here faces harder conditions than previous ones.

From this point to the end of the paper, all musical classes of the lowest hierarchical level in Figure 1 are called “genres”, while the divisions of higher levels are called “upper classes” or simply “classes”.

3.1. Classical

The songs of this class have the predominance of classical instruments like violins, cello, piano, flute, etc. This class is divided into two genres:

- instrumental: songs of this genre have no vocal elements;
- opera/chorus: this genre includes opera and classical songs where the orchestra is accompanied by a chorus.

3.2. Pop/Rock

This is the largest class of songs. The first division of this class is based in the presence or not of electronic elements, which are normally generated by synthesizers:

- if there is a predominance of electronic elements, the signals are classified as “electronic”;
- if there are no electronic elements, or such elements are very mild, the signals are classified as “organic”.

The subclass “electronic” is divided into the genres “pop” and “techno”. Techno songs have a faster and more repetitive beating than pop songs.

The subclass “organic” is still split one more time before reaching the lowest level:

- Songs classified as “rock” have a predominance of electrical guitars and drums. The final division of this class into genres is performed taking into account the rhythm and intensity of the song. Songs classified as “soft rock” are slow and soft; songs classified as hard rock have a more marked beating, stronger presence of drums and a faster rhythm; finally, songs classified as heavy metal are noisy, fast, and often have very aggressive vocals.

- Songs classified as “country” are quite related to rock. As in the case of rock, electrical guitars play an important role, but they have a particular sonority that is common in folk songs typical of southern United States. The final division of this class into “soft country” or “dancing country” is performed according to the rhythm, which is slow in the first case and fast in the second one.

3.3. Beating

The songs that compose this third and last musical class have strong percussive elements and a very marked beating. The first division of this class is as follows:

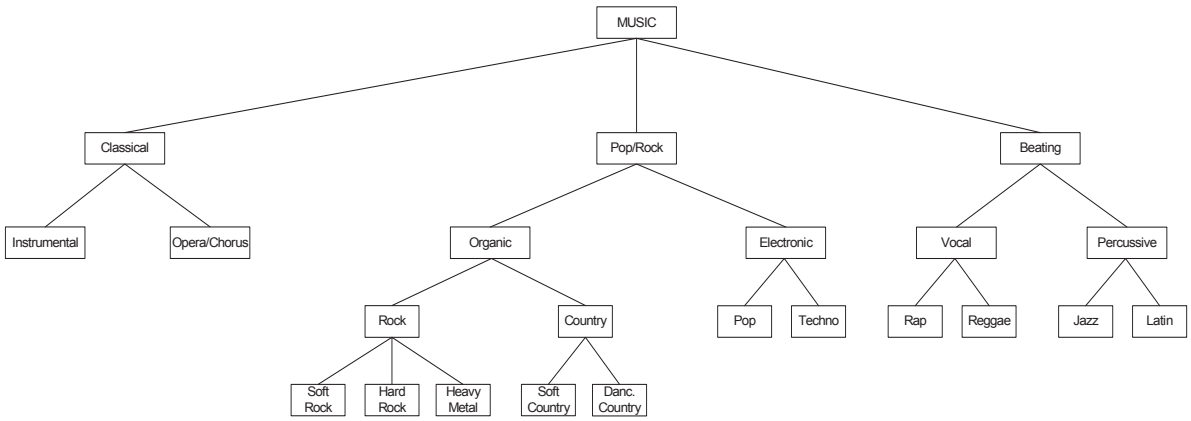


Figure 1 - Musical Genre Taxonomy.

- if the vocal elements are strong and dominate the song, the signal is classified as "vocal";
- if the percussive elements dominate the perception, the song is classified as "percussive".

The vocal class is further divided into two genres: rap, whose songs have really marked vocals, sometimes looking like actual speech, and reggae, the typical music of Jamaica. Some recent reggae songs are quite related to rap, situation that can cause some difficulties to differentiate such genres.

Finally, the percussive class is divided into two genres:

- "Jazz", which are songs dominated by piano and saxophone. Electric guitars and drums can also be present, especially in modern tendencies of jazz like Fusion; vocals, when present, are very characteristic and peculiar.
- "Latin", which is composed by Latin rhythms like salsa, mambo, samba and rumba; the songs of this genre have a very dancing and percussive rhythms, with strong presence of instruments of percussion and, sometimes, guitars.

4. FEATURE EXTRACTION

Before the feature extraction, the signal is divided into frames using a Hamming window of 21.3 ms, with 50 % superposition. The signals used in this work are sampled at 48 kHz, resulting in frames of 1,024 samples. The extraction of the features is performed individually for each frame. The description of each feature is presented in the following.

4.1. Zero-Crossing Rate

A zero crossing occurs whenever the amplitudes of two consecutive signal samples have opposed signs. The ZCR for a given frame is given by

$$zcr_i = 0.5 \cdot \sum_{n=1}^{N} \left| \operatorname{sgn}\left[x_i(n)\right] - \operatorname{sgn}\left[x_i(n-1)\right] \right|, \quad (1)$$

where $x_i(n)$ represents the samples of ith frame and $sgn[x_i(n)]$ is –1 or +1 as $x_i(n)$ is negative or positive respectively.

4.2. Spectral Roll-Off

This feature determines the frequency R_i for which the sum of the spectral line magnitudes is equal to 95% of the total sum of magnitudes, as expressed by

$$\sum_{k=1}^{R_i} X_i(k) = 0.95 \cdot \sum_{k=1}^{K} X_i(k), \quad (2)$$

where $|X(k)|$ is the magnitude of spectral line k resulting from a Discrete Fourier Transform with 1,024 samples applied to the frame i and K is half the number of spectral lines.

4.3. Loudness

The first step to calculate this feature is modeling the frequency response of human outer and middle ears. Such response is given by [12]

$$W(k) = -0.6 \cdot 3.64 \cdot f(k)^{-0.8} - 6.5 \cdot e^{-0.6 \cdot (f(k)-3.3)^2} + 10^{-3} \cdot f(k)^{3.6}, \quad (3)$$

where $f(k)$ is the frequency in kHz given by

$$f(k) = k \cdot d, \quad (4)$$

and d is the difference in kHz between two consecutive spectral lines (in this work, 46.875). The frequency response is used as a weighting function that emphasizes or attenuates spectral components according to the hearing behavior. The loudness of a frame is calculated according to

$$ld_i = \sum_{k=1}^{K} \left| X_i(k) \right|^2 \cdot 10^{\frac{W(k)}{20}}. \quad (5)$$

4.4. Spectral Centroid

This feature represents the mass center of the spectral energy distribution of the signals, and is given by

$$ec_i = \frac{\sum_{k=1}^{K} k \cdot \left| X_i(k) \right|^2}{\sum_{k=1}^{K} \left| X_i(k) \right|^2}. \quad (6)$$

The spectral centroid is given in terms of spectral lines. To obtain the value in Hz, ce must be multiplied by d.

4.5. Bandwidth

This feature determines the frequency bandwidth of the signal, and is given by

$$bw_i = \sqrt{\frac{\sum_{k=1}^{K} \left[(ce_i - k)^2 \cdot \left| X_i(k) \right|^2 \right]}{\sum_{k=1}^{K} \left| X_i(k) \right|^2}}. \quad (7)$$

Equation 7 gives the bandwidth in terms of spectral lines. To get the value in Hz, lb must be multiplied by d.

4.6. Spectral Flux

This feature is defined as the quadratic difference between the logarithms of the magnitude spectra of consecutive analysis frames and is given by

$$fe_i = \sum_{k=1}^{K} \left\{ \log_{10}\left[X_i(k)\right] - \log_{10}\left[X_{i-1}(k)\right] \right\}^2 . \quad (8)$$

The purpose of this feature is to determine how fast the signal spectrum changes along the frames.

4.7. Fundamental Frequency

This feature is based on the concept of multiple fundamental frequency detection. Since most audio signals are polyphonic (several sound sources), some kind of processing must be applied in order to accurately detect multiple fundamental frequencies. Most of the strategy described in the following is inspired in the multipitch analysis model presented in [13], as illustrated in Figure 2.

As can be seen, the input (signal frames) is divided into two bands by a filtering process. The high frequency portion of the input is obtained blocking frequencies below 1 kHz, while a 70-1000 Hz passband filter determines the low frequency portion. The high frequency portion is then submitted to a half-wave rectification. After that, it is also submitted to a 1 kHz lowpass filtering.

The periodicity detection, which results in x_2 in Figure 2, is given by

$$x_2 = \mathrm{IDFT}\left(\left|\mathrm{DFT}\left(x_{low}\right)\right|^c + \left|\mathrm{DFT}\left(x_{high}\right)\right|^c \right), \quad (9)$$

where DFT and IDFT represent the Discrete Fourier Transform and its inverse, respectively, and k is the compression factor to be used. The value of k is usually 2, which makes Equation (2) equivalent to the conventional calculation of the autocorrelation. In the present work, k was set to 1 after an optimization process.

The peaks of the autocorrelation given by x_2 are good indicators of potential fundamental frequencies present in the signal. However, since the signals are polyphonic and often very complex, x_2 shows lots of spurious information that can lead to wrong estimations. To reduce the amount of unwanted information, a peak pruning technique is applied. Firstly, a half-wave rectification is applied to clip negative values of x_2. The resulting function is time scaled (expanded in time) by a factor of two and subtracted from the clipped autocorrelation function. This procedure tends to eliminate all peaks whose time lags are twice the time lag of a stronger reference peak. It also removes near zero values of the autocorrelation. The procedure can be repeated for other multiples of each reference peak. In this work, all peaks with twice and three times the time lag of the reference peaks are eliminated.

The last step determines the time lag of the main remaining peak, whose inverse provides the corresponding fundamental frequency. The estimated frequencies are then converted to the MIDI scale, according to the procedure described in [1] and given by

$$m = 12\log_2\left(\frac{f}{440}\right) + 69 , \quad (10)$$

where f is the frequency in Hz and m is the MIDI number.

5. CLASSIFICATION STRATEGY

The features extracted for each frame are grouped into analysis segments corresponding to 1 s of the signal. Therefore, each group will have 92 elements, from which three parameters are extracted: mean, variance and main peak prevalence. This last parameter is calculated according to

$$p_{ft}(j) = \frac{\max\left[ft(i,j)\right]}{\frac{1}{I}\cdot\sum_{i=1}^{I} ft(i,j)}, \quad (11)$$

where $ft(i,j)$ corresponds to the value of feature ft in the frame i of segment j, and I is the number of frames into a segment. This parameter aims to infer the behavior of extreme peaks with relation to the mean values of the feature. High p_{ft} indicate the presence of sharp and dominant peaks, while small p_{ft} often means a smooth behavior of the feature and no presence of high peaks.

As a result of this procedure, each segment will lead to 12 parameters, which are arranged into a test vector to be compared to a set of reference vectors. The determination of the reference vectors is described next.

5.1. Determination of Reference Vectors

The reference vectors were determined according to the following steps:

a) Firstly, 80 signals with a length of 32 s were carefully selected to represent each one of the 13 genres adopted in this work, resulting in a training set with 1,040 signals. The signals were selected according to the subjective attributes expected for each genre, and were taken from the database described in Section 6.

b) Next, the parameter extraction procedure was applied to each one of the training signals. Since such signals have 32 s, 32 vectors of 12 parameters were generated for each signal, or 2,560 vectors representing each genre.

c) A comparison procedure was carried out taking two genres at a time. For example, the training vectors corresponding to the genres “pop” and “rap” were used to determine the 6 reference vectors (3 for each genre) that resulted in the best separation between such genres. Such reference vectors were chosen as follows. Firstly, a huge

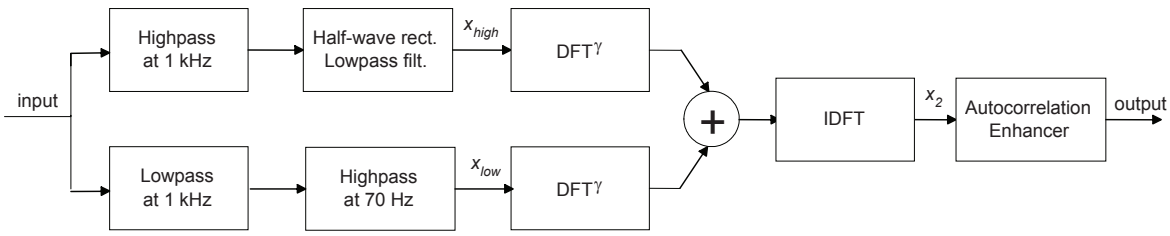


Figure 2 - Multipitch analysis scheme.

set of potential reference vectors was determined for each genre, considering factors as the mean of the training vectors and the range expected for the values of each parameter, discarding vectors that are distant from the cluster. After that, for a given pair of genres, all possible six-vector combinations extracted from both sets of potential vectors were considered, taking into account that each set must contribute with three vectors. For each combination, an Euclidean distance was calculated between each potential vector and all training vectors from both genres. After that, each training vector was labeled with the genre corresponding to the closest potential vector. The combination of potential vectors that resulted in the highest classification accuracy was taken as the actual set of reference vectors for that pair of genres.

d) The procedure described in item *c* was repeated for all possible pairs of genres (78 pairs for 13 genres). As a result, each genre has 12 sets of 3 reference vectors, resulting from the comparison with the other 12 genres. The number of reference vectors was fixed at 3 because this is the best compromise between accuracy and robustness. If less than 3 vectors were used, the set would not be general enough to represent the respective genre; on the other hand, if more than 3 vectors were adopted, the reference set would adapt too much to the set of training signals, losing robustness.

This pair-of-genres based comparison provides much better differentiation between the genres than using a single comparison considering all genres at a time. This is so because particular differences between the genres are much more stressed and explored in this way.

5.2. Test Procedure

Figure 3 illustrates the final classification procedure of a signal. The figure was constructed considering a hypothetical division into 5 genres (A, B, C, D and E) and a signal of 10 s, in order to simplify the illustrations. Nevertheless, all observations and conclusions are valid for the 13 genres and 32 s signals actually considered in this work. As can be seen in Figure 3, the procedure begins with the extraction of the parameter vector from the first segment of the signal (Figure 3A). Such vector is compared with the reference vectors corresponding to each pair of genres, and the smallest Euclidean distance indicates the closest reference vector in each case (gray squares in Figure 3B). The labels of such vectors are taken as the winner genres for each pair of genres (C). In the following, the number of wins of each genre is summarized, and the genre with most victories is taken as the winner genre for that segment (D); if there is a draw, the segment is labeled as "inconclusive". The procedure is repeated for all segments of the signal (E). The genre with more wins along all segments of the signal is taken as the winner (F); if there is a draw, the summaries of all segments are summed and the genre with more wins is taken as winner. If a new draw occurs, all procedures illustrated in Figure 4 are repeated considering only the reference vectors of the drawn genres; all other genres are temporarily ignored. The probability of a new draw is very close to zero, but if it occurs, one of the drawn genres is taken at random as winner. Finally, the winner genre is adopted as the definitive classification of the signal (G).

Normally, the last segment of a signal will have less than one second. In such cases, if the segment has more than 0.5 s, it is considered and the parameters are calculated using the number of frames available, which will be between 46 and 92. If such segment has less than 0.5 s, its frames are incorporated to the previous segment, which will then have between 92 and 138 frames.

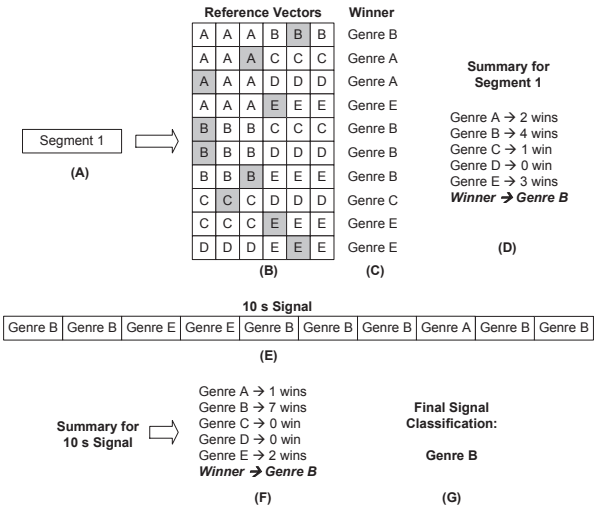


Figure 3 - Classification Procedure.

The classification is carried out directly in the lowest levels of the structure shown in Figure 1. This means that a signal is firstly classified according to the basic genres, and the upper classes are classified accordingly (bottom-up approach). This strategy was adopted because it was observed that as lower is the hierarchical layer in which the signal is directly classified the more precise is the classification of the signal into upper classes. In tests with a top-down approach, where the signals were classified layer by layer, starting with the topmost, the accuracy achieved was between 3 % and 5 % lower than that one achieved using the bottom-up approach.

Next section presents the results achieved by the proposal.

6. TESTS AND RESULTS

The database used in this work is composed by 2,103 music excerpts, which represent more than 20 hours of audio data (13.5 GB). The signals were sampled at 48 kHz and quantized with 16 bits. The audio material was extracted from Compact Discs, from Internet radio streaming and also from coded files (mp3, wma, ogg, aac). The music database was divided into a training set of 1,039 files, which was used to determine the reference vectors described in Section 5, and into a test set, which was used to validate the technique.

Figure 5 shows the confusion matrix associated to the tests. First column shows the target genres, and first row shows the genres actually estimated by the technique. Taking the first line as example, it can be seen that, from the 98 actual classical songs, 88 were correctly classified, 6 were classified as opera, and 4 were classified as jazz.

The main diagonal in Figure 4 shows the correct estimates, and all values outside the main diagonal are errors. Also, as darker is the shading of an area, the lower is the hierarchical layer. As can be seen, most of errors are concentrated inside a same class. Considering each layer separately, the accuracy was: 85.1 % for the 1st layer, 77.4% for the 2nd layer, 61 % for the 3rd layer and 58 % for

the 4th layer. Considering only the bottom genres, the accuracy achieved was 63.7 %.

	CL	OP	RO	RS	HM	CO	CD	PO	TE	RA	RE	JA	LA
CL	88	6	0	0	0	0	0	0	0	0	0	4	0
OP	11	50	0	0	0	0	0	0	0	0	0	5	0
RO	0	0	58	5	14	0	2	4	4	0	1	0	3
RS	1	0	6	50	0	5	1	7	0	1	2	2	3
HM	0	0	13	3	56	0	1	3	1	0	0	0	0
CO	1	0	3	8	0	30	10	0	0	0	0	2	7
CD	1	0	7	8	1	3	20	4	0	0	1	5	12
PO	0	0	7	4	1	0	0	59	11	3	7	3	4
TE	0	0	3	0	3	0	0	14	53	6	7	0	3
RA	0	0	0	0	0	0	0	3	4	58	15	0	2
RE	0	0	0	5	0	1	0	5	1	5	55	1	9
JA	1	5	2	6	0	7	2	5	0	0	3	50	7
LA	0	1	3	7	0	4	3	10	1	3	7	5	57

Figure 4 - Confusion matrix

As expected, the accuracy is higher for upper classes. The accuracy achieved for the first layer is above 85%, which is an outstanding result. The accuracy of 63.7 % for the basic genres is also excellent, especially considering that the signals were classified into 13 genres, which is more than any other previous work.

A direct comparison with previous techniques is very difficult, because the databases used in each case are different. However, some conclusions can be drawn. Most of previous works have achieved an accuracy of about 60 %, but using simple taxonomies. Taking specifically the results obtained in [1], the accuracy achieved was 61 % for a division into 10 genres. This indicates that the technique here proposed is, in terms of accuracy, at least at the same level of the best previous proposals.

Another aspect that must be considered is the performance of the technique when compared to a subjective classification. As discussed in Section 2, classifying musical signals in genres is a naturally fuzzy and tricky task, even when subjectively performed. The performance of humans in classifying musical signals into genres was investigated in [11]. In such research, it was asked for college students to classify musical signals into one of 10 different genres. The subjects where previously trained with representative samples of each genre. The students were able to correct judge 70 % of the signals. Despite a direct comparison is not possible due to differences in the taxonomy and databases, it can be concluded that the technique here proposed has achieved a performance very close to that obtained in the subjective tests, even with 3 more genres to consider.

Under the point-of-view of computational effort, the strategy has also achieved good results. The program, running in a personal computer with an AMD Athlon 2000+ processor, 512 MB of RAM and Windows XP OS, has taken a little more than 20 s to process an audio file of 32 s. This performance indicates that the procedure can be suitably used in real-time applications.

7. CONCLUSIONS AND FUTURE WORK

This paper presented a new strategy to classify music signals into genres. The technique uses 7 features, sets of reference vectors and a pair-of-genres based analysis to infer the classification of the signals.

The hierarchical approach has resulted in excellent performance in terms of accuracy, even when lower layers are considered. The results are comparable to the best techniques previously developed, and are very close to that ones observed in subjective tests with human listeners.

Although the good results achieved by the proposed techniques, further improvement is still possible. The first and more obvious direction for new research is the development of new features able to extract more useful information from the signals. Such new features could be based on psychoacoustic properties of human hearing, improving the correlation with the actual human perceptions. Another direction for future research is expanding the number of genres and the number of hierarchical levels, since it is expected that as deeper is the hierarchical structure, the more accurate is the classification of upper classes. Another interesting line of research is the extraction of features directly from the compressed domain of songs submitted to perceptual coders like MP3, WMA and Ogg-Vorbis.

Acknowledgements

Special thanks are extended to FAPESP for supporting this work under grant 04/08281-0.

References

[1] G. Tzanetakis and P. Cook, Musical Genre Classification of Audio Signals. *IEEE Trans. on Speech and Audio Processing*, 10(5): 293-302, 2002.

[2] G. Agostini, M. Longari and E. Pollastri, Musical Instrument Timbres Classification with Spectral Features. *EURASIP Journal on Applied Signal Processing*, 2003(1): 5-14, 2003.

[3] D. Pye, Content-based methods for the management of digital music. *In Proc. of ICASSP*, Istanbul, pp. 2437-2440, 2000.

[4] J. Saunders, "Real-Time Discrimination of Broadcast Speech/Music", *In Proc. of ICASSP*, Atlanta, pp. 993-996, 1996.

[5] L. Lu, H. -J. Zhang and H. Jiang, Content Analysis for Audio Classification and Segmentation. *IEEE Trans. on Speech and Audio Proc.*, 10(7): 504-516, 2002.

[6] E. Scheirer and M. Slaney, Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. *In Proc. of ICASSP*, Munich, pages 1331-1334, 1997.

[7] M. J. Carey, E. S. Parris and H. Lloyd-Thomas, A Comparison of Features for Speech/Music Discrimination. *In Proc. of ICASSP*, Phoenix, pages 149-152, 1999.

[8] E. Wold, T. Blum, D. Keislar, J. Wheaton, Content-Based Classification, Search, and Retrieval of Audio. *IEEE MultiMedia*, 3(3): 27-36, 1996.

[9] T. Zhang, C.-C. J. Kuo, Audio Content Analysis for Online Audiovisual Data Segmentation and Classification. *IEEE Trans. on Speech and Audio Processing*, 3(4): 441-457, 2001.

[10] F. Pachet, D. Casaly, A Taxonomy of Musical Genres. *In Proc. of Content-Based Multimedia Information Access (RIAO)*, Paris, 2000.

[11] J.-J. Aucouturier and F. Pachet, Representing Musical Genre: A State of the Art. *Journal of New Music Research*, 32(1): 83-93, 2003.

[12] T. V. Thiede, Perceptual Audio Quality Assessment Using a Non-Linear Filter Bank. PhD Thesis, Technical University of Berlin, 1999.

[13] T. Tolonen and M. Karjalainen, A Computationally Efficient Multipitch Analysis Model. *IEEE Trans. on Speech and Audio Processing*, 8(6): 708-716, 2000.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Fourier e Wavelets na Transcrição Musical do Sinal de Áudio

Josildo P. Silva, Frede O. Carvalho, Marcelo A. Moret
CEPPEV – Faculdade Visconde de Cairu
Salvador, Bahia, 40070-200, Brasil
josildo@terra.com.br, fredecarvalho@yahoo.com.br, moret@cairu.br

RESUMO

O presente trabalho objetiva apresentar um modelo computacional para identificar no sinal de áudio elementos fundamentais à transcrição musical, tais como a detecção de início dos eventos musicais e o cálculo do tom da nota. As análises de Fourier e wavelets foram utilizadas como ferramentas para o desenvolvimento de dois diferentes métodos. As particularidades das técnicas desenvolvidas foram discutidas e os resultados apresentados de forma que as vantagens e limitações dos mesmos ficassem evidentes.

1.INTRODUÇÃO

O objetivo final de um processo de transcrição musical é uma completa representação de todas as estruturas musicais e informações relacionadas, presentes em um sinal de áudio. Embora tenha aumentado o número de pesquisas neste campo, utilizar computadores para analisar e entender música tem sido uma tarefa de grande dificuldade e nem sempre bem sucedida ao longo dos últimos anos, sendo um problema complexo e ainda longe de ser resolvido [1,2].

Definindo sucintamente os elementos musicais, pode-se afirmar que melodias são seqüências de notas ao longo do tempo, usualmente contendo pausas alternadas; harmonia estaria determinada pelo relacionamento entre o tom das notas que são tocadas ao mesmo tempo e o ritmo seria determinado pelo início de tempo das notas (instante em que as notas são tocadas, do inglês "*onset time*") e do destaque com que estas notas são tocadas.

Desta forma é razoável afirmar que a análise do sinal de áudio de uma melodia de um único instrumento musical, é suficiente para estimar as notas musicais contidas no sinal, inclusive computando suas posições no tempo. Com base nessa assertiva, a proposta deste trabalho é realizar um estudo no campo da transcrição musical, utilizando como ferramentas as transformadas de Fourier e wavelets, a fim de extrair do sinal de áudio os componentes mínimos para geração de uma partitura musical, a saber: as notas musicais e suas localizações ao longo do tempo.

Para este propósito um programa de computador foi implementado com objetivo de realizar a transcrição musical dos instrumentos: flauta, piano, saxofone, trompete e violino. Estes instrumentos foram escolhidos por terem sido considerados objeto de estudo em outros trabalhos e principalmente porque cada um deles é representante de uma categoria acústica, a saber: a flauta é da categoria dos sopros com embocadura, o piano é da família das cordas percutidas, o saxofone é da categoria sopro com palheta simples, o trompete é da categoria sopro com bocal e o violino das cordas com arco. Algumas características do modelo proposto são:

- Implementação de algoritmos no ambiente MATLAB.
- Sinal de audio monofônico gerado a partir de arquivos MIDI.
- Ausência de qualquer treinamento prévio referente ao tipo de instrumento a ser analisado pelo sistema.

– Capacidade de processamento limitadas aos recursos de memórias do equipamento.

2. ANÁLISES DE FOURIER E WAVELETS

A transformada de Fourier é uma ferramenta fundamental para análise de sinais no domínio da freqüência. Neste ela opera de maneira global detectando o comportamento das sinusoidais presentes no sinal. Por conta disso, se o interesse é estudar mudanças abruptas em curtas regiões do sinal, a transformada de Fourier não produz resultados satisfatórios. Estes tipos de estudos carecem de um tipo de ferramenta adaptada as variações que ocorrem nas dimensões do tempo e da freqüência: A STFT (*Short-time Fourier Transform*). Esta ferramenta é o resultado dos estudos de Dennis Gabor (1946) que adaptou a transformada de Fourier utilizando uma técnica denominada "janelamento", de forma a mapear um sinal do domínio do tempo para as dimensões do tempo e da freqüência. A STFT pode ser definida pela equação 1

$$X_h(k,n) = \sum_{\mu=-N/2}^{N/2-1} h(\mu)x(n+\mu)e^{-2\pi k\mu/N} \qquad (1)$$

onde $n = 0 \ldots N-1$ pontos de x, $h(\mu)$ é janela de análise (geralmente uma janela "Hanning") e k é k-ésima componente de freqüência da FFT (*Fast Fourier Transform*).

A transformada wavelets pode ser definida como uma alternativa à clássica transformada janelada de Fourier – a STFT. Em linhas gerais a análise janelada de Fourier localiza no tempo as componentes de freqüências do sinal, enquanto análise de wavelets compara as variações de magnitude do sinal em várias resoluções ou escalas. Os blocos construtores da análise de Fourier são senos e cossenos multiplicados por uma janela deslizante de tamanho fixo. Na análise wavelets a janela é oscilante de tamanho variável, sendo chamada de "wavelet-mãe". Esta função "wavelet-mãe"', têm suporte compacto (rápido decaimento no infinito), é suave e satisfaz a condição fundamental $\int_{-\infty}^{\infty} \psi(t)dt = 0$, onde $\psi(t)$ é a "wavelet-mãe" [4].

A transformada contínua wavelet é definida por

$$W_f(a,b) = \int_{-\infty}^{\infty} f(x)\overline{\psi}_{(a,b)}(x)dx \qquad (2)$$

onde $\overline{\psi}_{(a,b)}(x) = \overline{\psi_{(a,b)}(x)}$. Uma alternativa à equação 2 é a que considera o cálculo dos coeficientes wavelets no espaço de Fourier [5], expressa por

$$W_n(s) = \sum_{k=0}^{N-1} \hat{x}_k \hat{\psi} * (s\omega_k) e^{i\omega_k n \delta t} \qquad (3)$$

onde $\hat{x}$ e $\hat{\psi}$ são as transformadas de Fourier de x e ψ.

3. ELEMENTOS DA TRANSCRIÇÃO MUSICAL

Os elementos essenciais para entendimento básico de uma partitura musical são as notas (sentido psico-acústico: grave ou agudo), quando estas se iniciam (*onset*) e por quanto tempo são "tocadas" [6]. Desta forma pode-se destacar dois elementos fundamentais no processo de transcrição musical:

– A detecção do início do evento musical no tempo (*onset time*), conseqüentemente sua duração;

– E uma vez detectado o evento no tempo, extrair dele a freqüência fundamental que irá caracterizar a nota musical.

O início do tempo da nota tem sido estudado em computação musical como condição de ajuda na tarefa de transcrição musical. Um sistema de transcrição inclui como seu primeiro processo um algoritmo para descobrir ao longo do sinal, palpites confiáveis que representem os reais eventos musicais que podem ser levados em conta para inferir informações musicais de alto nível (notação musical), sendo o algoritmo de detecção, o ponto inicial deste sistema. Contudo a implementação de um algoritmo que possa encontrar e nomear os reais eventos musicais é um problema complexo que exige uma grande atenção [7].

A figura 1 ilustra o método de detecção proposto em [1]. Este método localiza no tempo, a partir da informação de variações da fase do sinal, os instantes de transição ou mudanças no sinal. Estes instantes são evidenciados por meio de picos ou saliências que se destacam visivelmente na análise do processo do sinal de áudio, indicando possíveis *onset*.

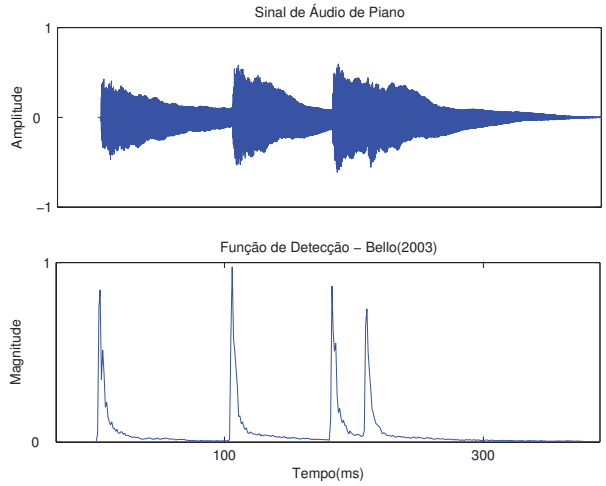


Figura 1 - Detecção de onset utilizando fase do sinal

O cálculo da freqüência fundamental, também conhecido como detecção do tom da nota, tem sido um tópico de exaustiva pesquisa por muitos anos que ainda é investigada hoje. O problema básico consiste em extrair do sinal sonoro a componente de menor freqüência, denominado de *F0* ou freqüência fundamental, isolando-a das outras componentes presentes no sinal [8].

Considerando-se $s(n)$ um segmento do sinal a ser analisado de tamanho N, onde $n = 0 \ldots N-1$, pela aplicação da FFT com uma janela "Hanning", pode-se obter $S(k)$, onde $k = 0 \ldots N-1$, que é a representação no domínio da freqüência de $s(n)$, com a resolução da freqüência dada por $\Delta f = f_s / N$, sendo f_s a freqüência de amostragem.

Uma observação importante a ser feita com relação a este método é que a periodicidade no domínio do tempo passa a ser representada por picos no domínio da freqüência. Naturalmente, o passo seguinte no cálculo da freqüência fundamental é a partir de $S(k)$, localizar todos seus picos e relacionar estes picos seguindo algum critério. Em [1] é utilizado o critério de que os tons musicais produzem padrões no domínio da freqüência, onde se espera que os lóbulos de cada pico obedeçam a relação $m \times f_0$, onde $m = 1 \ldots M$, M é o número de lóbulos e f_0 a freqüência fundamental.

4. MODELAGEM DO SISTEMA

Para efeito de simplificação, a metodologia apresentada nas próximas seções será denominada de Método Silva Para Transcrição Musical - MSTM. Duas versões do método foram implementadas visando confrontar e analisar os resultados. A primeira utiliza análise de Fourier (figura 2) como fundamento de implementação, enquanto que a segunda foi construída usando a análise wavelets (figura 3).

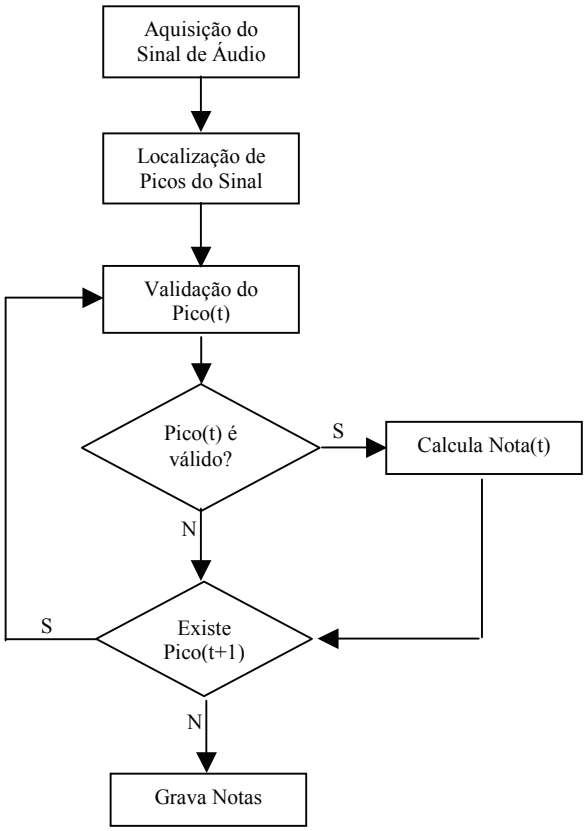


Figura 2 – Modelo utilizando Fourier

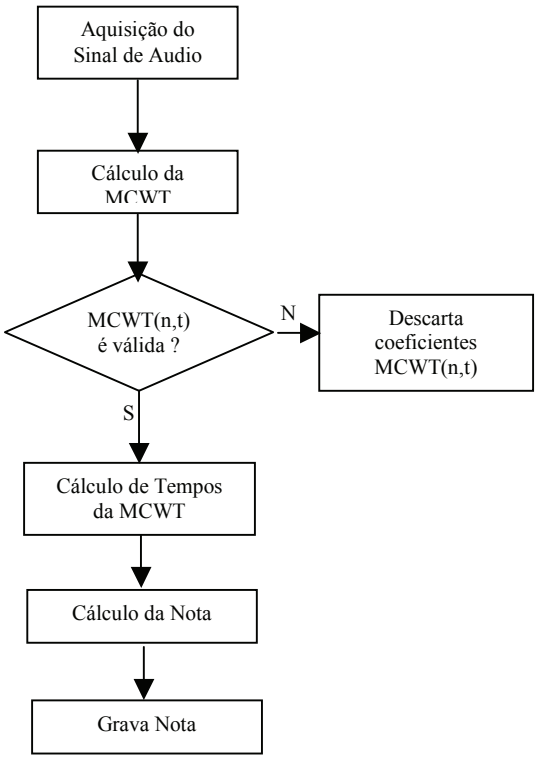


Figura 3 – Modelo utilizando wavelet

4.1 MSTM utilizando Fourier

Localização de picos do sinal

O método para o cálculo de início do tempo da nota de [1], localiza no tempo, a partir da informação de variações da fase, os instantes de transição ou mudanças no sinal. Estes instantes são evidenciados por meio de picos ou saliências que se destacam visivelmente na análise do processo do sinal de áudio. Esta técnica foi aplicada no MSTM Fourier para localizar os picos do sinal.

Cálculo da nota

Uma vez que um pico ou instante de início da nota é considerado válido, é iniciado o processo do cálculo da nota musical associada ao pico válido detectado. Este processo segue as técnicas propostas por [11], implementadas nos algoritmos desenvolvidos por [3], com adaptações realizadas no método de detecção da freqüência fundamental proposto por [1]. O cálculo consiste em analisar quadro a quadro a STFT entre picos válidos de forma a detectar que nota musical estaria associada a um referido pico.

Gravação da nota

Cada nota calculada nos quadros da STFT é acumulada quantitativamente. No final do processamento de cada pico é verificada a nota que mais se repetiu sendo esta escolhida como nota associada ao pico. Desta forma é obtido o início e o término da nota musical. Estes dados são gravados em arquivos ".mat", para posterior comparação com as informações do arquivo MIDI original.

4.2 MSTM utilizando wavelets

A idéia inicial para o método surgiu a partir do algoritmo para cálculo da transformada contínua de wavelet, proposto

em [3]. Este algoritmo utiliza como wavelet-base, a função complexa de Morlet. Um dos aspectos relevante desse algoritmo é que ele permite parametrizar a faixa de freqüência que se deseja trabalhar, o que é importante quando se deseja operar em uma banda de freqüência especifica, como por exemplo, a banda de freqüência dos instrumentos musicais. Foram realizadas alterações no algoritmo com o objetivo de torná-lo melhor adaptado à transcrição musical, quais sejam: a) A escala da wavelets foi calculada seguindo os valores da escala musical temperada; b) Os coeficientes wavelets foram limiarizados pela função de limiarização da audição.

Cálculo da MCWT

Nesta etapa, o conceito de multi-resolução de wavelets é aplicado juntamente com a teoria musical para processar o sinal em escalas wavelets à semelhança da escala musical temperada. O resultado desse processo é uma transformada wavelets contínua, adaptada ao processamento de sinais puramente musicais, denominada neste trabalho de Transformada Contínua de Wavelets para Música - MCWT. A figura 4 apresenta o resultado da MCWT com seus coeficientes validados do sinal de áudio do violino.

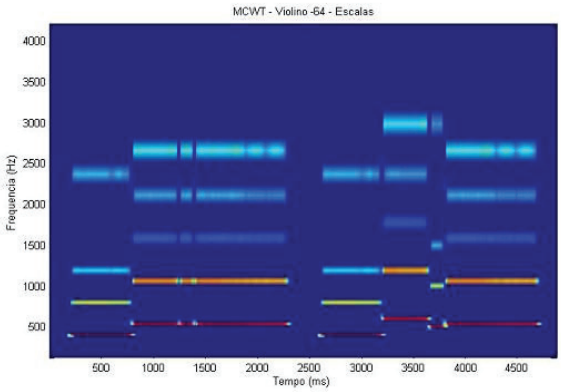


Figura 4 - MCWT de Violino

Cálculo de tempos da MCWT

Nesta etapa do processo, objetiva-se pela análise do comportamento dos coeficientes da MCWT, em cada escala válida, estabelecer os pontos de início e final da variação de magnitude. Estes pontos servirão de parâmetros para calcular o início do tempo das notas e suas respectivas durações. Para este fim foi construída a FMAG, que é a função que descreve o comportamento da magnitude dos coeficientes da MCWT. Os passos para cálculo da FMAG são:

- Suavização dos coeficientes da MCWT;
- Cálculo de pontos consecutivos;
- Limiarização de magnitude;
- Seleção de tempos.

Suavização dos coeficientes da MCWT

O processo de suavização consiste em unir pontos representativos entre as janelas de suavização, de forma que estes pontos mantenham a característica de continuidade. Na prática é selecionado um ponto de máxima magnitude na janela w e realizado interpolação linear com ponto de máxima da janela $w+1$.

Cálculo de pontos consecutivos

Na FMAG existem instantes t_n em que os valores da função saem de zero e retornam a zero. Para calcular os pontos consecutivos é marcado o instante t_1 em que a função é diferente de zero. Quando a função assume novamente valor igual a zero é marcado instante t_2. Desta forma a diferença entre t_2 e t_1 precisa ser maior que a duração mínima para que os pontos sejam considerados válidos. Na implementação desse trabalho a duração mínima foi estabelecida baseada na menor fração de tempo que o ouvido humano necessita para perceber uma transição de nota musical. Em [3] este valor é de aproximadamente 20 ms. Pontos consecutivos que não alcançam o mínimo de duração são descartados.

Limiarização de magnitude

Os diversos testes experimentais realizados com a MCWT revelaram que mesmo após eliminar aqueles coeficientes imperceptíveis à audição, coeficientes pequenos acabavam por interferir negativamente nos resultados, causando falsa detecção de eventos. A forma de eliminar essas interferências foi aplicar um processo de limiarização que consistiu em eliminar os coeficientes abaixo de um determinado valor. O valor de um terço da máxima magnitude foi encontrado após várias experimentações com os diversos instrumentos. Desta forma todos aqueles coeficientes inferiores a um terço do valor máximo da escala analisada foram descartados.

Seleção de tempos

Nesta etapa os pontos suavizados e limiarizados são percorridos e assinaladas as regiões com todas as posições onde a função sai de zero e retorna a zero. Estas posições se constituem nos instantes de início e final da variação de magnitude, sendo contadas como possíveis inícios de notas. Desta forma é construída uma matriz com todos os *onsets* das notas musicais através das várias escalas da MCWT, expressa por

$$PO_{ij} = \begin{bmatrix} t_{11} & t_{12} & w_{13} \\ \vdots & \vdots & \vdots \\ t_{i1} & t_{i2} & w_{i3} \end{bmatrix} \tag{5}$$

onde PO_{ij} é a matriz dos possíveis *onsets*, $i = 1 \ldots N$, N são os possíveis *onset* nas diversas escalas da MCWT e w_{i3} é a nota musical referente ao *onset* considerado.

Cálculo da Nota

Como visto, a FMAG produz nas diversas bandas válidas, os instantes de início e final de variação de magnitude, que pela MCWT correspondem a possíveis inícios de notas. É necessário processar estes valores a fim de detectar que nota está contida na informação das escalas ou bandas válidas. O método proposto para esta etapa é descrito nos seguintes passos:

- Seleção de candidatos à freqüência fundamental - são considerados candidatos à freqüência fundamental todos pontos da FMAG com valores de máximas superiores a 20% da máxima global. Este valor foi

encontrado após experimentações com os vários instrumentos.

- Para cada candidato à freqüência fundamental é verificado se ele possui o tempo suficiente para caracterizar uma nota musical.
- Para cada candidato selecionado é verificado seu grau de covariância temporal. Para este trabalho o grau de covariância é a medida do tempo de simultaneidade entre dois candidatos à freqüência fundamental. Isto é o mesmo que verificar se o tempo é dividido por dois harmônicos (candidatos à freqüência fundamental), ou ainda, quantos pontos de intersecção existem entre dois harmônicos.
- Quando dois candidatos possuem grau de covariância temporal maior que o mínimo de tempo para existência de nota, isto é, quando dois candidatos possuem uma região intersecção no tempo maior que 40 ms, é então verificado se eles são harmônicos entre si. Sendo eles harmônicos então o de menor freqüência é selecionado como freqüência fundamental, e, portanto como nota musical dominante naquela posição da FMAG.

5. RESULTADOS E DISCUSSÃO

A validação dos resultados apresentados neste trabalho seguiu a mesma metodologia utilizada em muitas referências da literatura consultada [7,9,10,12]. Segundo esta metodologia geralmente é calculada taxa relativa de acertos, utilizando uma expressão do tipo

$$PA = \frac{N_c - N_e}{N_t} \times 100 \qquad (6)$$

onde N_c é o número de evento corretamente detectados, N_e é número de erros, N_t é o total de eventos, de forma que PA representa o percentual de acerto do evento considerado (notas ou *onset*).

Para este trabalho foi aplicado especificamente a mesma metodologia de validação encontrada em [12], por ser mais completa em termos de avaliação dos eventos analisados. *Onsets* são considerados corretos se detectados dentro de uma janela de +/- 50 ms do *onset* real ou verdadeiro, isto é, *onsets* que se afastam mais do que 50 ms, tanto para esquerda (negativo), quanto para direita (positivo), são considerados incorretos. A expressão de cálculo da taxa de acerto, é dada por

$$PA = \frac{N_t - N_{ud} - N_{fd}}{N_t} \times 100 \qquad (7)$$

onde, N_t é o número total de eventos no sinal, N_{ud} representa o número de eventos não detectados e N_{fd} representa os eventos falsamente detectados.

Dez trechos de peças musicais foram analisados. A tabela 1 relaciona as músicas com as respectivas quantidades de notas e duração total do trecho em segundos.

Tabela 1 - Peças Musicais

Peça Musical	*Nº.Notas*	*Tempo(s)*
01. Marcha Nupcial	8	4,586
02. Pour Elise	35	6,989
03. Dança Húngara nº 5	19	9,764
04. Brasileirinho	49	9,555
05. Vassourinha	96	12,655
06. Garota de Ipanema	30	11,656
07. Hino Nacional	56	15,766
08. Um a Zero	37	9,102
09. Tico-tico no Fubá	102	15,999
10. Espinha de Bacalhau	99	18,344

Os resultados revelaram algumas situações comentadas a seguir:

- Notas repetidas foram motivos de falhas na detecção de *onsets*, principalmente no método utilizando wavelets;
- A natureza acústica do instrumento influência na detecção de *onset*. Instrumentos com acentuadas explosões de energia no instante de início da nota, como exemplo do piano, apresentam menor dificuldade de detecção que instrumentos ditos "suaves" no instante do início da nota, a exemplo do violino.
- O critério de utilizar o harmônico de maior magnitude como candidato à freqüência fundamental, em algumas situações levou ao erro de oitava. Nestas situações o harmônico de maior magnitude não coincidiu com o harmônico fundamental, sendo detectado uma oitava acima ou abaixo do verdadeiro tom da nota;
- O problema da resolução no domínio do tempo e da freqüência simultâneos, limitação da análise de Fourier, provocou detecção incorreta tanto em *onset* como em nota.
- Os métodos responderam bem nas situações de notas consideradas de curta duração, isto é, notas que duram entre 60 e 150 ms.

As figuras 5 e 6 mostram o gráfico do percentual de acerto de *onset* utilizando o MSTM Fourier e wavelet. A análise comparativa entre as duas figuras revela que de uma maneira geral, tanto o MSTM (Fourier) como o MSTM (wavelet) obtiveram uma resposta satisfatória na detecção do *onset*. Contudo, observa-se que o método segundo a análise de Fourier apresenta maior sensibilidade à natureza do sinal. Este fato encontra-se evidente no comportamento da curva descrita pelos resultados do sinal de áudio da flauta.

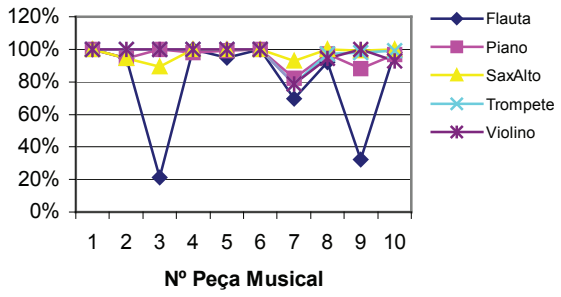


Figura 5 - Detecção de Onset MSTM(Fourier)

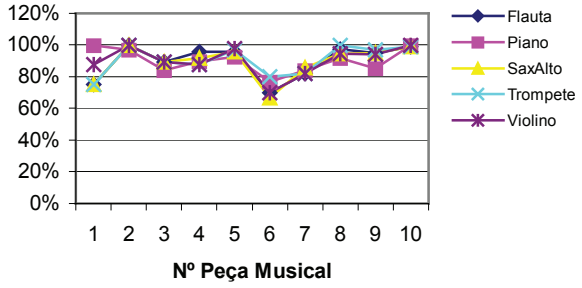


Figura 6 - Detecção de Onset MSTM(wavelet)

A figura 7 mostra o resultado dos tempos de execução de cada trecho de peça musical. Estes tempos representam a duração de processamento do MSTM nas amostras de áudio do piano. Como pode ser observado o método baseado em Fourier utiliza em média 30% do tempo necessário para o processamento com wavelets.

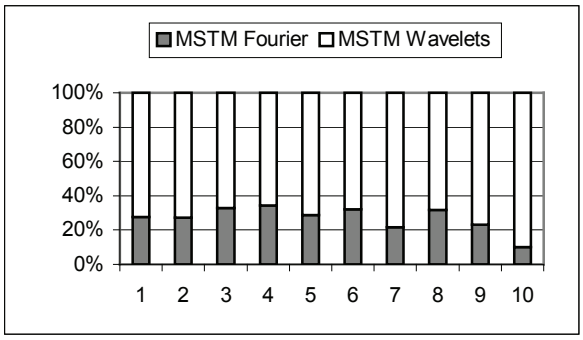


Figura 7 – Tempos de Execução do MSTM

6. CONCLUSÃO

Neste trabalho foi apresentado um modelo computacional com aplicação no campo da transcrição musical com resultados promissores.

O compromisso entre a resolução no domínio do tempo e da freqüência causado pela utilização de uma janela de tamanho fixo é a principal limitação dos métodos baseados na análise de Fourier, enquanto que uma das vantagens da análise de wavelets é atenuação da limitação análise de Fourier, devido a utilização de uma janela com tamanho variável. Esta evidência pôde ser comprovada nas análises dos trechos das peças musicais, onde os resultados com wavelets foram mais uniformes.

Desta forma, é plausível a conclusão de que uma metodologia desenvolvida utilizando o conceito de multi-resolução de wavelets está mais bem adaptada ao problema da transcrição musical, sendo indicada sua aplicação em trabalhos futuros com sinais reais polifônicos.

O tempo de processamento do método com wavelets demonstrou ser mais lento que o método implementado com Fourier. O desenvolvimento de um algoritmo wavelet especifico para sinais musicais, que considere a natureza inerente a este tipo de sinal é também um tópico desafiante para futuros trabalhos.

Os fontes do MSTM e todo material produzido nas análises dos trechos de peças deste trabalho podem ser solicitados pelo e-mail josildo@terra.com.br.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] BELLO, J. P. *Towards the Automated Anaylsis of Simple Polyphonic Music: A Knowledge Based Approach*. Tese de Doutorado. Queen Mary, University of London. London, 2003.

[2] HAINSWORTH, S. W. *Techniques for the Automated Analysis of Musical Audio*. Tese de Doutorado. University of Cambridge. Cambridge, 2003.

[3] JOHNSON , M. K. The Spectral Modeling Toolbox: A Sound Analysis/Synthesis System. Dissertação de Mestrado. Dartmouth College. Hanover, New Hampshire, 2002.

[4] FARIA, R. R. *Aplicação de Wavelets na Análise de Gestos Musicais em Timbres de Instrumentos Acústicos Tradicionais*. Dissertação de Mestrado. Escola Politécnica da Universidade de São Paulo. São Paulo, 1997.

[5] TORRENCE, C. E COMPO, G. P. *A Practical Guide to Wavelet Analysis*. Bulletin of the American Meteorological Society, 1998.

[6] KRUVCZUK, M. *Music Transcription for the Lazy Musician*. Relatório Técnico. New Hampshire, 2000.

[7] NAVA, G. P. E TANAKA, H. A convolutional-kernel based approach for note onset detection in piano-solo audio signals. ISMA2004. Nara, Japão, 2004.

[8] GERHARD, D. Pitch Extraction and Fundamental Frequency: History and Current Techniques. University of Regina, Canada, 2003.

[9] BELLO, J. P. E SANDLER, M. Phase-Based Note Onset Detection For Music Signals. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.

[10] DUXBURY,C. E BELLO, J. P. E SANDLER, M. A Combined Phase and Amplitude Based Approach to Onset Audio Segmentation. WSPCe. London, UK, 2003.

[11] SERRA, X. Musical Sound Modeling with Sinusoids plus Noise. Musical Signal Processing. Swets Zeitlinger Publishers, 1997.

[12] DUXBURY,C. SANDLER, M. E DAVIES, M. A Hybrid Approach to Musical Note Onset Detection. Proc. of the 5th Int. Conference on Digital Audio Effects (DAFX-02). Hamburg, Germany, 2002.

Índice de Autores

Author Index

Audio Engineering Society - Seção Brasil

4º Congresso de Engenharia de Áudio
10ª Convenção Nacional da AES Brasil

Patrocinadores:

Digidesign
Staner
FZ Audio
Ciclotron
Libor
Selenium

Expositores:

Ass. Brasileira dos Profissionais de Áudio
Acoustic Caixas Profissionais Ltda
Clínica Audiológica Audicare LTDA
H. Sheldon Serviços de Marketing Ltda
Spectral Balance Pro Audio Lighting
Ciclotron Ind. Eletrônica Ltda
CIS Group Corporation
Decomac Brasil Ltda
Digidesign
Feeling Estruturas Metálicas Ind. e Com. Ltda
Empresa Folha da Manhã S/A
FZ Indústria e Comércio Ltda
HMP Marketing Editorial Ltda
Hotsound Ind. e Com. Equipos. Eletrônicos Ltda
IATEC - Inst. de Artes e Técnicas em Comunicação
Instituto de Áudio & Vídeo
VD Ribeiro Epp
Leson Lab. de Engenharia Sônica Ltda
Libor Comércio e Importação Ltda
LJM Indústria e Comércio Ltda
JPF Ind. e Com. de Comp. Eletrônicos Ltda
MM-Rio Acessórios Musicais Ltda
Oversound Ind. e Com. Eletro Acústico Ltda
Pride Music Com. Imp. Distr. Ltda
Ferreira & Bento do Brasil Ltda
Quanta Brasil Imp. e Exp. Ltda
Editora Música e Tecnologia
Roland Brasil Imp. Exp. Com. Rep. e Serviços Ltda
Royal Instrumentos Musicais Ltda
Sabra Som Comercial Ltda
Eletrônica Selenium S/A
Sennheiser
SLM Sound Ligth M. Com. Ltda
Ookpik Amplicadores e Instrumentos Musicais
Staner Eletrônica Ltda
Studio R Eletrônica Ltda
Taw Equipamentos de Sonorização Ltda
Clever Luz e Som Comercial Ltda
Yamaha Musical do Brasil Ltda