AUDIO AES

10º CONGRESSO DE ENGENHARIA DE ÁUDIO

8 a 10 de Maio de 2012

Pravilhão Amarelo do Expo Centrer Norte - São Paulo - SP - Brasil

Foto: Fotolia

Anais do 10º Congresso de Engenharia de Áudio da AES-Brasil

Proceedings of the 10th AES Brazil Conference

8-10 Maio/*May*, 2012

Pavilhão Amarelo do Expo Center Norte – São Paulo – SP – Brasil

Copyright ©2012
Audio Engineering Society – Brazil Section

C76a Congresso de Engenharia de Áudio, 10., São Paulo, 2012.
Anais do 10º Congresso de Engenharia de Áudio AES Brasil e 16ª Convenção Nacional AES Brasil / editado por J. A. Mannis. – Rio de Janeiro: Sociedade de Engenharia de Áudio, 2012.

Disponível em: *www.aesbrasil.org/congressos*

1. Engenharia de áudio – Congressos. 2. Processamento de sinais – Congressos. 3. Computação musical – Congressos. 4. Acústica – Congressos. I. Mannis, José Augusto. II. Convenção Nacional AES Brasil, 16., São Paulo, 2012. III. Sociedade de Engenharia de Áudio. IV. Título.

CDD – 621.3893
- 621.38043
- 789.9
- 620.21

ISSN 2177-529X

Índices para Catálogo Sistemático

1. Engenharia de áudio – Congressos – 621.3893
2. Processamento de sinais – Congressos – 621.38043
3. Computação musical – Congressos – 789.9
4. Acústica – Congressos – 620.21

Os artigos publicados nestes anais foram editorados a partir dos originais finais entregues pelos autores, sem edições, correções ou considerações feitas pelo comitê técnico. A AES Brasil não se responsabiliza pelo conteúdo. Outros artigos podem ser adquiridos através da Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA, (*www.aes.org*). Informações sobre a seção Brasileira podem ser obtidas em (*www.aesbrasil.org*). Todos os direitos são reservados. Não é permitida a reprodução total ou parcial deste artigo sem autorização expressa da AES Brasil.

Impresso no Brasil.
Printed in Brazil.

Editoração e arte / *Publishing and art*
Bárbara Rossi Rosa

totum
marketing e comunicação

ORGANIZAÇÃO

APOIO

REALIZAÇÃO

Audio Engineering Society – Brazil Section

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

ADMINISTRAÇÃO

Presidente / *Chairman:*	Aldo Ricardo Soares
Vice- Presidente / *Vice- Chairman:*	Armando Baldassarra
Secretário / *Secretary:*	José Anselmo
Tesoureiro / *Treasurer:*	Joel Vieira de Brito
Coordenador Acadêmico / *Academic Coordinator:*	Júlio Lucchi
Conselheiros / *Counselors:*	Carlos Ronconi
	Dagoberto Almeida
	David Fernandes
	Flávia Calabi
	Framklim Garrido
	João Américo
	José Augusto Mannis
	Marcelo Claret
	Regis Rossi A. Faria

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

COMITÊ DE ORGANIZAÇÃO / *ORGANIZING COMMITTEE*

Coordenador da Convenção ***/ Convention Chair***
Aldo R. Soares (AES-Brasil)

Coordenadores do Programa Técnico ***/ Program Committee Chairs***
Regis Rossi A. Faria (USP)
Sidnei Noceti Filho (CTC-UFSC)

Coordenadores de Artigos ***/ Papers Chairs***
Francisco J. Fraga da Silva (CECS-UFABC)
Márcio Brandão (UnB)

Coordenador Editorial ***/ Editorial Chair***
José Augusto Mannis (IA-UNICAMP)

Coordenador de infra-estrutura ***/ Infrastructure Chair***
Júlio C. Lucchi (USJT / IMT / FEI / ITA)

Coordenadores de Comunicação ***/ Communications Chairs***
Paulo R. A. Marins (UnB)
Arcanjo Lenzi (LVA-EMC-UFSC)

COMITÊ DE PROGRAMA TÉCNICO (REVISORES)
TECHNICAL PROGRAM COMMITTEE (REVIEWERS)

Adolfo Maia Jr. (NICS-IME-IA-UNICAMP)
Alexandre Leizor Szczupak (COPPE-UFRJ)
André Dalcastagnê (Inst. Federal de Santa Catarina)
Anibal Ferreira (Universidade do Porto)
Bruno Fazenda (University of Salford, Manchester)
Carlos Ynoguti (INATEL)
Celso Kurashima (UFABC)
Diego Haddad (CEFET-RJ)
Fernando Pacheco (Inst. Federal de Santa Catarina)
Francisco Fraga (UFABC)
Gianpaolo Evangelista (Linköpings Universitet)
Guilherme Campos (Universidade de Aveiro)
Ivandro Sanches (FEI)
José Augusto Mannis (IA-UNICAMP)
Júlio Lucchi (USJT / IMT/ FEI / ITA)
Leonardo Nunes (UFRJ)
Marcelo Pimenta (UFRGS)
Marcelo Queiroz (IME-USP)
Márcio Brandão (UnB)
Miguel Ramírez (EP-USP)
Paulo Marins (UnB)
Phillip Burt (EP-USP)
Regis Rossi Faria (USP)
Sidnei Noceti Filho (LINSE-EEL-CTC-UFSC)
Walter Gontijo (UFSC)

EDITORIAL

A qualidade que se busca num sistema de áudio tem sempre por objetivo a qualidade da audição que ele proporciona, desde o conteúdo veiculado até sua fruição. O que ouvimos deve ser claro, coerente, inteligível e esteticamente estimulante. Todos os esforços assim mobilizados se justificam. Investe-se em tecnologia, métodos, equipamentos, aplicativos, acesso e veiculação de informação para atingir uma meta, um objetivo sonoro, um estado de audição, despertar uma sensibilidade perceptiva ou produzir um impacto em grupos ou segmentos sociais. Novas soluções decorrem dos desafios naturalmente trazidos por novas ideias e propostas. O avanço científico, tecnológico e produtivo no setor de áudio é movido, assim, por demandas decorrentes dos objetivos fim. Em todos os casos, o que conta mesmo é o que, como, quando e o quanto é sonoramente comunicado. O foco é, portanto, o conteúdo.

Quem adquire um disco, gosta de música. Movida pelo seu gosto, essa pessoa irá também a shows, concertos, ouvirá rádio, audio player e poderá parar durante um passeio para fruir as vibrações de um grupo musical em performance ao ar livre. Sua busca é sempre 'a música', independentemente dos meios, do suporte, da situação, do veículo, do sistema, do dispositivo. A adequação dos meios é, em última instância, o que resta a ser explorado pelos provedores de informação para atrair o público, dando-lhes acesso ao que buscam, neste caso, a música. Isso vale também para outros conteúdos: notícias, oportunidades, relacionamentos, interações utilitárias. Para cada um deles, inúmeras situações se associam às suas particularidades, ou seja, não basta ter 'o que ouvir', importa também 'como ouvir', podendo aí estar incluído 'onde e quando ouvir'. Para chegar a isso é preciso ter 'com o que ouvir', ou seja, recursos materiais ou imateriais, sistemas, através dos quais cada ouvinte poderá chegar a seu 'como ouvir'. A busca se move, portanto, pelo acesso aos produtos e às situações escolhidas, função do contexto cultural, dos hábitos, costumes e necessidades de cada indivíduo ou grupo social. Portanto, o objetivo fim de toda pesquisa, de todo desenvolvimento, de toda inovação, de todas as facilidades oferecidas, o último delta onde desemboca o último braço de fluxo dessa torrente, é a qualidade do conteúdo transmitido. Aquilo que se quer ouvir, com todas as particularidades que lhe são próprias. O que mobiliza, atrai e justifica esforços e investimentos, ou seja, 'o foco último' é, sempre, 'o conteúdo'.

Isso se observa na diversidade dos artigos apresentados nesta 10ª edição do Congresso de Engenharia de Áudio e 16ª Convenção Nacional da AES Brasil. Cada uma das sessões coloca também uma luz sobre conteúdos específicos. Na primeira sessão, temos eletrônica em áudio aplicada a transdutores, computação musical, análise e processamento do sinal, reconhecimento da fala aplicado a sistemas de diálogo automático, envolvendo assim música e facilidades de interatividade. Na segunda sessão, encontram-se aplicações para rádio, tv e produção fonográfica com espacialização, mixagem e monitoração em pós-produção em sistemas multicanal, bem como aplicações no campo da música, a saber: análise do sinal para reconhecimento de alturas musicais e acústica de instrumentos musicais, abrangendo música e comunicação oral. A terceira sessão, envolve restauração do sinal, melhoria do sinal de voz, marca d'água em áudio, relação entre percepção espacial sonora e imagem visual, análise do sinal para reconhecimento de andamentos musicais, ou seja, aplicações no campo da comunicação oral, audiovisual, interatividade, música e propriedade intelectual, esta última também remetendo às anteriores. Na comunicação oral e na interatividade, a atenção principal é o sinal sonoro da voz, a inteligibilidade da palavra e do sinal sonoro transmitido e ouvido. Já na música, contam a clareza das alturas, dos ritmos, dos timbres, do espaço, produzidos e escutados bem como a qualidade da imagem sonora produzida nos ouvintes. Nas artes acústicas, no audiovisual, nos produtos veiculados em sistemas de telecomunicação, chamados por Pierre Schaeffer de 'artes-relé', nas instalações multimídia com áudio, a inteligibilidade e a qualidade de inúmeros parâmetros referentes à fala e demais sinais sonoros, são determinantes para o conteúdo chegar com eficiência ao seu destino final.

Nesta edição comemorativa de 10 anos de congresso, os anais mostram na sua criação visual a valorização dos conteúdos na figura de uma criança escutando música, que poderia ser de séculos atrás, com fones de ouvido, talvez de anos atrás, porém uma escuta nova, atenta, indagadora, de descoberta, com olhos para cima, ou para um futuro que ela mesma representa. Ruben Alves nos sugere em sua Arte de ouvir que 'é preciso escutar as crianças para que sua inteligência desabroche [...] pois é dos sonhos que nasce a inteligência'. Na imagem da criança que em sua simplicidade inventa e descobre porque está livre para ouvir, podemos tentar nos posicionar de uma nova maneira para perceber que no áudio o mais importante pode também estar junto ao ouvinte.

José Augusto Mannis - Coordenador Editorial

PROGRAMAÇÃO DE PALESTRAS / *SCHEDULE OF LECTURES*

08 Maio/*May*

O ESTADO DA ARTE EM CIRCUITOS INTEGRADOS ANALÓGICOS PARA ÁUDIO
Terça-feira/*Tuesday,* 08 Maio/*May* 2012 11:30 - 13:00
Sala 3 / *Room 3*
Rosalfonso Bortoni (UFSC / Audio Precision)

MECANISMO DE AUDIÇÃO
Terça-feira/*Tuesday,* 08 Maio/May 2012 13:30 - 15:00
Sala 3 / *Room 3*
Marcelo Santos Portela (LVA-UFSC)

CARACTERIZANDO AMPLIFICADORES CLASSE D PARA P&D
Terça-feira/*Tuesday*, 08 Maio/*May* 2012 13:30 - 15:00
Auditório D / *Auditorium D*
Joe Begin (Audio Precision) Tradução: Rosalfonso Bortoni

SUBWOOFERS: DIMENSIONAMENTO E EMPILHAMENTO
Terça-feira/*Tuesday*, 08 Maio/*May* 2012 15:30 - 17:00
Sala 3 / *Room 3*
Homero Sette (Studio-R)

ACÚSTICA DE SALAS: SUBSÍDIOS PARA PROJETO
Terça-feira/*Tuesday,* 08 Maio/*May* 2012 15:30 - 17:00
Auditório D / *Auditorium D*
Stelamaris Rolla Bertoli (Unicamp)

EVOLUÇÃO DOS AUTOMIXERS
Terça-feira/*Tuesday,* 08 Maio/*May* 2012 17:30 - 19:00
Auditório D / *Auditorium D*
Dan Dugan (Yamaha)

PROGRAMAÇÃO DE PALESTRAS / *SCHEDULE OF LECTURES*

09 Maio/*May*

ÁUDIO 5.1 EM TELEVISÃO
Quarta-feira/*Wednesday,* 09 Maio/*May* 2012 13:30 - 15:00
Auditório D / *Auditorium D*
Rodrigo Meirelles (Rede Globo)

SISTEMAS DE PROGRAMAÇÃO MUSICAL NO SÉCULO 21
Quarta-feira/*Wednesday,* 09 Maio/*May* 2012 14:00 - 15:30
Sala 3 / *Room 3*
Victor Lazzarini (National University of Ireland)

ÁUDIO E RF: ONTEM, HOJE E AMANHÃ
Quarta-feira/*Wednesday,* 09 Maio/*May* 2012 15:30 - 17:00
Auditório D / *Auditorium D*
Tim Vear (Shure)

A EVOLUÇÃO DOS MIXERS ANALÓGICOS PARA OS MIXERS DIGITAIS
Quarta-feira/*Wednesday,* 09 Maio/*May* 2012 17:30 - 19:00
Auditório D / *Auditorium D*
Robert Scovill (Avid)

LA ACÚSTICA DEL TEATRO COLÓN DE BUENOS AIRES Y SU PRESERVACIÓN
Quarta-feira/*Wednesday,* 09 Maio/*May* 2012 18:00 - 19:30
Sala 3 / *Room 3*
Gustavo Jorge Basso (Universidad Nacional de La Plata)

10 Maio/*May*

MIXAGEM BASEADA EM OBJETO PARA SURROUND
Quinta-feira/ *Thursday,* 10 Maio/*May* 2012 13:30 - 15:00
Auditório D / *Auditorium D*
Jeff Levinson (ISONO Sound)

SOM AO VIVO E INSTALADO: GRANDES DIFERENÇAS
Quinta-feira/ *Thursday,* 10 Maio/*May* 2012 15:30 - 17:00
Sala 3 / *Room 3*
Aldo Soares (ARS Áudio e Acústica)

APLICAÇÕES PRÁTICAS DE SISTEMAS DE REALIMENTAÇÃO À REPRODUÇÃO DE BAIXAS FREQUENCIAS
Quinta-feira/ *Thursday,* 10 Maio/*May* 2012 15:30 - 17:00
Auditório D / *Auditorium D*
Mario Di Cola (Politecnico di Milano)

*Palestras confirmadas até o dia 03/05/2012

PROGRAMAÇÃO DE ARTIGOS CONVIDADOS / *SCHEDULE OF INVITED PAPERS*

09 Maio/*May*

ACÚSTICA ARQUITETÔNICA PARA SALA DE GRAVAÇÃO SONORA
Quarta-feira/*Wednesday,* 09 Maio/*May* 2012 11:30 - 13:00
Sala 3 / *Room 3*
Elcione Moraes (UFPA)

10 Maio/*May*

REVISÃO DE MÉTODOS DE AVALIAÇÃO DE QUALIDADE DE VOZ PARA SISTEMAS DE TELEPRESENÇA
Quinta-feira/ *Thursday,* 10 Maio/*May* 2012 13:30 - 15:00
Sala 3 / *Room 3*
Luiz Wagner Pereira Biscainho, Leonardo de Oliveira Nunes (UFRJ)

SESSÃO DE ARTIGOS / *PAPERS SESSIONS*

SESSÃO 1 / *SESSION 1*

AUDIO ELETRÔNICA; COMPUTAÇÃO MUSICAL; ANÁLISE E PROCESSAMENTO DO SINAL; RECONHECIMENTO DA FALA EM SISTEMAS DE DIÁLOGO AUTOMÁTICO

Terça-feira/*Tuesday,* 08 Maio/*May* 2012 09:00 - 11:00 Sala 3 / *Room 3*
Coordenador de sessão / *Chair:* Régis R. A. Faria

DESIGNING BODILESS MUSICAL INSTRUMENTS
FORNARI, J. (UNICAMP)

UTILIZAÇÃO DE UM SISTEMA COMPOSICIONAL ESPECTRAL NA CRIAÇÃO DE ANTICLEPSIDRA
SANTOS, R. S.; OLIVEIRA, L. J. P. (UFCG)

IMPLEMENTAÇÕES ATIVAS DE FILTROS PARA CONFORMAÇÃO DE RUÍDO ROSA
DALCASTAGNÊ, A. L.; NOCETI FILHO, S.; SILVA, G. C. K. (UFSC)

PROPOSTA DE UM SISTEMA DE DIÁLOGO AUTOMÁTICO BASEADO EM ALGORITMOS DE APRENDIZADO POR REFORÇO
AVILA, A.; FRAGA, F. J. (UFABC)

SESSÃO 2 / *SESSION 2*

ESPACIALIZAÇÃO, MIXAGEM E MONITORAÇÃO EM SISTEMAS MULTICANAL; ANÁLISE DO SINAL PARA RECONHECIMENTO DE ALTURAS MUSICAIS; ACÚSTICA DE INSTRUMENTOS MUSICAIS

Quarta-feira/*Wednesday,* 09 Maio/*May,* 2012 09:00 - 11:00 Sala 3 / *Room 3*
Coordenador de sessão / *Chair:* André Dalcastagnê

TRANSFORMADA Q CONSTANTE DE COMPRIMENTO DE JANELA FIXO
LIMA, J. A. S.; BRANDÃO, M. C. P.; FARIAS, M. C. Q. (UnB)

ESTÚDIO VIRTUAL PARA MONITORAÇÃO ATRAVÉS DE FONES DE OUVIDO DE SISTEMAS BASEADOS EM AMBISONICS DE 1ª ORDEM
SOUSA, F. (University of York)

DISCUSSÃO TEÓRICA SOBRE A INCLUSÃO DO ESCOAMENTO E PERFIS DE VELOCIDADE NA MODELAGEM FÍSICA DE INSTRUMENTOS DE SOPRO
OLIVEIRA, L. C.; GOLDEMBERG, R.; MENDES, R. S. (UNICAMP)

SÍNTESE POR AMOSTRAGEM DE PIANO BASEADA EM CODIFICAÇÃO WAVELET
ROQUE, T. R.; MENDES, R. S. (UNICAMP)

AUDIENCE PARA PD: UMA BIBLIOTECA PARA ÁUDIO ESPACIAL ORIENTADA A CENA SONORA
FARIA, R. R. A. (USP)

SESSÃO DE ARTIGOS / *PAPERS SESSIONS*

SESSÃO 3 / *SESSION 3*

RESTAURAÇÃO DO SINAL; MELHORIA DO SINAL DE VOZ; MARCA D'ÁGUA EM ÁUDIO; RELAÇÃO ENTRE PERCEPÇÃO ESPACIAL SONORA E IMAGEM VISUAL; ANÁLISE DO SINAL PARA RECONHECIMENTO DE ANDAMENTOS MUSICAIS EM REGISTROS SONOROS

Quinta-feira/*Thursday,* 10 Maio/*May,* 2012 09:00 - 11:00 Sala 3 / *Room 3*
Coordenador de sessão / *Chair:* Márcio Brandão

COMPARAÇÃO ENTRE MÉTODOS ESPECTRAIS PARA MELHORIA DE SINAIS DE VOZ SOB RUÍDO BRANCO
ANTUNES JÚNIOR, I.; BORDIN JUNIOR, C. J. (UFABC)

MARCA D'ÁGUA EM ÁUDIO: UM ESTUDO VOLTADO AO DIREITO DE PROPRIEDADE INTELECTUAL
JESUS, A. L. M.; ALVES, D. R. (PUC-MG)

INTERACTIONS BETWEEN AUDIO AND 3D STEREOSCOPIC IMAGE IN THE PERCEPTION OF DISTANCE AND SUITABLE RECORDING DISTANCE
KRUSZIELSKI, L. F.; KAMEKAWA, T.; MARUI, A. (Tokyo University of the Arts)

TEMPO ESTIMATION: EVALUATION OF DIFFERENT SPECTRAL FLUX COMPUTATION METHODS
NUNES, L. O.; BISCAINHO, L. W. P. (UFRJ)

ENCONTRO SOBRAC – AES BRASIL / *MEETING SOBRAC – AES BRAZIL*

ENCONTRO REUNINDO A SOCIEDADE BRASILEIRA DE ACÚSTICA E A SOCIEDADE BRASILEIRA DE ENGENHARIA DE ÁUDIO PARA O ESTABELECIMENTO DE UM PLANO DE COOPERAÇÃO.

Terça-feira / *Tuesday,* 08 Maio/*May* 2012 11:30 - 13:00
Auditório D / *Auditorium D*
Dinara Paixão (Presidente da SOBRAC) e Aldo Soares (Presidente da AES-Brasil)

SUMÁRIO / *CONTENTS*

Comitê de Organização / *Organizing Committee* IV

Comitê de Programa Técnico (revisores) / *Technical Program Committee (reviewers)* IV

Editorial V

Programação de Palestras / *Schedule of Lectures* VI

Sessões de Artigos / *Papers Sessions* VIII

PALESTRAS / *LECTURES*

OS MECANISMOS DA AUDIÇÃO

(08/mai 13:30 – 15:00 Sala 3)

Marcelo S. Portela (LVA-UFSC) 1

MUSIC PROGRAMMING SYSTEMS IN THE 21ST CENTURY

(09/mai 14:00 – 15:30 Sala 3)

Victor Lazzarini (National University of Ireland) 3

LA ACÚSTICA DEL TEATRO COLÓN DE BUENOS AIRES Y SU PRESERVACIÓN

(09/mai 18:00 – 19:30 Sala 3)

Gustavo G. Basso (Universidad Nacional de La Plata) 13

ARTIGOS CONVIDADOS / *INVITED PAPERS*

ACÚSTICA ARQUITETÔNICA PARA SALA DE GRAVAÇÃO SONORA

(09/mai 11:30 – 13:00 Sala 3)

Elcione Moraes (UFPA) 19

REVISÃO DE MÉTODOS DE AVALIAÇÃO DE QUALIDADE DE VOZ PARA SISTEMAS DE TELEPRESENÇA

(10/mai 13:30 – 15:00 Sala 3)

Leonardo O. Nunes, Luiz W. P. Biscainho (UFRJ) 28

SUMÁRIO / *CONTENTS*

ARTIGOS / *PAPERS*

DESIGNING BODILESS MUSICAL INSTRUMENTS

(08/mai 09:00 – 11:00 Sessão 1 Sala 3) (Artigo Completo / *Full Paper*)

José Fornari (UNICAMP) 41

UTILIZAÇÃO DE UM SISTEMA COMPOSICIONAL ESPECTRAL NA CRIAÇÃO DE ANTICLEPSIDRA

(08/mai 09:00 – 11:00 Sessão 1 Sala 3) (Artigo Completo / *Full Paper*)

Raphael S. Santos, Liduino J. Pitombeira Oliveira (UFCG) 47

IMPLEMENTAÇÕES ATIVAS DE FILTROS PARA CONFORMAÇÃO DE RUÍDO ROSA

(08/mai 09:00 – 11:00 Sessão 1 Sala 3) (Artigo Completo / *Full Paper*)

André Dalcastagnê, Sidnei Noceti Filho, Gabriel C. Kulevicz da Silva (UFSC) 55

PROPOSTA DE UM SISTEMA DE DIÁLOGO AUTOMÁTICO BASEADO EM ALGORITMOS DE APRENDIZADO POR REFORÇO

(08/mai 09:00 – 11:00 Sessão 1 Sala 3) (Artigo Curto / *Short Paper*)

Anderson Avila, Francisco Fraga (UFABC) 61

TRANSFORMADA Q CONSTANTE DE COMPRIMENTO DE JANELA FIXO

(09/mai 09:00 – 11:00 Sessão 2 Sala 3) (Artigo Completo / *Full Paper*)

Jonathan A. S. Lima, Márcio C. P. Brandão, Mylene C. Q. Farias (UnB) 66

ESTÚDIO VIRTUAL PARA MONITORAÇÃO ATRAVÉS DE FONES DE OUVIDO DE SISTEMAS BASEADOS EM AMBISONICS DE 1ª ORDEM

(09/mai 09:00 – 11:00 Sessão 2 Sala 3) (Artigo Completo / *Full Paper*)

Fábio Wanderley Janhan Sousa (University of York) 74

DISCUSSÃO TEÓRICA SOBRE A INCLUSÃO DO ESCOAMENTO E PERFIS DE VELOCIDADE NA MODELAGEM FÍSICA DE INSTRUMENTOS DE SOPRO

(09/mai 09:00 – 11:00 Sessão 2 Sala 3) (Artigo Completo / *Full Paper*)

Luís Carlos Oliveira, Ricardo Goldemberg, Rafael Santos Mendes (UNICAMP) 82

SÍNTESE POR AMOSTRAGEM DE PIANO BASEADA EM CODIFICAÇÃO WAVELET

(09/mai 09:00 – 11:00 Sessão 2 Sala 3) (Artigo Completo / *Full Paper*)

Thiago Rossi Roque, Rafael Santos Mendes (UNICAMP) 90

SUMÁRIO / *CONTENTS*

ARTIGOS / *PAPERS*

AUDIENCE PARA PD: UMA BIBLIOTECA PARA ÁUDIO ESPACIAL ORIENTADA A CENA SONORA

(09/mai 09:00 – 11:00h Sessão 2 Sala 3) (Artigo Completo / *Full Paper*)

Regis Rossi Alves Faria (USP) 98

COMPARAÇÃO ENTRE MÉTODOS ESPECTRAIS PARA MELHORIA DE SINAIS DE VOZ SOB RUÍDO BRANCO

(10/mai 09:00 – 11:00 Sessão 3 Sala 3) (Artigo Completo / *Full Paper*)

Irineu Antunes Júnior, Claudio J. Bordin Júnior (UFABC) 107

MARCA D'ÁGUA EM ÁUDIO: UM ESTUDO VOLTADO AO DIREITO DE PROPRIEDADE INTELECTUAL

(10/mai 09:00 – 11:00 Sessão 3 Sala 3) (Artigo Completo / *Full Paper*)

Ana Lúcia Martins de Jesus, Dorirley Rodrigo Alves (PUC-MG) 113

INTERACTIONS BETWEEN AUDIO AND 3D STEREOSCOPIC IMAGE IN THE PERCEPTION OF DISTANCE AND SUITABLE RECORDING DISTANCE

(10/mai 09:00 – 11:00 Sessão 3 Sala 3) (Artigo Curto / *Short Paper*)

Luiz Fernando Kruszielski, Toru Kamekawa, Atsushi Marui (Tokyo University of the Arts) 121

TEMPO ESTIMATION: EVALUATION OF DIFFERENT SPECTRAL FLUX COMPUTATION METHODS

(10/mai 09:00 – 11:00 Sessão 3 Sala 3) (Artigo Completo / *Full Paper*)

Leonardo O. Nunes, Luiz W. P. Biscainho (UFRJ) 125

Índice de Autores / *Authors Index* 133

Nesta edição comemorativa de 10 anos dos congressos de engenharia de áudio, os anais oferecem uma facilidade aos leitores: ao final de cada artigo, em seu rodapé, estão suas referências bibliográficas completas, prontas para serem copiadas e coladas nos trabalhos que o citarem.

PALESTRAS / *LECTURES*

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Os mecanismos da audição

Marcelo S. Portela
Laboratório de Vibração e Acústica, Universidade Federal de Santa Catarina
marcelo.portela@lva.ufsc.br

RESUMO

O sentido da audição é realizado por um processo conhecido como transdução auditiva. Nesta palestra será apresentada uma descrição do mecanismo que proporciona essa percepção. A orelha interna, a parte mais antiga do sistema auditivo, trata-se de uma evolução de um órgão sensorial dos peixes, responsável pela detecção de variações da pressão hidrostática. A mudança para ambientes terrestres tornou necessária uma adaptação para perceber as variações de pressão sonora no ar. Como a impedância acústica mais baixa no ar do que no fluido da orelha interna, a Natureza encontrou uma solução: a estrutura da orelha média. Com os ossículos, tornou-se possível transmitir energia de modo eficiente até a orelha interna. A cavidade timpânica veio aparecer somente nos amniotas, animais tetrápodes, há 220-250 milhões de anos. Ela teve grande impacto na capacidade auditiva dos animais, possibilitando a percepção de frequências sonoras mais altas. De modo geral, quanto maior o comprimento da papila auditiva, melhor a capacidade para ouvir sons de altas frequências. Para acomodar o epitélio auditivo cada vez mais longo, há cerca de 100 milhões de anos surgiu outra estrutura importantíssima: a cóclea, onde som é decomposto em frequências por cerca de 20.000 células sensoriais (células ciliares). As estruturas envolvidas na transdução auditiva serão detalhadas através de um vídeo que simula um passeio virtual pela orelha humana.

Serão explicadas as diferenças entre diferentes vias de percepção auditiva: a via aérea e a via óssea. A sensibilidade auditiva humana será comparada com a de diversos animais e será também esclarecido porque o ouvido humano não responde igualmente a todas as frequências nem a diferentes amplitudes, a partir de estudos que delimitaram o que hoje chamamos de limite de audibilidade e limite da dor. Estas pesquisas relacionam intensidade física da excitação e intensidade subjetiva da sensação e deram origem às curvas de ponderação, que compensam as variações de sensibilidade com a frequência e indicam atenuações a serem aplicadas a cada situação de exposição sonora.

Também serão demonstrados os efeitos do ruído sobre o aparelho auditivo em vários níveis, o mecanismo natural de defesa do ouvido (reflexo aural ou reflexo acústico) e o fenômeno da presbiacusia, processo de calcificação e perda da elasticidade das fibras no ouvido com o tempo. Dados de perda de audição por exposição prolongada ao ruído serão apresentados, bem como audiogramas típicos de pessoas com deficiência auditiva e curiosidades sobre o tema como, por exemplo, o fato do Brasil ser um dos líderes mundiais em nível de ruído urbano, ou que apenas 5% da população brasileira com problemas auditivos recorre a médicos.

Pretende-se discutir também algumas das muitas razões pelas quais raramente é possível garantir como um sistema de áudio vai soar apenas lendo suas especificações. Quando são feitas as medições acústicas de um

sistema desse tipo, tipicamente se utiliza um microfone calibrado, enquanto os ouvidos funcionam de forma diferente. Em primeiro lugar, eles trabalham normalmente em dupla em todos os momentos, o que pode ser chamado de audição biauricular ou binaural, conceito também a ser discutido na palestra. Será aproveitado o tema para divulgar a campanha INAD - Dia Internacional de Conscientização sobre o Ruído.

REFERÊNCIAS BIBLIOGRÁFICAS

HENRIQUE, Luís L. (2007), *Acústica Musical*, 2ª Edição, Lisboa, Fundação Calouste Gulbenkian.

BENADE, Arthur H. (1990), *Fundamentals of Musical Acoustics*, New York, Dover.

CHASIN, Marshall (2001), *Hear the Music - Hearing Loss Prevention for Musicians*, Toronto, Canadian Hearing Society.

EARGLE, John M. (1995), *Music, Sound and Technology*, New York, Van Nostrand Reinhold.

PLETSCH, Brandon (2003), *Auditory Transduction*, Animação Multimídia, primeiro lugar no Science and Engineering Visualization Challenge, Science Magazine and the National Science Foundation.

ROEDERER, Juan G. (1995), *The Physics and Psychophysics of Music: an Introduction*, New York, Springer-Verlag.

ROSSING, Thomas D. (1990), *The Science of Sound*, Reading, Addison-Wesley.

WRIGHTSON, T., KEITH, A. (1918), *An Inquiry into the Analytical Mechanism of the Internal Ear*, London, Macmillan.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Music programming systems in the 21st century

Victor Lazzarini[1]

[1] National University of Ireland, Maynooth, Sound and Digital Music Technology Group, Department of Music
Maynooth, Co.Kildare, Ireland

Victor.Lazzarini@nuim.ie

ABSTRACT

This paper discusses the development of music programming systems from the beginnings of Computer Music into the 21st century. It will trace the history and evolution of these languages and their main design characteristics, from the early off-line synthesis programs to realtime interactive systems. In particular, we will focus on Csound, highlighting its main internal aspects and its applications. The paper will also explore the various paradigms that have oriented the design and use of music programming systems. The discussion is completed in looking at the concept of computer music eco-systems and its pervasiveness in today's practice.

0 INTRODUCTION

Music programming systems have been at the core of the development of Computer Music since its inception in the 1950s. In fact, developments that were proposed by and implemented in such environments were essential for the ubiquitous presence of digital audio in the Music and Creative industries, and for its adoption by musicians. Innovative ideas introduced in these systems were adopted by instrument designers in successful products (the Moog and Buchla modular synthesizers, which borrowed the concept of unit generators; the Yamaha DX7 synthesizer, which implemented Frequency Modulation synthesis in hardware).

A music programming system, as implied by the name, is one that allows users to create music programmatically. Not only it provides the basis for defining the sequencing of events that make up a musical performance, but it allows ways to design, to a very fine degree, the processes that generate audio for it. In other words, it requires programmability at various levels. Some systems only offer this partially, for instance allowing users to create algorithms for score generation and event scheduling, but without providing synthesis capabilities to go with this. In addition, audio libraries, signal processing specification languages and programmable software synthesizers also only partially offer the capabilities that make up a music programming envi-

ronment.

This paper is organized as follows. First we will introduce the major music programming systems that were developed throughout the history of computer music. This section will discuss the major attributes and characteristics of these environments, from the perspective of system design and use. After introducing the major developments in the field, we will concentrate our attention in one particular system, Csound, and explore some of its internal aspects. Following this, we will consider a series of important paradigms that are central to our discussion. To conclude, we will look at the notion of eco-systems for music programming, and its importance for today's Computer Music practitioners.

1 THE DEVELOPMENT OF MUSIC PROGRAMMING SYSTEMS

In this section, we will trace the development of music programming languages from the early batch-computing systems to realtime interaction. The discussion will focus on technical aspects and characteristics of each of these systems, with an aim to gain insight on their design philosophies.

1.1 Early Systems

The path to modern music programming was developed along four decades, since the first direct digital synthesis program by Max Mathews, in 1957, MUSIC I, and its descendants [1]. The evolution of music software systems was characterized by increasing generality and flexibility, from simple synthesis programs to fully-fledged programming languages. Interestingly, some important technologies were introduced early on: the table-lookup oscillator, which is the single most important component in digital synthesizers, was introduced by Mathews in MUSIC II. The concepts of unit generator, a synthesis module and a compiler for computer instruments, which are still essential to much of modern music programming, were introduced in the next iteration of his system, MUSIC III [2] [3]. In particular, the original idea of the unit generator is one that has been applied almost universally, not only in software synthesis systems, but also in hardware devices, such as the classic analogue synthesizers (which were in fact preceded by MUSIC III). In MUSIC IV, these and other developments gave rise to a series of principles that, together, make up the MUSIC N paradigm.

1.2 MUSIC IV

Although some of the functionality seen in modern computer music synthesis systems was already present in MUSIC III [4], it is accepted that the first general model of a music programming system was provided by MUSIC IV, which ran on the IBM7094 computer. This was a substantially complex software package, described in some detail in its programmer's manual [5] and a tutorial [6]. It was composed of a number of elements that were to be executed in three distinct phases or *passes*. The first pass took control data, a computer *score* and associated function-table generation instructions, in an unordered form and stored in temporary tape files. This data was then input to the second pass, where data was sorted in time order and tempo transformations applied, producing another set of temporary files. Finally in the third pass a synthesis program was loaded, taking the data from the previous stage, and generating the sample data that would be stored in the output file for subsequent playback via a digital-to-analogue converter.

The core of the system was the synthesis program compiler, written in BE FAP (Bell Fortran Assembly Program, the Bell Telephone Labs IBM7094 assembler). This was effectively a block-diagram language compiler, used to connect the different unit generators available to the programmer. A number of basic conventions governed this language, such as the modes of connections allowed, determined by each unit generator and how they were organized syntactically. An incipient type system was present, defining U (unit generator outputs), C (conversion function outputs), P (note parameters), F (function tables) and K (system constants) data types. The basic code block in the synthesis program was the *instrument*, which made the written program an *orchestra*. The maximum number of concomitant instances of an instrument was fixed and each *note* in the prepared score for each program would have to be set to run on an specific (free) instance.

MUSIC IV made provisions for score processing subroutines, written in a high-level language, FORTRAN, which could handle some aspects of music programming for composition. Conversion functions used in the orchestra could also be written in that language. While not a fully-fledged music programming language in the modern sense, the system enabled a good deal of programmability, specially in the design of the synthesis program. MUSIC IV was recreated at Princeton University as MUSIC IVB [7] and then as MUSIC 4BF [8] [9], written in FORTRAN (and using that language for the programming of instrument definitions). It lives on as MUSIC 4C (M4C), which is a C-language port, still available on modern UNIX-derived operating systems.

M4C preserves the three passes of earlier versions, albeit with some modifications [10]. Instruments are written in C and compiled in an orchestra as before, which is now built into a single program that does all the processing stages. Pass 1 initializes the output soundfile, reads the score file, set parameters and allocates instances of instruments. Pass 2, as before, does he time sorting of the score. Pass 3 runs the sorted score through a scheduler that instantiates instruments at the correct times.

In M4C, similarly to MUSIC 4, distinct programs will be created for each orchestra, with maximum available instances predefined for each instrument. Such

orchestras can be written with instrument template files that combine some basic commands and C code blocks. From these, the INSDES compiler generates C code that can be built to create a M4C program that includes the orchestra and the three passes outlined above. To synthesize audio, this program is given a score file and an output soundfile. For orchestra design, M4C will require the C language toolchain to be present (preprocessor, compiler, linker), but as the system was designed for UNIX, this is taken for granted. In addition to the INSDES compiler, the package also includes an orchestra processor (NOTEPRO), which translates a textual score containing traditional staff notation music attributes (12 TET pitches, metric rhythms, etc) into a M4C numerical score. This system is still maintained for modern UNIX-like operating systems.

CMix (and later RTCMix), as discussed in [11], also employs a similar principle, where C-written instruments are compiled and linked into a main program that can be controlled by a score written in a scripting language, Minc. A main difference is that CMix does not employ an intermediary instrument specification as in M4C, but expects its orchestras to be written directly in C. For this, it provides a basic library of commonly-used unit generator functions.

1.3 MUSIC V and its Progeny

The culmination of Mathews' efforts at Bell Labs was MUSIC V [12], a new version of the system, mostly written in FORTRAN, which made it portable to other computer installations (in fact, it still can be run on modern operating systems). It still featured a three-pass process, however, the operation steps were more integrated than in MUSIC IV. The orchestra compilation step was now combined with pass 3, without the need to generate a separate synthesis program. FORTRAN conversion subroutines were also integral to the program code. Also the whole MUSIC V code was written in a single score, which contained both the note lists and the instruments. Unlike MUSIC IV, there was no maximum instance count for each instrument. Unit generators could be written either in FORTRAN or as separate machine-language subroutines.

Pass 1 of MUSIC V scans the score (which includes both the instrument definitions and the note list proper) and produces a numeric representation of it. The second pass sorts the note list in time order and applies the CONVT routine, which can be used to convert frequencies to sampling increments etc. Finally, pass 3 schedules the events, calls the unit generators and writes the output. MUSIC V, as its predecessor, provides simple orchestra data types: P (score parameters), V (scalar values), B (audio signal buffers) and F (function tables). Audio is generated in a series of sample blocks (or vectors), which by default held 512 samples. Vector-based processing became standard in most modern computer music systems (albeit with some notable options). Although no longer maintained, MUSIC V has been ported to modern systems using the gfortran compiler [13].

Moore's Cmusic [14], a component of the CARL computer music package [15], was largely modelled on MUSIC V, inheriting much of its orchestra syntax (for instance, its data types, unit generator names and parameter arrangement, among other things). It also extended the possibilities by offering a C preprocessor, with macro substitutions and include statements, as well as some conversion operators (like hz and db, to covert frequencies into sampling increments and decibels into linear amplitudes, respectively). Moreover, its score offered the possibility of direct expression evaluation in parameter fields, which was not present in any of the earlier systems. Cmusic was designed to work with other CARL programs in the context of a UNIX environment, making heavy use of interprocess communication such as pipes and IO redirection. Audio data synthesized by the software could be used directly as input to other processes, such as a reverb effect or a plotting program, or streamed to output. This enabled the augmentation of the music programming system with UNIX shell scripting. Cmusic is no longer being maintained, although a version of the system has been ported to modern UNIX-like operating systems [13].

Another notable successor to MUSIC V is Common Lisp Music (CLM) developed by Bill Schottstaedt from 1990 onwards, and still maintained, currently in its version 4 [16]. The system uses the Common Lisp (CL) language as a glue, with a lower-level backend, which is used for audio computation. On its early implementation, the backend was based on assembly-language code for a DSP microprocessor (the Motorola 56000), but later, the C language was used. Instruments are designed in CL, with the critical parts wrapped in a run macro, which compiles the Lisp code into the backend code. In modern systems, this is a C language module that is built as a dynamic library, loaded by CLM when audio processing is requested. Unlike MUSIC V and many of the comparable systems, the CLM-generated sound generating code does not employ vector-based processing of audio, but works on a sample (or sample-frame) by sample basis. The C-based CLM backend exists also as a complete C library (SndLib), which can be used independently in synthesis programs. CLM provides advanced support for manipulating note lists due to its use of the CL language. It can also be integrated with Snd soundfile editor that can serve as a frontend to the system.

1.4 Realtime-Oriented Systems

While music programming systems have traditionally been oriented to offline rendering of audio, the possibilities offered by modern general-purpose computing platforms for realtime audio processing have been explored by a number of systems. The earliest of these realtime-oriented systems, Max, was originally designed simply as a scheduler for an outboard synthe-

sizer [17]. It eventually incorporated its own signal processing capabilities and lives on in two widely-used modern systems, Pure Data (PD) [18], which is Free software, and MaxMSP [19], which is a closed-source, non-Free package. As it is not possible to examine the source code of the latter system, our discussion here will be limited to the former.

Aside from the superficial aspect of being graphically-edited, a significant difference between PD and the systems discussed in the previous sections is the absence of the concept of a score. Here, the idea of instances of compiled instruments being initiated and controlled by note events does not exist. In fact, instruments, which live in PD patches, are generally single-instance, unless copies of the patch are made. Also, given the realtime orientation, running patches can be directly modified, which is not possible with compiled instruments. This leads to different modes of interaction and of composition, which lean towards user interaction and improvisation.

The MUSIC N heritage lives on in PD with the concept of unit generators as the central objects in the system. In a sense, it embodies the most direct form of the modular synthesizer metaphor, whereby a program is made of various boxes that can be interconnected in various ways, with the possibility of various external controls being applied to them. Its author disputes the application of the concept of programming environment to it [17], citing the fact that, for economy reasons, the system lacks support for many aspects that are fundamental to creating programs. However, if we are considering the design of synthesizers or effects processors, PD is a legitimate music programming system, which can be used, of course, in context with other programming tools.

One of the essential aspects of operation in PD is its control scheduling. Triggering and passing data are unified in a single mechanism: simple triggers carry no data, while specific data types will work as specific triggers to object actions. This is embodied in PD's messaging system, which is used everywhere in the system to send sporadic control information from one object to another. It features a selector mechanism, where each message is headed by an identifier, that will trigger a method in a receiving object, producing an output.

In addition to these control messages, audio computation employs a completely separate system. This is effected by special unit generators, which react to a DSP message and implement vector-processing callbacks that are scheduled in a signal processing graph. The coexistence of an asynchronous messaging system and this ordered, sample-clock synchronous, audio system is uneasy. However, although it might offer difficulties in some situations, in practice, it can handle many use cases with no significant problems.

Another realtime-oriented music programming system is SuperCollider 3 (SC3) [20]. This is based on two components: an interpreted language (SCLang) and a separate Open Sound Control (OSC)-based software synthesizer (SCSynth). SCLang is a complete smallTalk-like object-oriented language that issue OSC commands over the network to SCSynth. It is designed to break the separation of orchestra (instrument) and score (note-lists), typical of MUSIC-N systems. Here, events and sound processing which can be programmed in an integrated way and the language features strong support for algorithmic composition. This is also a feature of another Lisp-based system, Nyquist [21], although it does not share the same realtime orientation as SC3.

2 CSOUND

Csound is possibly the longest running heir to the early MUSIC N systems. It was developed alongside a number of equivalent (at the time) systems, such as Cmusic, Cmix and M4C, in the 1980s using the by then well-established C language. In time, it developed into a much larger and multi-functional music programming environment and with the advent of its version 5 in 2006, it introduced a number of important concepts that were innovative for systems of the kind. In the next sections, we will try to trace its evolution from the beginning until today.

2.1 From MUSIC 360 to mit-ems Csound

MUSIC 360[22] was written at Princeton University by Barry Vercoe for the large IBM 360 computer [23]. It was directly derived from MUSIC IVB and MUSIC IVBF, which were also developed at Princeton, and thus equivalent to those systems as a MUSIC IV variant. This program was taken to other IBM 360 and 370 installations, including for instance the one at MIT. However, as it was tied in to those large computer installations, it did not suit smaller institutions, and was not widely available.

The structure of MUSIC 360 is very similar to its predecessors, utilizing the basic three passes discussed above. Unlike the original MUSIC IV program, all passes are combined into a single 'load module' (the program) after the orchestra is compiled and linked to the library subroutines. Pass 1 reads the score and implements the carry preprocessing (where values can be 'carried over' from one note statement to another). Pass 2 sorts the score in its correct temporal order and apply any tempo warping defined in it. Pass 3 calls the orchestra initializing its instruments and synthesizing the sound.

MUSIC 360 allowed any number of concomitant instances of instruments to be performed at the same time. The system defined two routines for each instrument, for the initialization-time and performance-time stages. This was an important feature of the system, evident in the separation of 'initialization' and 'performance' descriptions of each unit generator in the reference manual. The orchestra syntax is very close to assembly language, based on operator codes (opcodes)

representing the unit generators, something that will be inherited by MUSIC 11 and Csound.

An advanced aspect of the language was that arithmetic expressions of up to twelve terms could be employed in the code, with the basic set of operations augmented by a number of conversion functions. Data types included 'alpha'-types, which could hold references to unit generator results (both at I-time and P-time), K-types, used to hold values from a KDATA statement (that was used to hold program constants), P-types for note list p-fields and U-types, which could be used to reference unit generator results (as an alternative to 'alpha' variables). There was scoping control, as symbols could be global or local, which included a facility for accessing variables that were local to a given instrument. The language also supported conditional control of flow, another advanced feature when compared to equivalent systems. Some means of extendability were provided by opcodes that were able to call external FORTRAN IV subroutines at I- or P-time. In addition, to facilitate programming a macro substitution mechanism was provided. MUSIC360 was quite a formidable system, well ahead of its competitors, including MUSIC V.

At the MIT Experimental Music Studio, Vercoe developed MUSIC 11[24], a version of the MUSIC 360 system for the smaller PDP-11 minicomputer [25]. As the PDP 11 popularity grew in the 1970s, and with the introduction of the UNIX operating system, the program was used at various institutions both in the USA and elsewhere for over two decades. As we would expect, not only many of the innovative features of MUSIC 360 were carried over to the new system, but also import concepts were pioneered here.

One of the main design aspects first introduced in MUSIC 11 was the concept of control (k-) and audio (a-) computation rates. In this mechanism, for each control sample, a vector of ksmps audio samples was produced. This made the system possibly the most computationally efficient software for audio synthesis of its time. Together with the i-time concept, this established the main operation principle of the orchestra language, dividing instrument action times into initialization and performance at two rates, which was realized in the three basic data types: i,k (scalars) and a (vectors). Global, local and temporary variables were available (marked as g, l or t). MUSIC 11 also featured continuous garbage collection, as allocated instrument spaces, when free, could be taken over by new instances.

MUSIC 11 was superseded by the development of its C-language version, Csound, in 1986 (mit-ems Csound). This program inherited many aspects of the MUSIC 11, but now it integrated the orchestra compiler and loader into a single command. The original system included three separate commands, csound, scsort and perf. The latter was a just a convenience tool to call scsort and csound in sequence. These two commands would sort the score (scsort); compile and load the orchestra and run the score on it (csound). As with all programs of the era, mit-ems Csound was written in Kerninghan & Ritchie(K&R)-style C.

Csound was originally a very faithful port of MUSIC 11, so much that even today code for that language can be run on modern versions of the system. Some small differences existed in terms of a collection of new opcodes, and the removal of some others. Also, the separation between temporary and local variables was removed in Csound, and a means of extending the language with new C-code opcodes was provided. However, beyond these small differences, the central concepts were shared between the systems.

In the 1990s, the centre of development of Csound moved from MIT to University of Bath. Realtime operation had been introduced to the system [26] in the mit-ems version. From this, the system developed into an offline composition and realtime synthesis language with widespread applications explored in [27]. Examining the source code of the 1995 version, we see the system re-written in the established ANSI C89 form of the language (from the original K&R C dialect), supporting realtime audio on a number of hardware platforms running UNIX: Sun, NeXT, DEC, SGI and HP. The program had been ported to PC-DOS, also with realtime audio via soundblaster soundcards. Separately, at Mills College, a version of the system for the Macintosh platform had been developed [28]. In fact, at the end of the 1990s, the system had been ported to almost all modern general-purpose computing platforms with a C compiler.

In addition, separately it was ported to run on custom DSP hardware, in a closed-source version designated *extended Csound* [29]. This version eventually became part of a commercial project of Analog Devices, inc., to supply simple synthesizers for embedded applications [30]. Meanwhile, a large international developer community was involved in expanding the open-source system, which eventually came under the Lesser GNU Public License (and thus Free software). Many new unit generators were developed for it, culminating on the Csound 4.23 version in 2002. The consensus among the community was that the system required a major re-engineering to be able to move forward. A code freeze was established so that the new Csound 5 system could be developed.

2.2 Csound 5

The main goals of the development of Csound 5 were: to provide a clean, re-engineered system, away from the original monolithic program; to support re-entrant instances of the system in a dynamic-loadable library; and to provide a plugin mechanism for various aspects of the system (unit generators, utilities, function tables) [31]. A final goal, only achieved in later versions of the system, was to provide a new orchestra language parser, based on the bison compiler compiler

and the flex lexer, which would be maintainable and extendable. The main motivation behind this work was to adopt more up-to-date design paradigms (see section 3 below) and to allow for a variety of uses for the system. Csound 5.00 was eventually launched in 2006, twenty years after the first mit-ems release in 1986.

With this version, Csound became a programming library, which could be used to embedded the system as synthesis engine for a variety of applications. In fact, since version 4.21, Csound had been released as a library with a public application programming interface (API), however with a number of limitations (for instance, only a single copy of the engine could be used by hosts). In Csound 5, all of these issues were removed and the system became fully embeddable.

In the wake of these changes, many elements were introduced to provide a flexible use of the engine. It was made possible to use the system from a variety of programming languages and various third-party host/frontend applications were developed. These developments enabled, amongst other things, the porting of Csound to mobile operating systems (the Mobile Csound Platform, MCP [32]), and its use as plugin in a variety of formats, and in the One Laptop Per Child project, running on the XO computer [33]. In addition, the new parser allowed the exploration of multicore processors, with the parallel Csound code introduced by John ffitch [34], which depends on some of its features.

3 A FEAST OF PARADIGMS

The development of modern music programming systems relies on a number of established paradigms. Some of these are borrowed from general programming practice, such as object orientation, others emerged from the evolution of music systems themselves. In this next section, we will examine four of these paradigms that play an essential role in Computer Music.

3.1 The Music N Paradigm

The MUSIC N paradigm is an emergent set of properties from the main music programming systems discussed above. Although the concept has been used widely, there has never been a full definition for it. The essence of the paradigm is that an environment for Computer Music should be programmable. The existence of a compiler in MUSIC III, hailed by Mathews as its most important breakthrough [2], is something that survives in one shape or another in all of its descendants. This had the role of translating a symbolic description of a synthesis signal flow into a DSP graph, which would run in the computer. The term 'acoustic compiler', coined by Mathews, defines this element of the paradigm.

In some systems, instead of a compiler, we might find some sort of an interpreter, that would assemble the graph by creating and connecting instances of the processing boxes, the unit generators. These are also important components of the paradigm. Their existence is universal among music programming systems. They have also appeared elsewhere in electronic music instruments, such as the modular synthesizer, such is the pervasiveness of the concept. The table lookup oscillator, sometimes defined as the workhorse of digital synthesis, is an example of unit generator that is ubiquitously present in music programming. Given its importance, it can be argued that it also plays central role in the paradigm, together with the concept of function tables and their generation (GEN) routines.

A more disputed aspect of the MUSIC N paradigm is the presence of a separate score data specification language. While the most traditional systems, like the original MUSIC IV-V, Cmusic, CMix, MUSIC4C support and depend on the principle, other systems do not espouse it, at least directly (PD, SC3, Nyquist). Systems like CLM try to integrate it by making the score and orchestra language the same (Common Lisp), although maintaining a degree of separation between them. The case of Csound is one of a system where the score is separate but completely optional, having a variety of other methods for instrument instantiation.

We propose that the essential point of this aspect of the paradigm is not so much the existence of a separate score language, but the principle of instrument classes, from which objects can be created to produce audio. All of the systems described here would embrace this principle, even though in the case of PD, for instance, the separation between class and object can be virtually non-existent.

3.2 Object-Oriented Programming

From the perspective of programming, the MUSIC N paradigm is then realized through the principles of Object-Oriented Programming (OOP). This realization might be partial, with regards to what some view as the complete set of concepts involved in this paradigm. However, the essential ideas are present in all systems.

There is a general consensus that the first example of the OOP paradigm appeared in the Simula 67 language [35]. But it can be argued that systems like MUSIC IV already embodied the central aspects of object orientation, the idea of classes, as a description or template, and objects, the concrete instances of these. MUSIC IV provided these at two levels. Firstly, the unit generator can be thought of as a class, whose objects are instantiated in the instrument definition. Secondly, the instrument, as pointed out above, is also a class that can be instantiated a number of times by score events (the 'note cards' in MUSIC IV).

The OOP paradigm, it seems, serves the work involved in music programming very well, as it does in the case of Simula, the tasks of simulation. This is, in part, because much of the activities of music composition and performance can be modelled in object-oriented terms. For traditional music, we have the abstract idea of a note, of which instances a musical composition is made. This works even in extended concepts of music, where notes are not a valid concept anymore,

but where sound objects, textures, gestures, etc, can all be decomposed into objects, which we can describe in terms of types. So it is not a coincidence that the principle was discovered to fit the purposes of Computer Music very early on.

Other aspects of OOP are not supported in general by MUSIC N languages, but make an appearance in various systems. For instance, SC3 has a full set of features that have been modelled on the SmallTalk language, supporting various ancillary aspects of OOP. Csound 5 is a fully object-oriented programming library, which has both a C and a C++ APIs, as well as interface extensions and wrappers for various OOP languages. For the development of large music systems, this paradigm is essential. It has been found particularly useful as a means of support loadable modules, which use the concept of classes to provide a common interface between host systems and plugins. Such interfaces are found not only in music programming systems, such as PD, SC3 and Csound, but in architectures for Free and closed-source programs (Linux Audio Developer's Simple Plugin API, LADSPA; Virtual Studio Technology, VST; etc).

3.3 The Multi-language Paradigm

At the core of music programming lies the multi-language design paradigm, that of the use of a domain-specific language (DSL, in this case, the one provided by the music programming system) and, at different times, a system-implementation language (C/C++, FORTRAN) and/or a scripting language. This arrangement is very similar to what has been advanced by authors such as Ousterhout [36], where glue code is used to connect components implemented in languages that are closer to hardware. Such glue code can take the form of the instrument design language and any scripting that might be involved in, for instance, score generation.

This approach has been present in Computer Music since MUSIC IV, where the event generation for a given score was completely separated from the design of the synthesis program which would receive the score [5]. The MUSIC N languages themselves can be seen as glue code connecting unit generators implemented in a lower-level language [37]. This concept was extended to be the basis for the CARL system, where components where connected using shell scripting [15]. Similarly, it was used in Cmix [11], where synthesis programs written in C were glued by a score code written in MinC. More recently, it is the basis of the meta-programming principles embodied by Faust [38]. Since the advent of Csound 5, it has become an organic mode of operation, where Csound can be part of a system involving a number of other languages. In fact, Csound itself can embed other languages (such as Python and Lua).

3.4 The Open Source Dimension

The mechanism that propelled such momentous development in music programming, such as the one described in this paper, is indisputably the availability and exchange of source code. All the MUSIC N-derived systems were emphde facto open-source, even if the term and concept were never current at the time. However, we can see that these were given to any interested parties, with pleasure. These enabled the cross-fertilization of ideas, with the results clearly seen in the evolution of Computer Music as a research and artistic discipline.

The account of the earlier days of computer music perhaps demonstrate how the open source paradigm not only benefits the community of users, but also the evolution of the software itself. The larger its user base and developer community, the more it will thrive in the *software pool*. Csound is a good demonstration of this principle: although often derided by computer scientists, it grew to become one of the most used music programming systems. This was largely due to its availability, perhaps as much as to do with the existence of good documentation, performance and extendibility. Csound was one of the first music software to become fully available for ftp download when it was released in 1986.

Thanks largely to John ffitch's work on organising the source code into the *canonical* Csound, the software became the largest system around (in terms of its unit generator collection). This has been the effort of a large development community, which is evident when we look at the credit list, which has over thirty names on it. As with any other work of this type, people have come and gone, and the development has occurred in bursts and splits. A number of flavours, forks, of Csound were spawned, somewhat repeating the early developments in the MUSIC N family tree. Some of these survived, some died. But the development nevertheless went forward and things got changed and improved as needed.

It is interesting to compare the fate of two major systems of the same era: Cmusic and Csound. Twenty years after its first release, Csound was going strong with the release of the all-powerful version 5, whereas Cmusic had effectively disappeared. Some of the reasons for this were put forward very well by Richard Dobson, earlier on in 1999: "I suppose one significant difference between Cmusic and Csound, which might explain the relative lack of 'modern' opcodes in the former, is that while cmusic has largely remained the property and product of F.R. Moore, Csound has reaped the benefit of a large, skilful, enthusiastic and mostly unrestrained net-wide user and development group."[39]

The open-source paradigm has emerged from all of the experiences of developers in trying to evolve the sharing of ideas and the building of communities around software systems. Much of it has been put to the fore by the Free Software Foundation and its originator, Richard Stallman [40]. Although his principles go beyond the simple general guarantees for the copying, modification and distribution of source code, there is

much that is common between what is Free software and what is open source.

One of the important aspects that guarantees the status of open source software is the use of an appropriate license. In the case of Csound, the licensing issue was a question that took a while to be solved. The software had been originally released under the MIT License, which in the view of many people in the community was problematic and confusing. Csound was moved to the (Free-software) Lesser GNU Public License (GPL), for its version 5 release, after consultation with all its contributors and negotiations with the original copyright holders.

One of the major issues regarding the uncertain license status of certain open-source software was the fact that many developers did not really understand the need for licenses. For some, this was due to ignorance, and for others, like John ffitch himself and so many others, it was because they came from an era where the concept of open-source did not exist. As he explained, "in those days, the question was not there as everyone shared their software. In fact, when we managed to get a program to work we were so happy that we wanted to give it to everybody"[41].

With Csound, as with other systems, it is not only the case that the source code for the music programming system is available, but also that Csound-language code itself is freely distributed. Catalogues of Csound instruments were around for people to use, learn and modify. From Richard Boulanger's Toots to the well-known Amsterdam Catalogue, a wealth of code has been made available for the community. All these things are incredibly useful for people learning computer music, composition, signal processing, etc.. For this community, such resources are often more important than the software source code itself.

4 MUSIC PROGRAMMING ECOSYSTEMS

The model for music programming in the 21st century is that of an ecosystem of applications, which will provide support for a variety of tasks that are of value to Computer Music. Two of these can be identified as DSP programming/music application development, and composition. Music programming systems have to be able to flexible enough to allow for the emergence of this eco-system, either by participating in it, in a multi-application, multi-language environment, or by supporting (ie. providing the tools for) the development of cooperating applications. In this section, we examine these two these tasks and their support by modern systems.

4.1 DSP programming and application development

The translation of signal processing algorithms into running code has been a preoccupation of music programming since MUSIC III. Mathew's principle of unit generators as building blocks for a synthesis program has already been discussed above as a central aspect of computer music environments. Another important feature for the realization of DSP algorithms is support for elementary operations on signals, which was provided in terms of support for expressions in MUSIC360. Although this existed in MUSIC IV and MUSIC V via addition and multiplication, the possibility to use expressions, combined with some simple functions, allowed a compact way to translate formulae. In fact, Barry Vercoe states as one of the core aspects of a music programming language is the possibility to reproduce the operation of a unit generator using language primitives [25]. In this sense, the modern systems such as Csound, SC3 and PD enable DSP programming to fine degree of detail.

Some essential operations are awkward in certain-systems: for instance single-sample delays in PD are possible, but not intuitive to program. However, by taking advantage of the multi-language scenario, it is possible to use a system that is designed to translate DSP flowcharts and algorithms, such as Faust, to compile unit generators for a host music programming system (such as PD, Csound or SC3). That way, a large design problem can be broken down into separate components, utilizing the principle of separation of concerns [42].

This principle is even more evident when we come to the task of application development utilizing music programming systems. While environments such as SC3, PD and Csound can provide, to various extents, support for the authoring of music applications on their own, there is more flexibility and scope to using them in a multi-language scenario. Taking the case of Csound, for instance, it is very straightforward to create host applications in a variety of platforms, using either scripting or implementation languages, which use its music programming capabilities. With software developer kits (SDKs) such as the Mobile Csound Platform, Csound becomes the audio/music component in an agile development environment, in combination with the application-authoring languages (Objective-C for iOS; Java for Android and Web apps).

This scenario demonstrates a typical application ecosystem with various cooperating components. Here users can first prototype and develop their signal processing algorithms using Csound through one of its host frontends, using Faust (or C/C++) for creating new unit generators for the language if necessary. In complement, they can author their target application for a given system (desktop, mobile, web-based), add the Csound engine, which will load and run the processing code. This application can in turn be made to cooperate with other apps: for instance, a mobile app might be able to communicate with a desktop app, which can be used to design or make alterations to the Csound code, or share projects between the two in a variety of ways.

4.2 Composition

Composition has always been the prime area of music programming. Early non-realtime systems have been mainly composition environments, although they have also been used in signal processing research (a classic case is Chowning's development of FM synthesis [43]) and acoustics (e.g. J.C.Risset's studies of Shepard tones [44]). Thus, all modern music programming systems, continue in this tradition, even though some have been originally designed with performance in mind, as in the case of PD [17].

The tasks involved in the composition of Computer Music are a mix of instrument development, DSP design (as discussed above), and event programming/scheduling, which in many situations form a single continuum. Facilitating the integration of these tasks is important for some composers, whereas for others, a more clear separation between them is required. Music programming systems should be flexible enough to accommodate these two distinct needs, even if they are designed to give more support to one of them.

Even though some systems are designed to offer powerful support for algorithmic composition on their own as in SC3 and in CLM and Nyquist (due to their underlying language, Lisp), it is quite common to see the multi-language paradigm at play in this scenario. Composers will often feel at home using a well-designed general purpose language, such as for instance, Python, in music creation, which would in some way interface with their preferred music programming system. For this combination, the often criticized split between score and orchestra programming (which is also a type of separation of concerns as discussed above) is very welcome. It is very unwieldy to combine programming of events in a separate language with music environment like SC3, unless we are to discard completely its language component.

In fact, the simpler the score language, the easier it is to find means of integrating it within an ecosystem for composition. It allows for substantial customization by composers, which might be find awkward to write directly in any music programming system. For this purpose, Csound not only allows tight integration via its API, but also provides a simple mechanism to allow plugin score processors, written in any language, to be embedded in its code. These only need to output a legal Csound score as ASCII text to the standard output, and can be used to translate any code the composer writes in place of the usual score.

5 CONCLUSIONS

The modern music programming system is, therefore, one that allows users to develop musical applications and compositions of various forms, with full support for a variety of synthesis and processing methods. It needs to be extendable, and allow signal processing algorithms to be expressed in its language, but also capable of integration into an eco-system of applications and environments Support for object-oriented programming approaches is also a basic characteristic of these systems. The open source dimension is an important aspect, which facilitates the study and dissemination of ideas, and one which enabled the significant developments to date.

In the fifty-odd years of their existence, the software packages discussed here have brought an exceptional richness to all areas of Computer Music, from research to composition and performance. We can only hope that the next fifty years of development can bring just as many new possibilities as the ones provided so far.

REFERENCES

[1] M Mathews, "The digital computer as a musical instrument," *Science*, vol. 183, pp. 553–557, 1963.

[2] M Mathews, "An acoustical compiler for music and psychological stimuli," *Bell System Technical Journal*, vol. 40, pp. 553–557, 1961.

[3] J. O. Smith, "Viewpoints on the history of digital synthesis," in *Proc. 1991 Int. Computer Music Conf., Montreal*. 1991, pp. 1–10, Computer Music Association.

[4] T.H. Park, "An interview with max mathews," *Computer Music Journal*, vol. 33, pp. 9–22, 2009.

[5] M. Mathews and J. E. Miller, *MUSIC IV Programmer's Manual*, Bell Telephone Labs, 1964.

[6] J Tenney, "Sound generation by means of a digital computer," *Journal of Music Theory*, vol. 7, pp. 24–70, 1963.

[7] J. K. Randall, "A report from princeton," *Perspectives of New Music*, vol. 3, pp. 84–92, 1965.

[8] H. Howe, "A report from princeton," *Perspectives of New Music*, vol. 4, pp. 68–75, 1966.

[9] A. Roberts, "MUSIC 4BF, an all-FORTRAN music-generating computer program," in *Proceedings of the 17th Annual Meeting of the AES*. 1965, Audio Engineering Society, Preprint 397.

[10] J. Beauchamp, *Introduction to MUSIC 4C*, School of Music, University of Illinois at Urbana-Champaign, 1996.

[11] S. Pope, "Machine tongues xv: Three packages for software sound synthesis," *Computer Music Journal*, vol. 17, no. 2, pp. 23–54, 1993.

[12] M. Mathews, J. E. Miller, F. R. Moore, and J. R. Pierce, *MUSIC IV Programmer's Manual*, Bell Telephone Labs, 1964.

[13] R. Boulanger and V. Lazzarini, Eds., *The Audio Programming Book*, MIT Press, 2010.

[14] F. R. Moore, *Elements of Computer Music*, Prentice-Hall, 1990.

[15] G. Loy, "The CARL system: Premises, history and fate," *Computer Music Journal*, vol. 26, no. 4, pp. 23–54, 2002.

[16] W. Schottstaedt, *Common Lisp Music 4 Manual*, Centre for Computer Research in Music and Acoustics, Stanford University, 1996.

[17] M. Puckette, "Max at seventeen," *Computer Music Journal*, vol. 26, no. 4, pp. 31–43, 2002.

[18] M. Puckette, *The Theory and technique of computer music*, World Scientific Publ., New York, 2007.

[19] D. Zicarelli, "How i learned to love a program that does nothing," *Computer Music Journal*, vol. 26, no. 4, pp. 31–43, 2002.

[20] J. McCartney, "Rethinking the computer music language: Supercollider," *Computer Music Journal*, vol. 26, no. 4, pp. 61–68, 2002.

[21] R. Dannenberg, "Machine tongues xix: Nyquist, a language for composition and sound synthesis," *Computer Music Journal*, vol. 21, no. 3, pp. 50–60, 1997.

[22] B. Vercoe, *Reference manual for the MUSIC 360 language for digital sound synthesis*, Studio for Experimental Music, MIT, 1973.

[23] N. Lefford, "An interview with barry vercoe," *Computer Music Journal*, vol. 23, pp. 9–17, 1999.

[24] B. Vercoe, *MUSIC 11 Reference Manual*, Studio for Experimental Music, MIT, 1981.

[25] B. Vercoe, "Computer system and languages for audio research," *The New World of Digital Audio (Audio Engineering Society Special Edition)*, pp. 245–250, 1983.

[26] B. Vercoe and D. Ellis, "Real-time csound, software synthesis with sensing and control," in *Proc. 1990 Int. Computer Music Conf., Glasgow*. 1990, pp. 209–211, Computer Music Association.

[27] R. Boulanger, Ed., *The Csound Book*, MIT Press, Cambridge, Mass, 2000.

[28] M. Ingalls, "Improving the composer's interface: Recent developments to Csound for the Power Macintosh computer," in *The Csound Book*, R. Boulanger, Ed. 2000, MIT Press.

[29] B. Vercoe, "Extended csound," in *Proc. 1990 Int. Computer Music Conf., Hong Kong*. 1996, pp. 141–142, Computer Music Association.

[30] B. Vercoe, "Audio-pro with multiple dsps and dynamic load distribution," *British Telecom Technology Journal*, pp. 180–186, 2004.

[31] J. ffitch, "On the design of csound 5," in *Proceedings of 4th Linux Audio Developers Conference*, Karlsruhe, Germany, 2006, pp. 79–85.

[32] S. Yi and V. Lazzarini, "Csound for android," in *Proceedings of Linux Audio Developers Conference 2012*, Centre for Computer Research in Music and Acoustics, Stanford Univ., USA, 2012.

[33] V. Lazzarini, "A toolkit for audio and music applications in the xo computer," in *Proc. of the International Computer Music Conference 2008*, Belfast, Northern Ireland, 2008, pp. 62–65.

[34] J. ffitch, "Parallel execution of csound," in *Proceedings of the International Computer Music Conference*, Montreal, Canada, 2009, pp. 16–21.

[35] K. Nygaard and O.J. Dahl, "The development of the SIMULA languages," *ACM SIGPLAN Notices*, vol. 13, pp. 245–272, 1978.

[36] J. Ousterhout, "Scripting: higher-level programming for the 21st century," *IEEE Computer*, vol. 31, no. 3, pp. 23–30, 1998.

[37] M Mathews, F. R. Moore, and J. C. Risset, "Computers and future music," *Science*, vol. 183, pp. 263–268, 1974.

[38] Y. Orlarey, D. Fober, and S. Letz., "Faust: an efficient functional approach to dsp programming," in *New Computational Paradigms for Computer Music*. 2009, Edition Delatour.

[39] R. Dobson, "Computer music books (was re: cmusic)," e-mail to the music-dsp list: http://www.music.columbia.edu/pipermail/music-dsp/1999-June/034758.html, 1999.

[40] R. Stallman, *Free Software, Free Society: Selectd Essays of Richard M. Stallman*, GNU Press, 2010.

[41] J. ffitch, "On the open-source question," personal communication, 2005.

[42] R. Damasevicius and V. Stuikys, "Separation of concerns in multi-language specifications," *Informatica*, vol. 13, no. 3, pp. 255–274, 2002.

[43] 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.

[44] J. C. Risset, "Pitch control and pitch paradoxes demonstrated with computer-synthesized sounds," *Journal of the Acoustical Society of America*, vol. 146, pp. 88, 1969.

Sociedade de Engenharia de Áudio
Artigo de Congresso
Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

La acústica del Teatro Colón de Buenos Aires y su preservación

Gustavo Jorge Basso
Facultad de Bellas Artes - Universidad Nacional de La Plata
Calle 5 Nº 84, La Plata, Prov. de Buenos Aires, CP 1900, Argentina
basso@isis.unlp.edu.ar

RESUMEN

Entre los años 2004 y 2010 se realizaron trabajos en el Teatro Colón de Buenos Aires cuyo objetivo fue preservar su reconocida calidad acústica. En estre trabajo se describen los criterios generales de la intervención y se detallan las mediciones del estado acústico previo de la sala, las mediciones realizadas durante el desarme y las efectuadas durante su posterior rearme escalonado. Los resultados finales demuestran que la calidad acústica del Teatro Colón ha sido preservada.

0. INTRODUCCIÓN

El objeto de esta presentación es describir la metodología diseñada para proteger la reconocida calidad acústica original del Teatro Colón de la Ciudad de Buenos Aires, durante los trabajos de restauración que se desarrollaron entre los años 2004 y 2010.

El Teatro Colón, que se inauguró el 25 de mayo de 1908, se construyó a partir de un proyecto original del arquitecto Francisco Tamburini, continuado luego por los arquitectos Víctor Meano y Jules Dormal [1].

La sala principal cumple con las características generales del teatro italiano clásico en herradura. Tiene una capacidad total de 2.478 localidades, a los que se le pueden agregar unas 500 personas de pie. La planta está bordeada de palcos abiertos, en estilo francés, hasta el tercer piso. A partir de dicho nivel se ubican las localidades de Cazuela, Tertulia, Galería y Paraíso. En total siete niveles por encima de la platea, que presenta una suave pendiente. El foso de la orquesta, enmarcado por los palcos del arco del proscenio, tiene una capacidad para 100 músicos.

La idea de que el Colón posee una acústica extraordinaria circula desde hace mucho tiempo entre músicos y especialistas. En el año año 2000 se publicó un trabajo de Beranek e Hidaka [2], quienes encuestaron a personas que conocían una gran cantidad de teatros de ópera en todo el mundo. Los teatros fueron calificados en una escala de cinco valores: mala, pasable, buena, muy buena y "una de las mejores". El trabajo establece que el Colón no solamente figura entre los mejores teatros de ópera, sino que ocupa un primer lugar bien distanciado del resto.

En otro artículo, publicado en noviembre de 2003, Leo Beranek repite la metodología anterior y clasifica los auditorios para música sinfónica según su calidad acústica [3]. Aquí el Teatro Colón figura en el tercer lugar, luego de dos salas construidas específicamente para ese fin: la *Grosser Musikvereinssaal* de Viena y el *Symphony Hall* de Boston.

El Teatro Colón ocupa, por su acústica, un lugar único entre las salas para música del mundo. Es precisamente esta calidad acústica la que se debía preservar durante los recientes trabajos de restauración y puesta en valor [4] [5].

1. PRESERVACIÓN DE LA CALIDAD ACÚSTICA DEL TEATRO COLÓN

Como se trató de preservar el comportamiento acústico de una sala existente, el primer paso consistió en conocer tan profundamente como fuera posible la sala tal como funcionaba antes del comienzo de los trabajos de restauración. Para realizar esta tarea se emplearon diferentes fuentes de información:

-Planos de arquitectura actualizados.

-Archivos históricos del teatro.

-Estudios acústicos previos.

-Mediciones acústicas de la sala (modelo 1:1).

-Simulaciones digitales de la sala.

-Análisis auditivos realizados por músicos y especialistas.

A partir de este material se decidió sobre la factibilidad de cada una de las acciones previstas durante las tareas de restauración. Se decidió preservar en su totalidad la geometría de la sala, decisión que incluyó las inclinaciones de los pisos de platea y escenario y la estructura de la caja escénica original.

1.1 Metodología

La metodología empleada se puede dividir en las siguientes etapas [6]:

1. Diagnóstico del estado acústico previo al comienzo de las tareas de restauración. Realización de mediciones del campo acústico en base a la norma ISO 3382 [7].
2. Mediciones acústicas de la sala durante su desarme secuenciado
3. Medición en laboratorio de las características acústicas de los componentes y materiales retirados de la sala.
4. Medición en laboratorio de las características acústicas de los componentes y materiales a incorporar en reemplazo de los retirados de la sala.
5. Elaboración de un modelo acústico digital para control del proceso de desarme-armado de la sala.
6. Mediciones acústicas de la sala durante su rearmado secuenciado.
7. Medición final con la sala completamente equipada y puesta en valor.
8. Comparación de las mediciones mencionadas en la etapa 1 (condición inicial) y en la etapa 7 (condición final)

1.2 Mediciones

Se realizaron numerosas mediciones en la sala y mediciones en laboratorio. En lo que sigue se describen someramente dichas mediciones.

Mediciones del campo acústico de la sala antes del desarme

Durante el año 2006 se realizaron mediciones acústicas con el objeto de relevar las características acústicas del teatro de acuerdo con la normativa ISO-3382-1997. Fueron realizadas por el Instituto Argentino de Acústica, Electroacústica y Áreas Vinculadas bajo supervisión de los asesores acústicos responsables de las obras y certificadas por el *Instituto Argentino de Normalización y Certificación* (IRAM).

Se empleó una fuente omnidireccional normalizada ubicada en 4 posiciones diferentes, con el piso del foso bajo y a nivel del piso del escenario. Dicha fuente se excitó con barridos senoidales logarítmicos de 5,5 s de duración. Los micrófonos de medición se ubicaron en 21 posiciones diferentes dentro de la sala. El conjunto de mediciones permitió definir un mapa muy completo del comportamiento acústico del Teatro Colón antes de la intervención.

Figura 1 Medición de la sala en su estado previo

Secuencia de mediciones durante el desarme escalonado de la sala

Se realizaron mediciones parciales bajo norma ISO 3382 -Tiempo de Reverberación (T30) e intensidad relativa (Strength G)- cada vez que se completaba un paso en la secuencia de desarme de la Sala. Dicha secuencia fue la siguiente:

-Retiro de las butacas de platea

-Retiro de las alfombras de platea

-Retiro de las butacas de niveles superiores

-Retiro de las sillas de palcos y banquetas de los antepalcos

-Retiro de los cortinados de palcos

-Retiro de las alfombras de palcos

-Retiro de los cortinados de accesos

-Retiro de las alfombras del Paraíso

Figura 2 Mediciones en la sala durante el proceso de desarme

Mediciones en laboratorio

Las muestras de los elementos retirados de la sala se midieron en el *Laboratorio de Acústica y Luminotecnia* (LAL-CIC) de la Provincia de Buenos Aires. Se empleó en casi todos los casos el método de medición de la absorción acústica en cámara reverberante -Norma IRAM 4065/95 o ISO 354 [8]. En algunos casos también se empleó el método del tubo de ondas estacionarias -tubo de Kundt- para realizar mediciones preliminares.

Los materiales originales y los nuevos materiales de reemplazo fueron medidos en idénticas condiciones y debieron cumplir con las condiciones acústicas impuestas en el pliego del proyecto.

Figura 3 Medición en laboratorio de las butacas de platea

Figura 4 Medición con tubo de Kundt de las alfombras de paraíso

Secuencia de mediciones durante el armado escalonado de la sala

Una vez escogidos los materiales de reemplazo, se los ingresó en la sala en una secuencia inversa a la seguida durante le desarme. La secuencia fue:

-Colocación de la alfombras de Paraíso

-Colocación de Cortinados de accesos

-Colocación de las alfombras de palcos

-Colocación de los cortinados de palcos

-Colocación de las sillas de palcos y banquetas del antepalco

-Colocación de las butacas de los niveles superiores

-Colocación de las alfombras de la platea

-Colocación de las butacas de platea

Luego de la finalización de cada etapa parcial se midió la sala y se cotejaron los resultados con los obtenidos en la etapa correspondiente al desarme. Es decir, se comparó el comportamiento acústico de la sala con los materiales originales con su comportamiento con los materiales de reemplazo.

1.3 Modelo digital

Se elaboró un un modelo acústico digital de la sala a fin de responder a las propuestas realizadas por el equipo de arquitectos y asesores a cargo de la obra. Dicho modelo permitió simular virtualmente las modificaciones geométricas -eliminación o alteración de paredes, revestimientos, etc.- y de comportamiento acústico de los materiales propuestos -absorción, difusión, distribución, etc.- sin necesidad de intervenir la sala real. El modelo fue calibrado cuidadosamente a partir de los resultados de las mediciones 1:1 de la sala y de las mediciones en laboratorio de los textiles. Durante el test de verificación se pidieron reproducir con precisión los parámetros definidos por la norma ISO 3382.

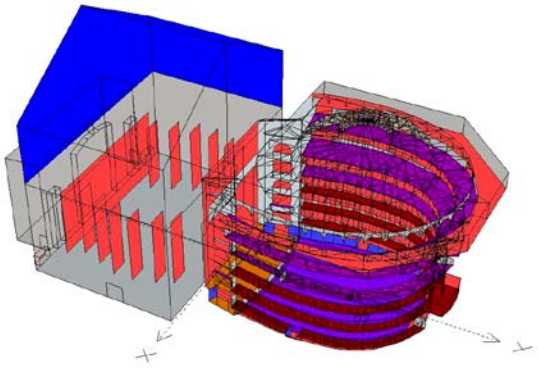


Figura 5 Imagen 3D del modelo digital

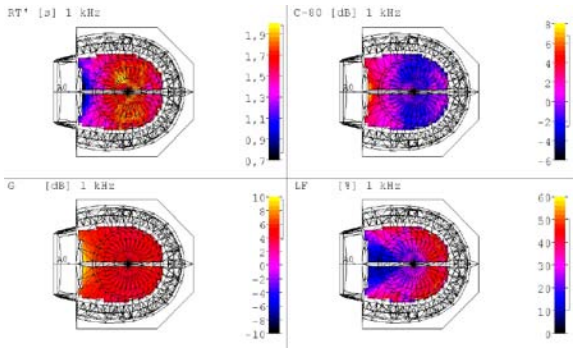


Figura 6 Ejemplo de resultados del modelo digital

A partir de los resultados de las simulaciones se elaboraron alternativas de intervención y se desestimaron numerosas propuestas que hubieran alterado la calidad acústica de la sala. A su vez, el modelo fue ampliamente utilizado para predecir el efecto en la sala del cambio de los textiles y decidir la factibilidad de cada reemplazo [9].

2. RESULTADOS

El 31 de marzo de 2010 se realizó la última serie de mediciones, con la sala en las mismas condiciones que presentaba en el año 2006 antes del comienzo de las obras. La metodología, el instrumental y el equipo de especialistas intervinientes fueron los mismos en todos los casos. Las condiciones ambientales similares -temperatura, humedad relativa, etc.- permitieron realizar una comparación efectiva de los resultados [10].

En el gráfico que sigue se aprecian los valores del parámetro testigo de la acústica de la sala, el *Tiempo de Reverberación T30*, medido en el año 2006 -antes del comienzo de las tareas de puesta en valor- y en marzo de 2010 -al finalizar dichas tareas. La diferencia entre ambas curvas está muy por debajo del límite de error aceptable definido por la asesoría en acústica, indicado en líneas de puntos, que fueron más estrictos que los exigidos en la norma ISO 3382.

Se obtuvieron resultados similares para el resto de los parámetros definidos en la norma ISO 3382 –en el segundo gráfico se comparan los valores de *Claridad C80* obtenidos. En todos los casos las diferencias quedaron dentro de la banda de error definido por las normas y, por lo tanto, por debajo de los límites de la percepción auditiva humana [11].

Figura 7 Promedios de Tiempo de Reverberación T30 medidos antes (2006) y después (2010) de los trabajos de restauración y puesta en valor. Los errores admitidos -líneas de puntos- son los establecidos por la asesoría en acústica.

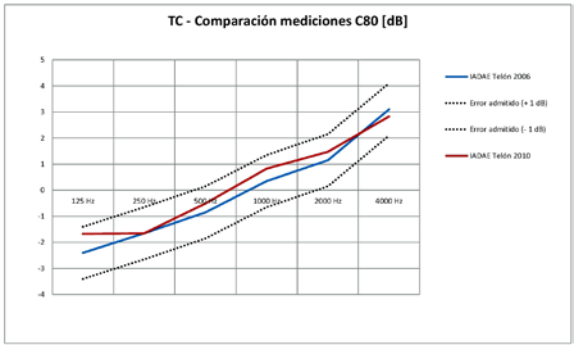


Figura 8 Promedios de Claridad C80 medidos antes (2006) y después (2010) de los trabajos de restauración y puesta en valor. Los errores admitidos -líneas de puntos- son los establecidos por la asesoría en acústica.

Aunque las condiciones de la sala fueron diferentes en cada caso, los valores anteriores son similares y comparables a los obtenidos en 1971 por el Ing. Federico Malvárez (con el telón cortafuego bajo) y en 1998 por el Instituto Takenaka (con cámara de orquesta). Estos últimos son los que empleó el Dr. Leo Beranek en los documentos en los que describe la acústica del Teatro Colón [12].

3. CONCLUSIONES

Como conclusión general, se puede decir que se adoptaron criterios extremadamente prudentes, que incluyeron la preservación de la geometría completa de la sala y un cuidadoso control de los elementos y materiales que reemplazaron, en los mismos lugares, a los originales.

El control del proceso descansó en una batería de mediciones avaladas con la presencia y actuación de instituciones de prestigio, cuyos valores muestran que no se modificaron las características acústicas de la sala durante los trabajos de puesta en valor. Como corolario, se pudo mantener el principio general que establecía que la totalidad de la Intervención acústica debía resultar, técnicamente, reversible.

Figura 9 Función sinfónica en el año 2010, luego de la reapertura

La opinión generalizada de los concurrentes a la sala, luego de una serie de conciertos y óperas ofrecidos desde su reapertura, se podría sintetizar en las palabras del músico Daniel Barenboim luego de su primera actuación en agosto de 2010: “La Acústica del Teatro Colon es sin dudas un milagro“.

4. AGRADECIMIENTOS

La totalidad de las tareas de acústica en el Teatro Colón fueron realizadas junto a los ingenieros Rafael Sánchez Quintana y Alberto Haedo. A ellos, por lo tanto, mi sincero agradecimiento.

5. REFERENCIAS BIBLIOGRÁFICAS

[1] Meano, Víctor. (1892). *El nuevo Teatro Colón*. Cap. Condiciones acústicas, Bs. As. Argentina. Ed. Kraft.

[2] Hidaka, T. y Beranek, L. (2000). *Objective and subjective evaluations of twenty-three operahouses in Europe, Japan, and the Americas*, Journal of the Acoustical Society of America, 107 (1), 368-383.

[3] Beranek, Leo (2003). *Subjective Rank-orderings and Acoustical Measurements for Fifty-Eight Concert Halls*, Acta Acustica, vol. 89, 494-508.

[4] Basso, G. y Canalis, I. (2008). *Evaluación de la calidad acústica del Teatro Colón de Buenos Aires*, en Actas del VI Congreso Iberoamericano de Acústica FIA 2008, Buenos Aires.
[5] Canalis, Ianina (2007). *Evaluación de la calidad acústica del Teatro Colón de Buenos Aires*, UNLa, Lanús, Argentina.
[6] Teatro Colón. *Pliego de Especificaciones Técnicas (PET) -Acústica* - Gobierno de la ciudad de buenos aires - Teatro Colón.
[7] Norma ISO 3382-1997.
[8] Norma IRAM 4065/95.
[9] Basso, G., Haedo, A. y Sánchez Quintana, R. (2009). *Preservación de la calidad acústica del Teatro Colón de Buenos Aires*, en la Revista de Acústica, Madrid, España, Vol. 40. Nos 3 y 4.
[10] Basso, G., Haedo, A. y Sánchez Quintana, R. (2010). *Preservation of the Acoustical Quality of the Teatro Colón of Buenos Aires*, en Actas del 2nd Pan-American/Iberian Meeting on Acoustics, Cancún, México.
[11] Basso, Gustavo (2006). *Percepción Auditiva*, Ed. de la UNQ, Bernal, Bs. As.
[12] Beranek, Leo (1996). *Concert Halls and opera houses: How they sound*, Acoustical Society of America, New York.

BASSO, G. J. La acústica del Teatro Colón de Buenos Aires y su preservación. In: CONGRESSO DE ENGENHARIA DE ÁUDIO AES BRASIL, 10., 2012, São Paulo, *Anais*... Rio de Janeiro: Sociedade de Engenharia de Áudio, 2012. p.13-17.

ARTIGOS CONVIDADOS / *INVITED PAPERS*

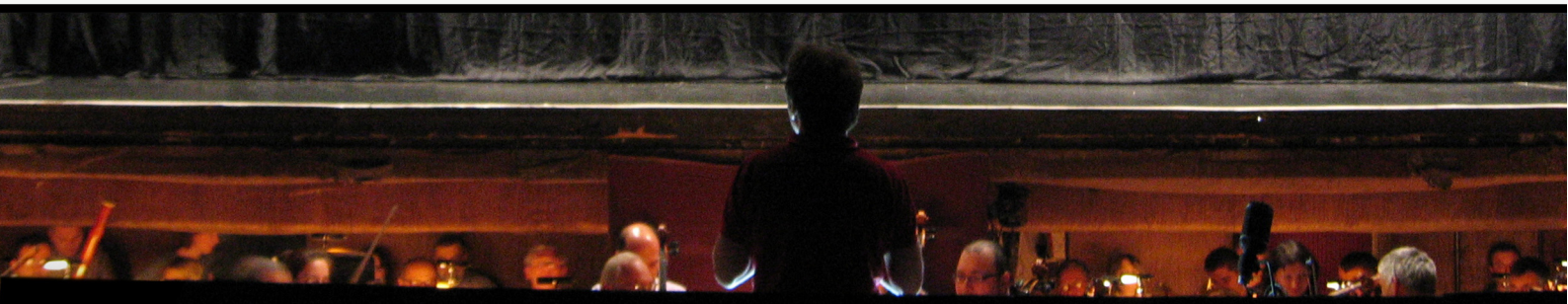


Foto: SXC

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Acústica arquitetônica para sala de gravação sonora

Elcione Moraes
Faculdade de Arquitetura e Urbanismo da Universidade Federal do Pará
Belém, Pará, Brasil. 66033-815
E-mail: elcione@ufpa.br

RESUMO
Este artigo apresenta os parâmetros e diretrizes de elaboração de projeto acústico para salas de gravação sonora. Descrevem-se soluções para o ideal isolamento e condicionamentos de diversos tipos de salas de gravação para música, televisão, cinema e radio. O objetivo do trabalho é mostra de forma direta e cautelosa os caminhos a serem percorridos para lograr a geração de um estúdio acusticamente eficaz. A partir da fundamentação teórica foi possível unir conhecimento específico e experiência prática em um documento teórico que ilustra exemplos de importantes estúdios para gravação sonora no Brasil e em outros países.

0. INTRODUÇÃO

O desenho de um estúdio de gravação está baseado na necessidade de um ambiente com qualidade acústica de excelência, assim como de sugerir condições de competitividade e elegância ambiental no momento da gravação e edição da produção seja musical ou locutora.

Aplicando os princípios básicos da Acústica Arquitetônica e da Sonorização se determina a geometria e estrutura das salas de gravação. Os cálculos dos parâmetros acústicos que definem a qualidade sonora de um recinto consideram de maneira geral o isolamento dos ruídos e vibrações desde o exterior e os fatores interiores influentes. Mediante as análises dos modos próprios de ressonância, de níveis de pressão sonora e do controle do ruído de fundo se estima o tempo de reverberação, as dimensões ótimas do recinto e a inteligibilidade da palavra. Este artigo descreve de modo geral os princípios técnicos e formais que levam ao desenho ideal de sala de gravação para a música e a palavra falada.

1. ASPECTOS GERAIS

Como norma geral deve-se seguir uma serie de passos no desenho de qualquer tipo de sala de gravação (música, rádio, Tv, etc.) no qual a acústica segue um papel importante: 1- Avaliação de todas as fontes de ruído e vibração nas proximidades (ruído de tráfego, aviões, trens, industriais, etc.); 2- A estrutura da edificação deve levar em conta prioritariamente o isolamento do ruído aéreo e estrutural (vibrações); 3- A adequada setorização do recinto é alcançada com a separação das zonas comuns com as áreas técnicas do estúdio (escritório, depósitos, serviços, etc.); 4- No caso de existir mais de um estúdio no mesmo edifício esses devem estar separados de modo a evitar paredes comuns entre eles. A utilização de superfícies flutuantes (teto e piso) contribui para a eficiência do isolamento das superfícies; 5- O consciente pré-dimensionamento dos espaços que compõem o estúdio evitará a perda ou carência de espaços para as instalações dos serviços de ar condicionado, eletricidade, luz, etc.

A dimensão e forma de um estúdio produzem diferentes efeitos sobre suas propriedades acústicas. A escolha incorreta das dimensões de um estúdio pode influir não só no uso ineficaz de seu volume, como também, na piora da uniformidade do campo sonoro, que pode, ainda, ser afetada pelo fato de que o número de frequência característica por unidade de largura de banda é significativamente menor em estúdios pequenos que em estúdios grandes.

Alguns cuidados quanto as dimensões do recinto podem evitar perda da qualidade acústica dos estúdios. A não existência de superfícies (paredes, piso e teto) de igual tamanho ou múltiplas entre si pode evitar a formação de modos (ondas estacionárias) no local. Assim como, a máxima e mínima relação de dimensão entre as superfície, numa escala limite de 3 para 1. Sempre que possível as dimensões do estúdio deve guarda certas proporções entre si.

De acordo com as necessidades assim como com a forma da planta do edifício, se consideram primeiramente o volume, forma e dimensões do local ou locais que correspondam ao projeto, levando em conta as considerações acústicas e de uso particularmente para cada tipo de recinto. O volume será definido em função do uso

que será dado ao estúdio e poderá depender do número de pessoas. Os recintos com forma de paralelepípedo são desenhados quando o tempo de reverberação for muito baixo e se a forma não tiver influência na distribuição da energia sonora (mais usual para escutar a palavra). Os grandes estúdios, seja para música ou palavra, têm desenho trapezoidal geralmente e com as superfícies laterais, teto e piso não paralelas (Fig.1). A escolha adequada das dimensões do estúdio permitir que a resposta em frequência do mesmo seja o mais neutra possível, sendo o número de modos próprios muito elevados e sua distribuição regular, caso contrário, algumas frequências serão mais realçadas que outras, colorindo, portanto, o recinto e o sinal sonoro.

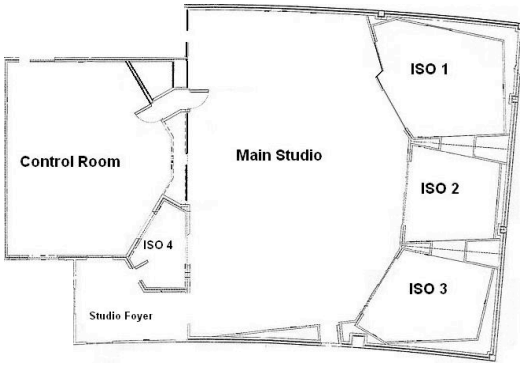


Figura 1 Plata baixa do Conservatory Recital Studio, Yong Siew Toh Conservatory of Music em Singapura. "<http://www.nus.edu.sg/music/>"

Evidente que os aspectos acústicos não são os únicos a serem considerados para a criação de recintos adequados, os condicionantes econômicos, adoção dos serviços, higienização, acessibilidade, são alguns dos fatores que deverão ser pensando conjuntamente com o projeto acústico. A qualidade da produção sonora, isto é, o ambiente acústico adequado para uma determinada gravação depende das exigência de uso do recinto, por exemplo, no estúdio, nas cabines de locução, nas salas de controle a escuta é mais crítica que nas sala de programação, sala de edição de vídeos, etc., por tanto, as exigência de soluções técnicas vai depender diretamente do tipo de uso do recinto.

Os condicionantes térmicos e seus sistemas de climatização, assim como os de iluminação são muito importantes devido ao incremento sonoro que oferecem ao recinto, devido a isso, é necessário o cuidado especial a fim de não permitir a introdução excessiva do ruído aéreo ao recinto.

2. CRITÉRIOS DE ISOLAMENTO E CONDICIONAMENTO ACÚSTICO

Os problemas mais importantes que se enfrentam ao tratar o desenho de uma sala de gravação são principalmente os referentes ao isolamento e condicionamento acústico.

2.1 Isolamento

A melhor defesa contra o ruído é a execução de um cuidadoso projeto de obra posto que qualquer intervenção numa obra já pronta traz inúmeros inconvenientes, além de elevar o custo da execução, portanto, o planejamento da obra a partir do projeto acústico detalhado reduz o risco de perda na qualidade acústico do recinto.

O principal objetivo consiste em obter um bom isolamento, tanto contra o ruído aéreo como contra o ruído estrutural, entre o estúdio de gravação e os locais adjacentes, para tal é necessário considerar no momento do desenho as leis fundamentais de isolamento acústico, consideram os mais e sistemas acústicos que se empregam para a construção de paredes, portas, visores, divisórias, teto, piso e mobiliário (quando for o caso), e os serviços básicos, tais como potencia elétrica e ar condicionado.

O isolamento é, portanto, o conjunto de medidas arquitetônicas que consiste em impedir a propagação de um sinal sonoro, mediante diferentes obstáculos refletores. Essa transmissão faz-se por vias distintas: pela passagem direta via parede; pela passagem indireta via flanqueamento, pelas estruturas vizinhas, pilares, lajes e/ou paredes confluentes; ou pelos vão (janelas, portas aberturas, dutos de ar, tubos de água, eletrodutos, ou por entreforro e entrepiso.

No caso da passagem direta pelas paredes, pode-se conseguir um considerável isolamento mediante paredes rígidas e pesadas, que refletem o som e não o absorva, podendo ser calculados a partir da Lei de Massa. Teoricamente sabe-se que o isolamento aumente em 6 dB ao duplicar-se a massa da superfície e manter a frequência fixa. De forma análoga, para um valor determinado de massa o isolamento aumenta 6 dB cada vez que se duplica a frequência.

Por outro lado também se pode amortecer o som por meio da absorção do mesmo. É muito difícil conseguir uma estrutura ou sistema que reflete totalmente o som sem que o absorva ainda que seja uma pequena parte dele, portanto, a perda por transmissão consiste na capacidade de uma superfície absorver a energia sonora.

Os materiais mais comumente empregados para isolar o ruído aéreo são: tijolo de diferentes tipos (com furos ou maciços); concreto; gesso; gesso acartonado; e fibras (lã de rocha, lã de vidro) de diferentes densidades utilizadas como "recheio" nas paredes duplas (tipo sanduíches).

Nos casos das portas, janelas ou visores entre a sala de gravação e de controle essas deverão ser convenientemente isoladas. O mercado nacional disponibiliza uma gama significativa de janelas, portas e visores acústicos que atendem às mais diversas exigência de isolamento dos vãos. É importante ressaltar que a execução em obra desses elementos isolantes é de difícil ajuste e pouca eficiência, de modo que para garantir o resultado calculado deve-se optar pelo uso dos elementos produzidos industrialmente.

Além da transmissão aérea, também o impacto e as vibrações precisam ser isolados, para isso é necessário impedir a propagação da energia sonora através das estruturas sólidas do recinto.

O impacto é geralmente originado por pisadas, quedas de objetos ou choques sobre a superfície do piso e podem ser ouvidos claramente pelos ocupantes internos e circunvizinhos. Isso ocorre devido a laje de piso ser um irradiador de energia sonora em ampla faixa de frequências, m função do movimento vibratório induzido pela excitação pontual. (BISTAFA, 2006).

O aumento da espessura da laje reduz o nível sonoro no recinto inferior, porém o elevado custo e o comprometimento no sistema estrutural do edifício torna essa ação inviável. A solução mais indicada consiste na utilização de material resiliente isolador (coxins de borracha, placas contínuas de lã de vidra ou lã de rocha, poliestireno expandido, cortiça, etc.) entre a laje estrutural e o contrapiso, formando o chamado piso flutuante. É importante para o bom desempenho do piso flutuante que a laje esteja isolada também dos fechamentos verticais (paredes).

Os princípios operacionais para o controle das vibrações não diferem substancialmente do ruído de impacto, embora seja mais frequente nesse caso que predominem as frequências puras em seu espectro, permitindo soluções mais especificas por não ter que estender sua eficácia a toda banda de frequência, como no caso do isolamento de impacto. (COLINA; MORENO, 2005)

As vibrações se amortecem muito pouco ao propagar-se em superfícies sólidas, motivo pelo qual seu isolamento implica, geralmente, na interrupção do sólido, intercalando capas elásticas. É frequênte, também, em sólidos laminados, adicionar capas de materiais pesados com

grande viscosidade e módulo de cisalhamento muito baixo. Pode-se lograr, ainda, um bom resultado implantando sistemas que vibrem em concordância de fase, ou seja, com elementos de dimensões pequenas com relação ao comprimento de onda do som. A frequência limite entre o som e a vibração de um corpo sólido gira em torno dos 100 Hz.

2.2 Condicionamento acústico

Tão importante quanto o próprio processo de desenho da sala de gravação é a comprovação e avaliação dos parâmetros acústicos do recinto, tarefa essa que deve ser realizada tanto durante a fase de construção como ao final da mesma. Uma vez coletos todos os elementos pode-se particularizar às características próprias do recinto. Deve-se assegurar um adequado condicionamento acústico da sala de gravação e de controle do som.

Para isso é necessário a redução (controle) do ruído de fundo; um tempo ótimo de reverberação, adequada inteligibilidade do som e um grau de difusão uniforme em todos os pontos das salas, considerando que suas propriedades acústicas estão diretamente relacionadas as reflexões das ondas sonoras em todas as superfícies limites (paredes, piso e teto), de modo a que o tempo de reverberação seja idôneo em cada caso.

O ruído de fundo pode ter diversas origens é aquele que se encontra no interior do recinto quando não há atividade sendo desenvolvida no recinto, e pode ter diversos tipos de origem, instalação de ar condicionado, elétrica/hidráulica, ou mesmo o ruído proveniente do exterior da sala.

A escolha de um tempo ótimo de reverberação é muito variável, já que depende fundamentalmente do volume do recinto, da absorção presente no mesmo e da frequência da fonte sonora. A diferença entre tempos de reverberação (TR) a baixas e altas frequências são em algumas salas de ordem de 5 a 6 segundos, e as diferença entre os TR ótimos para diferentes tipos de música, estão entre 0,5 e 2 segundos. O problema está quando se pretende um tempo de reverberação correto tanto para palavra como para música, e ainda para diferentes tipos de músicas, na maior parte dos casos a solução que se adota é estabelecer um compromisso entre os tempos de reverberação esperado. Em alguns caso é possível variar o tempo de reverberação em função do tipo de uso da sala ou corrigi-lo com a adoção de meios artificiais.

A importância da absorção acústica se manifesta, sobretudo, no controle da reverberação do recinto, magnitude que condiciona a inteligibilidade da palavra e a qualidade da escuta da música. Para uma absorção variável nos estúdios, os materiais absorventes com coeficiente de absorção muito diferentes podem ser colocados sobre uma superfície móvel, e podem estar posicionadas paralelamente as paredes de fechamento recinto e mover-se sobre elas, ou retiradas do estúdio caso seja necessário. A adoção de painéis ou colunas giratórias revestidas em cada fase com materiais sonoros de diferentes características é uma estratégia muito válida para o aumento ou diminuição do tempo de reverberação com base unicamente na absorção sonora. Os ressonadores contribuem para o tempo ótimo de reverberação de frequências muito concretas.

A inteligibilidade dos sons é um parâmetro de quantificação subjetivo que avalia em percentagem, ou numa escala de 0 a 1, a quantidade de sons inteligíveis relativamente à totalidade de sons emitidos. A quantificação rigorosa deste parâmetro pode ser conseguida de forma experimental, recorrendo-se à medição do índice RASTI (Índice de Transmissão Rápida de Linguagem), embora os valores de inteligibilidade variem entre 0 (inteligibilidade nula) e 1 (inteligibilidade perfeita) estes valores nunca são atingidos, para a maioria das situações considera-se uma inteligibilidade aceitável entre 0,5 e 0,64, boa entre 0,65 e 0,85 e excelente acima de 0,85. (MATEUS, 2008)

Um campo difuso é o mesmo que considerar um campo reverberante, ou seja, o ambiente tem a mesma pressão sonora em todos os pontos da sala, e radiação sonora idêntica em todas as direções.

Figura 2 Paredes difusoras do estúdio para gravação musical do Crystal Clear Studios em Dallas –TX (EUA). “<http://www.crystalclearstudios.com/recordingstudio.php>”

Um dos problemas mais importante nos locais para gravação sonora é o sistema de ar condicionado, como processo que tende ao controle simultâneo, dentro de um ambiente delimitado, da pureza, da umidade, temperatura e circulação do ar.

O desenho acústico do sistema de ar condicionado em sala para gravação sonora e controle, tem que ser realizado com muita cautela, de modo a que seu incremento ao nível de ruído ambiental seja mínimo. Para tal devem ser cuidados os desenhos dos condutores principais e secundários, grade de ventilação, difusores, etc., para que a velocidade de deslocamento do ar seja a adequada.

3. ESTÚDIOS DE GRAVAÇÃO DE MÚSICA

Os estúdios de gravação musical adquiriram uma grande importância no século XX como consequência do desenvolvimento do mundo da eletricidade e da eletrônica, a partir do surgimento dos modernos sistemas eletrônicos que permitem gravações musicais (por meio de mixagem, sintetizadores, etc.) vem evoluindo e divulgando novos conceitos musicais. Graças às técnicas de gravação musical é possível recordar, estudar e analisar as grandes orquestras e músicos por diferentes gerações.

Igualmente importante como a interpretação musical em um estúdio de gravação é a qualidade acústica da sala para que a fidelidade da gravação seja lograda. Para se produzir uma gravação de qualidade baixa ou média pode-se recorrer a qualquer tipo de estúdio (caseiro ou não) convenientemente preparado, entretanto, a escassez das propriedades acústicas devidamente pensadas será notada com pouco esforço ou conhecimento técnico.

Os cuidados gerais exigidos para um eficiente estúdio de gravação musical devem assegurar que:

1. Todo o recinto (estúdio-controle) estará montado sobre uma estrutura própria flutuante;
2. Não haverá superfícies paralelas, incluindo o par piso-teto;
3. Haverá controles digitais com tendência a incorporação de computadores;
4. Os microfones serão de delicada sensibilidade e usos específicos;
5. Os painéis refletores e absorvedores sejam móveis;
6. Os espaços pensados para bateria e percussão serão afastados e com adoção de “aquários”;
7. Serão dimensionados dutos isolantes para os condutores de ar condicionado e cabos; e
8. Operador técnico bem qualificado.

Devido a que o tipo de trabalhos que se realizada, assim como o tamanho e a natureza dos grupos musicais, é muito variado, para garantir umas condições acústicas ótimas,

para a gravação de música, planteia-se um recinto em que possa variar o tempo de reverberação. Assim como, as exigências acústicas dos estúdios para gravação de música através de um sistema eletroacústico não são sempre as mesmas, embora a condições ótimas em grandes estúdios (volume > 2000 m^3) são aquelas em que o tempo de reverberação independe do volume.

Figura 3 Um dos quatro estúdios do Centro Cultural de Melbourne, Austrália. Foto da autora, 2010.

Entretanto, ao determinar o valor apropriado do tempo de reverberação para diferentes tipos de música, deve-se adotar uma estratégia que responda às exigências do alto tempo de reverberação da música sinfônica moderna como de baixo valor do TR solicitado para música romântica.

Em estúdios musicais de tamanho reduzido pode-se supor com bastante probabilidade que o tempo de reverberação dependa do volume.

Anos atrás se utilizam vários tipos de estúdios para gravação musical, dependendo do tipo de música a qual se desejasse gravar. Hoje, é mais comum, prático e menos custoso, criar uma única sala de gravação para todos os estilos musicais, mesmo que para isso tenha-se que construir um grande estúdio com subdivisões internas. A maior dificuldade nesse processo é ajustar o tempo de reverberação a cada tipo de música e número de músicos ou instrumentos, sem que isso acarrete em perda da eficiência do isolamento sonoro da sala. Em alguns casos há de recorrer-se aos reverberadores artificias.

A natureza da variação do tempo ótimo de reverberação com a frequência baseia-se em proposições teóricas geralmente confirmadas na prática. Por exemplo, resulta conveniente ter um realce no tempo ótimo de reverberação nas baixas frequências (até 20% ou 30% na frequência de 125Hz), frente as frequências médias. Também é fato que existe uma caída notável do tempo ótimo de reverberação nas altas frequências, devido a absorção do som pelo ar, que aumenta quase se eleva a frequência do som e o tamanho do estúdio. (RECUERO, 1993).

O volume de um estúdio de gravação musical deve ser escolhido considerando primeiramente a necessidade de se criar umas condições acústicas ótimas. É importante, desde esse ponto de vista, que se mantenha no estúdio uma sonoridade suficiente, guardando uma relação específica entre a potência sonora e o volume do estúdio.

Para manter equilibrado o balance entre os níveis de pressão sonoras dos diferentes instrumentos musicais, é conveniente que a intensidade sonora (volume) não seja muito forte, já que isso pode vir a desorientar o operador e o diretor da gravação. Como o nível de pressão sonora da musica está relacionado com numero de instrumentos ou músicos, o volume do recinto estará, também, variando com esse número.

A forma dos estúdios para gravação musical tem grande importância para garantir o bom desempenho do recinto. Devido a exigência de um tempo de reverberação relativamente alto, as muitas reflexões dentro do campo sonoro, inclusive entre as paredes paralelas, podem causar um decaimento lento do sinal sonoro, e conduzir a uma maior desigualdade do campo sonoro.

A utilização de superfícies não paralelas pode reduzir o número de ondas axiais, que tem uma natureza mais ordenada, contribuindo, também, a desigualdade do campo sonoro.

Assim, a forma menos ariscada para um estúdio de gravação musical é a trapezoidal e não retangulares, com ângulo de desviação que formam as superfícies do recinto é aproximadamente 10º. Por estas razões as superfícies laterais e algumas vezes o teto de alguns estúdios para gravação musical estão inclinados a um ângulo de cerca de 5º. Em vista da dificuldade de desenho de grandes estúdios com superfícies inclinadas, essas são frequentemente irregulares, ou ainda, têm uma forma não regular somente nas áreas onde se posicionam os músicos.

Ou fator importante para a gravação de música é a difusão do campo sonoro que deve ser a mais alta possível. Esta difusão pode ser lograda com a distribuição regular dos materiais absorventes dentro do estúdio, ou ainda, com o uso de boas superfícies dispersoras, geralmente apresentam formas convexas ou estão dispostas de forma que os elementos formam ângulos entre si. Para que isso ocorra, as superfícies de dispersão e absorção devem ter dimensões comparáveis com o comprimento da onda sonora e estar disposta de maneira alternada ao longo de toda a parede da sala de gravação. (COX; D'ANTONIO, 2004).

Figuras 4 e 5 Estúdios com forma irregular e paredes difusoras, no Fenix Recording Studios em Varnhem, Suécia. "<http://fenixrecording.com/studios>"

As gravações estereofônicas diretas, isto é, as que não são gravados os instrumentos em separado e logo mesclados, agregam uma problemática adicional, já que a naturalidade do som implica em uma difusão e certo tempo de reverberação, enquanto que uma boa localização espacial se consegue com mínima difusão e baixa reverberação. Uma forma prática de minimizar o problema é gravar com um tempo de reverberação baixo e logo aumenta-lo com meios eletrônicos.

3.1 Grandes estúdios para gravação de orquestras

Os estúdios projetados para pequenas e grandes orquestras, assim como para coral vocal, possuem volume

em torno de 2000 m^3 e o tempo de reverberação pode variar entre 1,2 s e 1,5 s, dependendo dos materiais absorventes utilizados. Pode haver mudanças do tempo de reverberação para médias e altas frequências. O nível de pressão sonora máximo não deve ultrapassar os 95 dB. (COX; D'ANTONO, 2004)

Os estúdios com volume em torno de 5000 m^3 empregam-se para as grandes orquestras, podendo ser ajustada para as orquestras pequenas e inclusive para um solista. Nesse caso o volume da sala é determinado em função de duas situações, os momentos de ensaio da orquestra (volume de 50 m^3 por músico), e com a presença do público (volume de 15 m^3 por pessoa). O tempo de reverberação pode variar de 2,4 s a 1,2 s, sem o com público respectivamente, e com diferentes materiais absorventes, indicam os mesmo autores.

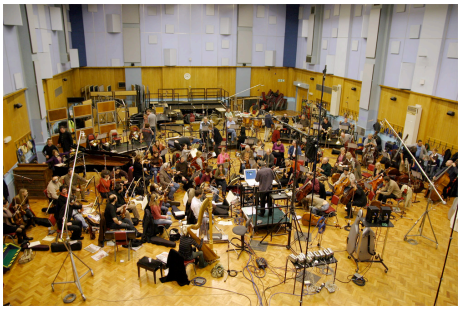
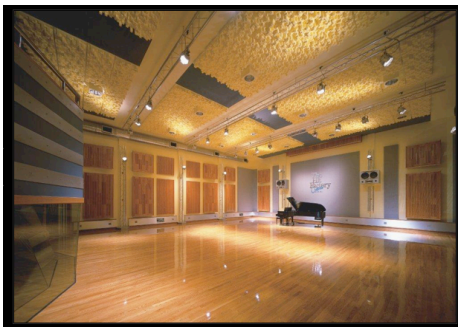


Figura 6 Estúdio para gravação de orquestra do famoso Abbey Road Studios, Londres.
"<http://www.nus.edu.sg/music/>"

No caso de música de câmera e pequenas orquestras, o volume desses é de uns 1000 m^3 e seu tempo de reverberação pequeno, em torno de 0,9 s com uma orquestra de câmara presente, e de 1,2 s quando só estiverem os solistas.

Quando seja necessário determinadas zonas do estúdio com uma absorção sonora específica, para separa certos instrumentos musicais individuais ou um grupo de instrumentos de outros, por exemplo, pode usar cortinas, bambolinas, grandes e móveis, etc., sem que haja a necessidade de construir áreas exclusivas para essas atividades.



Figuras 7 Estúdios adequado para música de câmera. Hit Factory Studios, Nova Yoork.
"<http://fenixrecording.com/studios>"

3.2 Estúdios para gravação de rock e pop

Normalmente esses estúdios são pequenos e carecem de uma acústica natural. A particularidade dos estúdios para gravação de rock ou música pop é o utilização das cabines parcialmente fechadas, com conjuntos de microfones, para gravação de instrumentos de percussão, teclado, etc. É conveniente evitara as reflexões para que possa ouvir com maior nitidez os diferentes instrumentos musicais. O isolamento das portas e visores deve ter uma atenção especial dado que os níveis de pressão sonora são muito elevados e não devem diminuir a percepção dos operadores. O visor e porta podem ser um mesmo elemento, ou seja, o visor pode abrir-se e permitir a passagem ao estúdio, e facilita a gravação direta de instrumentos eletrônicos que estejam situados próximos à ela. O controle pode ser ampliado ao abrir a porta de comunicação com o estúdio.

Para se proporcionar uma melhor viveza ao local se pode permite algumas reflexões individuais procedentes de determinadas superfícies que irão realçar o nível sonoro do instrumento.

Cuidados especiais devem ser considerados quanto ao posicionamento dos instrumentos no estúdio. Determinados instrumentos situados perto da porta podem ser afetados e gerar o fenômeno da realimentação, isso devido a que o monitoramento no controle se realiza com elevado nível de pressão sonora (em torno de 100 dB).

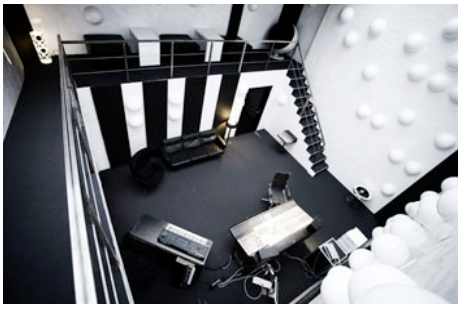


Figura 8 Delta Lab Recording Studios o mais moderno e versátil estúdio da Dinamarca .
"<http://www.deltalab.dk>

4. CARACTERÍSTICAS ACÚSTICAS DE ESTÚDIOS DE RADIO E TELEVISÃO

Os grandes estúdios de televisão embargam um complexo emaranhado de luzes, movimento de câmara e microfone, organização de atores, realizadores, somados as exigências acústicas específicas. Um estúdio de TV pode ser utilizado por diferentes grupos de atores, realizadores, decoradores, técnicos de diferentes áreas e apoiadores, simultaneamente. O projeto acústico tem que dar conta de ajustar todas essas manifestações sem interferir na produção ou realização do programa a ser gravado.

Tanto o produtor como o operador de som deve ver e ouvir desde suas respectivas cabines e através dos monitores de televisão, na maioria das vezes em circuito fechado, todos os "atores" envolvidos no processo de gravação. Em alguns casso, em se tratando de estúdio não muito grandes, é possível permitir que as cabines de controle permaneçam abertas para a comunicação direta entre todos.

Estes estúdios podem ser classificados como pequenos, médios e grandes. Os primeiro normalmente estão designados a trabalhos que não requerem movimento de câmara. Os últimos são destinados a trabalhos que reúnem justo o contrário, grande movimento das câmaras e microfones, e diferentes tipos de cenários.

Tanto a mobilidade das câmaras como do sistema de manejo e controle dos focos implicam na necessidade de amplos espaços com alturas consideráveis nas zonas de trabalho. O desperdício de espaço e necessidade de equipamentos específicos fazem com que, por motivos econômicos, se tente aproveita-los ao máximo, adotam desenhos que sejam aptos a diferentes tipos de programas.

Desde o ponto de vista puramente econômico, os estúdios destinados a um único propósito são absolutamente impraticáveis. No entanto, na grande maioria dos casos a tempo de reverberação deve ser muito menor que no caso de um estúdio para gravação de musica

com as mesmas dimensões. Este fato ocorre devido a que muitos programas devem manter os microfones fora do alcance das câmaras, por motivo estético. Isto significa que estarão muito longe dos atores e implica na necessidade de ter um tempo de reverberação muito baixo para que a relação som direto – som refletido se mantenha em uma margem aceitável.

Ao buscar a generalidade funcional, os estúdios deverão admitir o uso de qualquer tipo de "cenário acústico". Se o estúdio for muito reverberante as ondas refletidas em diferentes superfícies romperiam qualquer possibilidade de intimidade no local. Assim, o estúdio deve ser tratado com intensidade uniforme e diminuir as reflexões a um valor suficiente.

O movimento das câmaras, microfones, luzes e cenários, geram um ruído considerável. Ainda que esses sons se produzam longe dos microfones, podem ser captados por eles se a reverberação do local amplificar sua intensidade.

O solo deve permitir o movimento das rodas das câmaras, cenários e operadores, por tanto, deverá ser liso, resistente e não poroso, e consequentemente bons refletores sonoros.

Os diferentes tipos de equipamentos ocuparam cerca de ¼ da superfície das paredes, o resto dessas superfícies deverão ser tratadas com absorvedores e/ou difusores sonoros.

Em estúdios de gravação da palavra falada, para a boa compreensão e identificação da pessoa que fala, a transmissão e recepção do sinal sonoro oral através de um sistema de comunicação, deve ser medido mediante o índice de "inteligibilidade da fala", dado por SIL (Speech Interference Level) e PSIL (Preferred Speech Interference Level). A inteligibilidade depende do tempo de reverberação do recinto, que deve ser baixo para evitar interferências e aumentar a clareza, caso contrario, o microfone receberá além do sinal direto as reflexões procedentes do sinal anterior, gerando assim, queda da inteligibilidade.

É importante que o nível do sinal falado seja pelo menos 50 dB, e que o nível dos sons secundários esteja 20 ou 30 dB por debaixo do nível do sinal falado

Com o fim de assegurar a igualdade do campo sonoro em um pequeno estúdio para a palavra, assim como para fazer desaparecer o efeito de cansaço no locutor, os estúdios para locução devem ter um volume não inferior a 20 m^3. Geralmente a forma mais adequada é a paralelepipédica retangular, pois não introduz distorções consideráveis por seu elevado amortecimento.

Já que o volume desses estúdios é pequeno e necessitam criar um campo sonoro uniforme, é preciso usar materiais absorventes eficazes e de igual dimensão e igualmente distribuído por todo o recinto, com o fim de criar os tempos de reverberação desejados. Por isso, a forma desse tipo de estúdio tem pouco influencia na sua eficácia acústica.

Figura 9 Estúdio da Rádio Nacional da Espanha (RNE), Madri. http://www.aspa.net/webaspa/imagenes/instalaciones/gran_RNE_madrid_6.jpg\

Salas de gravação e filmagem sincronizada, como nos estúdios de televisão e cinema, o som é gravado simultaneamente com a imagem. No caso de filmagem sincronizada o som que se escuta na sala de projeção deve corresponder acusticamente com a imagem que se vê na tela, de modo que com a mudança do volume do recinto que está sendo visualizado na tela, na prática mudaria também o tempo de reverberação do local e, portanto, muda também a sonorização. Por esses motivos é necessário criar nos estúdios de gravação condições próximas as que se vê para o acompanhamento das palavras. Isto é conseguido variando o posicionamento do grupo de microfones com relação a fonte sonora, para o controle das característica acústicas dentro dos limites reduzidos, com equipamentos especiais, variar a reverberação até obter o valor desejado.

Pela conveniência das câmaras de televisão e de cinema, estes estúdios têm forma retangular. Esta forma não tem um feito real na qualidade da gravação sonora, pois os estúdios estão fortemente amortecidos. Além disso, os equipamentos técnicos ao estabelecimento do campo sonoro difuso, que compensa, até certo ponto, o efeito das condições acústicas que resultam da forma retangular do recinto.

As paredes e tetos desses estúdios são revertidos com materiais absorventes muito eficazes, somente na parte inferior das paredes, cerca de 3 metros, é recomendado o uso de superfícies rígidas (absorventes sonoros ressonantes, painéis ressonadores com enchimento de fibra mineral), que garantem a não deformar das paredes com a mudança constante de cenários.

Não se recomenda deixar amplas zonas da superfície total sem tratar já que podem produzir ecos intoleráveis. Em um estúdio de grande dimensão, uma pequena superfície refletantes, por exemplo, pode produzir um eco muito notável em comparação com a pequena reverberação do resto do estúdio.

É necessário nos estúdios de televisão e cinema que o isolamento acústico seja muito cuidadoso, pois os microfones, que necessitam estar posicionados afastados do ator por questão estética, são muito sensíveis ao ruído ambiental. O isolamento deve produzir uma apreciável redução do nível de ruído que recebe o microfone.

Os estúdios de gravação de televisão destinados a obras teatrais ou outros tipos de programa de grande logística, não diferem dos estúdios de filmagem cinematográfica, quanto ao seu uso e as condições acústicas exigidas.

Figura 10 Estúdio de TV do Canal Público do Panamá. "<http://www.panoramaaudiovisual.com/wp-content/uploads/2010/12/EUROCOM_PANAMA_VIRTUAL.jpg>

Em vista que a gravação sonora não tem lugar nos mesmos estúdios, mas sim em grandes plataformas montadas sobre eles, há de considerar a possibilidade de estabelecer as condições acústicas precisa mediante o uso das mencionadas plataformas.

Esta plataforma montada no estúdio e ocupando parte dele é em essência um recinto conectado acusticamente com o resto do estúdio. As únicas circunstâncias sob as

quais as condições acústicas na plataforma não estão influencias pelos dados acústicos do estúdio são quando o valor da área das aberturas é muito pequenas. O processo de caída da energia sonora em recintos acoplados acusticamente está determinado basicamente pelo recinto mais reverberante, inclusive nos caso em que o recinto "vivo" é pequeno com relação ao outro recinto mais absorvente. Portanto para se criar essas plataformas independentemente, o tempo de reverberação do estúdio deverá ser inferior a 0,5 segundos. Entretanto a independência acústica dessas plataformas sob estas condições é muito difícil de conseguir, já que é complicado obter neles um tempo de reverberação curto. Portanto, as condições acústicas no cenário montado sobre essas plataformas dependem diretamente do tempo de reverberação do estúdio, e para controlar esse tempo será necessário trabalhar a absorção do recinto ou alteração do volume.

Uma estratégia de reduzir o tempo de reverberação a baixas frequências a valores inferiores as medias e altas frequências, é a utilização dos ressonadores de Helmholtz. O uso dessa custosa operação justifica-se pelo grande uso dos microfones direcionais. Estes microfones se tornam omnidirecionais as frequências cujo comprimento de onda é grande em comparação com as dimensões do recinto, com isso, tendem a recorrer excessiva informação reverberante em baixas frequências, colorando o sinal, efeito este compensado com a utilização dos ressonadores de Helmholtz.

Os estúdios médios são normalmente usados em televisão para programas estáticos, como os telejornais. O tratamento que se deve aplicar a esses estúdios é o mesmo que se aplica nos grandes estúdios. Entretanto, suas restrições na reverberação não são tão grandes já que as dimensões não permitem retardos elevados. Os ruídos produzidos são facilmente percebidos por não haver movimento dos equipamentos, o que exige um grande cuidado com o isolamento sonoro do recinto.

Com base em alguns estúdios típicos para gravação de reportagens, colóquios, revistas, jornalismo, etc., Recuero (1993) recomenda estúdios com capacidade de 20 – 50 pessoas e equipe técnica sem publico, as dimensões expressas nas tabelas 1 e 2 abaixo:

Tabela 1 Dimensões recomendada para estúdios de informação

Área do piso (m^2)	Altura (m)	Volume (m^3)	Área total (m^2)	Nº de pessoas
100	6,5	650	400	20
150	6,5	975	600	35
200	7,5	1500	800	50

Fonte: Recuero, M., 1993.

Tabela 2 Dimensões recomendada em estúdios de televisão.

Tipo de estúdio	Área do piso (m)	Altura (m)	Volume (m^3)	Nº de pessoas no interior
Informativo	12/15	2,8/3,2	34/48	1/2
Entrevista	50/80	4/5	200/400	10/15
Pequenos	100	5/6	600	20
Pequenos	150	6	900	35
Pequenos	200	7	1400	50
Grandes	300	8	2400	100
Grandes	450	10	4500	200
Grandes	600	11	6600	250
Grandes	1000	15	15000	400

Fonte: Recuero, M., 1993.

Nesses tipos de estúdios deve-se resguardar um tempo de reverberação de 0,40 a 0,60 segundos, de acordo com a frequência, (que permite gravar fala e música), e a inteligibilidade da palavra superior a 85%.

Figura 11 Estúdio de televisão para jornalismo da Universidade da Flórida. "<http://www.siebeinacoustic.com/portfolio/project%20types/tv_stations_and_broadcast_facilities.html>"

Os estúdios pequenos são comumente usados para gravação de entrevistas ou anúncios com locutor único. As dimensões recomendadas por Recuero (1993) são as da tabela 3.

Tabela 3 Dimensões recomendada em estúdios de televisão

Área do piso (m^2)	Altura (m)	Volume (m^3)	Área total (m^2)	Nº de pessoas
12	2,8	34	70	1
15	3,2	48	100	2
50	4	200	250	10
80	5	400	350	15

Fonte: Recuero, M., 1993.

Tendo em conta que se usa para gravação de palavra uma margem dinâmica não muito grande (para o caso de gravação de dramática), um bom tempo de reverberação é de 0,35 segundos. Pela mesma razão, a inteligibilidade deve ser de 95%. A absorção será a corresponde ao volume para obter o tempo de reverberação correspondente.

Este tipo de estúdio deve ter a forma de paralelepípedo retangular, muito conveniente para estúdios pequenos por não introduzir distorções consideráveis.

5. SALA DE CONTROLE

Nos últimos anos a evolução dos equipamentos e sistemas de gerenciamento e controle dos sons nas cabines vem sofrendo importante transformação. Equipamentos cada vez mais sofisticados permitem compensar os problemas acústicos que podem acorrem nas salas de controle dos estúdios.

O feito principal dessas salas é permitir com que o técnico de som possa fazer uma avaliação precisa do som que está sendo registrado. Para isso é necessário a existência de transdutores de alta fidelidade e eficácia, além de um equipamento eletrônico de qualidade e um ambiente acusticamente bem tratado, de modo a não permitir que não haja coloração do som percebido. Qualquer falha nesses elementos repercutirá, sem dúvida, no som escutado pelo técnico e, por tanto, nas decisões que ele tomara a cerca das gravações sonora, podendo produzir uma degradação da qualidade do conteúdo gravado.

Existem diferentes teoria quanto ao condicionamento acústico das salas de controle dos estúdios. Há os que defendem a tese da necessidade da máxima absorção possível, porém aplicada uniformemente por todas as superfícies. Por outro lado, há quem proponha que a parte frontal do controle devem ter alguns painéis refletores, enquanto que a parede de fundo deve ser tratada com materiais absorventes. Existem também os que defendem que a absorção deve estar na parte frontal do controle, as

demais superfícies devem ser tratadas com materiais não absorventes.

Pode-se observar que essas proposições são completamente diferentes uma das outras. A prática nos mostra que as estratégias nem sempre podem ser generalizadas, especialmente quando há variação das dimensões do recinto.

Figura 12 Sala de controle grande com alta reflexão e difusão. Estúdio Cherry Beach Sound, Toronto, Canadá. "<http://www.cherrybeachsound.com/control-room-A.html>"

As salas com grande volume, por exemplo, as reflexões das paredes alcançam o técnico de áudio com um retraso considerável com relação ao som direto, devido a distancia entre o técnico e as paredes de reflexão. Por outro lado, em uma sala com volume pequeno, as reflexões chegam ao ouvido do técnico com tempo muito próximo ao som direto. Essas diferenças entre os tempos de separação dos sons diretos e refletidos são muito importantes desde o ponto de vista do mecanismo da audição e da capacidade de integrar os sons. Quando um som direto e refletido chega com uma separação igual ou inferior a 16 metros, o som os soma em um único som, ou seja, não se percebe o eco. Porém, se essa separação for superior a 22 metros e o som refletido chega receptor de uma forma isolada, o ouvinte perceberá o eco com nitidez, porém se existem muitas reflexões o fenômeno será percebido como reverberação.

Não se pode negar que as reflexões dão vida e corpo ao som nas pequenas salas de controle. Um entorno muito absorvente para as caixas/monitores não é totalmente certa, como consequência da presença inevitável do vidro do visor acústico, isto ocasiona um mau funcionamento das caixas, já que a 100Hz, o direcionamento é praticamente esférico, quer dizer, independentemente da orientação que se tem das caixas, a baixa frequência, quase todo o som estará dirigido ao visor acústico e ao operador. Esta energia a baixa frequência refletida pelo visor chega ao operador sem nenhum acompanhamento das reflexões próximas a ele. Isso distorce o som de tal forma que o técnico sente a necessidade de atenuar os sons graves, quando na realidade o sinal real não o necessita.

Como consequência do exposto anteriormente vê-se a necessidade de desenhos que proporcione reflexões curtas na sala de controle. O uso de elementos construtivos nas superfícies de parede ou teto e o correto posicionamento do operador podem encurtar o caminho percorrido pelo som refletido. É conveniente evitar paredes paralelas para que não se produzam ondas estacionárias. O atraso do som direto e refletido pela parte posterior da sala permite ao técnico escutar o tempo de retardo inicial do estúdio sem mascaramento. Ademais, as reflexões desviadas por distintos caminhos produzem a impressão de se estar em um recinto maior, já que aumento o tempo de reverberação. Todos esses efeitos combinados produzem uma maior clareza do som, aproximando o operador do que em realidade está sendo gravado.

O grau de absorção e reflexões das salas de controle dos estúdios em muito vai depender das exigências quanto ao tempo de reverberação do recinto e do retardo dos percursos feito pelas primeiras reflexões, que diretamente estará em consequência da forma e posicionamento das superfícies internas. Para cada caso em particular, seja uma grande sala de controle, que pode estar monitorando dois estúdios simultaneamente, assim como, uma sala individual de controle de um estúdio de tamanho reduzido, as soluções acústicas se assemelham, guardadas suas proporções.

Em linhas gerais as salas de controle que dispõem das superfícies de parede isolantes tendem a ter um abaixa coloração do recinto, já que nível sonoro pareça menos importante que o comportamento acústico da sala, podendo ser compensado com o aumento da energia sonora.

Estas características de salas de controle parecem indicar que a localização da imagem sonora, a relativa coloração e a perspectiva do tratamento acústico, podem ser consideradas a princípio como um suporte parcial de conceito de condicionamento acústico de salas de controle. Embora, obter uma boa absorção por baixo de 125 Hz de frequência é bastante difícil, a absorção a baixa frequência é mais efetiva nas paredes e nas esquinas posterior, sob o ponto de visto acústico.

As salas de controle e o estúdio estão acoplado eletroacusticamente por meio do microfone, amplificador e caixas sonoras. Neste caso o tempo de reverberação do recinto secundário está dominado pelo tempo de reverberação do recinto principal, e maior. Entretanto, como os tempos de reverberação dos dois recintos são similares, a magnitude desse efeito é reduzida.

Não existem normas que determinem o volume das salas de controle, porém um número elevado desses projetos reúnem volumes em torno de 30 e 150 m^3. Como o padrão de reflexão desejado tem relação com a distância entre as paredes posteriores, as paredes laterais e a posição do console, por conseguinte vai influir no tamanho e na forma da sala. Em caráter geral se desenha a planta com forma trapezoidal, mais estreita para os fundo e mais larga para parte do visor acústico, além do teto inclinado, com superfícies angulares (compressão ou expansão), mediante um desenho cuidadoso, pode proporcionar um maior número de reflexões úteis.

Uma técnica de tratamento acústico consiste em que a parede posterior e as laterais levarem uma grade de material acusticamente transparente, chapa ou madeira perfurada, por exemplo, isso ajuda a justar acusticamente o controle, sem grandes variações.

As reflexões procedentes da grade podem ter um caráter tonal, o espaço uniforme das perfurações atua como uma rede de difração às frequências cujo comprimento de onda seja comparável com a separação das perfurações, assim como com as características delas. As perfurações podem não ser transparentes ao som em baixas frequências.

O tratamento acústico nas salas de controle afetará o valor do tempo de reverberação no recinto. Esses valores costumam variar entre 0,13 e 0,2 segundos, embora em algumas salas de maior dimensão esse valor pode chegar a 0,47 segundos. Sendo que nesse caso o tempo de reverberação não deve varia com a frequência, a fim de que as diferentes componentes do sinal desapareçam gradativamente com a mesma velocidade, e a gravação sonora não estará influenciada pelas características acústicas da sala de controle.

No caso de sala de controle de estúdio de televisão, existem três salas de controle necessárias: imagem, som e luz. No caso de imagem o tratamento acústico desse tipo de sala geralmente está enfocado a facilitar a comunicação entre o produtor e os demais membros da equipe de produção. Para isso o primeiro passo é reduzir a reverberação que poderia causar confusão se falasse mais de uma pessoa ao mesmo tempo. Os tempos típicos de reverberação rondam entorno dos 0,25 segundos.

Os recintos de controle do som devem proporcionar boas condições para o desenvolvimento da função, já que é o responsável por todas as questões referentes à seleção e

disposição dos microfones, uso de reverberadores artificiais, etc. Por tanto, essas salas devem cumprir os mesmo requisitos que as exigidas para o projeto acústico de estúdio de som.

As salas de controle de luz não requerem tratamento acústico especial.

6. CONCLUSÃO

São imprescindíveis para o bom desempenho acústico de uma sala de gravação sonora as análises preliminares, a localização do estúdio, a identificação dos pontos de maior debilidade, o isolamento dos sons e ruídos provenientes do exterior e gerados no próprio recinto, assim como, o correto condicionamento acústico que favorecer o desempenho eficaz do recinto.

Cada etapa do projeto do estúdio tem uma parcela importante no resultado final da qualidade acústica do recinto, e deverá ser pensado conjuntamente de modo a não permitir que uma determinada estratégia de projeto seja prioritária em detrimento de outra.

O desenho do local, a partir da forma e volume é responsável por grande parte do bom funcionamento acústico do estúdio. Por tanto, deve ser pensado conjuntamente com os demais critérios do projeto. É fato que um projeto corretamente coordenado e integrado todas as suas fases será mais eficiente e nos custoso que no caso de não haver flexibilidade para o ajuste dos diferentes sistemas. Um projeto acústico é muito mais valioso e econômico se pensando conjuntamente com o projeto arquitetônico e estrutural.

Os estúdios para gravação sonora dispõem hoje de sofisticados equipamentos eletrônico que em grande parte camuflam, ou mesmo eliminam, alguns dos problemas decorrentes das falhas acústicas dos recintos. Ainda assim, a responsabilidade pelo adequado funcionamento acústico do estúdio é de responsabilidade da ciência. Cabe aqui salientar que o êxito de um projeto acústico do estúdio de gravação sonora está na multidisciplinariedade entre os especialistas e os sistemas por eles adotados.

7. REFERÊNCIAS BIBLIOGRÁFICAS

[1] RECUERO, Manuel. Acústica de Estudios para gravación sonora. 2ª ed. Madrid: Instituto Oficial de Radiotelevision Española, 1993.

[2] BISTAFA, Silvio. Acústica aplicada ao controle do ruído. São Paulo: Ed. Edgard Blücher, 2006.

[3] COLINA, Carlos; MORENO, Antonio. Acústica de la edificación. 5ª ed. Madrid: UNED, 2005.

[4] SILVA, Péricles. Acústica arquitetônica e condicionamento do ar. 5ª ed. Belo Horizonte: Edital E. T. Ltda, 2005.

[5] SAN MARTÍN, Juan Eugenio. Acústica Arquitectónica para salas de grabación. Seminário de acústica arquitectónica para salas de gravación. Buenos Aires, sd.

[6] CARRIÓN, Antoni. Diseño acústico de espacio arquitectónico. Barcelona: UPC, 1998.

[7] FOODY, S. Optimización de forma en acústica arquitectónica. V congresso Iberoamericano de Acústica. Santiago: SCA, 2006.

[8] COX, Trevor; D'ANTONO, Peter. Acoustic Absorbers and Diffusers. Theory, design and application. Londor: Taylor y Francis Group, 2005.
.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Revisão de métodos de avaliação de qualidade de voz para sistemas de telepresença

Leonardo O. Nunes[1] e Luiz W. P. Biscainho[1]

[1] DEL/Poli & PEE/COPPE, Universidade Federal do Rio de Janeiro
Caixa Postal 68504 – Rio de Janeiro, RJ, Brasil

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

RESUMO

As características inerentes aos sistemas modernos de telepresença, como transmissão de áudio em banda completa e uso de microfones e alto-falantes desacoplados em salas de média dimensão com diversos dispositivos eletrônicos, propõem novos desafios para a avaliação automática da qualidade da fala por eles transmitida—em especial pela necessidade de lidar com degradações de origem acústica como ruído de fundo, reverberação e eco acústico. Este artigo faz uma revisão de trabalhos de pesquisa recentes nessa área desenvolvidos no Brasil, envolvendo: cenários para a avaliação de qualidade, sistemas de avaliação automática e classificadores de tipos de degradação.

0 INTRODUÇÃO

A ubiquidade dos sistemas de comunicação por fala trouxe uma demanda por serviços de alto desempenho, que por sua vez requerem métodos confiáveis para medir a qualidade da voz recebida. De uma forma geral, o método mais adequado para se medir qualidade de áudio é o teste subjetivo, em que se pede aos participantes que julguem a qualidade de um dado sinal ao escutá-lo. Infelizmente, tais testes são muito custosos, tanto financeiramente quanto em termos de tempo [1], além de necessitarem de um planejamento cuidadoso para que seus resultados sejam significativos [2]. Os chamados métodos objetivos de avaliação de qualidade, por sua vez, foram desenvolvidos para operar de forma automática, sem a necessidade de recorrer a ouvintes; isso os torna mais baratos e rápidos de executar, porém lhes confere acurácia inferior à dos métodos subjetivos.

Métodos de avaliação de qualidade automáticos podem ser classificados como paramétricos ou baseados no sinal [3]. Os primeiros procuram avaliar a qualidade através de parâmetros que podem ser obtidos diretamente da rede de transmissão (como, por exemplo, o tipo de codificador). Um exemplo de avaliador paramétrico é o E-Model, descrito em [4]. Já ferramentas baseadas no sinal procuram estimar a qualidade diretamente do sinal de fala e podem utilizar apenas o si-

nal sob teste (que foi degradado pelo sistema de transmissão) ou, além deste, o sinal de referência (sua versão sem degradações). Um exemplo de método que utiliza os sinais degradado e de referência é o chamado PESQ (do inglês *Perceptual Evaluation of Speech Quality*), descrito na Recomendação P.862 do ITU [5]; já a Recomendação P.563 da ITU [6] descreve um método de avaliação que se vale apenas do sinal degradado.

Outra forma de classificação de métodos de avaliação de qualidade diz respeito a como eles são utilizados nos sistemas de comunicação. Um avaliador de qualidade pode ser empregado durante o projeto, a manutenção ou a operação do sistema [7], e cada uma dessas etapas requer um tipo adequado de ferramenta. Durante o projeto e a manutenção, o operador tem completo controle sobre o sistema e pode realizar quantos testes forem necessários; nesse caso, podem ser utilizadas ferramentas intrusivas e que requeiram o sinal de referência. Já a avaliação de qualidade durante a operação do sistema requer ferramentas que operem de forma não-intrusiva; tais ferramentas de monitoração [3] devem medir os sinais ou parâmetros da rede internamente e inferir a qualidade a partir dessas medidas.

Outro fator importante para avaliação de qualidade é a aplicabilidade dos métodos. Normalmente, cada avaliador só pode ser utilizado para sinais com largura de banda específica e corrompidos por determinados tipos de degradações. Por exemplo, o algoritmo PESQ [5] foi desenvolvido para lidar com sinais de banda telefônica (de 300 Hz até 3,5 kHz) e degradações provocadas por codificação, erros de transmissão e ruído de fundo. O W-PESQ [8] (do inglês *Wideband PESQ*), por sua vez, estende a banda de atuação do PESQ até 7 kHz. No entanto, ambos foram considerados inadequados para lidar com problemas relacionados a audibilidade [9] e eco [10]. Nos últimos anos, foram propostas diversas modificações aos algoritmos da família PESQ de forma a estender a sua aplicabilidade. Por exemplo, o PESQ é modificado para lidar com eco em [11] e com saturação em [12]. Houve também tentativas de tornar o método mais abrangente, como a proposta em [13]. O novo padrão de qualidade da ITU P.863 (chamado POLQA) [14], além de lidar com sinais com largura de banda até 24 kHz, leva em consideração outros tipos de distorção como, por exemplo, reverberação, além de lidar naturalmente com a combinação de diferentes degradações.

Recentemente, a demanda por sistemas de telepresença de alta qualidade motivou a busca por avaliadores específicos para este problema [15]. Em tais sistemas, a qualidade pode ser afetada por uma combinação de degradações oriundas da rede de transmissão (como ruído de codificação, *jitter* e perda de pacotes) e de origem acústica (como ruído de fundo, eco e reverberação). As degradações acústicas apresentam um especial desafio a sistemas de telepresença, pois em geral não há controle sobre elas, que dependem do ambiente em que o usuário está localizado. Além disso, em sua maioria, os avaliadores de qualidade desenvolvidos para sinais de fala não são preparados para lidar com voz em banda completa (até 20 kHz ou mais).

Esta revisão irá abordar um conjunto de soluções recentes para o problema de avaliação de qualidade de sistemas de telepresença. O foco será nas degradações acústicas e em sinais com alta taxa de amostragem. Inicialmente, na Seção 1, serão descritos modelos para geração de sinais degradados e determinação de pontos de medida em redes de telepresença. A seguir, na Seção 2, serão tratados dois métodos de avaliação de qualidade, voltados um para sinais degradados por ruído de fundo e outro para sinais degradados por eco. Na Seção 3, será revisto um identificador e classificador das degradações acústicas presentes em sinais de fala. Por fim, será feita uma discussão geral na Seção 4.

1 MODELOS PARA SISTEMAS DE TELEPRESENÇA

Nesta seção, são descritos os cenários propostos em [16] e [17] para modelagem de degradações acústicas em sistemas de telepresença. Tais modelos podem ser utilizados para: identificar possíveis pontos de medidas para os sinais submetidos a avaliadores de qualidade, detalhar quais degradações estão presentes em cada ponto de medida e auxiliar a simulação de sinais degradados para a geração de bancos de sinais.

1.1 Modelo Geral

Ponto de partida para a criação dos cenários que serão detalhados a seguir, o modelo geral para sistemas de telepresença mostrado na Figura 1, descrito originalmente em [16], considera apenas o caso de uma conversação entre dois participantes. Na figura, pode-se notar o Participante 1 na Sala 1 em conversação com o Participante 2 na Sala 2. A emissão sonora saindo da boca do Participante 1 é recebida pelas suas próprias orelhas[1] e pelo microfone na Sala 1 modificado pela resposta da sala $H_1(z)$. Cada sinal de ruído N_1^i de uma dada fonte i como, por exemplo, um condicionador de ar ou a ventoinha de um computador dentro da Sala 1, chega às orelhas do Participante 1 e aos microfones da Sala 1, após as modificações por $H_{N^i_{1_{\text{in}}}}(z)$ e $H_{N^i_{1_{\text{out}}}}(z)$ em seus respectivos percursos. O sinal que chega da Sala 2 através dos alto-falantes na Sala 1 atinge as orelhas do Participante 1 modificado pela resposta $H_1'(z)$ e é acusticamente realimentado para o microfone na Sala 1 através de $M_1(z)$. Uma descrição análoga pode ser feita para a Sala 2. A caixa 'Rede' resume as degradações causadas pela rede de transmissão, como, por exemplo, atrasos de transmissão, atenuação, perda de pacotes e codificação.

O modelo descrito pode ser simplificado, gerando o que se vê na Figura 2, pelas seguintes modificações:

[1] É considerada apenas a transmissão direta, sem multipercurso, por questões de simplicidade.

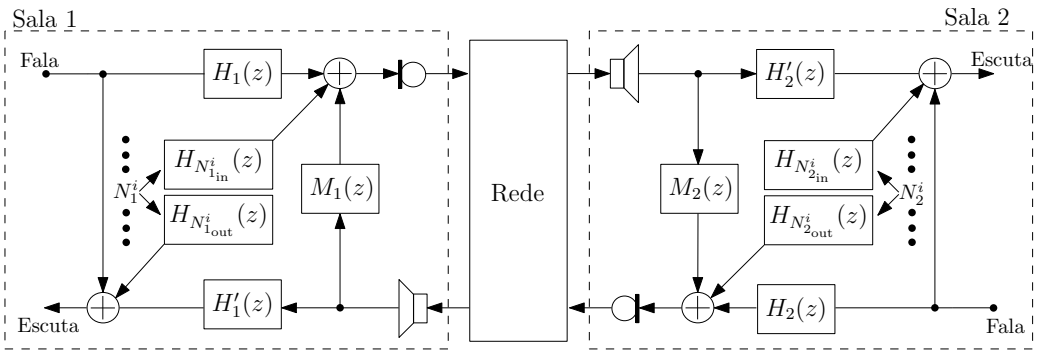


Figura 1: Modelo geral para sistemas de telepresença. Figura adaptada de [16].

- Apenas o ciclo de transmissão e recepção relacionado ao Participante 1 foi considerado;
- Uma única fonte de ruído aditivo, modelada como a combinação de todas as fontes anteriores, foi utilizada;
- Uma única resposta $H_{N_j}(z)$ representa as modificações sofridas pelo sinal de ruído na sala j, para $j \in \{1,2\}$;

Nesse modelo mais simples, assume-se que apenas um participante está falando. Assim, qualquer método de avaliação que seja desenvolvido utilizando este modelo deve estar associado a um detector de fala simultânea [18]. Os novos sinais de ruído podem ser descritos como

$$\overline{N}_1(z) = \sum_{i=0}^{M_1-1} H_{N^i_{1_{\text{in}}}}(z) N^i_1(z) \tag{1}$$

e

$$\overline{N}_2(z) = \sum_{i=0}^{M_2-1} H_{N^i_{2_{\text{in}}}}(z) N^i_1(z), \tag{2}$$

onde M_1 e M_2 são os números de fontes de ruído nas Salas 1 e 2, respectivamente.

Conforme mencionado anteriormente, esse modelo foi utilizado para o desenvolvimento dos cenários intrusivos e não-intrusivos que são apresentados nas próximas seções.

1.2 Modelo Intrusivo

Os modelos intrusivos possuem dois objetivos [16]: (1) representar as degradações de forma a permitir a geração controlada de bancos de sinais degradados para o auxílio ao desenvolvimento de ferramentas de avaliação de qualidade; (2) elencar pontos de medidas do sinal de referência e do sinal sob teste que sejam representativos para testes objetivos e subjetivos.

O modelo exibido na Figura 2 pode ser simplificado de forma a caracterizar de forma independente degradações por eco, ruído e reverberação em cada sala. Isto é feito através da separação do modelo geral em duas partes: um cenário local e outro remoto.

A Figura 3 mostra uma representação do cenário local, que permite caracterizar degradações acústicas geradas no lado do falante: ruído de fundo provocado por fontes locais ($\overline{N}_1$) e reverberação devido à resposta ao impulso da sala local.

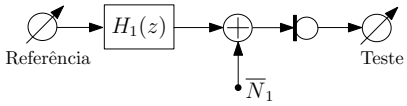


Figura 3: Cenário local [16].

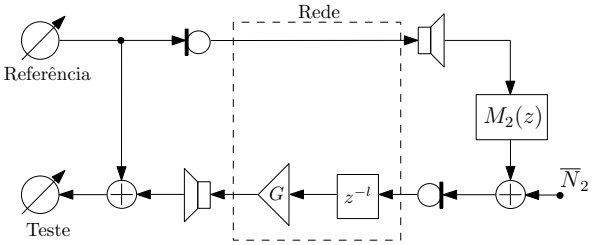


Figura 4: Cenário remoto [16].

A Figura 4 contém o diagrama do cenário remoto, que permite caracterizar degradações acusticamente geradas na sala em que o falante não está: majoritariamente, eco acústico resultante do acoplamento caixa-microfone combinado com os atrasos da transmissão. São descartados os efeitos da sala local ($H_1(z) = 1$ e $\overline{N}_1 = 0$ na Figura 3); possíveis distorções causadas pela rede de transmissão são desconsideradas; e o efeito da transmissão do sinal é modelado como apenas um atraso l e um ganho G. A inclusão do ruído de fundo $\overline{N}_2$ e da resposta $M_2(z)$ da sala remota permite uma modelagem mais realística.

1.3 Modelo Não-Intrusivo

Em [17], foi apresentado um modelo com os mesmos objetivos do descrito na seção anterior, porém talhado para monitoração e avaliação não intrusiva de sistemas de telepresença. De forma similar ao que foi feito no caso intrusivo, foram gerados dois cenários a partir do modelo simplificado mostrado na Figura 2: o do percurso direto e o do percurso com realimentação.

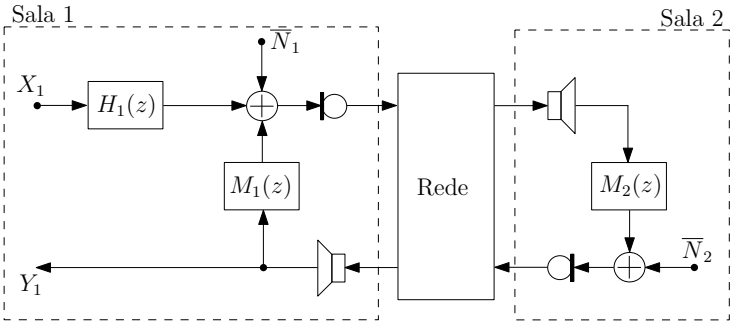


Figura 2: Modelo simplificado para sistemas de telepresença. Apenas são considerados os trajetos de transmissão e recepção para sinais emitidos pelo participante na Sala 1. Figura adaptada de [17].

O cenário do percurso direto, mostrado na Figura 5, modela isoladamente as degradações que podem ocorrer independentemente de qualquer realimentação do sinal entre as duas salas, isto é, as modificações do sinal entre sair da boca do falante na Sala 1 e ser reproduzido pelo alto-falante da Sala 2.

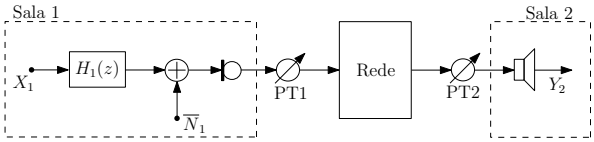


Figura 5: Cenário do percurso direto. PT no diagrama signifca "Ponto de Teste". Figura adaptada de [17].

Dois possíveis pontos de testes, PT1 e PT2, podem ser identificados neste cenário. O primeiro permite a avaliação de sinais que serão transmitidos para a Sala 2, enquanto o segundo permite a avaliação dos sinais que nela chegam. Podem ser avaliadas em:

- PT1: degradações acústicas acontecendo na Sala 1 (reverberação e ruído de fundo) e possíveis distorções não-lineares causadas pelo equipamento de captura de áudio (microfone e conversor analógico-digital) como, por exemplo, saturação.
- PT2: todas as degradações de PT1 mais as distorções causadas pela rede, como ruído de codificação.

O PT1 do cenário do percurso direto é similar ao ponto de teste do cenário local da seção anterior. Já o PT2 é mais abrangente, incluindo degradações acústicas e de transmissão.

No cenário do percurso com realimentação, mostrado na Figura 6, são consideradas apenas as degradações relacionadas com a realimentação do sinal entre as duas salas devida ao acoplamento caixa-microfone. Neste caso, o cancelador de eco acústico (AEC, do inglês *Acoustic Echo Canceller*) da Sala 2 é explicitamente representado. Foram desconsideradas as degradações acústicas ocorrendo na Sala 1 e o acoplamento caixa-microfone da Sala 1. Os dois pontos de teste do modelo são

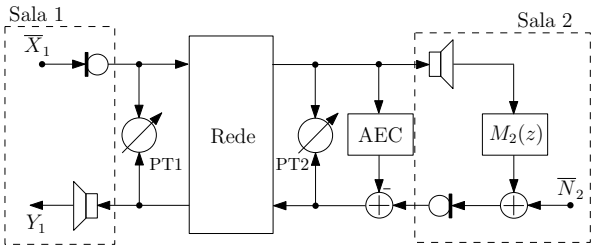


Figura 6: Modelo do percurso com realimentação. PT no diagrama significa "Ponto de Teste". Figura adaptada de [17].

- PT1: que permite a avaliação do eco acústico e do ruído originados na Sala 2.
- PT2: que permite a avaliação do AEC da Sala 2.

Ambos possuem duas terminações, permitindo o acesso aos sinais de referência e degradado. O PT1 é muito similar ao par (Referência, Teste) do cenário remoto da seção anterior, se a rede de transmissão for simplificada para um atraso e um ganho. O PT2, por outro lado, permite a avaliação isolada da operação do AEC.

1.4 Utilização dos Modelos em Testes Subjetivos

Uma das principais aplicações dos modelos apresentados nas seções anteriores é no desenvolvimento de banco de sinais de fala para serem utilizados em testes subjetivos e, posteriormente, no processo de desenvolvimento de ferramentas para avaliação automática de qualidade. Neste seção, é feita uma breve discussão de como tais banco de sinais podem ser gerados utilizando-se os modelos descritos.

Idealmente, deveriam ser levados em conta durante o desenvolvimento dos bancos os seguintes passos [16]:

1. Especificação das características dos sinais de referência;
2. Especificação das degradações envolvidas (definido-se o modelo e o tipo de teste mais adequado para o problema) e dos níveis das degradações;

3. Geração do conjunto de sinais degradados com base no modelo escolhido utilizando os níveis de degradações selecionados.

A Recomendação P.800 da ITU trata da escolha de sinais de fala de referência para testes subjetivos [19]. Segundo este padrão, cada sinal a ser testado deve ser composto por frases curtas com duração entre 2 e 3 s cada. Estas devem ser retiradas de publicações não técnicas ou jornais e não devem fazer sentido umas com as outras, embora idealmente devam ser escolhidas de forma a garantir variabilidade fonética no conjunto. Devem, ainda, ser emitidas por falantes de ambos os sexos. Dependendo do tipo de teste subjetivo, devem-se utilizar de 2 até 5 frases num mesmo sinal, com intervalos de 0.5 s de silêncio entre elas. Também é preciso ter o cuidado de equalizar as audibilidades de todas as frases entre si antes da composição dos sinais. Pode ser encontrado em [16] um procedimento para ajuste da audibilidade das frases.

Uma vez gerados os sinais de referência, devem ser definidas as degradações que serão analisadas e qual será o tipo do teste subjetivo utilizado. De forma geral, deve-se optar por testes do tipo ACR (do inglês *Absolute Category Rating* [19]) para o modelo não-intrusivo e testes do tipo DCR (do inglês *Degradation Category Rating* [19]) para o modelo intrusivo. As degradações a serem avaliadas determinarão a escolha da versão dos modelos:

- Ruído de fundo e reverberação: cenário local (intrusivo) ou PT1 do cenário do percurso direto (não-intrusivo);
- Eco e ruído de fundo: cenário remoto (intrusivo) ou PT1 do cenário do percurso com realimentação (não-intrusivo).

Nos dois cenários do modelo não-intrusivo, o PT2 deve ser utilizado quando se deseja avaliar o efeito da rede de transmissão.

Para os cenários do modelo intrusivo, os sinais degradados seriam medidos nos pontos indicados como Teste nas Figuras 3 e 4. Para o cenário do percurso direto do modelo não-intrusivo, os pontos PT1 e PT2 podem ser utilizados para se obter os sinais degradados. Já para o cenário do percurso com realimentação do modelo não-intrusivo, os sinais em PT1 e PT2 não podem ser utilizados diretamente em testes subjetivos, porque o efeito do eco só é percebido quando o sinal de retorno é somado acusticamente com o sinal da voz do falante; nesse caso, o sinal a ser utilizado no teste deveria ser $Y_1 + \overline{X}_1$.

2 AVALIADORES DE QUALIDADE

Nesta seção, são descritos dois avaliadores intrusivos de qualidade desenvolvidos para sistemas de telepresença: o primeiro, originalmente apresentado em [20], se destina a sinais contaminados por ruído de fundo; o segundo, originalmente apresentado em [21], é voltado para a sinais degradados por eco acústico. Em ambos, os sinais sob teste são de banda completa (até 20 kHz ou mais).

2.1 Avaliador para Sinais Contaminados com Ruído

Em sistemas de telepresença, é comum a contaminação do sinal por ruído de fundo aditivo de origem acústica, devido à presença de fontes de ruído nas salas em que estão localizados os participantes. Devido ao fato de as características espectrais do ruído variarem conforme a sua fonte e a sua percepção depender do próprio sinal de fala ao qual ele se somou, um sistema que utilize os sinais para medir a qualidade será mais eficaz do que um sistema puramente paramétrico baseado na estimativa da SNR, por exemplo.

Em [20], é proposto um método de avaliação da qualidade de sinais de voz de alta qualidade degradados por ruído que estende os métodos PESQ e W-PESQ. Estes são brevemente descritos na próxima seção, seguidos de uma revisão do método de [20].

2.1.1 PESQ e W-PESQ

O método de avaliação de qualidade conhecido como PESQ foi projetado para avaliar a qualidade de sinais de fala de banda estreita (de 300 Hz até 3.4 kHz) em sistemas telefônicos. Na Figura 7, pode ser visto o diagrama de blocos do PESQ que é divido em três partes: pré-processamento, modelo perceptivo e modelo cognitivo. Cada parte será detalhada a seguir.

O pré-processamento ajusta os sinais de entrada para que possam ser submetidos ao modelo perceptivo. A primeira modificação aplicada aos sinais é o ajuste do nível do sinal de teste para que este fique igual ao nível do sinal de referência. Em seguida, os sinais são filtrados de forma a simular o efeito de um aparelho telefônico. Por fim, os sinais são alinhados temporalmente entre si. Podem ser encontradas em [22] maiores informações sobre estes procedimentos.

A modelagem perceptiva procura gerar uma representação do sinal que simula o sinal na saída do ouvido interno de um ser humano. Ela é composta por uma divisão do sinal em blocos, aplicação da transformada de Fourier em cada bloco, mapeamento para a escala Bark [23] e transformação para uma escala de audibilidade [23].

A partir das saídas fornecidas pelo modelo perceptivo para os sinais de referência e degradado, o modelo cognitivo calcula duas medidas de distorção. A primeira medida computa a diferença entre o que é ouvido no sinal de teste e o que é ouvido no sinal de referência. Essa diferença leva em consideração critérios da cognição humana; por exemplo, o sumiço de um trecho do sinal é considerado mais danoso à qualidade do que a adição de informação. A segunda medida leva em consideração a quantidade de quadros que estão fora do

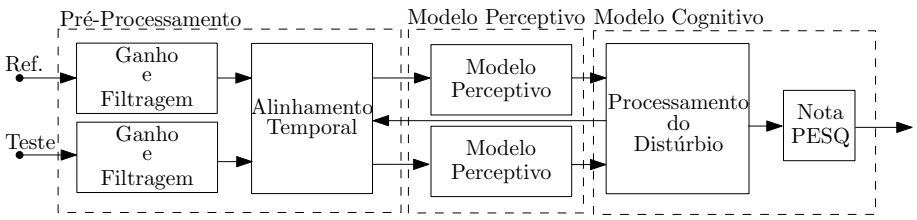


Figura 7: Diagrama em blocos do PESQ. Adaptado de [20].

alinhamento temporal devido a, por exemplo, problemas de sincronia. Essas duas medidas, são, então, combinadas para gerar a nota PESQ final.

Para ser comparada diretamente com o MOS, a nota PESQ ainda precisa ser mapeada para uma escala adequada. É descrita em [24] uma curva de mapeamento padronizada pelo ITU.

O padrão W-PESQ [8] é similar ao PESQ, porém foi projetado para avaliação de sinais de banda larga (de 50 Hz até 7 kHz) que não foram distorcidos pelo uso de monofones. Por isso, a filtragem inicial do PESQ é substituída por um filtro passa-altas com frequência de corte em 100 Hz.

2.1.2 EW-PESQ

Nesta seção, é brevemente descrito o método proposto em [20] para avaliar a qualidade de sinais de fala em banda completa (até 20 kHz ou mais) que sofreram a incidência de ruído de fundo. Este novo algoritmo utiliza como ponto de partida o W-PESQ e estende sua faixa de atuação na frequência, tendo sido chamado de EW-PESQ (do inglês *Extra-Wideband PESQ*).

A principal modificação a ser realizada no W-PESQ é sobre o seu modelo perceptivo, que agora precisa descrever a audição na região de frequências mais altas (acima de 7 kHz). Duas possíveis modificações no modelo cognitivo foram propostas em [20]. Na primeira, o modelo do W-PESQ é diretamente estendido, dando origem ao EW-PESQ(E); nesse caso, os resultados obtidos pelo modelo perceptivo do EW-PESQ para regiões abaixo de 7 kHz são idênticos aos resultados pelo do W-PESQ. Na segunda modificação, todo o modelo perceptivo é substituído pelo do método PEAQ [25, 26] (do inglês *Perceptual Evaluation of Audio Quality*), um avaliador automático da qualidade de sinais de áudio em geral na taxa de amostragem de 48 kHz, que descreve a audição em frequências até 18 kHz, dando origem ao EW-PESQ(R).

A outra modificação realizada sobre o W-PESQ se deu no mapeamento da saída do modelo cognitivo, detalhado mais a frente. O EW-PESQ preservou todos os demais blocos do W-PESQ.

2.1.3 Base de sinais e Testes Subjetivos

Para auxiliar o desenvolvimento e medir o desempenho do EW-PESQ, foi criado um banco de sinais de fala degradados por ruído de fundo. Para isto, um conjunto de sinais de referência foi gerado através da concatenação de 2 frases (selecionadas de um conjunto de 200 [27]), com um intervalo de 0,5 s entre elas.

Os sinais degradados foram obtidos adicionando-se ruído aos sinais de referência. Três tipos de ruído foram empregados: ruído de ar-condicionado, ruído de computador e ruído branco gaussiano. Foram utilizadas razões sinal-ruído (SNR, do inglês *Signal-to-Noise Ratio*) variando de 25 até 55 dB em passos de 5 dB. Foram gerados 5 diferentes sinais para cada possível combinação de SNR e tipo do ruído para 4 diferentes falantes (2 homens e 2 mulheres). Ao todo, foram criados 112 sinais, todos amostrados a 48 kHz com 24 bits de precisão.

A avaliação subjetiva desses sinais foi feita através de testes tipo DCR, seguindo as recomendações de [19]. As notas eram fornecidas numa escala de 1 até 5, com resolução de 0,1. No total, 21 participantes julgaram cada sinal, para o qual se calculou, então, o MOS (do inglês *Mean Opinion Score*).

2.1.4 Mapeamento entre nota PESQ e a escala MOS

O modelo cognitivo do PESQ (e, por consequência, do W-PESQ e do EW-PESQ) gera como saída uma nota numa escala diferente da escala MOS. Para compatibilizar as duas escalas, a ITU recomenda a utilização da função

$$y = 0.999 + \frac{4}{1 + e^{\alpha_1 x + \alpha_2}}, \tag{3}$$

onde x é a nota na escala PESQ e y o valor correspondente na escala MOS. Os parâmetros α_1 e α_2 foram encontrados para o PESQ e o W-PESQ de forma a se obter um bom mapeamento entre as duas escalas sobre uma extensa base de sinais.

De forma similar, esses dois parâmetros foram otimizados para o EW-PESQ. Para isso, é utilizada metade dos sinais descritos na seção anterior. De posse do MOS e da nota EW-PESQ atribuídos a esses sinais, uma rotina de otimização procurou os valores de α_1 e α_2 que minimizariam o erro quadrático médio entre as notas MOS preditas pela equação (3) e as notas obtidas nos testes subjetivos. Como ponto de partida da busca, utilizaram-se os valores do W-PESQ, $\alpha_1 = -1{,}3669$ e $\alpha_2 = 3{,}8224$. Ao fim da etapa de otimização, encontraram-se os valores: $\alpha_1 = -1{,}9902$ e $\alpha_2 = 6{,}6007$ para o EW-PESQ(E); e $\alpha_1 = -1{,}9027$ e $\alpha_2 = 6{,}1833$ para o EW-PESQ(R).

2.1.5 Avaliação de Desempenho

Os sinais que não foram utilizados na obtenção da função de mapeamento foram utilizados para medir o desempenho do EW-PESQ. Para isso, calculou-se entre os valores preditos pelo EW-PESQ (nas versões E e R) e os valores de MOS obtidos nos testes subjetivos o coeficiente de correlação, que mede quão similares são dois conjuntos de dados—um coeficiente de 100 % indica que os dados são exatamente iguais, a menos de uma transformação afim. Para o EW-PESQ(E), o coeficiente de correlação foi de 97,0 %, enquanto que para o EW-PESQ(R) foi de 96,8 %. Nas Figuras 8 e 9 podem ser vistos os valores medidos e preditos para cada um dos sinais utilizados na avaliação.

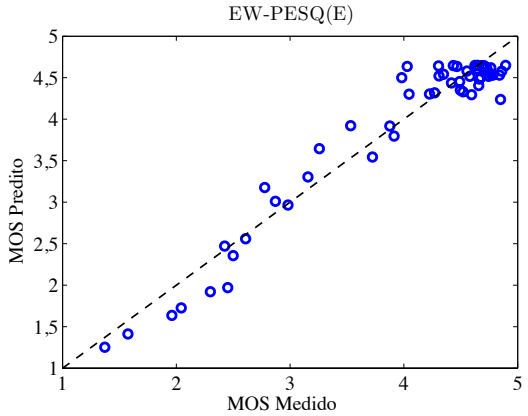


Figura 8: EW-PESQ(E): valor medido x valor predito. Cada sinal é denotado por 'o'. Adaptado de [20].

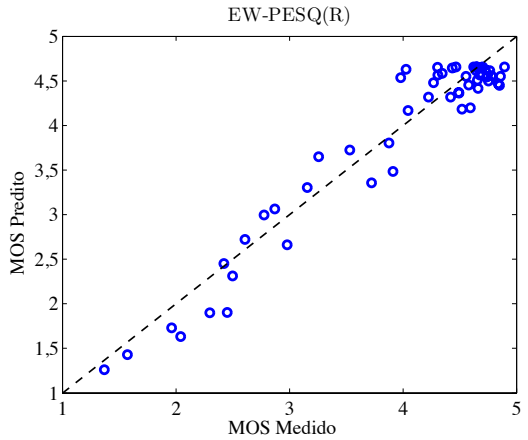


Figura 9: EW-PESQ(R): valor medido x valor predito. Cada sinal é denotado por 'o'. Adaptado de [20].

Como pode ser observado, ambas as versões do EW-PESQ tiveram um bom desempenho, sendo capazes de predizer com alta acurácia o MOS de sinais degradados por ruído. Considerando que o EW-PESQ(E) mantém compatibilidade com o PESQ e W-PESQ sem perda de desempenho em relação ao EW-PESQ(R), aquele mostra-se preferível a este.

2.2 Avaliador para Sinais Degradados por Eco

Em [21], é apresentado um avaliador objetivo para sinais de fala de banda completa degradados por eco. Em particular, foram consideradas condições de eco encontradas em sistemas de telepresença. A seguir, são descritos os passos utilizados para se gerar a ferramenta de qualidade.

2.2.1 Sinais e Teste Subjetivos

Inicialmente, foi necessária a criação de um banco de sinais de fala degradados por eco. Para isso, foi utilizado o cenário remoto apresentado na Seção 1.2 com $M_2(z) = 1$ (desconsiderando, portanto, o efeito do percurso caixa-microfone na sala remota). Os sinais de referência foram montados a partir de um conjunto de 200 frases [27] gravadas por quatro falantes (dois homens e duas mulheres), conforme as recomendações [19, 28]. Para simular uma conversa, os sinais de referência são compostos de 4 frases faladas por 2 falantes alternadamente: frase 1 por falante 1, frase 2 por falante 2, frase 3 por falante 1 e frase 4 por falante 2. Foi inserido um trecho de silêncio entre as frases de forma que o sinal de eco de uma frase não se misturasse ao início da frase seguinte. Os sinais degradados foram gerados degradando-se por eco as frases 1 e 3 com atrasos de {100, 200, 300 ou 400} ms e atenuações de {40, 35, 30, 25 ou 20} dB. Ao todo, foram gerados 108 sinais degradados por todas as combinações possíveis de atraso e atenuação, repetidas pelo menos cinco vezes cada. Todos os sinais são amostrados em 48 kHz com precisão de 24 bits.

De posse dos sinais, conduziu-se um teste subjetivo de avaliação de qualidade do tipo DCR seguindo as recomendações de [19]. Cada sinal foi avaliado por 20 participantes. Foi pedido que os participantes avaliassem apenas a perda de qualidade pela presença de eco. As opiniões eram marcadas numa escala de 0 (imperceptível) até 5 (muito incômodo). Ao fim do teste, a nota MOS foi calculada para cada sinal degradado como a média das opiniões individuais.

2.2.2 QAfE

Para se gerar o avaliador de qualidade, chamado QAfE (Quality Assessment for Echo) foi utilizado como ponto de partida o método PEAQ (ver Seção 2.1.2). O PEAQ pode ser dividido em duas partes: uma etapa de cálculo de medidas intermediárias (MOVs, do inglês *model output variables*) e um modelo cognitivo (na forma de uma rede neural) que agrupa as MOVs numa única nota. Considerando que o modelo cognitivo do PEAQ foi treinado para sinais de áudio degradados por ruído de codificação, ele não é adequado para o problema em questão. Pareceu possível, entretanto, examinar e selecionar MOVs individuais, agregando-as depois segundo um novo modelo cognitivo.

Assim, numa primeira etapa do desenvolvimento, foi estudado o relacionamento entre as MOVs do PEAQ

e o MOS obtido no teste subjetivo. A Figura 10 mostra este relacionamento para a MOV chamada *average distorted blocks* (ADB, MOV 5 do PEAQ). Como pode ser visto, há uma forte tendência de se obter valores altos de MOS para baixos valores da MOV, e vice-versa. Com isso, essa MOV foi selecionada para servir individualmente de base para uma ferramenta para avaliação da qualidade de sinais de fala degradados por eco.

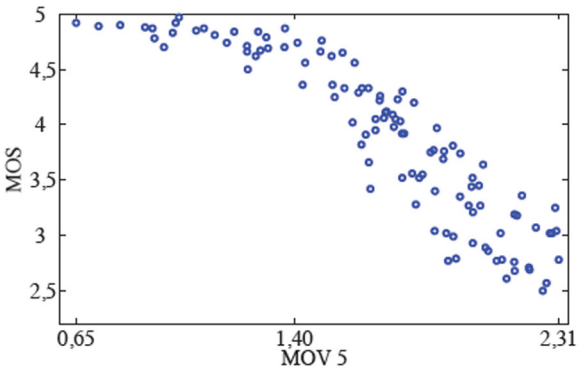


Figura 10: Valor assumido pela MOV 5 (*average distorted blocks*) versus o MOS para os sinais degradados por eco. Adaptado de [21].

Observando-se a Figura 10, pode-se ver que a ADB progride de 0 (alta qualidade) a 2,5 (baixa qualidade) de forma não-linear em relação ao MOS. Portanto, para que esses valores sejam utilizados para predizer o MOS, ainda é necessário realizar o respectivo mapeamento. Para esta finalidade, foi utilizada a mesma expressão empregada no EW-PESQ, mostrada na equação (3). Para encontrar os parâmetros α_1 e α_2, levaram-se os pares de valores de MOS medidos e respectivos valores da MOV 5 atribuídos a metade dos sinais degradados (conjunto de treinamento) a um algoritmo de otimização. Os valores que minimizaram o erro quadrático médio entre os valores preditos e medidos de MOS foram $\alpha_1 = 2{,}7$ e $\alpha_2 = -5{,}741$.

Com a segunda metade dos sinais degradados (conjunto de teste), fez-se a avaliação do desempenho do QAfE, obtendo-se um fator de correlação de 94 % entre valores preditos e valores medidos de MOS. A Figura 11 relaciona esses dois valores para os conjuntos de treinamento e de teste. Como se pode observar, o QAfE apresenta bons resultados para o problema abordado.

2.3 Outros Avaliadores

Além dos dois métodos descritos nessa seção, podem ser encontrados na literatura outros avaliadores de qualidade específicos para sistemas de telepresença. Por exemplo, é apresentada uma solução em [29] para avaliação intrusiva de sinais contaminados por reverberação. Já em [30], é descrita uma solução para avaliação não-intrusiva de sinais reverberados. Ainda

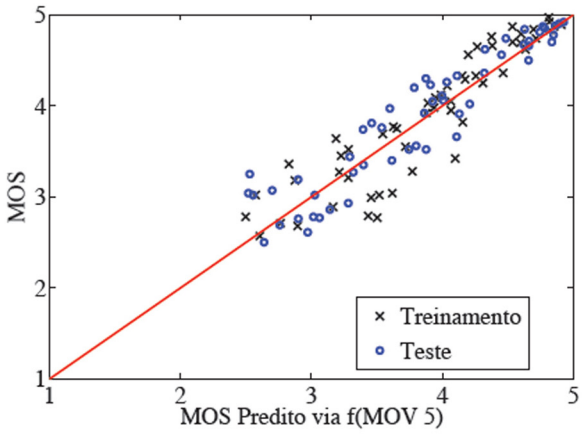


Figura 11: Valores preditos para o MOS x MOS para as duas partições dos dados. Adaptado de [21].

para o caso de avaliação de qualidade de sinais degradados por eco, é detalhada em [31] uma solução alternativa à apresentada na seção anterior. Nesta nova solução não-intrusiva, a qualidade do sinal é predita utilizando-se apenas dois parâmetros extraídos a partir dos sinais degradado e de referência: o atraso e o ganho.

3 CLASSIFICAÇÃO DE TIPOS DE DEGRADAÇÃO

Em geral, avaliadores objetivos de sinais de fala são desenvolvidos para um tipo específico de degradação, como é o caso dos avaliadores descritos nas seções anteriores. Na prática, no entanto, podem ocorrer várias degradações concomitantemente, ou então o tipo de degradação que afeta o sinal pode ser desconhecido. Em [32], foi apresentado como uma possível solução para este problema o esquema mostrado na Figura 12, em que uma etapa de classificação seria responsável por identificar os tipos de degradações presentes no sinal e com isso selecionar o método de avaliação de qualidade adequado. Isso exige, então, o desenvolvimento de um classificador de tipos de degradação (DTC, do inglês *Degradation Type Classifier*).

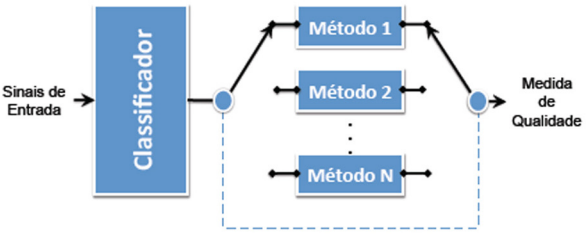


Figura 12: Esquemático de um sistema de avaliação de qualidade para diversos tipos de degradações. Figura adaptada de [32].

Soluções similares também foram propostas em [33, 34], sendo que nesses trabalhos o classificador

foi utilizado para selecionar diferentes modelos cognitivos; os resultados por eles relatados mostram que a etapa de classificação tende a melhorar o desempenho de métodos de avaliação de qualidade. Também é utilizada uma etapa de classificação na Recomendação P.563 da ITU [6] para avaliação não-intrusiva da qualidade de sinais de fala.

Em [35], foi descrito um classificador de degradações acústicas comumente encontradas em sistemas de telepresença, com referência, capaz de identificar até 5 possíveis degradações: ruído de fundo (N), eco acústico (E), reverberação (R) e combinações de eco acústico com ruído de fundo (E+N) e de reverberação com ruído de fundo (R+N). O sistema opera em duas etapas, detalhadas nas próximas seções: na primeira, a partir dos sinais de referência e degradado, é obtido um vetor de atributos; em seguida, este é submetido à segunda etapa, de identificação das degradações presentes no sinal.

3.1 Extração de Atributos

Nesta seção, são descritos os atributos utilizados em [35]. Era desejado que os atributos extraídos contivessem informações relevantes sobre o tipo de degradação afetando o sinal sob teste, e a principal motivação para as escolhas feitas foi explorar as diferentes formas como os sinais são degradados: aditivamente no caso de ruído e convolutivamente no caso de reverberação e eco. No total, foram propostos 10 atributos; 4 deles, brevemente descritos a seguir, mostraram-se mais relevantes para o problema em questão.

O primeiro atributo é o mínimo da envoltória do resíduo. Para chegar a esse atributo, inicialmente subtrai-se do sinal degradado o sinal de referência. Em seguida, calcula-se a envoltória de energia deste resíduo, e por fim encontra-se o menor valor por ela assumido. Este atributo é apropriado para a detecção de eco, cuja ocorrência isolada produz um valor próximo de zero.

O segundo atributo é a curtose [35] do resíduo (o mesmo obtido para a obtenção do atributo anterior). A curtose mede o achatamento da distribuição do resíduo. O resíduo de sinais com ruído tende a se concentrar numa região relativamente larga em torno do zero, enquanto o de outras degradações tem um pico proeminente na origem. Em razão disso, esse atributo auxilia a detecção de ruído no sinal degradado.

O terceiro atributo é calculado a partir da divisão ponto a ponto entre as transformadas de Fourier do sinal degradado e do sinal de referência. O atributo é obtido como a razão entre as médias geométrica e aritmética da 'função de transferência' obtida, da qual mede a planura espectral (em inglês, *spectral flatness*). No caso de reverberação, a função de transferência é efetivamente um modelo da sala em que o sinal estava presente, e produzirá valores baixos para o atributo.

O quarto e último atributo é a diferença relativa entre as normas euclidianas do sinal degradado e do sinal de referência. No caso de reverberação, as duas normas costumam ser próximas; no caso de sinais degradados por ruído e por eco, a norma do sinal degradado costuma ser maior.

3.2 Banco de Sinais

Para treinar e avaliar os classificadores apresentados na próxima seção, foi necessário um banco de sinais degradados. Para gerar os sinais, foram utilizados os cenários intrusivos descritos na Seção 1.2. Especificamente, o cenário local foi utilizado para gerar sinais degradados por ruído, reverberação e reverberação com ruído; e o cenário remoto foi empregado na geração de sinais degradados por eco e eco mais ruído. Os sinais de referência foram escolhidos como frases isoladas (tiradas de um banco de 200 [27]) emitidas por 4 diferentes falantes (dois homens e duas mulheres), gravadas num estúdio profissional a uma taxa de amostragem de 48 kHz com 24 bits de precisão.

No total, foram gerados 1000 sinais corrompidos para cada tipo de degradação, com diferentes parâmetros. No caso de reverberação, foram utilizados diferentes respostas ao impulso de um conjunto de salas. Para ruído, utilizaram-se diferentes tipos de ruído e impuseram-se diversos valores de razão sinal-ruído. Já para eco, a sala remota foi variada em conjunto com o atraso e o ganho da rede de transmissão.

3.3 DTC

Dispondo-se dos atributos, restava identificar a degradação presente no sinal. Para isto, foram desenvolvidos dois classificadores supervisionados em [35]. Sendo necessário seu treinamento, dividiu-se o banco de sinais descrito na seção anterior aleatoriamente (de diversas formas) em duas partes: uma para treinamento e outra para validação. Os resultados descritos a seguir são as médias (e desvios-padrões) para 10 diferentes partições.

Nesta revisão, são mostrados apenas os resultados para o classificador *Random Forest* (RF), que é uma composição de árvores de classificação. Esse classificador foi treinado e avaliado para diferentes número de árvores utilizando os 4 atributos descritos, obtendo bons resultados independentemente do número árvores. É mostrada na Tabela 1 a matriz de confusão[2] para o classificador com 5 árvores. Como pode ser observado, no pior caso, foram erroneamente classificados apenas 0,2 % dos sinais. Resultados similares foram obtidos utilizando outro esquema de classificação.

4 DISCUSSÃO FINAL

Foi apresentada neste artigo uma breve revisão de modelos e métodos para avaliação da qualidade de sinais de fala em sistemas de telepresença. O foco foi na avaliação de sinais acusticamente degradados, já que,

[2] A matriz de confusão mede quantos sinais degradados por x são classificados como sendo degradados por y.

Tabela 1: Matriz de confusão obtida para o classificador RF com 5 árvores. Os valores exibidos estão em % e desvio padrão é exibido entre parênteses. (Verd. = Verdadeiro, Ident. = Identificado)

Verd. \ Ident.	E	R	N	E+N	R+N
E	100,0 (0,1)	0,0 (0,0)	0,0 (0,0)	0,0 (0,1)	0,0 (0,0)
R	0,0 (0,0)	100,0 (0,0)	0,0 (0,0)	0,0 (0,0)	0,0 (0,0)
N	0,0 (0,0)	0,0 (0,0)	99,9 (0,0)	0,0 (0,0)	0,1 (0,0)
E+N	0,0 (0,0)	0,0 (0,0)	0,0 (0,0)	100,0 (0,0)	0,0 (0,0)
R+N	0,0 (0,0)	0,2 (0,2)	0,0 (0,0)	0,0 (0,0)	99,8 (0,2)

em geral, esse tipo de degradação é mais difícil de ser controlada e mitigada. Também foi descrito um identificador das degradações que podem estar afetando um sinal de fala, que pode ser utilizado em conjunto com os métodos de avaliação de qualidade ou como uma ferramenta de diagnóstico.

Considerando as publicações recentes, percebe-se ainda a necessidade do desenvolvimento de métodos que operem de forma não-intrusiva sobre o sistema, permitindo a monitoração em tempo real da qualidade do sinal transmitido. Outra lacuna é o estudo de como a combinação de diferentes degradações afeta a qualidade global do sinal. E uma questão ainda em aberto é a avaliação de qualidade multimídia, i.e., como a qualidade da fala e do vídeo podem ser combinadas numa única medida da qualidade do sistema de telepresença.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] S. Bech and N. Zacharov, *Perceptual Audio Evaluation – Theory, Method and Application*, Wiley, Chichester, England, 2007.

[2] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests – A review," *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, June 2008.

[3] ITU-T, *Rec. P.10 & G.100: Vocabulary and effects of transmission parameters on customer opinion of transmission quality*, International Telecommunication Union, Geneva, Switzerland, July 2006.

[4] ITU-T, *Rec. G.107: The E-model, A Computational Model for Use in Transmission Planning*, International Telecommunication Union, Geneva, Switzerland, December 2011.

[5] ITU-T, *Rec. P.862: Perceptual Evaluation of Speech Quality (PESQ): Objective Method for End-to-end Speech Quality Assessment of Narrow Band Telephone Networks and Speech Codecs*, International Telecommunication Union, Geneva, Switzerland, February 2001.

[6] ITU-T, *Rec. P.563: Single-ended Method for Objective Speech Quality Assessment in Narrowband Telephony Applications*, International Telecommunication Union, Geneva, Switzerland, May 2004.

[7] S. Möller and A. Raake, "Telephone speech quality prediction: Towards network planning and monitoring models for modern network scenarios," *Speech Communication*, vol. 38, no. 1, pp. 47–75, January 2002.

[8] ITU-T, *Rec. P.862.2: Wideband Extention to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, International Telecommunication Union, Geneva, Switzerland, November 2007.

[9] N. Côté, V. Gautier-Turbin, and S. Möller, "Influence of loudness level on the overall quality of transmitted speech," in *123rd AES Convention*, New York, USA, October 2007, Preprint 7175.

[10] J. G. Beerends, A. Hekstra, A. Rix, and M. Hollier, "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part II – Psychoacoustic Model," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, October 2002.

[11] R. Appel and J. G. Beerends, "On the quality of hearing one's own voice," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 237–248, April 2002.

[12] J. G. Beerends and J. M. van Vugt, "An extension of PESQ for assessing the quality of speech degraded by severe time clipping and linear frequency response distortions," Presented at the Joint Conference of the German and French Acoustical Society, March 2004.

[13] J. G. Beerends, B. Busz, P. Oudshoorn, J. van Vugt, K. Ahmed, and O. Niamut, "Degradation decomposition of the perceived quality of speech signals on the basis of a perceptual modeling approach," *Journal of the Audio Engineering Society*, vol. 55, no. 12, pp. 1059–1076, December 2007.

[14] ITU-T, *Rec. P.863: Perceptual objective listening quality assessment*, International Telecommunication Union, Geneva, Switzerland, January 2011.

[15] H. Aoki and A. Takahashi, "Analysis of relationship betweeen overall quality and psychological factors affecting high-quality speech communication services," in *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, March 2005, vol. 1, pp. 57–60.

[16] F. R. Ávila, L. W. P. Biscainho, L. O. Nunes, A. F. Tygel, B. Lee, A. Said, T. Kalker, and R. W. Schafer, "A teleconference model with acoustic impairments suitable for speech quality assessment," in *Anais do XXVII Simpósio Brasileiro de Telecomunicações*, Blumenau, Brazil, October 2009, SBrT, Paper 58318.

[17] L. O. Nunes, F. R. Ávila, L. W. P. Biscainho, B. Lee, A. Said, and R. W. Schafer, "Monitoring scenario for non-intrusive quality assessment of teleconference systems," in *Anais do XXIX Simpósio Brasileiro de Telecomunicações*, Curitiba, Brazil, October 2011, SBrT, Paper 85448.

[18] ITU-T, *Rec. P.56: Objective Measurement of Active Speech Level*, International Telecommunication Union, Geneva, Switzerland, December 2011.

[19] ITU-T, *Rec. P.800: Methods of Subjective Determination of Transmission Quality*, International Telecommunication Union, Geneva, Switzerland, August 1996.

[20] B. C. Bispo, P. A. A. Esquef, L. W. P. Biscainho, A. A. de Lima, F. P. Freeland, R. A. de Jesus, A. Said, B. Lee, R. W. Schafer, and T. Kalker, "EW-PESQ: A quality assessment method for speech signals sampled at 48 kHz," *Journal of the Audio Engineering Society*, vol. 58, no. 4, pp. 251–268, April 2010.

[21] L. W. P. Biscainho, P. A. A. Esquef, F. P. Freeland, L. O. Nunes, A. F. Tygel, B. Lee, A. Said, T. Kalker, and R. W. Schafer, "An objective method for quality assessment of ultra-wideband speech corrupted by echo," in *Proceedings of the 2009 IEEE International Workshop on Multimedia Signal Processing*, Rio de Janeiro, Brazil, October 2009, Paper 149.

[22] A. Rix, M. Hollier, A. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part I – Time-Delay Compensation," *Journal of Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, October 2002.

[23] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, Emerald, Bingley, UK, 5th edition, 2003.

[24] ITU-T, *Rec. P.862.1: Methods for Objective and Subjective Assessment of Quality: Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO*, International Telecommunication Union, Geneva, Switzerland, November 2003.

[25] ITU-R, "Rec. BS.1387.1: Method for objective measurements of perceived audio quality," November 2001.

[26] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ—the ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, January/February 2000.

[27] J. A. Moraes, A. Alcaim, and J. A. Solewics, "Freqüência de ocorrência dos fones e lista de frases foneticamente balenceadas no português do Rio de Janeiro," *Revista da Sociedade Brasileira de Telecomunicações*, vol. 7, no. 1, pp. 23–41, Dezembro 1992.

[28] EBU Technical, *EBU Tech.3276: Listening conditions for the assessment of sound programme material*, European Broadcast Union, Geneva, Switzerland, 2nd edition, May 1998.

[29] A. A. de Lima, T. de M. Prego, S. L. Netto, B. Lee, A. Said, R. W. Schafer, T. Kalker, and M. Fozunbal, "On the quality-assessment of reverberated speech," *Speech Communication*, vol. 54, no. 3, pp. 393–401, March 2012.

[30] T. de M. Prego, A. A. de Lima, S. L. Netto, B. Lee, A. Said, R. W. Schafer, and T. Kalker, "A blind algorithm for reverberation-time estimation using subband decomposition of speech signals," *Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2811–2816, April 2012.

[31] L. O. Nunes, F. R. Ávila, A. F. Tygel, L. W. P. Biscainho, B. Lee, A. Said, and R. W. Schafer, "A parametric objective quality assessment tool for speech signals degraded by acoustic echo," *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, To appear.

[32] P. A. A. Esquef, L. W. P. Biscainho, L. O. Nunes, B. Lee, A. Said, T. Kalker, and R. W. Schafer, "Quality assessment of audio: Increasing applicability scope of objective methods via prior identification of impairment types," in *Proceedings of the 2009 IEEE International Workshop on Multimedia Signal Processing*, Rio de Janeiro, Brazil, October 2009, Paper 148.

[33] H. Yuan, T. H. Falk, and W.-Y. Chan, "Classification of speech degradations at network endpoints using psychoacoustic features," in *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, Vancouver, Canada, April 2007, pp. 1602–1605.

[34] H. Yuan, T. H. Falk, and W.-Y. Chan, "Degradation-classification assisted single-ended quality measurement of speech," in *Proceedings of the INTERSPEECH*, Antwerp, Belgium, August 2007, pp. 1689–1692.

[35] L. O. Nunes, L. W. P. Biscainho, B. Lee, A. Said T. Kalker, and R. W. Schafer, "Degradation type classifier for full band speech contaminated with echo, broadband noise, and reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2516–2526, November 2011.

NUNES, L. O.; BISCAINHO, L. W. P. Revisão de métodos de avaliação de qualidade de voz para sistemas de telepresença.
In: CONGRESSO DE ENGENHARIA DE ÁUDIO AES BRASIL, 10., 2012, São Paulo, *Anais*... Rio de Janeiro: Sociedade de Engenharia de Áudio, 2012. p. 28-39.

SESSÃO 1 / *SESSION 1 - PAPERS 1*

Foto:SXC

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Designing bodiless musical instruments

Jose Fornari
Núcleo Interdisciplinar de Comunicação Sonora – NICS / UNICAMP
Campinas, São Paulo, 13083-872, Brasil

tutifornari@gmail.com

ABSTRACT
A common feature found in most musical instruments is the usage of a physical interface to retrieve musician gestures. This is found from traditional musical instruments (e.g. a piano that retrieves gestures throughout keyboards and pedals) to electronic musical instruments, with their digital interfaces, (e.g. MIDI controllers). This article introduces a new type of computer music instrument that does not require a physical (touchable) interface to mediate gestural retrieval. Rather, it is controlled directly by musician's movements without any physical part to be touched, grasped or attached. They are here named as Bodiless Musical Instruments (BMIs). This article presents two BMIs and discusses their computational implementation, artistic performance and aesthetic achievement.

0. MOTIVATION AND BACKGROUND

Every musician knows the challenge involved in adapting his/her body to the mechanics of a musical instrument. For some, this can turn into an ordeal. There are countless clinical cases of injuries among musicians, associated with the repeated (and often erroneous) practice during extended periods of time, specially when the musician is eager to overcome a technical challenge quickly. While most of them will find their way to interact with the musical instruments without undergoing any major physical conditions, some will inevitably quit practicing, feeling unable to fully adapt to their musical instrument of choice. This is studied by Ergonomics; the field concerned with the understanding of interactions among humans and other elements of its quotidian environment. Ergonometrics aims to decrease deleterious effects of Repetitive Strain Injuries (RSI), which musicians and other performers can develop over time, and eventually lead to long-term disabilities.

Nonetheless, musical proficiency will inevitably require endless hours of repetitive practice. This involves a cognitive process of gestural learning. Some researchers suggest that it takes about ten thousand hours of practice to become an expert in any subject, such as mastering a musical instrument [1]. It is possible that great part of this time is required to develop the specific gestures for the musician to physically adapt to the musical instrument.

There are also cases of people with physical disabilities, acquired or inborn, that are unable to develop their musicality in traditional musical instruments, in a traditional manner. Of course, there are people, such as the musician Mark Goffeney[1] that, in spite of having no arms, found a way of fluently playing a traditional guitar. However, amazing cases like that one are very rare. For most people with a physical or cognitive disability, playing a traditional musical instrument is out of reach, Being able of making music is known as an activity that delivers entertainment, joy and stress relieve. It would certainly be enticing for disabled people to access a type of musical instrument that adapts to their range o gestures, instead of the other way around. For this purpose, a musical instrument without a "body" of its own (as further described) can be helpful. As each physical and cognitive restriction is unique, such new type of musical instruments has necessarily to be physically malleable or even bodiless – which means, without a touching gesture interface – seems to be a good choice.

Computer technology has allowed the creation of gestures interfaces (GI), interfaces able to capture body movements in real-time and retrieve them as a stream of

1 http://footguitar.com/

digital data. The entertainment industry has lately provided inexpensive GIs, such as the Wii remote-control; a wireless control that retrieve gamers' movements retrieved through motion sensors, such as: a tridimensional accelerometer, an infra-red camera, and accessories containing gyroscopes. Some of these equipments, have already been explored as musical interfaces, to control sound synthesis parameters in real-time [2]. However, these GIs are also material pieces of hardware. They have the similar requirements and limits, as any traditional musical instrument. The users still have to manipulate them and, by doing so, they need to adapt their gestures to control the GI. A successful experiment, first-prize awarded, in February of 2012, at Georgia Tech's 2012 Guthman Musical Instrument Competition[2] is: "*Xth Sense*", defined as a biophysical, wearable technology that reacts to the vibrations of muscle tissue [3]. *Xth-Sense* sensors are microphones attached to the performer's forearms that read muscular vibrations (5 – 100 Hz), used to control in real-time a sonification process [4]. This is still a virtual musical instrument using touchable interfaces. On the other hand, a GI that does not require any type of physical contact with the performer's body may help the artist to freely explore their own personal gestures. It can even enable the creation of virtual musical instruments that adapt to the gesture of the performer, instead of requiring the performer to adapt to the instrument.

This article describes the technical aspects of two performances of sonic art . They use motion capture algorithms as Touchless Interfaces (TIs) to control sound synthesis models in real-time, as a mean of creating Bodiless Musical Instruments (BMI); virtual musical instruments that can be played without any contact with the performer's body. Here is presented two performances that were developed with this philosophy. The first one is "tIRAtEIMAS"[3] presented at EIMAS 2010 (*Encontro Internacional de Música e Arte Sonora*). The second one is "A Pedra", performed at SBCM 2011 (*13º Simpósio Brasileiro de Computação Musical*). Each one uses a different strategy of motion capture to retrieve the performer's gestural data in real-time. Such data is used to control the dynamic creation of sonifications, throughout two real-time sound synthesis models.

BMIs depend on TIs, as their interfaces to capture gesture data without the need of having the performer to grasp, hold, or wear any piece of hardware. Nevertheless, there is a sensor that the performer is interacting with, even though at distance. Probably the first akin, and most famous one, is Theremin [5], the electronic musical instrument created at the beginning of the twentieth century by Lev S. Termen. This is an electronic musical instrument with two antennas that independently control – normally by the proximity of hands – a sinewave amplitude and frequency. The performer is supposed to avoid touching the antennas, while keeping the attention focused on the distance with them. Theremin is now an icon of the avant-garde electronic music.

There have been several studies on TIs, such as the system to recognize shapes, positions and rotations of hands, through Computing Vision (CV) [6], a TI for general gesture data acquisition [7] and a real-time visual recognition system that enables a human conductor to control a virtual orchestra using gestures retrieved from the tracking of traditional conductor's baton [8].

This works aims to introduce the concept of BMI through the example of two performances. One of BMI implementation goals is to release the performers from wearing a sensor, which necessarily holds part of their attention to the retrieving gestural data process, instead of letting the performer to focus on gestural parameters, such as the variation and control on the effort of forearms muscles (as in the case of Xth Sense), or any sort of touchable interface. BMIs may free the performers attention to sensors bodily feedbacks, so they can explore new gestures as if "playing into the air", without focusing on the gestural data acquisition process. This, of course, is an ideal goal, which is most surely not fully achieved by the performances presented here. This article solely intend to the bring attention to this subject and present two introductory performances whose goal is towards the creation of BMIs.

Both motion capture algorithms and sound synthesis models presented in the performances, were programmed in PureData (PD)[4], the well-known open-source software for real-time processing and control of sound and image. PD is a multi-platform software, designed to enable the development of multimedia real-time performances. This work presents two motion capture strategies. They are here used for the real-time control of the sound synthesis models, so the performer can intuitively control complex sound generation processes by practicing and learning new gestures, without being restricted or guided by any physical interface.

The next section describes two methods of touchless gestural data retrieval. Following that, section 2 describes the real-time sound synthesis models that are controlled by the gestural data retrieved. Section 3 describes the type of touchless interaction for each performance. Section 4 describes this two performances in details. Section 5 concludes this work with a discussion on the achieved results and future developments in the BMI subject.

1. BODILESS RETRIEVAL OF GESTURES

The human brain is able to perform extremely complex tasks at ease. One of them is the coordination of body movements. Gestures are here defined as movements with a particular intention. Through meticulous gestures, musicians express their art. During performance, this gestural communication occurs dynamically and in real-time. They can be guided and triggered to other senses cues. For instance, it is usual that, during improvisations, musicians anticipate the actions of their colleges just by a glimpse in their movement patterns or facial expressions. Classical conductors use to overdo body expressions as a strategy of broadcasting, to the musicians in the orchestra (but also perceived by the audience), the musical meaning of a passage.

There are several examples in the literature of studies using different types of interactions and movement sensors applied to musical expressivity. They are known by different names, such as: Human Computer Interface (HCI), Haptic Interaction (HI), Input Devices (ID) and, the already mentioned, Gesture Interfaces (GI). To name a few, in [2] a virtual "Berimbau" (the single-string African-Brazilian percussion instrument) was implemented, together with other virtual percussions (shaker and bell). They were implemented in PD and controlled by a real-time streaming of gestural data retrieved by two motion-sensors: an Infra-Red tracking camera and a tridimensional accelerometer, both from the Wii remote-control. This virtual Berimbau allowed the exploration of extended musical techniques (impossible to be achieved by a real Berimbau). In [9], BodyMusic is introduced; a cyber instrument that enables musicians to interactively and intuitively control musical tonality and melody, through body motions. In [10] is presented a computer model that uses the same Wii remote-control as a GI to capture movements data in real-time of a conductor.

This work describes two TI implementation methods for real-time gestural retrieval, implemented as PD algorithms (patches) and using as movement sensor the laptop webcam. The first one uses, as method, the frame

2 http://www.gtcmt.gatech.edu/guthman2012

3 http://www.youtube.com/watch?v=T4o32cG-A5Q

4 http://www.puredata.info

differentiation of consecutive frame subtraction. The second method uses real-time image filtering. Both methods use PD objects from the Graphics Environment for Multimedia (GEM) library. The first method uses the object "pix_movement". This one subtracts consecutive frames from the digital video streaming, acquired in real-time by the webcam. This object has a threshold rate argument that goes from zero to one, and sets the pixel blob size of the frame subtraction, which sets the movement detection sensibility for image noise. The second object is "pix_blob". This one calculates the centroid for the biggest pixel blob, in the image plan of coordinate (x, y). Figure 1 shows the patch for this algorithm. This one was used in the first performance here mentioned, "*tIRAtEIMAS*".

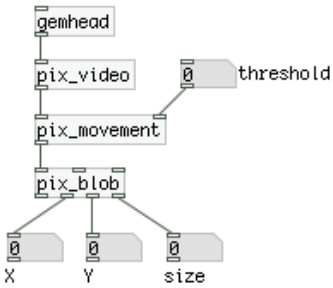


Figure 1. PD patch for single track movement detection.

The second algorithm for movement retrieval uses the PD/GEM object "pix_movement2". This one embeds a timebase IIR-filter algorithm for tracking the movement of multiple pixel blobs simultaneously. It detects these pixel variations for each frame, with respect to the corresponding two previous frames, and a chosen background image. In order to minimize the disturbance of image noise, during motion tracking, this object ignores changes between two corresponding pixels blobs that are smaller than a certain size, specified by its "threshold" parameter. This is initially set by the user but adjusts itself dynamically along its operation, moving towards the rate specified in the second threshold; the "lo_threshold", that sets the absolute minimum threshold for this object. Figure 2 shows the PD patch for this algorithm. This patch is part of the second performance here presented: "*A Pedra*".

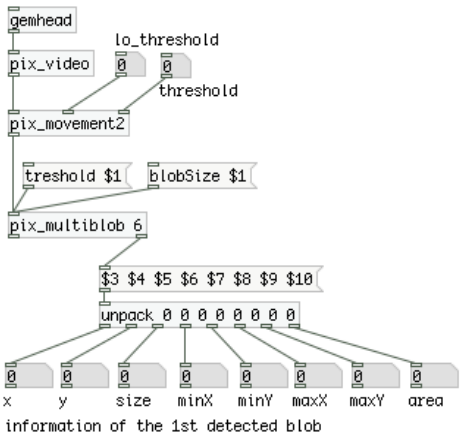


Figure 2. PD patch for multiple track movement detection

2. REAL-TIME SOUND SYNTHESIS MODELS

Sound synthesis methods can be classified in two categories: Linears and Nonlinears. Linear are the ones based on the Fourier series. Their generated sound can be mathematically represented by a system of linear (independent) equations. Nonlinear methods are given by nonlinear equations, which are not reducible to independent expressions. Some nonlinear models may also be described as systems with feedback. Linear methods of sound synthesis normally allow full and independent control over all partial components of sounds. However, its complexity – as well as the required computer processing – increases in proportion with the complexity of the sound generated. Nonlinear methods, on the other hand, can generate complex waveforms with less computational processing, but without the ability of independently controlling each parameter in the sound generation process, which may lead to unpredictable perceptual results. The design of sounds using nonlinear methods are mostly empirical; developed by means of trial-and-error. Classical types of Linear methods are the Additive and the Subtractive syntheses. Examples of Nonlinear methods are: the FM (frequency modulation) and PM (physical modeling) syntheses [11].

The performance "*tIRAtEIMAS*" uses several FM synthesis models that are parametrized by gestural data retrieved by a movement detected patch similar to the one presented in Figure 1. Figure 3 shows the patch of the basic structure for one of the eleven FM synthesis models used in this performance.

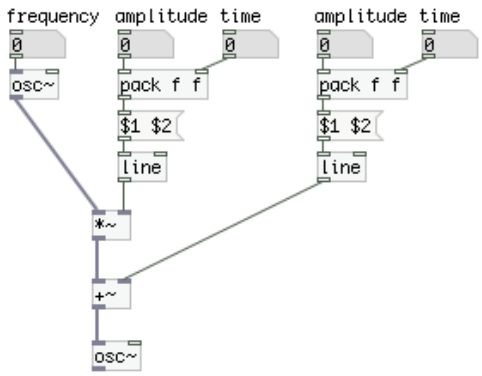


Figure 3. PD patch of one of the FM syntheses models used intIRAtEIMAS.

The great richness of sounds generated by Nonlinear sound synthesis methods, together with its low computational processing requirements, make them particularly interesting to be used in real-time computer music performances. However, its non-intuitive controllability is sometimes a constrain for its usage in sound design, specially the ones aiming to achieve a specific sonification or intending to emulate a known sound source (e.g. the sound of a known musical acoustic instrument). This is probably the reason why nonlinear methods, such as the FM synthesis, were lately replaced by editing synthesis methods, such as the Wavetable [12].

Computer technology has unlocked the access for the control of several and simultaneous aspects of sound synthesis. However, sounds have a myriad of components to be controlled, and usual HCI devices (as the computer mouse and keyboard), are only able to control a very limited amount of these parameters simultaneously. In fact, any gestural interface will somehow limit the amount of

parameters that can be simultaneously controlled, which is, by far, smaller than the number of actual parameters that can be accesed to be controlled, in any sound synthesis model. There are several researches on the study of adaptive systems to control the non-intuitive parameters of nonlinear sound synthesis. There are studies using Evolutionary Algorithms to find optimal parametrizations of FM synthesis models [13], [14]; as well as the method known as Evolutionary Sound Synthesis, that generates a dynamic population of waveforms (its individuals) that undergo processes of Reproduction and Selection [15]. However, these methods of parametrization are not necessarily intended for real-time HCI sound design, which seems to be the reason why they are still rarely used in musical performances. On the other hand, the brain ability of controlling all body movements parameters, simultaneously and in real-time, is a successful example of a natural process able to dynamically control a large number of complex parameters effortlessly. For most of our routine movements, regardless their complexity, its control is performed by the motor cortex brain region, that simultaneously handle the position and movements of several body parts, such as: head, spine and limbs, so our body can even express artistic meaning throughout fine and precise gestures [16].

This work aims to use this brain capacity of movement coordination to control sound synthesis models throughout the data retrieved by TIs, thus creating a BMI. By not requiring contact with the performer's body, this virtual musical instrument let its performer to freely explore gestures that, intuitively and ludic, can control a large number of parameters of real-time sound syntheses models. This way, it is possible to not only use nonlinear sound synthesis method, but also linear ones, with a large number of parameters, such as the ones based on the Fourier series, to generate real-time sonifications. Figure 4 shows the PD patch for the Linear sound synthesis used by the second performance here presented: "*A Pedra*".

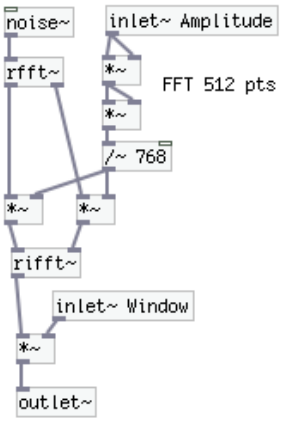


Figure 4. PD patch of a FFT decomposition of white noise, followed by IFFT resynthesis, creating a Linear Subtractive synthesis model.

3. PLAYING WITH PURE GESTURES

Once upon a "computer music" time, sound synthesis models would require several hours of a mainframe computer to be processed. Nowadays, sounds can be easily generated in real-time by computer models running in laptops. Software environments, such as PD, allow the exploration of many possibilities of sonifications, by connecting a few audio processing objects (as seen in Figures 3 and 4). In consequence, the constriction has migrated from the "ability of processing" to the "ability of controlling". The more complex is a sound synthesis model, the more controlling parameters it will have. For instance, in "*tIRAtEIMAS*" there are 11 real-time FM synthesis models. As seen in Figure 3, each model has 5 parameters of control. Thus, to fully control the entire patch, it is needed to control 55 parameters. Together, they can generate rich and unexpected sonifications, but their control can not be done merely by a computer mouse and keyboard. For this model, the performer control these parameters through the gestures of hands, retrieved by the webcam, and learn, by repetition, the correlation between gestures and the resulting sonic textures, so the artist can intuitively control the sonification process to create the intended result. Figure 5 shows the mapping of regions in the space (frontal plan) done by the PD patch of the first performance. For each region, there is a FM synthesis model associated with it.

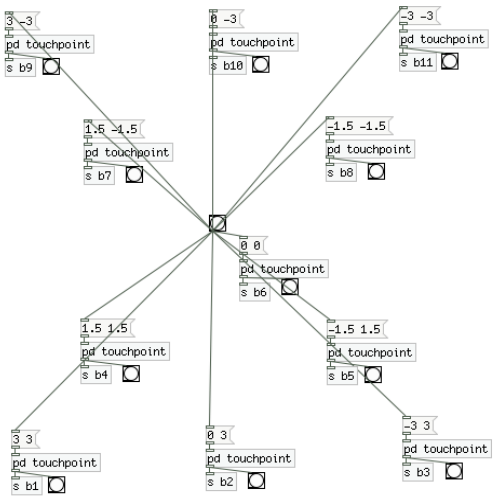


Figure 5. Space mapping of gestural data, in "*tIRAtEIMAS*".

In the second performance, the patch uses a Subtractive synthesis, as seen in Figure 4. This one has a white-noise generator, given by the object "noise~". This is the source of acoustically rich but perceptually meaningless sonifications. This raw material is manipulated be the gestural data retrieved by algorithms such as the one presented in Figure 2. Throughout it, the white noise is "sculpted" in real-time. Gestural data control the dynamic filtering of its spectral magnitude, done by the product of its source with a cosine window, given by the object "inlet~ Window". The dynamic control of these parameters, controlled by the gestures of the performer, allow the intuitive and ludic exploration of new sonorities. Figure 6 shows a moment of the dynamic spectral gain manipulation, that is used to sculpt the white-noise, thus generating the sonifications of these performance.

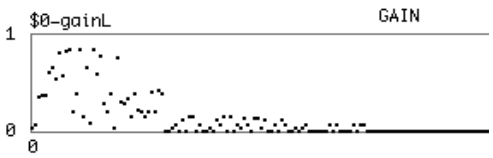


Figure 6. A moment of the spectral gain array, during the performance "*A Pedra*".

4. THE PERFORMANCES

These two performances were implemented in PD-extended, version 0.42. The PD patches run in a laptop with webcam, which is used as the real-time gestural data retrieving sensor. In tIRAtEIMAS, the patch uses the webcam to map 11 regions in the space, in front of the performer, each one controlling the parameters of a two-oscillators FM synthesis models shown in Figure 3. The patch also process video in real-time, as captured by the webcam, that was projected on the roof, during this performance. The real-time video shows images acquired by the webcam, of the performer's gestures, and displays it with dynamic visual effects, together with a moving circle, representing the single-point movement tracking of the gesture. Each time the performer reached a mapped region of the space, as shown in Figure 5, the circle moves towards it, in the projection, underlining which FM synthesizer is currently active. The performer plays this patch by two TIs controls: 1) the webcam, that captures the movement of the hands across these mapped region, in the virtual plane of Figure 5, and 2) the microphone, a sonic input control, that captures percussive onsets (i.e. the performer clapping hands or snapping fingers) to reindex the 11 synthesizers parameters, according to these onset sonic characteristics.

The second performance here presented, "*A Pedra*", uses the algorithm shown in Figure 2, to retrieve multiple movements threads simultaneously. Figure 7 shows a moment of the real-time animation of this performance. The two small objects here seen (a green circle and a "cloud of squares") represent two different movements being tracked simultaneously, which are, in this case, each hand of the performer. Their movement is associated with the shape of the spectral gain windows, as shown in Figure 6. There are two of these windows; one for the left hand and another for the right one. The major object shown here keeps slowly rotating, resembling a planetary movement, while its surface flutters according to the recorded voice declaiming the poem described below.

Figure 7. A moment in the dynamic animation of "A Pedra"

This performance was inspired in a poem written by the Brazilian poet Carlos Drummond de Andrade (1902-1987), entitled "*No Meio do Caminho*", translated by Elizabeth Bishop as "In the Middle of the Road". This is an important poem, considered as foundational for Brazilian modernism and, as stated by Charles Bernstein "beyond that, for the poetics of the Americas". This performance starts with the voice recording of Drumond declaiming his own poem. It follows a sonification process controlled by the performer's gestures that creates an artificial soundscape as an interactive sonic artwork.

Besides the creation of BMIs, these performances also advocate for a new concept for Musical Notation. It is here argued that a computer model, such as the ones presented in Figures 1 to 5, can also be seen as a form of notating music, specially for computer music and/or any form of computer-aided sonic art. Any type of Musical Notation is in fact a protocol to create a musical algorithm that, once run (by a machine or a human performer) will bring about the actual music piece, as an aural experience. Therefore, musical notation is not music in itself, but a procedural programming for its creation. In this sense, PD can be seen as a canvas to notate music. Creating a PD patch is similar of composing music (or any sonic art). Here, considering the patch as notation, the gestural control corresponds to its performance, which is bound by a human interpretation , to certain extent, similar to a guided improvisation. One of the first researches, in the records, on the usage of a computer code as musical notation is the "Visual Listening Scores", by Rainer Wehinger, developed for the musical piece "Artikulation" of György Ligeti [17]. Since then, many other artists have developed code as musical notation, even without being aware of (or mentioning it), such as ome of the works known by the term "Algorithmic Composition". Some of them also created computer models of real-time processing to represent the structure of their musical pieces. A typical example is shown in Figure 8, which presents an excerpt of the piece *Solitude2*, by Hans-Christoph Steiner, composed in PD, in 2004.

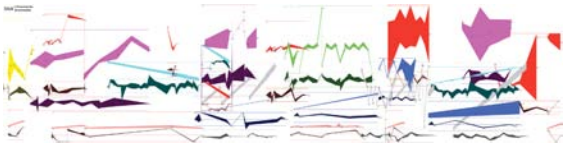


Figure 8. Excerpt of the graphical musical notation Solitude, from Hans-Christoph Steiner, written in PD.

The graphical score shown in Figure 8 resembles the natural extension of the musical notation, as proposed by Wehinger, but with two advantage: 1) Being fully compounded of this new form of music notation (i.e. the computer code given by graphic notation), 2) Being written and displayed in a computer screen, which gives full dynamism and plasticity for such structure (as opposed to a score written in paper, which can not be changed in real-time). This type of score provides the procedural steps of the computational model of the digital processing of a pre-recorded sound material. The steps are described as PD structures, using different colors, to represent distinct procedural transformations of the sound source. The timeline is given from left to right, following the same orientation of any typical magnitude spectrogram, or traditional (western) musical notation score. It is possible that the cooperation of BMIs, as the ones presented here, with new forms of virtual musical notations, made of computer code, may represent the future developments of music, still to be achieved, as a contemporary art that is now, more than ever, composed, analyzed and processed with computers.

5. CONCLUSION

This article explained that many (real or virtual) musical instruments and/or electronic gestural interfaces has to be in contact with the performer's body, in order to capture gestures, retrieving the data that will control a real-time sonification or music making process. For those, the performer's (human) body has to be in contact with the (mechanical) body of a traditional musical instrument, or a piece of hardware equipment, such as the computer mouse, keyboard, touchpads, and so forth. This type of musical human-machine interaction requires the performers to train their gestures to adapt to the machine, instead of having the machine adapting to the artist's natural gestural behavior. Such condition may limit or inhibit the performance of

many, specially the ones unable to fit their gestures style to the requirements of the musical interface. Nevertheless, there are other forms of motion capture that do not depend on body contact. They were called here "Touchless Interfaces" (TIs). This article presented two performances using TIs that aimed to create interactive sound art presentations. Both were based on real-time motion capture algorithms that retrieve gestures from the performer without any physical contact. This way, the performer is not required to wear, hold or touch any material piece. The retrieved gestural data control real-time sound synthesis models, together creating a touchless computer musical instrument. They were named "Bodiless Musical Instruments" (BMIs). The performances presented here were produced by two simple BMIs where the gestures were improvised to create dynamic sound textures for interactive artistic performances.

These two Sound Art performances use simple gestural retrieval algorithms of Computer Vision (CV), to control in real-time computer models of sound synthesis. The goal was to introduce the concept of a Bodiless Musical Instrument (BMI); a computer music virtual instrument controlled by the performer's gestures without any physical contact to material interfaces. There are several successful efforts to acquire in real-time, gestures, by means of CV, using software environments such as: Processing, or OpenCV. However, this work does not focus on the techincal details of CV subject, but in its usage for the creation of BMIs. Although CV is in fact an efficient strategy to implement TIs, this is only one of the possible approaches to create a TI. Other works have successfully implemented TIs with other techniques, (such as Theremin, described in the introductory section of this article). Here, the goal was to introduce and exemplify the creation of a computational environment for the artistic exploration of bodiless interfaces controlled through gestural sonic art installations. Further BMIs implementations shall aim to create an environment that allow a performer of any artistic field to explore and develop gestures associated with sound objects that allow this artist the free exploration of sonifications, through his/her own set of personal gestures, thus creating soundscapes out of performances, led by the intuitive development of a dynamic process of multimodal performance.

Although TIs may free the performer to extend the gestures repertoire, they normally lack to deliver touch sense, or tactile feedback. This can make the interaction initially awkward, specially for performers used to handle typical musical instruments, with lots of tactile feedback (e.g. a trumpet player feels the vibrations patterns of the instrument on fingers, lips, throat, and so forth) . In BMIs, the absence of vibration, pressure and the constant feeling of weightless movements, can make some performers feeling adrift. However, they resemble the gestural interaction of a conductor (maestro) with the orchestra. During performance, the conductor interacts with the orchestra mostly by means of touchless gestures, that don't follow a rigorous pattern. During rehearsals, the conductor train the musicians in the orchestra to understand and recognize this/her gestural repertoire, and relate them with specific musical meanings. Further works can explore adaptive methods, such as the one in [15], to emulate the recognition of gestures acquired by TIs, so the computer model is able to learn the gestural intention of the performer and adapt to it. Both performances presented here are still to fully explore tridimensional gestural information retrieval, which, in order to be achieved, might require the usage of two (or more) webcams (to retrieve gestural data from different angles, thus creating 3D data). Further works on this subject shall develop the retrieval of tridimensional movements, specially to interact with artists from other fields, such as dancers, and actors, thus controlling in real-time the dynamic multimodal processes of sonifcation for their visual art performances, created by bodiless musical instruments.

6. ACKNOWLEDGEMENTS

This work was financially supported by FAPESP, process: 2010/06743-7.

7. REFERENCES

[1] Levitin D. This is your brain on music: The science of a human obsession. Dutton/Penguin Books. 2006.

[2] Fornari, J., Gestural Controlled Virtual Percussion Instruments. SBGames 2010. IX Simpósio Brasileiro de Jogos e Entretenimento. 08-10 November. Florianópolis - SC. 2010. ISSN: 2179-2259.

[3] Donnarumma M. (2011) "Xth Sense: researching muscle sounds for an experimental paradigm of musical performance" Proceedings of the Linux Audio Conference (LAC'11).

[4] Hunt, A., Hermann, T., Neuhoff, J. G. (2011) The Sonification Handbook. Logos Publishing House, Berlin. 2011, 586 pages. ISBN 978-3-8325-2819-5.

[5] Glinsky, A., Theremin: Ether Music and Espionage. Urbana, Illinois: University of Illinois Press. 2000.

[6] Tarabella, L., Gestural and Visual Approaches to Performance, in Wanderley, M., & Baffier, M. (Eds.), Trends in Gestural Control of Music CD-ROM, IRCAM, Paris, 2000, pp. 604—615.

[7] T. Ishikawa, Y. Horry and T. Hoshino, Touchless input device and gesture commands, Proc. Int'l Conf. on Consumer Electronics, 2005, pp. 205-206.

[8] Sebastien Schertenleib, Mario Gutierrez, Frederic Vexo, and Daniel Thalmann. 2004. Conducting a Virtual Orchestra. IEEE MultiMedia 11, 3 (July 2004), 40-49. DOI=10.1109/MMUL.2004.5.

[9] Horace, H. S., Kwong, B., Law, K. C. K., 2005. Bodymusic: a novel framework design for body-driven music composition. In ACE'05: Proceedings of the 2005 ACM SIGCHI. International Conference on Advances in Computer Entertainment Technology, pages 342-345, New York, NY, USA. 2005.

[10] Peng, L., Gerhard, D., A Wii-based gestural interface for computer-based conducting systems. International Conference on New Interfaces for Musical Expression. 2009.

[11] Cook, P. R., Real sound synthesis for interactive applications". A K Peters, Ltd. 263 pages. ISBN: 978-1-56881-168-0. 2002

[12] Moore, F. R., Elements of Computer Music, pp. 159-169, Prentice-Hall, Englewood Cliffs, NJ. 1990.

[13] Garcia, R., Growing Sound Synthesizers using Evolutionary Methods. Proceedings of ALMMA 2002 Workshop on Artificial Models, 2001.

[14] Horner, A., Beauchamp, J., Haken, L., Machine Tongues XVI: Genetic Algorithms and Their Application to FM Matching Synthesis. Computer Music Journal, Vol. 17, No. 4, pp. 17-29. Winter 1993.

[15] Manzolli, Fornari, J., Maia Jr., A., Damiani, F., The Evolutionary Sound Synthesis Method. MULTIMEDIA '01 Proceedings of the ninth ACM international conference on Multimedia. ACM New York, NY, USA. ISBN:1-58113-394-4. 2001.

[16] Hermann, T., and Andy Hunt. The Importance of Interaction in Sonification. Proceedings of ICAD Tenth Meeting of the International Conference on Auditory Display, Sydney, Australia, July 6–9, 2004.

[17] Lerdahl, Fred (1988). Cognitive Constraints on Compositional Systems, Contemporary Music Review 6 (2), pp. 97–121.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Utilização de um sistema composicional espectral na criação de Anticlepsidra

Raphael Sousa Santos[1] e Liduino José Pitombeira de Oliveira[1]

[1] Universidade Federal de Campina Grande, Unidade Acadêmica de Arte e Mídia
Campina Grande, PB, 58429-140, Brasil

raphaelss@gmail.com, pitombeira@yahoo.com

RESUMO

Este artigo tem como objetivo definir e descrever um sistema composicional que se utiliza de procedimentos espectrais para a geração de estruturas harmônicas e duracionais, as quais são empregadas no planejamento de *Anticlepsidra*, para violino, clarinete e piano. Essas estruturas são geradas a partir de aplicativos computacionais escritos em C++: uma calculadora de modulação em anel e um manipulador de funções, que além de gerar essas estruturas é também capaz de aplicar-lhes uma evolução temporal.

0 INTRODUÇÃO

Um sistema composicional é "um conjunto de diretrizes, formando um todo coerente, que coordenam a utilização e interconexão de parâmetros musicais, com o propósito de produzir obras musicais" [1]. Essas diretrizes são subsistemas, estabelecidos a partir de uma série de definições, que modificam os dados de entrada e se relacionam entre si a partir de operações pré-determinadas pelo compositor, bem como a partir de decisões eventuais que alteram a operação do sistema. O sistema apresentado nesse trabalho foi empregado no planejamento de *Anticlepsidra*[1], para violino, clarinete e piano, e foi projetado a partir de dois subsistemas baseados em procedimentos utilizados na música espectral.

A música espectral surge na França no começo dos anos 70 com os compositores envolvidos no Ensemble l'Intineráire, mais notadamente Gérard Grisey e Tristan Murail [2]. Os espectralistas colocam o timbre como elemento principal em sua estética e tomam a estrutura interna do som como base para a composição [3]. De fato, a construção de harmonias a partir de modelos espectrais, que podem ter origem em análises de sons concretos ou baseados em técnicas oriundas da música eletroacústica é uma das mudanças mais básicas introduzidas pelos espectralistas [2].

[1] A anticlepsidra é um sólido geométrico formado a partir de um cilindro equilátero, do qual subtraímos dois cones opostos pelos vértices cujas bases coincidam com as bases do cilindro e cuja altura seja igual ao raio da base.

Os dois subsistemas utilizados no planejamento de *Anticlepsidra* foram uma calculadora de modulação em anel e um manipulador de funções, implementados como aplicativos computacionais em C++, os quais são capazes de gerar estruturas harmônicas e duracionais, além de, no caso do manipulador de funções, ser capaz de aplicar-lhes uma evolução temporal. A automatização dos cálculos, realizada por esses aplicativos, tem unicamente a função de tornar transparentes os processos formais e disponibilizar o acesso destes procedimentos a um maior número de compositores que queiram realizar experimentações espectrais. É necessário enfatizar que o sistema não realiza decisões composicionais específicas, mas apenas gera estruturas a serem trabalhadas de acordo com a visão estética particular de cada compositor.

1 MODULAÇÃO EM ANEL

A modulação em anel é uma técnica da música eletrônica e, mais especificamente, uma forma de modulação de amplitude [4]. O espectro resultante desta técnica contém, para cada par de componentes senoidais entre os sons originais, um par de senoides cujas frequências são a soma e a subtração das frequências dos parciais que as geraram. Assim, a modulação de uma senoide por outra senoide, de 50 Hz e 120 Hz de frequência, por exemplo, gera um sinal com apenas dois parciais: a soma das frequências das originais, 170 Hz, e a subtração destas mesmas frequências, 70 Hz. A Figura 1 mostra o exemplo da modulação de um sinal de 50 Hz e seus dois primeiros harmônicos por um de 300 Hz e seus dois primeiros harmônicos. Um fato a ser notado é que os componentes de frequência 450 Hz e 750 Hz são gerados cada um por dois pares de parciais distintos, resultando, assim, em uma maior intensidade em relação aos demais.

Curiosamente, diferente da modulação de amplitude tradicional, os sinais originais não estão diretamente presentes no espectro resultante e, não há, necessariamente, uma distinção entre onda portadora e onda moduladora [2]. A técnica é utilizada tanto como forma de síntese pura, como também de alteração de sons concretos. Na música espectral, a modulação em anel aparece já em obras como *Partiels* e *Modulations* de Grisey [2].

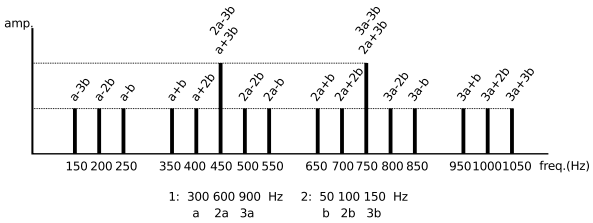


Figura 1: Espectro produzido por Modulação em Anel

2 ESTRUTURAS BASEADAS EM FUNÇÕES

2.1 Construção de Estruturas de Frequências Através de Funções

A construção de estruturas de frequências[2] pode ser realizada através de funções da forma: $y = f(x)$, onde f é a função, x é a ordem ou número do parcial e pertence ao conjunto dos números inteiros e y é a frequência resultante componente da estrutura gerada.

A Figura 2 traz exemplos de funções e seus respectivos espectros gerados – nelas, x é a ordem ou número do parcial, pertencente ao conjunto dos números inteiros, e as demais letras representam constantes reais. O primeiro exemplo traz uma função da forma $f(x) = ax$, que sempre gera um espectro harmônico cuja fundamental é a. O segundo exemplo mostra uma função da forma $f(x) = ax + b$, que gera espectros semelhantes aos gerados por síntese FM, isto é, espectros não necessariamente harmônicos[3], mas com uma distância constante (a, no caso da função) entre as frequências. O terceiro exemplo traz uma função da forma $f(x) = ax^b$, que gera espectros cuja distância entre os parciais diminui, se $b < 1$, ou aumenta, se $b > 1$, ao longo do espectro.

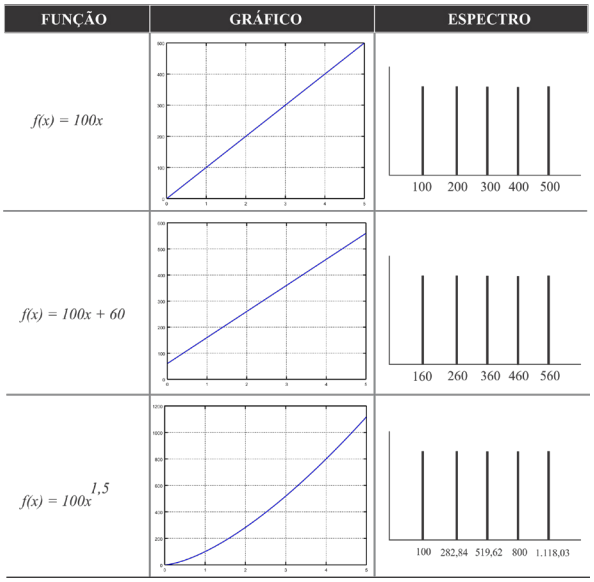


Figura 2: Espectros produzidos por três funções distintas

2.2 Construção de Estruturas de Durações Através de Funções

O mesmo procedimento utilizado na construção de espectros através de funções pode ser aplicado analogamente na construção de estruturas de durações, sejam

[2] O conceito de estrutura de frequências é largamente utilizado na literatura dedicada à música eletroacústica, como por exemplo em Windsor[5] e Wannamaker[6].

[3] Nos espectros harmônicos, os parciais são múltiplos da fundamental.

de ritmos propriamente ditos ou de durações de seções formais. Segundo Murail, no domínio das durações é fácil organizar a aparição de elementos em termos de funções. Duas abordagens básicas podem ser tomadas: a posição temporal em função do número ou ordem do evento ou a duração em função do número ou ordem do evento [7]. Utilizaremos aqui a primeira abordagem, isto é, funções $y = f(x)$ onde x é um número inteiro e representa o número ou ordem do evento e y é um número real que representa um ponto no tempo. Assim, as durações são dadas pela diferença entre estes pontos no tempo. Por exemplo, a função $f(x) = 10x$ para x de 0 a 5, resulta nos valores 0, 10, 20, 30, 40 e 50. Estes valores são pontos no tempo. As distâncias entre eles é que representam as durações. Neste caso, foram geradas 5 seções de 10 segundos, pulsos, compassos, ou qualquer outra unidade de referência. Generalizando a partir do exemplo: funções da forma $y = ax$, que no domínio das alturas gerariam espectros harmônicos, aqui geram ritmos periódicos.

Com funções simples é possível gerar *accelerandos* e *rallentandos*. Tornando essas funções mais complexas, podem ser descobertos vários tipos de flutuações para introduzir surpresas nestes processos ou para descrever padrões de organização ou desorganização duracional [7]. Grisey propõe uma escala de complexidade, da ordem à desordem, para durações [8]. A Figura 3 apresenta a escala proposta por Grisey associada a exemplos de funções que produzem estruturas duracionais pertencentes às respectivas categorias.

O primeiro nível da escala, periódico, pode ser descrito por uma função da forma $f(x) = ax$ e consiste basicamente de uma série de eventos de mesma duração. Este nível apresenta o maior grau de previsibilidade.

A segunda categoria, contínuo-dinâmico, se refere à *accelerandos* e *rallentandos* graduais e possui previsibilidade média. Grisey afirma que curvas logarítmicas são tão importantes para a percepção de durações como para a percepção de alturas [8]. Assim, curvas desse tipo produziriam as sensações mais uniformes de *accelerandos* e *rallentandos*. Isto implica que entre durações mais longas haverá diferenças maiores que entre durações mais curtas. O mesmo se observa no âmbito das alturas, entre intervalos equivalentes que produzem a mesma sensação em registros diferentes: por exemplo, a distância entre o Lá2 (110Hz) e o Lá3 (220Hz) é de 110, enquanto a distância entre o Lá4 (440Hz) e o Lá5 (880Hz) é de 440. Esses dois intervalos produzem a mesma sensação de oitava.

A categoria descontínuo-dinâmico possui baixa previsibilidade e diz respeito a *accelerandos* e *rallentandos* não graduais. Grisey coloca duas formas em que isto pode acontecer: quando há elisões, isto é, saltos, no processo ou quando há uma aceleração (positiva ou negativa) estatística [8]. Este segundo caso é o demonstrado na Figura 3. Nele, observa-se uma aceleração média ao longo do tempo com imprevisibilidade nos movimentos imediatos. Nesse exemplo, o R é uma variável aleatória, obtida por uma operação randômica, definida uniformemente em um espaço amostral que vai de -3 a 3.

A categoria estatística é totalmente imprevisível e implica em desordem máxima. Nesse caso, utilizamos uma variável aleatória definida uniformemente entre 0 e 10.

O último caso é uma categoria especial na escala de complexidade, que Grisey associa à ausência de divisão rítmica, ao completo silêncio ou a um único som [8]. Representamos essa categoria com uma função constante, onde todos os eventos iniciam em um único ponto no tempo. Como exemplo, citamos a obra *Composition 1960 #7*, de La Monte Young.

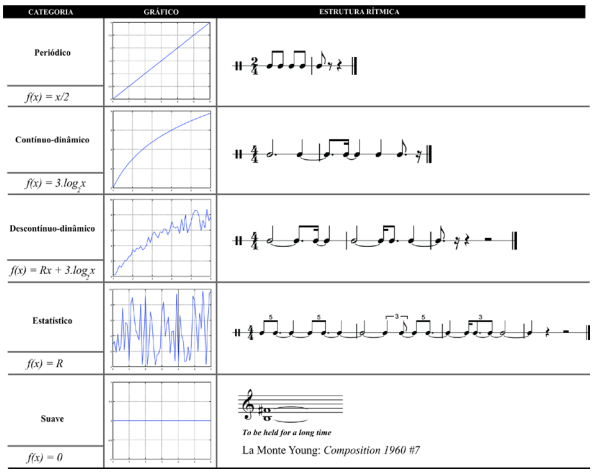


Figura 3: Escala de complexidade de Grisey

3 DEFINIÇÃO DO SISTEMA COMPOSICIONAL

Segundo Mesarovic, um sistema pode ser descrito como a transformação de um conjunto de entradas (estímulo) em um conjunto de saídas (resposta) na abordagem *input-output* [9]. Através desta abordagem, o sistema composicional aqui discutido é melhor entendido como um conjunto de subsistemas interligados. Um relativo à modulação em anel e outro a geração de estruturas de frequências e/ou de durações através de funções, e um terceiro subsistema melhor visto no contexto do sistema composicional como uma lista de definições e/ou restrições ao uso deste sistema como um todo.

3.1 Subsistema 1: Modulador em Anel

Este sistema é uma simples máquina de calcular estruturas harmônicas baseadas em modulação em anel. Dados dois conjuntos de alturas ou duas alturas a serem utilizadas como fundamentais e o número de harmônicos a serem adicionados a cada uma destas é gerada uma estrutura harmônica baseada na modulação em anel das estruturas de frequências dos conjuntos dados como entrada.

3.2 Subsistema 2: Manipulador de Funções

Este sistema tem três finalidades: a construção de estruturas harmônicas, de estruturas de duração e de progressões (tanto de estruturas harmônicas como de duração). O sistema utiliza, como parâmetros de entrada, funções, intervalos de valores sobre os quais a função deve ser calculada e o número de estruturas a serem geradas.

O caso mais básico é a geração de uma única estrutura, harmônica ou de durações. Neste caso deve ser fornecida uma única função, doravante denominada função principal, e um intervalo de números inteiros que serão utilizados para calcular as frequências ou durações. Por exemplo, supondo que se forneça a função $f(x) = 65x$ e os valores de 1 a 4 para x, a estrutura de frequências resultante (o conteúdo espectral) consistirá de 65, 130, 195 e 260 Hz, ou seja, os quatro primeiros parciais da série harmônica construída sobre 65 Hz, que corresponde ao Dó2 (considerando-se o Dó4 como o Dó central do piano). Este resultado é mostrado na Figura 4.

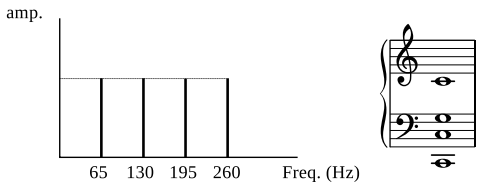


Figura 4: Estrutura harmônica gerada a partir da função $f(x) = 65x$ para $x \in \{1, 2, 3, 4\}$

Para o caso da geração de estrutura de durações, o procedimento é o mesmo, mas vale ressaltar que as durações são as distâncias entre os valores no eixo das ordenadas dos resultados das funções. Por exemplo, tomando a função $f(x) = 30x$ como função principal e os valores de 0 a 5 para x para o cálculo do tamanho de seções formais de uma composição, têm-se como resultado os valores 0, 30, 60, 90, 120 e 150. Estes valores são apenas "pontos no tempo" demarcando o começo e/ou fim das seções. O que se tem são 5 seções de 30 unidades. Outro exemplo é a geração de ritmos, considerando a função $f(x) = (x/2)^2$ como função principal e os valores de 0 a 4 para x, têm-se como resultado os valores 0, 0.25, 1, 2.25 e 4. Tomando as diferenças entres estes "pontos no tempo" temos os valores rítmicos 0.25, 0.75, 1.25 e 1.75. A Figura 5 mostra este ritmo considerando a semínima como unidade de tempo.

Figura 5: Estrutura rítmica gerada a partir da função $f(x) = (x/2)^2$ para $x \in \{0, 1, 2, 3, 4\}$

Outro caso um pouco mais complexo é a geração de uma progressão de estruturas. À função principal podem ser adicionados parâmetros a serem calculados por funções secundárias. Desta forma, as funções secundárias não geram frequências ou eventos diretamente, mas apenas geram novos valores para alterar a função principal a cada nova estrutura calculada. Além disso, a variável de funções secundárias, ao contrário da função principal, pode tomar o valor de qualquer número real. Por exemplo, supondo a geração de 5 simultaneidades com a função principal $f(x) = 100x + a$, com os valores de 1 a 4 (conteúdo espectral) para x e a função secundária $f(y) = 50y$, associada ao parâmetro a da função principal, com um intervalo de 0 a 2. A função secundária tem a finalidade de descrever a evolução do parâmetro a ao longo das cinco simultaneidades. Assim, o intervalo fornecido para y, de 0 a 2, é equitativamente dividido em cinco pontos: 0, 0.5, 1, 1.5 e 2. Aplicando esses valores à função secundária, resultam os valores 0, 25, 50, 75 e 100 para o parâmetro a na função principal. Desta forma, serão geradas 5 estruturas a partir da função principal cada uma com um dos valores calculados para a (Tabela 1).

	Simultaneidade				
Parcial	*1*	*2*	*3*	*4*	*5*
1	100	125	150	175	200
2	200	225	250	275	300
3	300	325	350	375	400
4	400	425	450	475	500

Tabela 1: Simultaneidades geradas a partir da função primária $f(x) = 100x + a$ para $x \in \{1, 2, 3, 4\}$ e da função secundária $f(y) = 50y$ para $y \in [0, 2]$

O mesmo princípio também pode ser aplicado à geração de progressões de estruturas duracionais. Supondo que se queira gerar uma progressão de 3 estruturas rítmicas a partir da função principal $f(x) = (x/2)^a$ com x entre 0 e 4 e com função secundária $f(y) = y$ com y de 1 a 3 associada ao seu parâmetro a. O intervalo fornecido para y é dividido equitativamente em 3 pontos, gerando os valores 1, 2 e 3, que, aplicados na função, resultam nestes mesmos valores. Assim como no exemplo anterior, cada um destes valores é utilizado na função principal no lugar do parâmetro a. O resultado é apresentado na Figura 6.

3.3 Sistema Composicional

O Sistema composicional consistirá em uma série de definições que controlarão a utilização dos subsistemas supramencionados e fornecerão diretrizes gerais que possam engendrar distintos planejamentos composicionais a partir dos mesmos princípios.

1. Estruturas harmônicas podem ser obtidas a partir de:

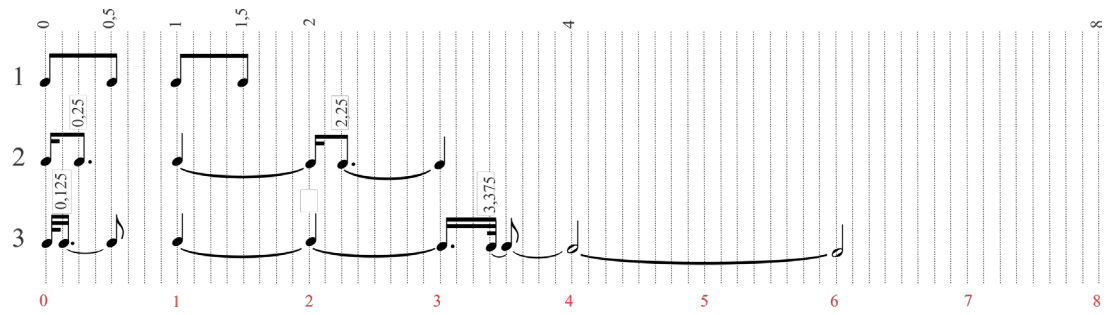


Figura 6: Três estruturas rítmicas geradas a partir da função primária $f(x) = (x/2)^a$ para $x \in \{0, 1, 2, 3, 4\}$ e função secundária $f(y) = y$ para $y \in [1, 3]$

(a) Subsistema Modulador em Anel, onde os valores de entrada consistem em dois conjuntos de alturas.

(b) Subsistema Manipulador de Funções, que calcula um espectro a partir de funções fornecidas pelo compositor.

2. As estruturas duracionais serão obtidas a partir pelo Subsistema Manipulador de Funções, que calcula durações a partir de funções fornecidas pelo compositor

3. A evolução das estruturas harmônicas é obtida:

   (a) Pela reaplicação de estruturas pré-existentes no Subsistema Modulador em Anel.

   (b) Pelo uso de mais de uma função no Manipulador de Funções. Uma função principal descreverá o espectro e funções secundárias descreverão a evolução espectral.

4. A evolução de estruturas duracionais será obtida pelo uso de mais de uma função no Manipulador de Funções. Uma função principal descreverá a estrutura inicial e funções secundárias descreverão sua evolução.

5. Técnicas composicionais espectrais tradicionais, sugeridas por Murail poderão ser livremente utilizadas na fase de planejamento composicional:

   (a) *Splitting*: somente uma região do espectro é utilizada.

   (b) Filtragem: reforço ou atenuação de certos parciais.

   (c) Exploraçao espectral: transformação do espectro em linha melódica [7].

4 IMPLEMENTAÇÃO

Ambos os subsistemas de modulação em anel e de geração de estruturas através de funções foram implementados em C++. O segundo utilizou-se também da biblioteca muParser, de código livre, para lidar com expressões matemáticas. Ambos são acompanhados de um arquivo de texto com instruções gerais de uso.

4.1 Modulador em Anel

A Figura 7 mostra a interface do aplicativo. Nela, a região demarcada indica os campos onde se especificam os dois espectros necessários ao cálculo da modulação em anel. Se o campo "Harm" estiver marcado, deve-se indicar apenas uma altura a ser tomada como fundamental e o número de harmônicos a serem considerados como componentes do espectro. Se o campo "Harm" estiver desmarcado deve-se indicar uma lista de alturas, separadas por espaço, a serem consideradas como componentes do espectro. Na porção inferior da interface do aplicativo encontram-se dois campos para indicar os nomes dos arquivos a serem gerados. Acima, o campo e seu respectivo botão "Gerar" referem-se a geração de arquivos do Lilypond para a visualização do espectro em notação tradicional e também audição através de arquivo MIDI também gerado através do Lilypond. Abaixo, o campo e seu respectivo botão "Gerar" referem-se a geração de arquivos do Csound para a audição do espectro. O aplicativo fornece ainda as opções de aproximar as alturas resultantes para quartos de tom, ao invés de semitons, e a opção de limitar a extensão do espectro resultante.

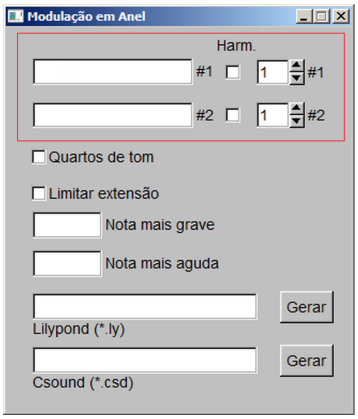


Figura 7: Interface do Modulador em Anel

4.2 Manipulador de Funções

A Figura 8 mostra a interface do aplicativo. A marcação em vermelho indica os campos onde se intro-

duzem a função principal, as duas funções secundárias e seus respectivos parâmetros. As operações permitidas e instruções gerais para a entrada das funções são listadas em um arquivo de texto que acompanha o programa. A única variável presente nas funções deve ser representada por x. A função principal aceita ainda duas constantes, a e b, que tomarão os valores produzidos por suas respectivas funções secundárias, conforme descrito e exemplificado na seção 3.2. Nos campos "x" se indicam os intervalos a serem usados como valores de x nas funções, que na função principal pertence aos números inteiros e nas secundárias pertence aos números reais. O aplicativo ainda oferece a possibilidade de "mapear" os resultados de uma função a um novo dado intervalo, isto é, o menor e o maior valor gerados pela função serão substituídos pelo menor e o maior valor, respectivamente, do intervalo fornecido e os demais valores intermediários serão calculados proporcionalmente.

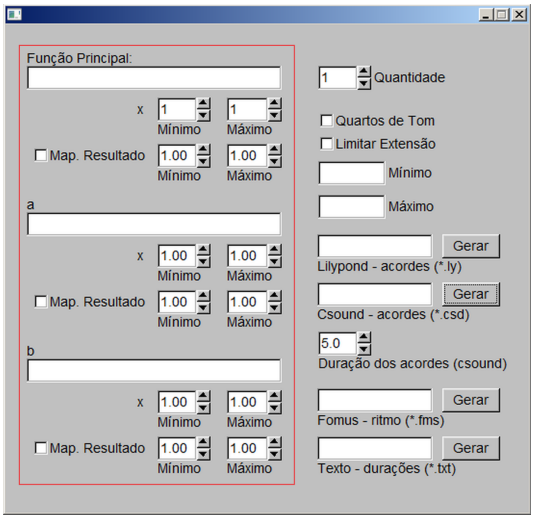


Figura 8: Interface do Manipulador de Funções

A direita da seção demarcada encontra-se o campo "Quantidade" onde é indicado o número de estruturas a serem geradas, a opção de aproximar as alturas resultantes para quartos de tom ao invés dos semitons tradicionais e a opção de limitar a extensão das alturas resultantes, eliminando tudo que estiver acima ou abaixo das alturas indicadas como mais aguda e mais grave. Abaixo destes se encontram os campos para a entrada dos nomes dos arquivos a serem gerados. Para a produção de simultaneidades há a opção de gerar arquivos do Lilypond ou do Csound. No segundo caso, há a necessidade de indicar também a duração de cada acorde. Para a produção de estruturas de durações, há a opção de gerar arquivos de texto com as durações simplesmente listadas numericamente, o que é útil para o cálculo da duração de seções, ou a de gerar arquivos do Fomus. Uma vez neste formato ele pode ser convertido por este software, Fomus, para Lilypond ou MusicXML para ser aberto em programas de notação com suporte a este formato.

Alguns detalhes da implementação de caráter prático úteis a serem mencionados são: As frequências são tomadas em absoluto, isto é, frequências negativas geradas pelas funções são consideradas iguais às positivas sem nenhuma consideração de inversão e cancelamento de fase. Caso sejam gerados "tempos negativos" o menor deles é considerado o ponto zero e as durações são calculadas a partir dele. A quantização das durações para ritmos em notação tradicional é realizada pelo Fomus. Um arquivo de nome props.fms acompanha o aplicativo e nele podem ser inseridos comandos do Fomus para manipular o resultado da quantização. A documentação deste software pode ser consultada para este fim.

5 PLANEJAMENTO COMPOSICIONAL

O primeiro passo no planejamento de *Anticlepsidra* envolveu o cálculo da quantidade e duração das sessões que iriam compô-la. Para isso foi utilizada a função $f(x) = 2^{x/2}$, para x de 0 a 13, no Manipulador de Funções, para a geração de durações. Assim, os seguintes valores foram gerados: 0.58, 0.82, 1.17, 1.65, 2.34, 3.31, 4.68, 6.62, 9.37, 13.25, 18.74, 26.5, 37.49 e 53.01. Cada um desses valores, em segundos, corresponde à duração de uma seção. Esta sequência de durações foi retrogradada e justaposta à original tendo o menor valor como eixo de simetria conforme mostrado na Figura 9, que é um palíndromo no formato de uma ampulheta deitada. Observa-se que as durações das seções diminuem à medida que nos aproximamos do centro da peça, a tal ponto que as seções centrais se confundem com figurações rítmicas, como se observa na Figura 10.

Nas quatro seções iniciais e finais, o material harmônico é oriundo da manipulação de funções. Por exemplo, na primeira seção, utilizam-se duas funções: 1) $y = 38.89x^a$ e 2) $y = 1.15{-}x$, para $x \in 0, 0.1, 0.2, 0.3, 0.4$. Da quinta à nona (e também da décima-nona à vigésima-terceira) seções, utiliza-se modulação em anel a partir do material das seções precedentes. Da décima à décima-oitava seções, utiliza-se um único espectro horizontalizado, produzindo uma linha melódica . O piano das quatro primeiras seções varia continuamente de uma textura em blocos para uma textura mais fragmentada. Essa textura, por sua vez, é totalmente monofônica nas seções centrais (décima à décima-oitava).

As figurações rítmicas aceleram até o centro do palíndromo e depois *rallentam*. Do início ao centro da peça, predominam configurações periódicas sobrepostas a *accelerandos* descontínuo-dinâmicos e uma pequena proporção de configurações estatísticas, de acordo com a escala de complexidade de Grisey (Figura 3). Do centro para o final da peça, as mesmas configurações se mantêm, com exceção da descontínuo-dinâmica que aparece como *rallentando*. A página inicial de *Anticlepsidra* é mostrada na Figura 11.

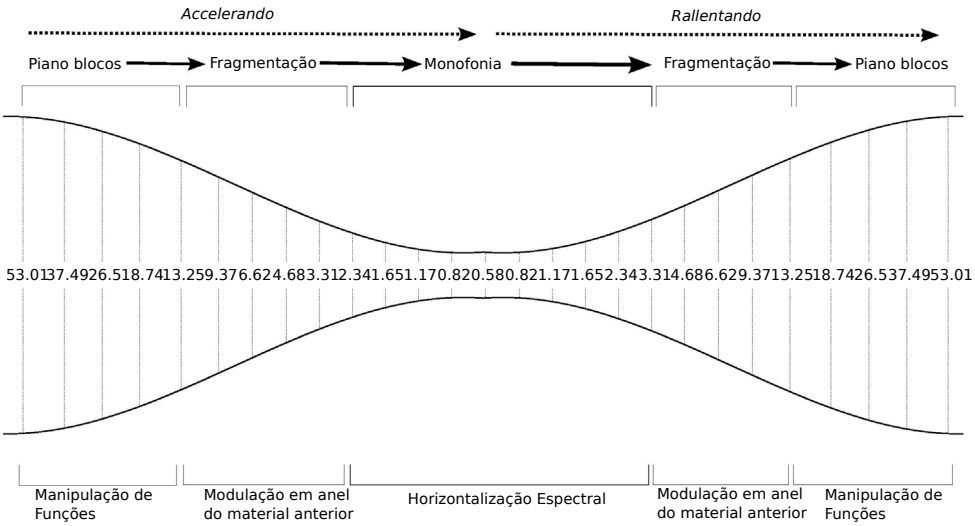


Figura 9: Planejamento composicional de *Anticlepsidra*

Figura 10: Figurações rítmicas das nove seções centrais de *Anticlepsidra*

6 CONCLUSÃO

Neste trabalho, utilizamos procedimentos espectrais coordenados por um sistema composicional, com intuito de produzir uma obra para violino, clarinete e piano, intitulada *Anticlepsidra*. Os cálculos envolvidos no processo foram tornados transparentes ao compositor através do desenvolvimento de dois aplicativos em C++, que manipulam as decisões estéticas composicionais com intuito de gerar estruturas harmônicas e duracionais. Estes aplicativos, de código livre, estarão brevemente disponíveis gratuitamente aos compositores no site do GAMA (Grupo de Análise Musical), da Universidade Federal de Campina Grande (PB), através do link www.musica.ufcg.edu.br/gama.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] F. Lima, "Desenvolvimento de sistemas composicionais a partir de intertextualidade," M.S. thesis, Universidade Federal da Paraíba, 2011.

[2] J. Fineberg, *Sculpting Sound: an Introduction to the Spectral Movement – its ideas, techniques and music*, Ph.D. thesis, Columbia University, 1999.

[3] F. Rose, "Introduction to the pitch organization of french spectral music," *Perspectives of New Music*, vol. 34, no. 2, pp. 9–36, 1996.

[4] C. Roads, *Computer Music Tutorial*, MIT Press, Cambridge, MA, 1996.

[5] W. Windsor, "Frequency structure in electroacoustic music: ideology, function and perception," *Organised Sound*, vol. 2, pp. 77–82, 1997.

[6] R. Wannamaker, "The spectral music of james tenney," *Contemporary Music Review*, vol. 27, no. 1, pp. 91–130, 2008.

[7] T. Murail, "Spectra and pixies," *Contemporary Music Review*, vol. 1, no. 4, pp. 157–170, 1984.

[8] G. Grisey, "Tempus ex machina: A composer's reflection on musical time," *Contemporary Music Review*, vol. 2, pp. 239–275, 1987.

[9] M. D. Mesarovic and Y. Takahara, *General Systems Theory: Mathematical Foundations*, Academic Press, New York, 1975.

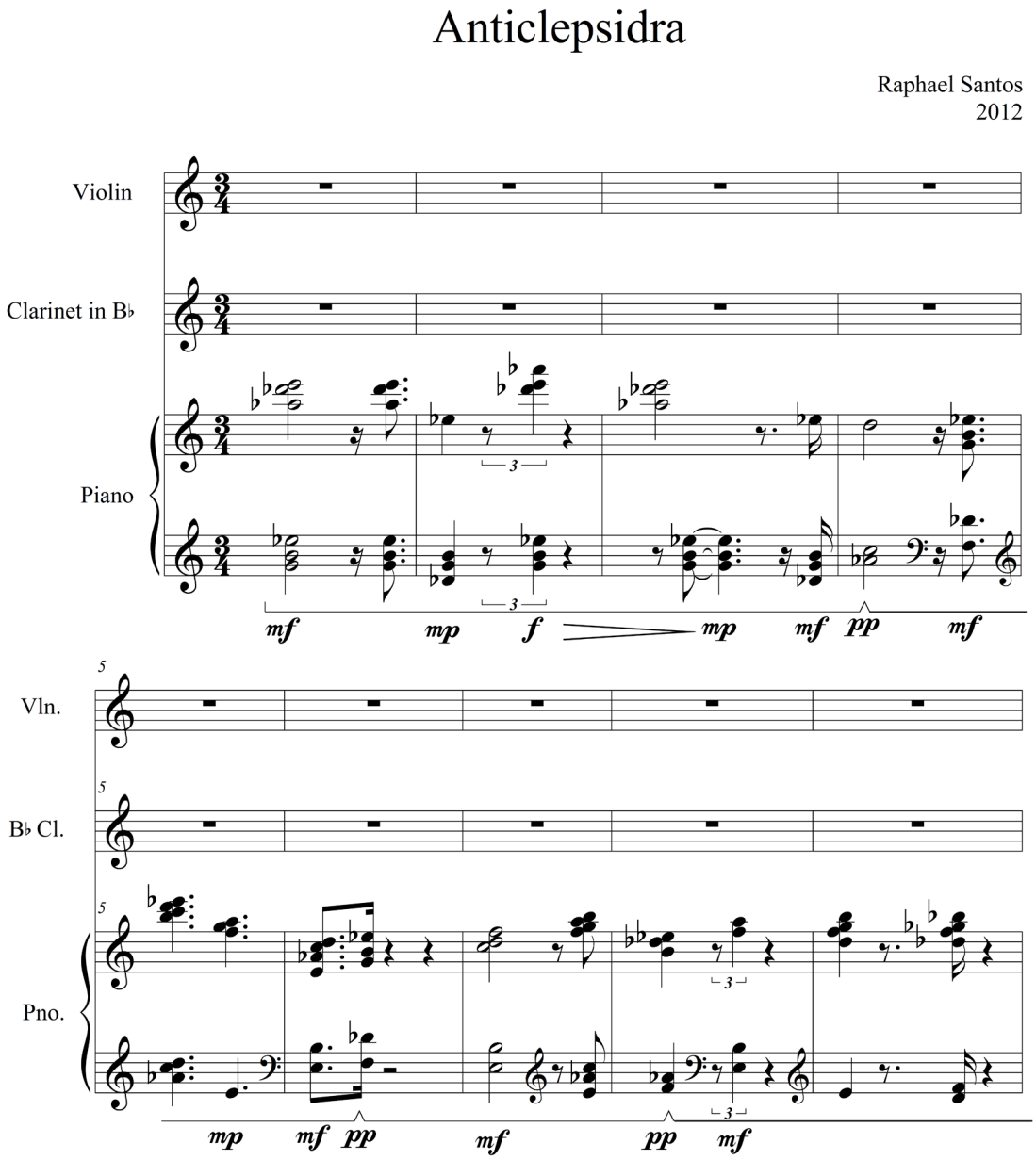


Figura 11: Página inicial de *Anticlepsidra*

Sociedade de Engenharia de Áudio
Artigo de Congresso
Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Implementações ativas de filtros para conformação de ruído rosa

André Luís Dalcastagnê[1], Sidnei Noceti Filho[2] e Gabriel C. Kulevicz da Silva[2]
[1]Departamento Acadêmico de Eletrônica
Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina, Campus Florianópolis
Florianópolis, SC, 88020-300, Brasil
[2] LINSE – Laboratório de Circuitos e Processamento de Sinais
Departamento de Engenharia Elétrica, UFSC
Florianópolis, SC, 88040-900, Brasil
andreld@ifsc.edu.br, {sidnei,gabriel}@linse.ufsc.br

RESUMO
Este trabalho apresenta um método de projeto de uma rede ativa para conformação de ruído rosa com resposta em freqüência similar à da rede passiva apresentada na norma NBR 10303. A topologia proposta possui o menor número possível de componentes. Além disso, os pareamentos das singularidades bem como o cascateamento das redes de primeira e segunda ordem são feitos de forma a se obter a menor sensibilidade possível. Os resultados obtidos mostram que a versão ativa possui uma sensibilidade menor do que a passiva na região central da sua faixa plana. Entretanto, no início e no fim da faixa plana, a versão passiva é a que apresenta menor sensibilidade.

0. INTRODUÇÃO

A norma NBR 10303:1988 [1] (cancelada em 09/06/2011), da Associação Brasileira de Normas Técnicas (ABNT), indica o método de ensaio para a determinação da potência elétrica RMS admissível em alto-falantes. Ela apresenta um filtro passivo passa-faixa de quinta ordem para conformação do ruído rosa, uma vez que esse tipo de ruído é o que melhor simula um programa musical típico. Tal filtro possui uma característica inadequada para os dias de hoje: a sua freqüência de corte inferior é em torno de 65 Hz, não contemplando assim o teste de *subwoofers*, comumente utilizados em sistemas de som automotivo e de *home cinema*, bem como em sistemas de som profissional. Para suplantar tal deficiência, em [2] foram obtidos via otimização três filtros com a mesma topologia do filtro da NBR 10303, mas com freqüências de corte inferiores de 20, 30 e 40 Hz. Na realidade, foi considerado como ponto de partida o filtro apresentado na norma DIN 45573 [3], de topologia idêntica ao filtro da NBR 10303, mas com componentes de valores diferentes, resultando em uma resposta em freqüência ligeiramente diferente e uma impedância de entrada superior.

O objetivo deste trabalho é apresentar o método de projeto de filtros ativos para conformação do ruído rosa, que apresentem a mesma resposta em freqüência das redes passivas com freqüências de corte inferiores de 20, 30 e 40 Hz propostas em [2]. As redes ativas apresentam algumas vantagens sobre as passivas: i) nas redes ativas, o ganho na "faixa plana" pode ser arbitrado pelo projetista; nas redes passivas, o ganho na "faixa plana" é da ordem de −3,9 dB; ii) a rede ativa é menos influenciada pelas perdas

nos capacitores, pois, devido a sua topologia, ela utiliza capacitores de menores valores, o que torna possível o uso de capacitores de melhor qualidade (capacitores de polipropileno em vez de capacitores de poliéster); iii) nas redes ativas, o usuário pode arbitrar facilmente todos os capacitores com valores comerciais; na rede passiva, tal procedimento é impraticável, sendo necessário associar valores comerciais de capacitores. Como desvantagem, a rede ativa necessita de fonte de alimentação para os amplificadores operacionais.

O restante deste artigo está organizado da seguinte forma. A Seção 1 apresenta as redes passivas bem como as suas funções de transferência, desenvolvidas em [2]. A Seção 2 apresenta as redes ativas utilizadas neste trabalho, bem como a estratégia adotada para os pareamentos das singularidades da função de transferência e para o cascateamento das redes de primeira e segunda ordem utilizadas, além das expressões de projeto dessas redes. A Seção 3 detalha o projeto de uma rede ativa para conformação de ruído rosa, usada neste trabalho para a comparação com a respectiva versão passiva. A Seção 4 apresenta um estudo de sensibilidade, comparando as topologias passiva e ativa. Além disso, ela apresenta os resultados práticos obtidos com o filtro ativo projetado na Seção 3. Finalmente, a Seção 5 apresenta as conclusões deste trabalho.

1. FILTRO PASSIVO PARA CONFORMAÇÃO DE RUÍDO ROSA

A topologia do filtro passivo utilizado para conformação do ruído rosa proposto em [1] é a mostrada na Figura 1. Os valores dos seus componentes, determinados em [2], são mostrados na Tabela 1, para as três opções de freqüência de corte inferior disponibilizadas. A Figura 2 apresenta a magnitude da resposta em freqüência de cada filtro. Note que somente para freqüências abaixo de 40 Hz, há uma diferença significativa na magnitude da resposta em frequência dos três filtros.

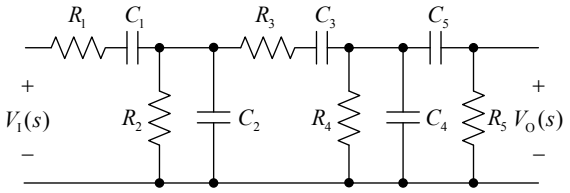


Figura 1 Filtro passivo utilizado para conformação de ruído rosa.

O filtro da Figura 1 é de quinta ordem e apresenta uma característica passa-faixa. Ele possui três zeros na origem e dois no infinito. Logo, a sua função de transferência é do tipo

$$H(s) = \frac{Ks^3}{s^5 + a_4 s^4 + a_3 s^3 + a_2 s^2 + a_1 s + a_0} \qquad (1)$$

Para obter os coeficientes de (1) em função dos dez componentes da rede, utilizou-se neste trabalho o *toolbox* Symbolic do *software* Matlab®. Tais expressões literais não são aqui mostradas por serem muito extensas. Substituindo os valores da Tabela 1 nessas expressões, obtêm-se os coeficientes de (1) apresentados na Tabela 2.

Tabela 1 Valores dos componentes dos filtros passivos para conformação do ruído rosa, para as três opções de freqüência de corte inferior

Componente	$f_{3dB} = 20$ Hz	$f_{3dB} = 30$ Hz	$f_{3dB} = 40$ Hz
R_1(kΩ)	4,32	4,32	4,32
C_1(nF)	680	390	100 + 4,7
R_2(kΩ)	23,2	24,3	76,8
C_2(nF)	12 + 1,2	15	2,7
R_3(kΩ)	4,42	6,49	88,7
C_3(nF)	560 + 22	470 + 56	27
R_4(kΩ)	37,4	54,9	523
C_4(nF)	3,3	2,2	0,82
R_5(kΩ)	121	110	402
C_5(nF)	150 + 15	82 + 10	22

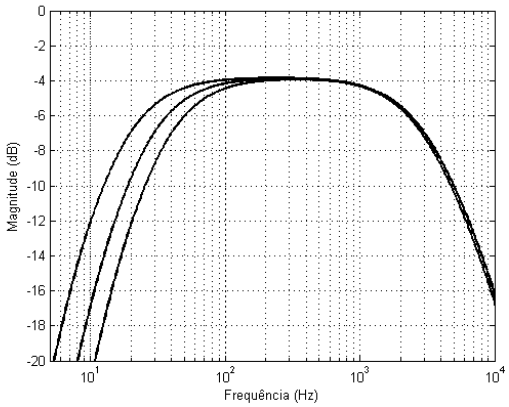


Figura 2 Magnitudes das respostas em freqüência dos filtros para conformação de ruído rosa passivos com freqüências de corte inferiores de 20, 30 e 40 Hz.

Tabela 2 Coeficientes da função de transferência (1) para as três opções de freqüência de corte inferior

Coeficiente	$f_{3dB} = 20$ Hz	$f_{3dB} = 30$ Hz	$f_{3dB} = 40$ Hz
K	$1{,}20228\times10^{9}$	$1{,}08083\times10^{9}$	$1{,}17873\times10^{9}$
a_4	$1{,}17886\times10^{5}$	$1{,}11882\times10^{5}$	$1{,}16587\times10^{5}$
a_3	$1{,}88060\times10^{9}$	$1{,}70249\times10^{9}$	$1{,}86926\times10^{9}$
a_2	$3{,}26548\times10^{11}$	$4{,}33853\times10^{11}$	$6{,}42123\times10^{11}$
a_1	$1{,}45765\times10^{13}$	$2{,}80557\times10^{13}$	$5{,}64232\times10^{13}$
a_0	$1{,}75365\times10^{14}$	$3{,}90254\times10^{14}$	$1{,}17379\times10^{15}$

Independentemente da freqüência de corte inferior considerada, a função de transferência (1) possui apenas polos reais. Desse modo, ela pode ser reescrita da seguinte forma:

$$H(s) = \frac{Ks^3}{(s+\sigma_1)(s+\sigma_2)(s+\sigma_3)(s+\sigma_4)(s+\sigma_5)} \qquad (2)$$

onde σ_i representa o módulo de cada polo real p_i. Os valores de σ_i, em radianos por segundo, para as três possíveis opções de freqüência de corte inferior são apresentados na Tabela 3.

Tabela 3 Singularidades da função de transferência (2) para as três opções de freqüência de corte inferior

Parâmetro	$f_{3dB} = 20$ Hz	$f_{3dB} = 30$ Hz	$f_{3dB} = 40$ Hz
K	$1{,}20228\times10^9$	$1{,}08083\times10^9$	$1{,}7873\times10^9$
σ_1(rad/s)	19,84941	19,16128	30,37960
σ_2(rad/s)	41,92578	73,59593	93,97777
σ_3(rad/s)	113,29358	165,34822	224,83116
σ_4(krad/s)	18,80569	17,84731	18,75935
σ_5(krad/s)	98,90577	93,77723	97,47904

2. FILTRO ATIVO PARA CONFORMAÇÃO DE RUÍDO ROSA

A função de transferência (2) pode ser decomposta em duas funções de segunda ordem e uma de primeira ordem, existindo diversas possibilidades de pareamento das suas singularidades. Visando minimizar a sensibilidade das redes ativas, opta-se neste trabalho por: i) fazer o pareamento entre um zero na origem com o polo de menor magnitude ($-\sigma_1$), gerando uma função passa-alta de primeira ordem; ii) fazer o pareamento dos outros dois zeros na origem com os polos $-\sigma_2$ e $-\sigma_3$, gerando uma função passa-alta de segunda ordem; iii) fazer o pareamento dos dois zeros no infinito com os polos de maior magnitude ($-\sigma_4$ e $-\sigma_5$), gerando uma função passa-baixa de segunda ordem. A fim de facilitar a implementação e minimizar a sensibilidade da rede ativa, os ganhos no infinito das funções passa-alta e o ganho na origem da função passa-baixa são arbitrados unitários. A opção i foi escolhida porque o polo de menor magnitude é o que exige o capacitor de maior valor para ser implementado, o que torna necessário o uso de um capacitor de poliéster em vez de um de polipropileno. Assim, a perda no capacitor é maior. Apesar disso, na rede de primeira ordem, tal característica é irrelevante: mesmo usando um capacitor de grande perda, o desvio da resposta obtida em relação à ideal é irrelevante. No caso da rede passa-alta de segunda ordem, isso não é verdade. Mas como neste trabalho tal rede implementa polos de maior magnitude ($-\sigma_2$ e $-\sigma_3$, opção ii), ela usa capacitores de menor valor, tornando possível a adoção de capacitores de polipropileno, de baixas perdas. A opção iii é natural: parear os polos de maior magnitude com os zeros no infinito. Com esse pareamento, a função de transferência da rede ativa é

$$H(s) = \frac{\sigma_4\sigma_5}{(s+\sigma_4)(s+\sigma_5)} \cdot \frac{s^2}{(s+\sigma_2)(s+\sigma_3)} \cdot \frac{s}{s+\sigma_1}. \quad (3)$$

A Figura 3 ilustra a forma de implementação de (3) através da rede ativa adotada neste trabalho. São utilizadas redes de segunda ordem Sallen-Key [4] e uma rede passa-alta de primeira ordem. Existem diversas possibilidades de cascateamento. Neste trabalho, é utilizada a rede passa-baixa de segunda ordem na entrada do sistema, com o intuito de minimizar problemas oriundos do *slew rate* finito dos amplificadores operacionais. As redes passa-alta são usadas na sequência a fim de minimizar possíveis problemas de *offset*. É importante utilizar a rede passa-alta de primeira ordem na saída porque ela é a que possui o menor carregamento, pois a sua corrente de saída é praticamente nula devido à realimentação unitária, tornando desprezível o efeito da resistência de saída do amplificador operacional. Ainda que seja conectada uma carga na saída do circuito, a rede passa-alta de primeira ordem é a que apresenta a menor influência da resistência de saída do amplificador operacional.

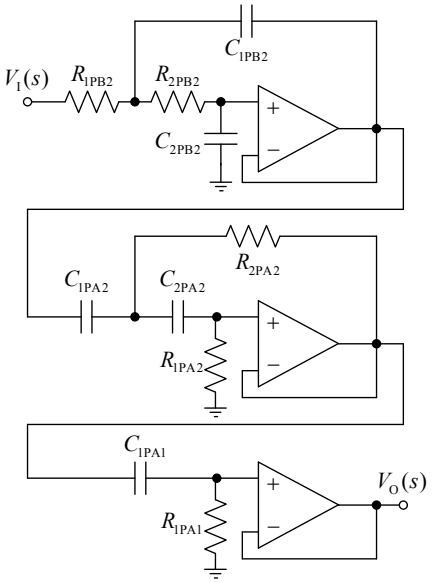


Figura 3 Filtro ativo para conformação de ruído rosa adotado.

2.1 Projeto da Rede Passa-Baixa de Segunda Ordem

A função de transferência da rede passa-baixa de segunda ordem da Figura 3 é

$$H_{PB2}(s) = \frac{\sigma_4\sigma_5}{(s+\sigma_4)(s+\sigma_5)} = \frac{a_{0PB2}}{s^2 + a_{1PB2}s + a_{0PB2}}. \quad (4)$$

Substituindo os parâmetros da Tabela 3 em (4), os coeficientes obtidos para as três opções de frequência de corte inferior são os apresentados na Tabela 4.

Tabela 4 Coeficientes da função de transferência (4) para as três opções de freqüência de corte inferior

Parâmetro	$f_{3dB} = 20$ Hz	$f_{3dB} = 30$ Hz	$f_{3dB} = 40$ Hz
a_{0PB2}	$1{,}85999\times10^9$	$1{,}67367\times10^9$	$1{,}82864\times10^9$
a_{1PB2}	$1{,}17711\times10^5$	$1{,}11624\times10^5$	$1{,}16238\times10^5$

Com os valores de a_{0PB2} e a_{1PB2} definidos, escolhem-se valores comerciais para os capacitores C_{1PB2} e C_{2PB2} e os valores dos resistores são definidos por

$$R_{1PB2} = \frac{a_{1PB2}C_{1PB2} + \sqrt{(a_{1PB2}C_{1PB2})^2 - 4a_{0PB2}C_{1PB2}C_{2PB2}}}{2a_{0PB2}C_{1PB2}C_{2PB2}} \quad (5)$$

e

$$R_{2PB2} = \frac{1}{a_{0PB2}R_{1PB2}C_{1PB2}C_{2PB2}}. \quad (6)$$

Para que os resultados de (5) e (6) sejam positivos, os valores dos capacitores da rede devem ser escolhidos de modo a satisfazerem a seguinte condição:

$$C_{2PB2} < C_{1PB2}\left(\frac{a_{1PB2}^2}{4a_{0PB2}}\right). \quad (7)$$

2.2 Projeto da Rede Passa-Alta de Segunda Ordem

A função de transferência da rede passa-alta de segunda ordem da Figura 3 é

$$H_{\mathrm{PA2}}(s) = \frac{s^2}{(s+\sigma_1)(s+\sigma_2)} = \frac{s^2}{s^2 + a_{1\mathrm{PA2}}s + a_{0\mathrm{PA2}}}. \quad (8)$$

Substituindo os parâmetros da Tabela 3 em (8), os coeficientes obtidos para as três opções de frequência de corte inferior são os apresentados na Tabela 5.

Tabela 5 Coeficientes da função de transferência (8) para as três opções de freqüência de corte inferior

Parâmetro	$f_{3\mathrm{dB}} = 20$ Hz	$f_{3\mathrm{dB}} = 30$ Hz	$f_{3\mathrm{dB}} = 40$ Hz
$a_{0\mathrm{PA2}}$	$4{,}74992\times10^3$	$1{,}21689\times10^4$	$2{,}11291\times10^4$
$a_{1\mathrm{PA2}}$	$1{,}55219\times10^2$	$2{,}38944\times10^2$	$3{,}18808\times10^2$

Com os valores de $a_{0\mathrm{PA2}}$ e $a_{1\mathrm{PA2}}$ definidos, adota-se o mesmo valor comercial para os capacitores da rede, ou seja, $C_{1\mathrm{PA2}} = C_{2\mathrm{PA2}} = C$. Feito isso, os valores dos resistores são definidos por

$$R_{1\mathrm{PA2}} = \frac{2}{a_{1\mathrm{PA2}}C} \quad (9)$$

e

$$R_{2\mathrm{PA2}} = \frac{a_{1\mathrm{PA2}}}{2a_{0\mathrm{PA2}}C}. \quad (10)$$

2.3 Projeto da Rede Passa-Alta de Primeira Ordem

A função de transferência da rede passa-alta de primeira ordem da Figura 3 é

$$H_{\mathrm{PA1}}(s) = \frac{s}{s+\sigma_1} = \frac{s}{s+a_{0\mathrm{PA1}}}. \quad (11)$$

Substituindo os parâmetros da Tabela 3 em (11), o coeficiente obtido para cada uma das três opções de frequência de corte inferior é o apresentado na Tabela 6.

Tabela 6 Coeficiente da função de transferência (11) para as três opções de freqüência de corte inferior

Parâmetro	$f_{3\mathrm{dB}} = 20$ Hz	$f_{3\mathrm{dB}} = 30$ Hz	$f_{3\mathrm{dB}} = 40$ Hz
$a_{0\mathrm{PA1}}$	$1{,}98494\times10^1$	$1{,}91612\times10^1$	$3{,}03796\times10^1$

O projeto da rede é bastante simples: adota-se um valor comercial para o capacitor da rede e o resistor é dado por

$$R_{1\mathrm{PA1}} = \frac{1}{a_{1\mathrm{PA1}}C_{1\mathrm{PA1}}}. \quad (12)$$

3. EXEMPLO DE PROJETO

Neste trabalho, para servir como base de estudo e comparação com a rede passiva que implementa a mesma função de transferência, é usada uma rede ativa com frequência de corte inferior de 40 Hz.

O projeto da rede ativa da Figura 3 é feito através das expressões (4) a (12). Existem infinitas opções de valores de componentes da rede da Figura 3 que satisfazem (3). Neste trabalho, foram adotados os seguintes valores comerciais da Série E-12 para os capacitores do circuito: $C_{1\mathrm{PB2}} = 2{,}2$ nF; $C_{2\mathrm{PB2}} = 1$ nF; $C_{1\mathrm{PA2}} = C_{2\mathrm{PA2}} = 22$ nF; $C_{1\mathrm{PA1}} = 180$ nF. Feito isso, os resistores foram determinados através de (5), (6), (9), (10) e (12), considerando nessas expressões os coeficientes da rede com $f_{3\mathrm{dB}} = 40$ Hz (Tabelas 4, 5 e 6). Por fim, os resistores são aproximados para valores comerciais da Série E-96 (1 % de tolerância). O conjunto de valores dos componentes da rede aqui projetada é apresentado na Tabela 7.

Tabela 7 Componentes da rede ativa para conformação de ruído rosa projetada, com freqüência de corte inferior de 40 Hz

Componente	Valor
$R_{1\mathrm{PB2}}$	59 kΩ
$R_{2\mathrm{PB2}}$	4,22 kΩ
$C_{1\mathrm{PB2}}$	2,2 nF
$C_{2\mathrm{PB2}}$	1 nF
$R_{1\mathrm{PA2}}$	287 kΩ
$R_{2\mathrm{PA2}}$	340 kΩ
$C_{1\mathrm{PA2}}$	22 nF
$C_{2\mathrm{PA2}}$	22 nF
$R_{1\mathrm{PA1}}$	182 kΩ
$C_{1\mathrm{PA1}}$	180 nF

4. ANÁLISE DE SENSIBILIDADE E RESULTADOS EXPERIMENTAIS

O primeiro objetivo desta seção é comparar a sensibilidade das redes para conformação de ruído rosa abordadas neste trabalho: topologias passiva (Figura 1) e ativa (Figura 3). Por simplicidade e sem perda de generalidade, usam-se apenas a rede passiva com frequência de corte inferior de 40 Hz (valores dos componentes descritos na Tabela 1) e a rede ativa com frequência de corte inferior de 40 Hz projetada na seção anterior (valores dos componentes mostrados na Tabela 7). As conclusões obtidas para o caso considerado podem ser estendidas para redes com outros valores de freqüência de corte inferior.

As simulações realizadas neste trabalho consideram componentes de valores comerciais, resistores com 1 % de tolerância, capacitores com 5 % de tolerância (na prática, usam-se capacitores com tolerância de 10 % medidos em uma ponte RLC, a fim de se obter uma tolerância de 5 %), perdas nos capacitores, amplificadores operacionais com GB de 3 MHz (tolerância de 20 %) e resistência de saída de 200 Ω. Tais simulações foram realizadas através do *software* SG2, que está disponível para *download* em http://www.linse.ufsc.br/~sidnei.

A Figura 4 apresenta os desvios estatísticos [4], em decibéis, das redes passiva e ativa com freqüência de corte inferior de 40 Hz. A faixa de freqüência considerada corresponde à faixa plana, que vai de 40 Hz até aproximadamente 3 kHz (na realidade, em torno de 2,95 kHz, valor difícil de ser visualizado em uma escala logarítmica). Pode-se notar que a rede passiva apresenta um desvio menor na faixa de freqüência entre 40 e 58,5 Hz bem como na faixa que se estende de 1,25 a 3 kHz. Já na faixa de freqüência compreendida entre 58,5 Hz e 1,25 kHz (parte mais central da faixa plana do filtro, em que ele apresenta o maior ganho), a rede ativa apresenta desvios

menores. A fim de facilitar a comparação desses valores, a Tabela 8 apresenta os desvios estatísticos no início e no fim da faixa plana (40 Hz e 3 kHz, respectivamente), bem como na freqüência central do filtro, em torno de 342,4 Hz.

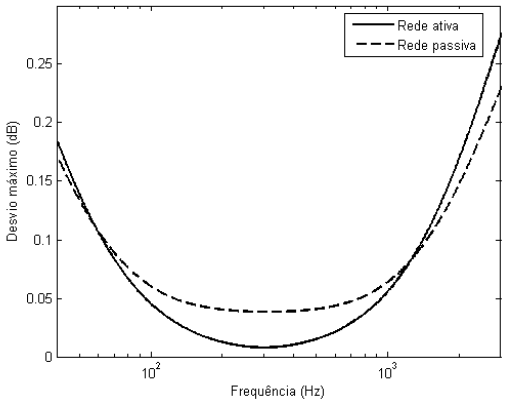


Figura 4 Desvios estatísticos das redes passiva e ativa com freqüência de corte inferior de 40 Hz.

Tabela 8 Desvios estatísticos das redes passiva e ativa com freqüência de corte inferior de 40 Hz, no início, no meio e no fim da banda de passagem

Frequência	Desvio da Rede passiva	Desvio da Rede ativa
40 Hz	0,1728 dB	0,1860 dB
342,4 Hz	0,03833 dB	0,008316 dB
3 kHz	0,2290 dB	0,2737 dB

A rede ativa projetada foi montada a fim de se medir a magnitude da sua resposta em freqüência. Para isso, utilizou-se um osciloscópio digital com erro máximo de medição de ±3 %. Na montagem, foram utilizados resistores com tolerância de 1 %, capacitores com tolerância de 5 % (na realidade, 10 %, mas medidos em uma ponte RLC a fim de se obter 5 % de tolerância) e o amplificador operacional TL082, que possui GB de 3 MHz e resistência de saída de 200 Ω. A Figura 5 apresenta os resultados obtidos na faixa de 20 Hz a 10 kHz, em que os círculos representam os valores medidos. Na Figura 6, é destacada a banda de passagem do filtro, que se estende de 40 Hz a 3 kHz. Pode-se notar que várias medidas ficaram fora do desvio máximo esperado em relação à curva nominal, limites estes calculados em função das tolerâncias dos componentes.

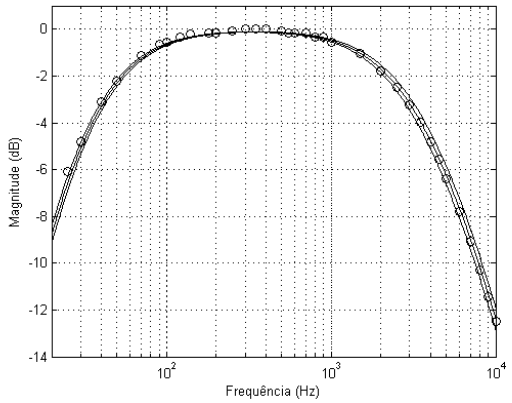


Figura 5 Curvas de magnitude nominal e com desvios estatísticos da rede ativa com freqüência de corte inferior de 40 Hz, juntamente com os valores medidos entre 20 Hz e 20 kHz (círculos).

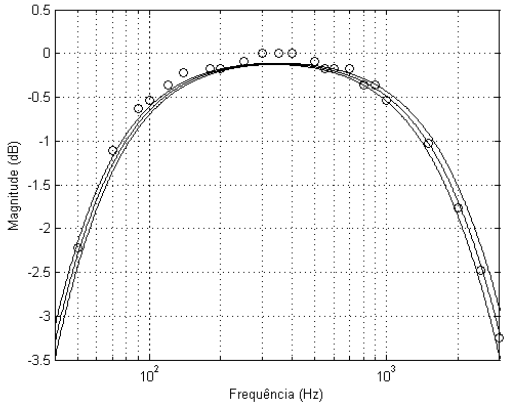


Figura 6 Ampliação da Figura 5 na banda de passagem do filtro, entre 40 Hz e 3 kHz.

Tal fato deve-se quase exclusivamente ao erro do instrumento de medição, no caso o osciloscópio digital. Para se estudar isso, acrescentou-se aos desvios máximos os piores casos possíveis: erro de +3 % no sinal de saída e de −3 % no sinal de entrada e erro de −3 % no sinal de saída e de +3 % no sinal de entrada. As Figuras 7 e 8 apresentam os resultados. Com essa consideração, as medidas ficam dentro da faixa esperada.

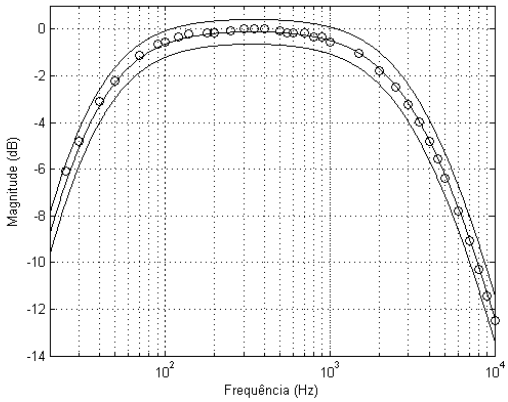


Figura 7 Curvas de magnitude nominal e com desvios estatísticos da rede ativa acrescidos do erro de medição do osciloscópio com freqüência de corte inferior de 40 Hz, juntamente com os valores medidos entre 20 Hz e 20 kHz (círculos).

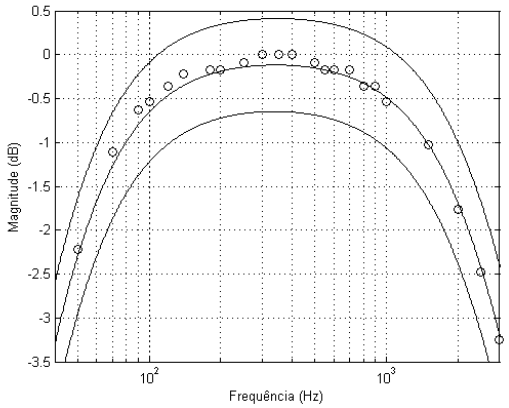


Figura 8 Ampliação da Figura 7 na banda de passagem do filtro, entre 40 Hz e 3 kHz.

5. CONCLUSÕES

Este trabalho apresentou um método de projeto de redes ativas para conformação de ruído rosa, com uma resposta em freqüência de magnitude similar a da rede apresentada na norma NBR 10303, que define o método de ensaio para a determinação da potência elétrica RMS admissível em alto-falantes. A topologia proposta possui o menor número possível de componentes passivos. Além disso, o pareamento das singularidades bem como a forma de cascateamento das redes de primeira e segunda ordem ativas foram feitos de forma a obter a rede ativa com a menor sensibilidade possível. Os resultados obtidos mostram que a versão ativa possui uma sensibilidade menor do que a passiva na região central da sua faixa plana. Entretanto, no início e no fim da faixa plana, a versão passiva é a que apresenta menor sensibilidade. Na rede montada, a principal dificuldade foi o erro de medição do osciloscópio digital utilizado que, se não for levado em conta, faz com que as medidas fiquem fora dos desvios estatísticos máximos.

6. REFERÊNCIAS BIBLIOGRÁFICAS

[1] Associação Brasileira de Normas Técnicas, *NBR 10303*, 1988.

[2] A. L. Dalcastagnê, S. Noceti Filho and H. S. Silva, "A Numerical Method to Modify the NBR 10303 Filter Frequency Response," presented at the AES 115th Convention, New York, United States of America, 2003, October 10-13.

[3] Deutsches Institut für Normung e.V., *DIN 45573*.

[4] S. Noceti Filho, Filtros Seletores de Sinais, Editora da UFSC, Florianópolis, 3ª edição, 2010.

DALCASTAGNÊ, A. L.; NOCETI FILHO, S.; SILVA, G. C. K. Implementações ativas de filtros para conformação de ruído rosa.
In: CONGRESSO DE ENGENHARIA DE ÁUDIO AES BRASIL, 10., 2012, São Paulo, *Anais*... Rio de Janeiro: Sociedade de Engenharia de Áudio, 2012. p. 55-60.

Sociedade de Engenharia de Áudio
Artigo de Congresso
Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Proposta de um sistema de diálogo automático baseado em algoritmos de aprendizado por reforço

Anderson Avila[1] e Francisco J. Fraga[1]
[1]Universidade Federal do ABC (UFABC)
Santo André, SP, 09210-170, Brasil
{anderson.avila@gmail.com,francisco.fraga@ufabc.edu.br}

RESUMO

Neste artigo é apresentada uma proposta de desenvolvimento de um sistema de diálogo automático que servirá como um dispositivo de auxílio para consulta a respeito de informações turísticas. O objetivo é oferecer ao usuário uma interface de interação homem-computador baseada em um sistema automático de entendimento de fala e que seja, sobretudo, amigável. O foco do trabalho está em utilizar técnicas de aprendizado por reforço associadas ao processo de decisão de Markov para aperfeiçoar as ações realizadas pelo sistema com base nas sentenças obtidas na saída do subsistema de reconhecimento da fala.

0. INTRODUÇÃO

Um dos objetivos de um sistema de atendimento automático baseado no entendimento da fala é possibilitar o diálogo entre homem e computador usando a língua natural. São diversas as vantagens oferecidas por um sistema como este. Dentre elas, podemos citar as seguintes: facilitar o uso do computador por parte de pessoas com deficiência, diminuir os riscos inerentes à realização de tarefas em paralelo como, por exemplo, dirigir e manusear o celular, aumentar a produtividade e diminuir os custos. Na aplicação visada por este trabalho, a finalidade é criar uma ferramenta que possibilite a qualquer usuário acessar informações turísticas por meio de um sistema de diálogo homem-máquina.

Um sistema de atendimento automático tradicional, em geral oferece ao usuário um menu contendo uma lista de opções, e um *prompt* que o usuário utiliza para se comunicar com o sistema. Nesse tipo de aplicação a interação pode acontecer através de toque em tela (*touchscreen*) ou da fala. Quando realizada através do toque, o usuário pode ter que passar por uma experiência entediante e frustrante, sendo muitas vezes forçado a navegar em uma estrutura extensa e complexa de opções. Quando isso ocorre, é comum que o usuário opte pelo atendimento humano quando este está também disponível. Da mesma forma, quando a estrutura de menus é demasiadamente curta e simples, pode vir a ser difícil para o usuário encontrar a categoria de serviços ou atendimento adequados para solucionar o seu problema. Encontrar o equilíbrio no número de opções e detalhes presentes em um menu não é uma tarefa trivial. Além disso, a experiência do usuário com sistemas em que a interação acontece através da fala pode ser igualmente frustrante. Isso se dá pelo fato de que, em geral, os softwares utilizados obrigam o usuário a aprender o que a máquina espera que ele diga.

O objetivo deste trabalho é utilizar técnicas de Aprendizado de Máquina (**ML** - *Machine Learning*) que permitam que o sistema se adapte ao usuário. Com isso pretende-se que estruturas rígidas de menu sejam substituídas por um gerenciador de diálogo, eliminando-se, assim, a necessidade do usuário se comunicar com o

sistema através de frases pré-determinadas. Espera-se que o software reconheça e compreenda a vontade do usuário de maneira que este possa acessar as informações e recursos do sistema de forma mais natural e amigável.

1. ETAPAS DO DESENVOLVIMENTO

Um sistema de diálogo automático pode ser obtido com a integração de três módulos: reconhecimento da fala, compreensão de língua natural e um gerenciador de diálogo, conforme mostra a figura 1.

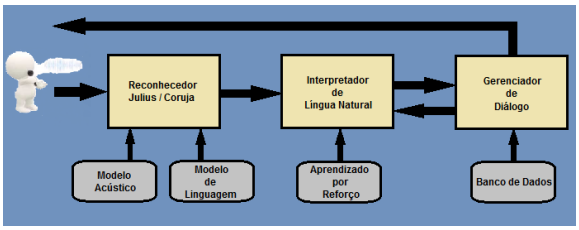


Figura 1 – Arquitetura para um sistema de atendimento automático baseado em entendimento de fala.

A proposta deste trabalho é adaptar o reconhecedor utilizado em um trabalho anterior que visava a tradução automática do português falado para o inglês, em um contexto de turismo [1], de maneira a integrá-lo a duas novas tecnologias: um interpretador de língua natural, baseado em aprendizado por reforço, e um gerenciador de diálogo.

A primeira parte da implementação, i.e, gerar um texto escrito a partir da fala, consistiu no reaproveitamento da parte do sistema de tradução referente ao de fala, na qual foi utilizado como base o decodificador *Julius* [2]. Este decodificador, em princípio, é capaz de realizar o reconhecimento de frases faladas em qualquer língua, desde que seja adequadamente alimentado com os modelos acústicos apropriados, bem como com os correspondentes modelos de linguagem, incluindo os respectivos dicionários de pronúncia.

Para facilitar esta parte do desenvolvimento, foi utilizado o software *Coruja* 0.2, API desenvolvida pelo grupo *FalaBrasil* [3][4][5], da Universidade Federal do Pará, que permite aos programadores controlarem as funcionalidades do *Julius* com maior rapidez e agilidade.

Ao interpretador caberá realizar o *parsing* da sentença reconhecida no módulo anterior, apresentando-o dentro de uma estrutura semântica que seja compreendida pelo gerenciador de diálogo e que possibilite que este execute uma ação específica. De posse do significado, pretende-se que o gerenciador de diálogo realize uma consulta à base de dados e apresente, através de um sintetizador de voz, as informações de interesse do usuário. O gerenciador de diálogo deverá também enviar ao interpretador um valor numérico, cuja finalidade é a de avaliar se a qualidade daquilo que foi interpretado foi boa ou ruim.

2. RECONHECIMENTO DA FALA

Conforme mencionado na seção anterior, para o subsistema de reconhecimento da fala foi utilizado como plataforma-base o *Julius*, que é uma ferramenta open-source utilizada no desenvolvimento de sistemas de reconhecimento de fala contínua de vocabulário extenso (**LVCSR** – *Large Vocabulary Continuous Speech Recognition*), que permite a conversão de fala em texto. O software, usando modelos de linguagem baseados em gramáticas *N-gram* e modelos ocultos de Markov (**HMM** – *Hidden Markov Models*) na parte acústica, apresenta excelente desempenho quando roda em PCs [2].

Uma das maneiras de se construir modelos acústicos para o *Julius* é através da utilização do software HTK (*Hidden Markov Model Toolkit*). O software possibilita estimar parâmetros para modelos ocultos de Markov, realizando treinamento com sentenças (ou frases) e suas respectivas transcrições fonéticas.

Num sistema de reconhecimento de fala, assume-se que o sinal da fala é, na verdade, uma sequência de símbolos codificados (fonemas) que deverão ser apresentados na forma escrita. A operação inversa do reconhecimento, a síntese de fala, consiste em gerar uma sequência de símbolos (grafemas) convertidos em trechos de um sinal de fala (forma de onda). Nos dois casos é possível reconhecer padrões através do treinamento de algoritmos baseados nos modelos ocultos de Markov.

Os modelos acústicos e de linguagem usados pelo *Julius* podem ser definidos em qualquer língua, o que oferece aos desenvolvedores a possibilidade de construir sistemas de reconhecimento de fala adaptados para as suas necessidades específicas. Como modelo acústico, utilizamos os recursos para o Português Brasileiro (PB) presentes no site do projeto *FalaBrasil*[1]. Em relação ao modelo de linguagem, uma vez que o nosso objetivo principal era avaliar o comportamento dos modelos utilizados para reconhecimento e síntese da fala, e queríamos criar um sistema para situações específicas em um hotel, criamos frases funcionais comumente utilizadas em um contexto de turismo [1]. A gramática foi desenvolvida seguindo recomendações da documentação do *Julius* [6].

A montagem de uma gramática escrita para o Julius é feita, em uma primeira etapa, através de dois arquivos: ".grammar" e ".voca" [6], conforme ilustrado no exemplo de uma gramática livre de contexto escrita para o Julius, apresentada na Tabela 1.

Em seguida, faz-se necessário converter os arquivos ".grammar" e ".voca" para os respectivos formatos ".dict" e ".voca", que são os realmente utilizados pelo *Julius*. Utilizamos o script *mkdfa.pl*, presente no pacote Julius-4.1.5-win32bin [10] para realizar a conversão da gramática criada no nosso projeto. Por ser um decodificador da fala, capaz de processar sentenças de qualquer idioma, a precisão e desempenho da ferramenta *Julius* depende muito da qualidade dos modelos utilizados. Neste projeto, utilizamos o modelo acústico *LaPSAM* v1.5 [3], disponibilizado pelo grupo *Fala Brasil*.

No caso dos modelos acústicos para o PB, um dos principais desafios é superar a falta de bases de dados disponíveis para treino [3]. O *LaPSAM* v1.5, produzido pelo grupo de pesquisa *Fala Brasil,* se apresenta como uma alternativa aos corpus não gratuitos para PB encontrados no mercado atualmente. Este modelo, arquitetado com a utilização do software *HTK* [8], foi construído a partir de *audiobooks* (livros falados para o português brasileiro).

O *Coruja* 0.2 [9] é uma API desenvolvida pelo grupo *Fala Brasil*, com interface bem simplificada, que permite aos programadores que desejam desenvolver aplicações de

[1] http://www.laps.ufpa.br/falabrasil/

reconhecimento da fala exercer um controle mais prático e ágil sobre a *engine* do *Julius*. Esta ferramenta esconde boa parte da complexidade do decodificador, tornando mais rápido o desenvolvimento de aplicações de reconhecimento de fala em plataforma *.NET*.

Exemplo.grammar	exemplo.voca
S : NS_B HMM QUESTION NS_E S : NS_B QUESTION NS_E GREET : ADJECTIVE_1 PERIOD_1 GREET : ADJECTIVE_2 PERIOD_2 GREET : PEDIDO_1 PEDIDO_2 QUESTION : PEDIDO_1 PREP PEDIDO_2 QUESTION : REFEICAO QUESTION : PLACE_1 DESTINO_1 QUESTION : PLACE_2 DESTINO_2 QUESTION : PAGAMENTO QUESTION : RESPOSTA QUESTION : NUM QUESTION : SEMANA QUESTION : GREET QUESTION : COMANDO	% NS_B <s> sil % NS_E </s> sil % HMM FILLER sil FILLER sil % ADJECTIVE_1 bom b o~ % ADJECTIVE_2 boa b o a % PERIOD_1 dia dZ i a % PERIOD_2 tarde t a R dZ i noite n o j tS i % PEDIDO_1 Reserva R e z e R v a % PEDIDO_2 quarto k w a X t u hotel o t E w % PREP de dZ i

Tabela 1 : Exemplo de gramática definida para o *Julius*

Para criar de uma gramática livre de contexto para o *Julius* é necessário definir uma correspondência entre os grafemas e os fonemas das palavras utilizadas. Isso é feito através do arquivo ".voca", conforme ilustrado na Tabela 1. Para obter o dicionário de pronúncia, i.e., os fonemas correspondentes às palavras existentes nas categorias que criamos, utilizamos a ferramenta *Conversor Grafema-Fone1.5* [10]. Este software é mais um dos recursos disponibilizados pelo grupo *Fala Brasil* e que foi utilizado durante a criação do nosso modelo de linguagem. Posteriormente, foi preciso corrigir manualmente algumas das sequências fonêmicas geradas automaticamente pelo software, pois, embora esta ferramenta seja de grande utilidade prática e proporcione um grande ganho no tempo de desenvolvimento, ela não é perfeita e sempre ocorrem alguns erros de conversão.

3. APRENDIZADO POR REFORÇO

Aprendizado por Reforço é um conceito que foi originalmente estudado, nos anos 60, por pesquisadores ligados à psicologia e que, inicialmente, envolveu a análise de como os animais reagiam a estímulos caracterizados como recompensa negativa (dor, fome) e positiva (ingestão de alimentos) [11]. Sistemas de aprendizado por reforço estão baseados no paradigma de tentativa e erro, sucesso e fracasso, sendo que, em linhas gerais, o objetivo é que o sistema aprenda, através da experiência, quais ações levam as melhores recompensas e quais devem ser evitadas por não produzirem bons resultados.

Formalmente, um sistema de aprendizado por reforço pode ser descrito como um agente que interage com o ambiente através de ações e de maneira estocástica [12]. Tal sistema pode ser descrito como tendo um número finito ou infinito de estados $\{S_i\}$ para um dado número de ações $\{a_j\}$ que podem ser realizadas pelo agente. Para cada estado é associada uma ação com uma probabilidade de transição $P(s'|s,a)$. Ou seja, estando no estado s, no instante t, e realizando a ação a, o agente é levado para o estado s'. Esta transição é associada a uma recompensa representada por $R(s,a)$. O mapeamento de todos os estados $\{S\}$ para todas as ações $\{A\}$ é definido como política π, sendo que o objetivo de um algoritmo de aprendizado por reforço é encontrar uma política ótima (π^*), ou seja, aquela que maximize a recompensa R.

A proposta neste trabalho é considerar o interpretador de língua natural como o agente que aprende a política ótima, uma vez que ele definirá a estrutura semântica que será utilizada pelo gerenciador de diálogo, de forma que este possa escolher a melhor ação a ser executada.

3.1 Interpretador de Língua Natural

Uma vez reconhecidas as palavras pronunciadas pelo usuário, o próximo passo consiste em compreender o significado do que foi falado. Para o contexto proposto neste trabalho, ou seja, o de permitir que o usuário obtenha informações turísticas através de um sistema de diálogo automático, três atributos deverão ser definidos durante o processo de interpretação: atividade, local e período.

A estratégia será garantir que o interpretador de língua natural, em conjunto com o gerenciador de diálogo, obtenha estes atributos. Para definir os atributos o interpretador irá realizar um *parsing* da frase reconhecida, considerando as estruturas gramaticais definidas no módulo de reconhecimento da fala. Desta maneira, uma frase do tipo "Eu gostaria de visitar o Jardim Botânico" seria representada da seguinte maneira:

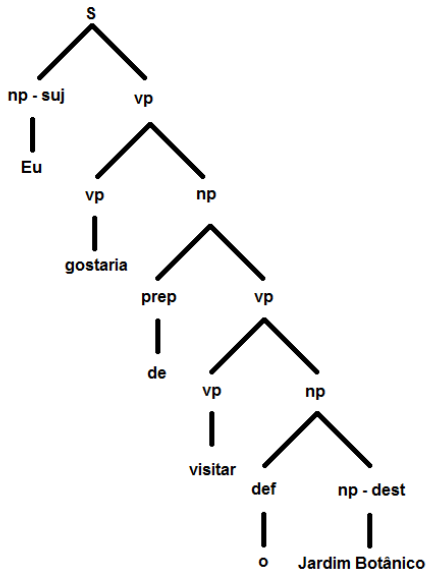


Figura 2 – Estrutura gramatical definida pelo interpretador

3.2 Modelagem e Treinamento

O sistema será modelado definindo-se um conjunto de ações (*queries* e apresentação de informação ao usuário) para cada estado. Cada ação estará associada a uma recompensa $R(s,a)$, enviada pelo gerenciador de diálogo, sendo que $R(s,a) = 0$ para uma confirmação negativa e $R(s,a) = 1$ para uma confirmação positiva. O que se pretende, inicialmente, uma vez que não se conhece a política que maximize a recompensa ao final do processo, é

executar, de forma aleatória, uma sequência de ações (ou diálogos), que representarão um conjunto de treinamento dado por $\{(s1,a1); (s2,a2), (s3, a3),...,(sn,an)\}$. Através dessa sequência será possível estimar a probabilidade, $P(s',a')$, de uma ação ocorrer em determinado estado. Dessa maneira, pretende-se que o agente aprenda, através da experiência, a política ótima π^*. Para tanto, será utilizado um algoritmo supervisionado, ou seja, que realize o cálculo da função de utilidade $Q(s,a)$ com base num resultado final previamente conhecido. A função de utilidade é calculada considerando-se não apenas a recompensa obtida pela ação *a* no estado s, mas também a somatória de todas as recompensas de ações futuras, $Q(s',a')$, que maximizem o resultado final, dada a probabilidade de transição $P(s',a')$. O algoritmo para a função de utilidade é dado pela seguinte fórmula [11]:

$$Q(s,a) = R(S,a) + \Sigma P(s'|s,a) \max Q(s',a') \qquad (1)$$

4. GERENCIADOR DE DIÁLOGO

Após o *parsing* e execução do algoritmo de aprendizado por reforço, uma estrutura semântica contendo três atributos é enviada para o gerenciador de diálogo. O principal papel do gerenciador de diálogo é transformar a estrutura recebida em uma ação que, consequentemente, irá resultar na transição para um novo estado.

Além disso, o cabe ao gerenciador a responsabilidade de interação com o usuário. Portanto, perguntas, confirmações, apresentação de informações da base de dados, verificação de satisfação do usuário, cálculo e envio da recompensa associada à determinadas ações para o módulo interpretador de língua natural são típicas tarefas que estarão atribuídas ao gerenciador de diálogo. Haverá uma base de dados que deverá ser alimentada com as informações turísticas e que receberá as consultas definidas pelo próprio gerenciador de diálogo.

5. DISCUSSÃO E CONCLUSÃO

Um dos maiores problemas em se utilizar aprendizado por reforço em problemas de processamento de língua natural é a dificuldade em se encontrar um *corpus* adequado para treinamento [12]. Mesmo quando se tem um *corpus* que atende às necessidades do domínio envolvido, o processo de aprendizagem é muito custoso tendo em vista a quantidade de possibilidades (ação, estado) com que o algoritmo precisaria ser treinado. Por isso, optamos por definir uma gramática que, de certa forma, restringe a flexibilidade plena do sistema, sendo, no entanto, uma solução de compromisso em face das dificuldades acima relatadas. Futuramente queremos encontrar maneiras de expandir o corpus e aumentar a capacidade de treinamento. Isso tornaria o sistema mais próximo de um atendimento semelhante ao humano, do ponto de vista da percepção do usuário.

Uma das questões mais difíceis em processamento de fala e de língua natural é fazer o computador reconhecer e entender adequadamente a fala e, principalmente, de encontrar os meios adequados de interagir com o usuário humano por meio da sua própria língua. O sistema ainda se encontra em fase embrionária de desenvolvimento; o que se pretendeu com esse trabalho foi apresentar de maneira sintética, porém abrangente uma proposta de arquitetura, combinando técnicas de aprendizado de máquina e processos de decisão de Markov, para se obter um sistema de atendimento automático baseado no entendimento da fala.

6. REFERÊNCIAS BIBLIOGRÁFICAS

[1] AVILA, A. R.; FRAGA, F. J. Integração de Sistemas de Reconhecimento, Tradução e Síntese Automática de Fala para Facilitar a Comunicação de Turistas. In: LAC 2011 – The Fourth Latino American Conference of the Audio Engineering Society, 2011, Montevidéu.

[2] Lee, A. The Julius book. Nagoya Institute of Technology, 2010.

[3] Silva, P., Neto, N., Klautau, A. Novos Recursos e Utilização de Adaptação de Locutor no Desenvolvimento de um Sistema de Reconhecimento de Voz para o Português Brasileiro. XXVII Simpósio Brasileiro de Telecomunicações, 2009.

[4] Silva, P., et al. Speech Recognition for Brazilian Portuguese using the Spoltech and OGI-22 Corpora. XXVI Simpósio Brasileiro de Telecomunicações, 2008.

[5] Neto, N., Patrick, C., Klautau, A., Trancoso, I.. Free tools and resources for Brazilian Portuguese speech recognition. Journal of Brazilian Computer Society, v. 17, n. 1, p. 53-68, Nov.2010.

[6] How to write a recognition Grammar for Julius. URL: <http://julius.sourceforge.jp/en_index.php?q=en_grammar.html>, acessado em: 21/06/2011.

[7] Get Julius for Windows API. URL: <http://julius.sourceforge.jp/en_index.php?q=index-en.html#sapi>, acessado em: 22/06/2011.

[8] HTK. URL: <http://htk.eng.cam.ac.uk>, acessado em: 22/06/2011.

[9] Neto, N., et al. Um Reconhecedor de Voz Livre para Português Brasileiro com Interface de Programação. URL: <http://www.laps.ufpa.br/falabrasil/files/coruja-doc-pt.pdf>, acessado em: 23/06/2011.

[10] Reconhecimento de Voz Para o Português Brasileiro – Conversor Grafema-Fone v1.5. URL: <http://www.laps.ufpa.br/falabrasil/downloads.php>, acessado em: 23/06/2011.

[11] RUSSELL, S.; NORVING, P. Artificial Intelligence. Prentice Hall, 2009.

[12] LEMON, O.; PIETQUIN, O. Machine Learning for Spoken Dialogue System. INTERSPEECH, 2007.

SESSÃO 2 / *SESSION 2 - PAPERS 2*

Fotos: SXC

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Transformada Q constante de comprimento de janela fixo

Jonathan Alis Salgado Lima, Márcio C.P. Brandão e Mylène C.Q. Farias[1]

[1] Universidade de Brasília, Instituto de Ciências Exatas, Departamento de Ciência da Computação
Brasília, DF, Brasil

jonathanalis@gmail.com, brandao@unb.br, mylene@unb.br

RESUMO

Neste artigo, apresentamos uma proposta de transformada de análise tempo-frequência, baseada na transformada Q constante (Constant Q Transform – CQT), com tamanho fixo de janela temporal. A transformada é voltada para aplicações em análise de sinais de música. Assim como a CQT, a transformada proposta possui uma resolução de frequência exponencial, o que apresenta vantagens na análise de sinais de música. Diferentemente da CQT, a técnica proposta utiliza um banco de filtros passa-faixa com comprimento fixo de janela temporal, o que resulta em uma resolução temporal fixa nas frequências de análise. Devido a independência entre as resoluções de tempo e frequência da técnica proposta, obtivemos melhores resultados do que os resultados obtidos utilizando a CQT. A técnica se mostrou adequada para aplicações como reconhecimento de alturas musicais em sinais de música.

0 INTRODUÇÃO

Na análise de sinais de música é conveniente utilizar intervalos de frequências espaçados geometricamente, uma vez que a frequência se relaciona exponencialmente com a altura musical na escala musical bem temperada, a mais utilizada no ocidente [4]. Esta característica dos sinais de música dificulta a utilização das transformadas mais populares em sua análise, como por exemplo a Transformada de Fourier de Tempo Curto (*Short Time Fourier Transform* – STFT) [6], onde as frequências de análise se encontram linearmente espaçadas.

A transformada Q constante (*Constant Q Transform* – CQT) é uma transformada frequentemente utilizada para a análise de sinais de música [1, 2], sendo implementada utilizando uma série de filtros com espaçamento logarítmico. Mais especificamente, cada filtro possui uma largura de banda que é múltipla da largura de banda do filtro anterior, conforme detalhado na Seção 4. Desta forma, a CQT apresenta o espaçamento geométrico de frequência adequado para a análise de sinais de música. Por outro lado, no domínio do tempo os filtros da CQT possuem tamanho de janela que varia

de acordo com a frequência analisada. Portanto, quanto maior a frequência, menor é a janela de análise. Esta característica da CQT compromete a resolução temporal do conteúdo de frequências baixas dos sinais.

Neste trabalho, propomos uma transformada para análise de sinais de música que, além de apresentar uma resolução de frequência exponencial, possui uma resolução fixa no tempo: a Transformada Q Constante de Comprimento de Janela Fixo (*Constant Q Transform with Fixed length Window* – CQT-FW). A transformada é implementada utilizando filtros passa-faixa Chebychev tipo I e janelas de tamanho fixo nas frequências de análise da CQT.

1 Transformada de Fourier de Tempo Curto

A STFT consiste na aplicação da Transformada Discreta de Fourier (*Discrete Fourier Transform* – DFT) individualmente em vários segmentos curtos de um sinal. Para aplicar a STFT em um sinal, primeiramente multiplica-se o sinal por um conjunto de janelas deslocadas no tempo e não sobrepostas, de maneira a 'recortar' o sinal em vários segmentos curtos. Em seguida, aplica-se a DFT a cada segmento do sinal [6].

A fórmula da STFT é dada pela seguinte equação:

$$X(m,\omega) = \sum_{n=\infty}^{\infty} x[n]w[n-m]e^{-j\omega n}, \tag{1}$$

onde $x[n]$ é o sinal amostrado no tempo, $w[n-m]$ é a janela temporal e m é o deslocamento da janela. Observa-se que $X(m,\omega)$ é uma função tanto da frequência (ω) quanto do deslocamento no tempo (m). O espectrograma é obtido calculando-se a magnitude $|X(m,\omega)|$ para cada frequência e deslocamento no tempo.

Uma das características da DFT é que um aumento da resolução na frequência gera uma diminuição da resolução no tempo, ou seja, se aumentarmos a janela temporal (diminuindo assim a resolução no tempo), teremos um maior número de frequências analisadas. Assim, teremos uma boa resolução de frequências e passos pequenos no domínio do tempo. Porém, esta abordagem leva à uma superposição das janelas temporais, causando um certo espalhamento no tempo.

2 Transformada Q Constante

Enquanto a STFT tem as frequências espaçadas de forma linear, a Transformada Q Constante (CQT) permite escolher as frequências de análise com larguras de banda com crescimento exponencial [1, 2]. A razão entre uma frequência de análise e a frequência imediatamente anterior é sempre constante e igual a $2^{\frac{1}{m}}$, ou seja, $f_k = 2^{\frac{1}{m}} f_{k+1}$. Sendo assim, as frequências de análise da CQT são dadas pela seguinte equação:

$$f_k = 2^{\frac{k}{m}} f_{min}, \tag{2}$$

na qual f_{min} é a frequência mínima e m é a resolução de frequência relativa, correspondente ao número de frequências de análise entre f_0 e $2f_0$ (uma oitava musical).

Na STFT, a largura de banda (Δf) é a razão entre a taxa de amostragem e o tamanho da janela. Definimos o fator de qualidade Q como sendo a razão entre a frequência e a largura de banda, ou seja:

$$Q = \frac{f_k}{\Delta f}. \tag{3}$$

Na STFT, Q varia com a frequência de análise. Na CQT, o valor de Q é sempre constante para todas as frequências de análise. Para isto, o tamanho da janela temporal deve variar inversamente com a frequência de análise, ou seja, o comprimento da janela da k-ésima frequência $N[k]$ deve ser igual a

$$N[k] = \frac{f_s}{\Delta f_k} = \frac{f_s}{f_k}Q, \tag{4}$$

onde f_s é a frequência de amostragem. Aplicando essas condições à fórmula da DFT, obtemos a fórmula da CQT:

$$X[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} W[k,n]x[n]e^{\frac{-jk\pi Qn}{N[k]}} \tag{5}$$

na qual $W[k,n]$ é a janela, que varia com n e com k, e $N[k]$ é o tamanho da janela.

Para a análise de música, utilizamos a CQT com janelas deslizantes, como na STFT. Mas, como o tamanho das janelas varia com a frequência de análise, o passo deve ser escolhido com cuidado. Se o passo for muito grande (por exemplo, do tamanho da janela correspondente à frequência inferior f_{min}), as janelas das frequências mais altas não conterão toda a informação do sinal. Se, por outro lado, o passo for pequeno (por exemplo, do tamanho da janela correspondente à frequência superior f_{max}), haverá superposição entre as janelas de forma diferente para cada frequência de análise.

3 Transformada Q Constante de Comprimento de Janela fixo

Conforme apresentado na Seção 1, na STFT as frequências de análise são linearmente espaçadas, o que não é o cenário ideal para a análise musical. Além disso, quanto maior a resolução na frequência, menor é a resolução no tempo, ou seja, para uma alta resolução de frequência precisaríamos de uma janela de tempo maior. Esse problema pode ser contornado tomando passos de menor tamanho que o próprio tamanho da janela.

Já a CQT produz uma resolução de frequência adequada para a análise musical, porém ao custo de uma variação no tamanho das janelas temporais de análise ao longo das frequências. Em outras palavras, a desvantagem da CQT é que as janelas correspondentes às

baixas frequências são maiores que as janelas correspondentes às altas frequências. Sendo assim, se o passo for grande, as janelas das frequências mais altas serão menores que o passo, fazendo com que elas não contenham toda a informação do sinal. Se, por outro lado, o passo for pequeno, as frequências mais baixas conterão informação das janelas vizinhas, gerando espalhamento que varia de acordo com a frequência.

Neste trabalho, implementamos uma transformada que preserva as frequências de análise em escala exponencial, mas que garante uma resolução temporal constante para todas as frequências de análise, a qual nomeamos Transformada Q Constante de Comprimento de Janela fixo (*Constant Q Transform with Fixed length Window* – CQT-FW). O projeto da CQT–FW é realizado utilizando um banco de filtros centrado nas frequências da escala temperada. As frequências de análise utilizadas na CQT–FW são as mesmas da CQT, ou seja:

$$f_k = (2^{\frac{k}{p}}) f_{min}, \tag{6}$$

na qual p é o número de frequências que serão analisadas a cada oitava. Este número deve ser um múltiplo de 12, uma vez que em uma oitava existem 12 notas musicais. Desta forma, as frequências de análise corresponderão exatamente às notas musicais. Com este objetivo, construímos um banco de filtros passa-faixa com bandas de passagem centradas nestas frequências de análise. Nesta seção, descreveremos os detalhes do projeto da transformada proposta.

3.1 Projeto do banco de filtros

Para implementar a transformada proposta, utilizamos um banco de filtros com banda de passagem proporcional à distância entre as sucessivas frequências de análise. Em outras palavras, a banda da esquerda deve ser proporcional à diferença $(f_k - f_{k-1})$ e a banda da direita proporcional à diferença $(f_{k+1} - f_k)$. Dessa forma, a banda de passagem cresce com a frequência de análise f_k, mantendo o valor Q constante.

Ao contrário do que acontece na STFT, a frequência f_k não está localizada no centro da banda de passagem. Isto se deve ao espaçamento exponencial das frequências que é necessário para análise do sinal de música. Entretanto, se considerarmos a escala musical igualmente temperada, a frequência f_k está centralizada nessas frequências. A banda de rejeição também segue a mesma regra.

Definimos, então, um filtro passa-banda Chebychev tipo I com banda de passagem igual a:

$$[f_k - c_1(f_k - f_{k-1}), f_k + c_1(f_{k+1} - f_k)] \tag{7}$$

e banda de rejeição igual a:

$$[f_k - c_2(f_k - f_{k-1}), f_k + c_2(f_{k+1} - f_k)], \tag{8}$$

onde c_1 e c_2 são constantes, de tal forma que $c_2 > c_1$ para que a banda de rejeição seja maior que a de passagem. Esses coeficientes devem ser escolhidos de tal forma que a soma das magnitudes das respostas em frequência de todos os filtros seja o mais próximo possível de 1, para que a energia seja preservada.

A partir das equações 6 e 7, obtemos a largura de banda Δf_k. Logo, o fator de qualidade Q do filtro, definido como sendo a frequência de análise dividida pela largura de banda, é dado pela seguinte equação:

$$Q = \frac{f_k}{\Delta f_k} = \frac{f_{min} \cdot 2^{k/p}}{c_1 \cdot f_{min}\left(2^{\frac{k+1}{p}} - 2^{\frac{k-1}{p}}\right)}$$

$$Q = \frac{1}{c_1} \frac{1}{\left(2^{1/p} - 2^{-1/p}\right)}. \tag{9}$$

Como p e c_1 são constantes, o fator de qualidade Q é constante para todas as frequências. Além disso, garantimos que a banda de rejeição de um filtro termina onde começa a banda de rejeição do próximo filtro.

Como para o nosso objetivo é importante que a resposta na banda de rejeição seja monotonicamente decrescente, o tipo de filtro utilizado foi o Chebychev tipo I [6]. Outra vantagem é que a ordem obtida para o filtro Chebychev é menor do que a ordem obtida com o filtro Butterworth para as mesmas especificações. A Figura 1 ilustra a resposta em frequência do filtro projetado. O gráfico superior da figura mostra a amplitude da resposta em dB para todo o espectro, enquanto que o gráfico inferior mostra a amplitude da resposta para o intervalo entre $f_{k-1} = 362,1$ e $f_{k+1} = 367,3$.

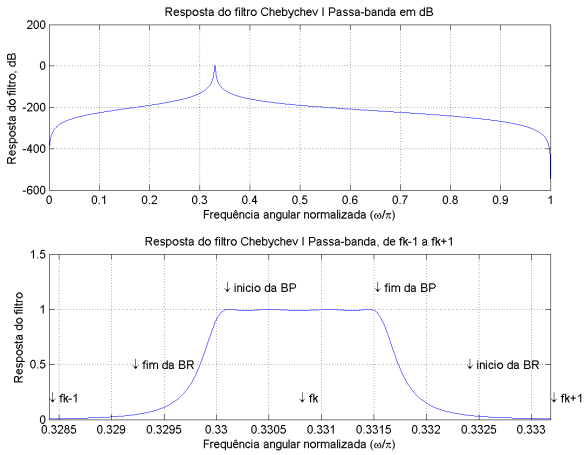


Figura 1: Amplitude em dB da resposta em frequência do filtro Chebychev I passa-banda de ordem 8, com frequência central $f_k = 364,7$ e frequência de amostragem 2.205 Hz. O gráfico superior mostra a resposta para todo o espectro, enquanto que o inferior a resposta para o intervalo entre $f_{k-1} = 362,1$ e $f_{k+1} = 367,3$.

3.2 Fator de Reamostragem

Uma desvantagem da técnica descrita na Seção 3.1 é que a banda de passagem pode ficar bastante estreita se o número de frequências de análise por oitava (resolução de frequência) for muito alto ou se a

frequência mínima de análise for muito baixa. Isto torna o projeto do filtro passa-banda mais complicado. Uma possível solução para este problema consiste em sub-amostrar o sinal segmentado com frequência de amostragem f_s. Desta forma, a banda de passagem torna-se mais larga em relação à nova frequência máxima do sinal ($f_s/2$ ou π radianos). Para evitar *aliasing*, o sinal é pré-filtrado utilizando um filtro passa-baixas com frequência de corte igual $f_s/2$.

Como exemplo, considere um sinal com frequência máxima de 44.100 Hz. Se sub-amostrarmos este sinal utilizando uma taxa de 4.410 Hz, é possível obter uma banda de passagem 10 vezes maior do que a original. Isto simplifica muito o projeto do banco de filtros. Evidentemente, este processo limita a frequência de análise superior a $f_s/2$.

3.3 Análise no Domínio da frequência

A relação de Parseval para a transformada de Fourier de Tempo Discreto (*Discrete Time Fourier Transform* – DTFT) é dada por:

$$\sum_{-\infty}^{\infty} |x[n]|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |X(e^{j\omega})|^2. \tag{10}$$

Se considerarmos as restrições do nosso sinal, temos que ele é real e foi filtrado em torno de uma única frequência central $\omega_k = 2\pi f_k$. Podemos, então, considerar que ele é similar a uma senóide com apenas uma componente de frequência, ou seja,

$$\begin{aligned} x[n] &= A_k cos(\omega_k n + \phi) \\ &= \frac{A_k}{2}(e^{j(\omega_k n+\phi)} + e^{-j(\omega_k n+\phi)}). \end{aligned} \tag{11}$$

Como a DTFT do sinal dentro do intervalo $[-\pi, \pi]$ é dada por

$$X(e^{j\omega}) = \pi A_k e^{j\phi}(\delta(\omega - \omega_k) + \delta(\omega + \omega_k)),$$

podemos reescrever a eq. 10 como

$$\int_{-\pi}^{\pi} |X(e^{j\omega})|^2 = A_k^2/2. \tag{12}$$

Combinando as eqs. 10 e 12, obtemos a seguinte relação:

$$A_k = \sqrt{2\sum_{-\infty}^{\infty} x[n]^2}. \tag{13}$$

Este resultado indica que a energia de um sinal $x[n]$ filtrado em ω_k é proporcional à raiz quadrada das somas quadráticas de $x[n]$. Levando em consideração os limites da janela análise ($l \leq n \leq l + L - 1$), podemos reescrever a eq. 13 expressando A_k em função de l. Desta forma, podemos obter o valor de A_k para cada janela temporal. Ou seja, obtemos a energia de cada janela temporal e frequência de análise f_k, gerando o espectrograma.

Uma vez que estamos interessados apenas nas energias relativas entre as frequências e não nos seus valores absolutos, por uma questão de simplicidade eliminamos o fator $\sqrt{2}$ da eq. 13, obtendo a seguinte expressão:

$$A_k(l) = \sqrt{\sum_{n=l}^{l+L-1} x[n]^2}. \tag{14}$$

Isso é feito para a saída de cada um dos filtros. O coeficiente $A_k(l)$ é associado à frequência de análise f_k e à l-ésima janela do sinal original. Com isso, obtemos um espectrograma que mostra a magnitude para cada frequência de análise e para cada janela temporal do sinal.

3.4 A janela

Utilizamos nesse trabalho a janela Blackman [6], uma vez que esta janela apresentou melhores resultados para a nossa aplicação. Para manter a energia constante, a soma das janelas deslocadas deve ter um valor constante. Atingimos um valor constante de amplitude na soma das janelas utilizando um passo de deslocamento temporal igual a $1/3$ do tamanho da janela Blackman utilizada. A Figura 2 ilustra esta característica para sete janelas.

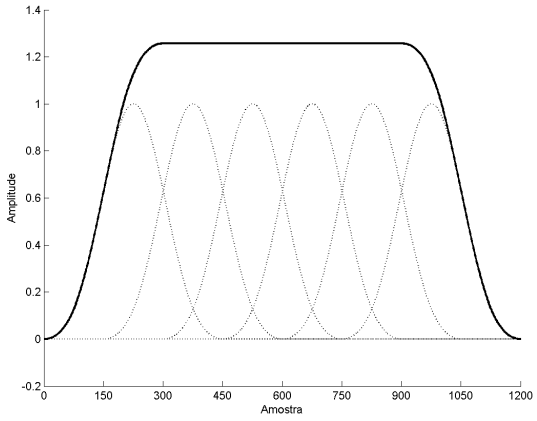


Figura 2: Amplitude de 7 janelas do tipo Blackman de igual comprimento. O passo utilizado é $1/3$ do tamanho das janelas. As linhas tracejadas representam as sete janelas Blackman, enquanto que a linha cheia representa o somatório das amplitudes destas janelas.

3.5 Ganho na amplitude

Resultados iniciais mostraram que, após aplicação da técnica descrita, as baixas frequências de todos os sinais testados apresentavam menor energia do que seria esperado. Foi necessário, então, aplicar um ganho ao resultado final, de forma a compensar esse 'desequilíbrio' entre as energias das diferentes frequências

de análise. Para obter estes ganhos, utilizamos um ruído branco (contendo todas as frequências analisadas) como sinal de entrada. A partir da energia média obtida para cada frequência de análise, obtemos o fator de ganho necessário para que todas as frequências tenham energia equivalente. Desta forma, foi possível compensar as energias aplicando este ganho à saída de cada uma das frequências.

4 RECONHECIMENTO DE ALTURAS MUSICAIS

A técnica proposta pode ser utilizada para o reconhecimento das alturas de notas musicais em um sinal de áudio. Uma possível aplicação é a obtenção automática de partitura de músicas. Nessa seção, utilizamos a CQW–FW para reconhecimento de alturas musicais. Para isto, utilizamos uma alteração da técnica *Gathered log spectrum* [5, 3] com o objetivo de obter os melhores candidatos às frequências fundamentais e, portanto, boas estimativas das alturas musicais presentes no áudio.

4.1 *Gathered log Spectrum*

Na técnica *Gathered log Spectrum* [5] [3], o espectro de frequência é manipulado algebricamente de tal forma que os picos remanescentes dessa manipulação nos fornecem os melhores candidatos para as frequências fundamentais do sinal musical. A técnica consiste de três passos. O primeiro passo é o cálculo do $\rho_0(f_0)$, que atribui à frequência uma média do logaritmo dos seus harmônicos. Desta forma, a energia das notas musicais que possuem muitos harmônicos é intensificada.

Uma estimativa $\rho_0(f_0)$ pode ser obtida utilizando a seguinte equação:

$$\rho_0(f_0) = \frac{1}{N_H} \sum_{i=0}^{N_H} log|S(i \cdot f_0)|, \tag{15}$$

onde $S(i \cdot f_0)$ é a magnitude do sinal na frequência $i \cdot f_0$, ou seja, o i-ésimo harmônico de f_0. O valor N_H é uma estimativa do número de harmônicos múltiplos da frequência f_0. Quando não se conhece ao certo o valor de N_H, esse valor é estimado como sendo o número de múltiplos de f_0 que preenchem o espectro até a frequência máxima $\frac{f_s}{2}$. Caso os múltiplos de f_0 não estejam no espectro, pode-se usar uma interpolação para obter uma estimativa da magnitude do sinais nas frequências desejadas. Essa técnica utiliza o logaritmo para evidenciar as parciais de frequências mais fracas, minimizando o efeito do ruído.

Ao aplicarmos esse primeiro passo, reforçamos as frequências fundamentais, mas geramos submúltiplos das fundamentais. Portanto, o segundo passo do algoritmo consiste em realizar um pós-processamento para retirar estes múltiplos. Isto é feito subtraindo-se, para cada frequência, a energia da frequência submúltipla que tem a maior energia, conforme expresso pela seguinte equação.

$$\rho_1(f_0) = \rho_0(f_0) - \max_{q \in \mathbb{N}} \rho_0 \left(\frac{f_0}{q} \right) \tag{16}$$

Este procedimento atenua os candidatos às frequências fundamentais que são múltiplos das verdadeiras frequências fundamentais. Quando se trata de sons monofônicos este procedimento é suficiente para o reconhecimento da freqência fundamental. A frequência de pico após este pós-processamento é a estimativa da frequência fundamental, ou seja $\hat{f}_0 = \max(\rho_1(f_0))$. Os picos dos sub-múltiplos não afetam a estimativa, uma vez que as suas amplitudes são necessariamente menores do que a amplitude da verdadeira frequência fundamental.

No caso polifônico, os picos de submúltiplos devem ser removidos. Picos de submúltiplos de uma frequência fundamental geralmente não serão maiores que a amplitude da frequência fundamental que os gerou, mas eles podem ser maiores do que as outras frequências fundamentais do sinal. Logo, o terceiro passo do algoritmo consiste na remoção dos picos de submúltiplos. Para isso, a função $\rho_1(f_0)$ é processada outra vez a fim de remover os submúltiplos de f_0. A fórmula para remoção do $(k-1)$-ésimo submúltiplo é a seguinte:

$$\rho_2(f_0) = \rho_1(f_0) - a_k \rho_1(k f_0), \tag{17}$$

onde a_k é um fator de atenuação.

A Figura 3 mostra a aplicação dos três passos do algoritmo *Gathered log spectrum* em um sinal musical. O gráfico superior da figura mostra o cálculo de ρ_0, o gráfico do centro mostra a atenuação dos múltiplos dos f_0's e o gráfico inferior mostra a atenuação dos submúltiplos dos f_0's. O trecho de música testado tem 3 frequências fundamentais. Os asteriscos no gráfico representam uma das frequências fundamentais, seus múltiplos e seus submúltiplos. Os círculos e quadrados representam duas outras frequências fundamentais e seus múltiplos e submúltiplos correspondentes. Podemos observar dos gráficos que a técnica ajuda a remover a ambiguidade na detecção das frequências fundamentais. Conseguimos obter melhores estimativas das frequências fundamentais selecionando apenas os picos.

4.2 A Técnica Utilizada

De acordo com Cancela [3], um dos problemas da técnica *Gathered log spectrum* e do seu pós-processamento é que, nos casos polifônicos, a atenuação dos sub-harmônicos não é adequada. Isto se deve ao fato que, mesmo após o pós-processamento, a energia destes sub-harmônicos é frequentemente maior do que a energia de outras frequências fundamentais.

Optamos, então, por não realizar o cálculo de $\rho_0(f_0)$ para não gerar os sub-harmônicos, e nem rea-

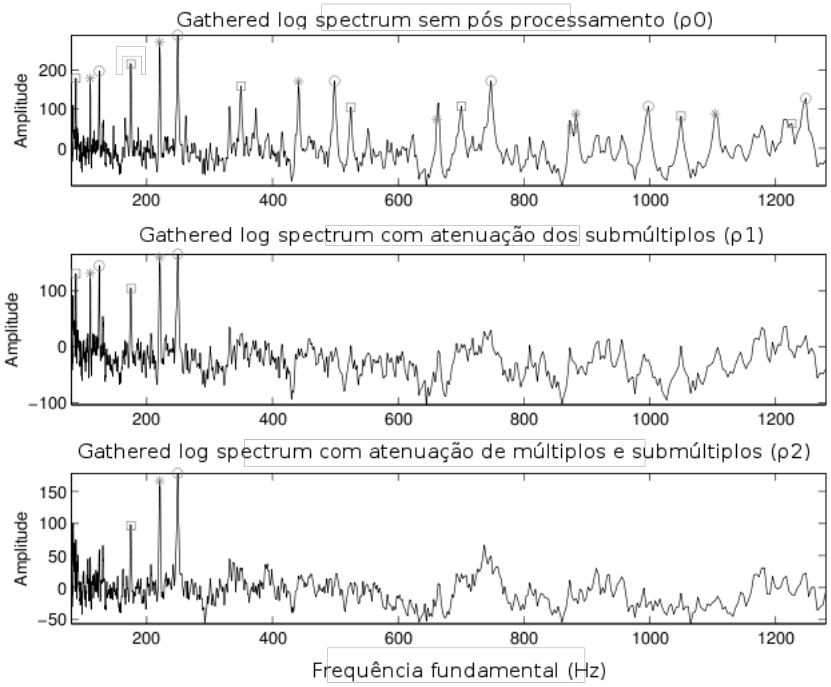


Figura 3: Resultados dos três passos da aplicação da técnica Gathered log spectrum, adaptado de [3]. Superior: Cálculo de ρ_0. Centro: Atenuação dos múltiplos dos f_0's. Inferior: Atenuação dos submúltiplos dos dos f_0's.

lizar o cálculo de $\rho_2(f_0)$, pois não precisamos eliminá-los. Portanto realizamos apenas o cálculo do $\rho_1(f_0)$, que permite eliminar os harmônicos, deixando apenas a frequência fundamental. Com isto, para podermos cancelar as frequências harmônicas dos sinais analisados, a frequência fundamental de cada nota da polifonia deve estar presente.

A equação que define ρ_1 depende de ρ_0. Logo, adaptamos esta equação (eq. 16) para depender apenas do espectro sem processamento e limitamos o resultado a valores positivos. A nova equação adaptada para ρ_1 é:

$$\rho_1(f_0) = \max\left[S(f_0) - \max_{q\in\mathbb{N}} S\left(\frac{f_0}{q}\right), 0\right], \qquad (18)$$

onde $S(f_0)$ é a magnitude do sinal na frequência f_0, segundo a CQT-FW. Seguindo este procedimento, obtemos melhores candidatos para as frequências fundamentais. Estes candidatos, por sua vez, são boas estimativas para as alturas musicais, considerando que o sinal de entrada tenha alturas musicais bem definidas.

5 Resultados

Os testes neste trabalho foram todos realizados utilizando a linguagem computacional **MATLAB**, na qual fizemos uma implementação da CQT-FW[1]. Primeiramente, apresentamos os espectrogramas do sinal chirp obtidos utilizando a CQT e a CQT-FW. Em seguida, utilizamos a técnica de reconhecimento de alturas musicais descrita na Seção 4.2 para analisar o conteúdo de um trecho de uma música.

[1]Disponível em http://sourceforge.net/projects/cqtfw/files/

5.1 Sinal de teste – Chirp

Como sinal de teste, utilizamos um sinal *chirp* cujas frequências variam linearmente de 0 a 1.000 Hz em um intervalo de 5 segundos, considerando uma frequência de amostragem de 44.100 Hz. Utilizamos esse sinal de teste porque ele cobre toda a faixa de frequências considerada. Além disso, é um sinal não-estacionário, o que é frequente para o caso de sinais de áudio. Nos espectrogramas gerados o eixo x corresponde ao tempo, enquanto que o eixo y corresponde às notas *MIDI*, que estão relacionadas com a frequência pela função $MIDI = 69 + 12\log_2(f/440)$. O valor da energia e da frequência em cada instante de tempo é dado pela luminância dos pixels no gráfico, ou seja, quanto mais escuro, maior a energia.

Primeiramente, analisamos o sinal chirp utilizando a CQT, com parâmetros $f_{min} = 110$, $f_{max} = 1000$ e passo igual ao período da frequência mínima f_{min}. As resoluções de frequência utilizadas foram 12, 24, 48 e 96 frequências de análise por oitava. Nas Figuras 4 (a) e (b), são apresentados os espectograma do sinal chirp utilizando a CQT para as resoluções com 12 e 96 frequências de análise, respectivamente. Nestas figuras, podemos observar que o aumento na resolução de frequência causa uma diminuição na resolução temporal (ver parte inferior esquerda do gráfico). Isto se deve ao fato de nas baixas frequências a janela de análise da CQT é maior, causando um espalhamento do sinal.

Fizemos o mesmo teste com CQT-FW, utilizando os mesmos limites do intervalo de frequência e das resoluções de frequência. Utilizamos uma subamos-

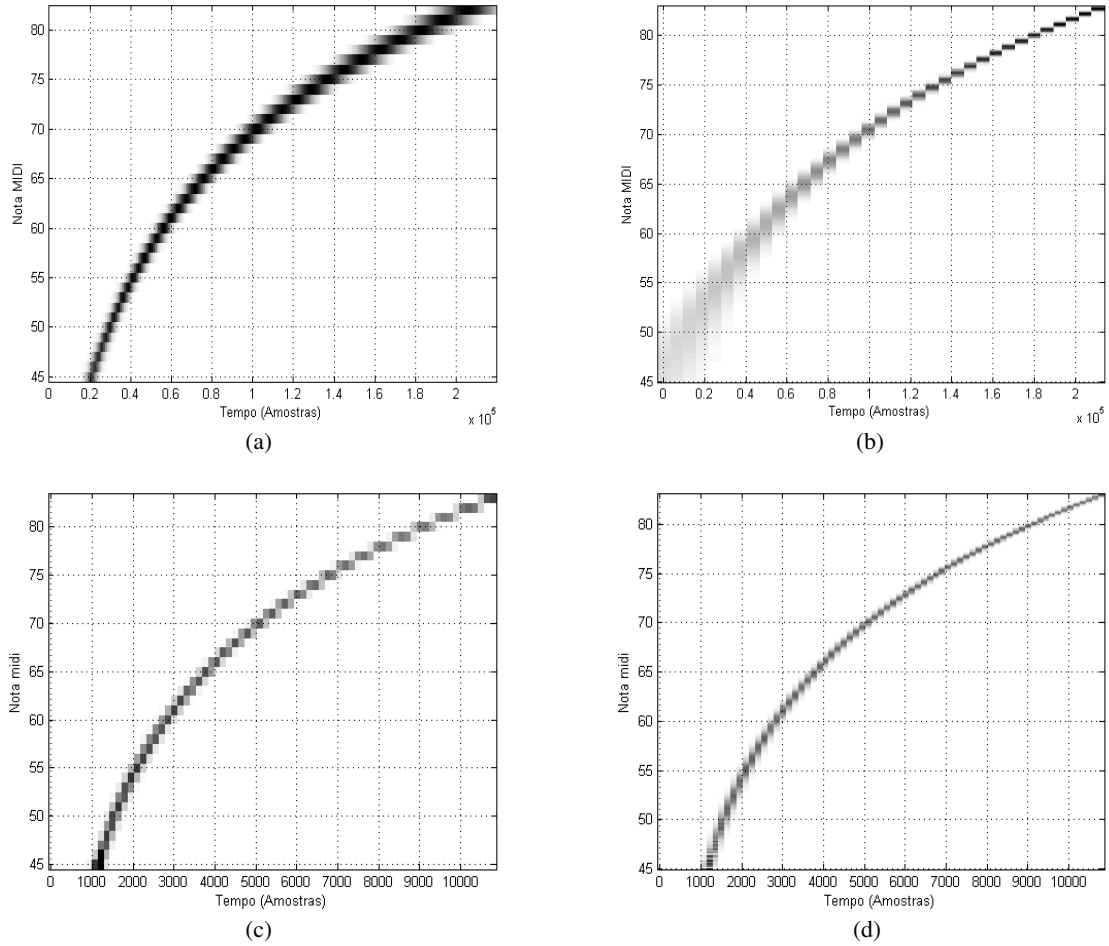


Figura 4: Magnitude do sinal chirp obtida com: (a) CQT com resolução de 12 frequências de análise por oitava, (b) CQT com resolução de 96 frequências de análise por oitava, (c) CQT-FW com resolução de 12 frequências de análise por oitava e (d) CQT-FW com resolução de 96 frequências de análisepor oitava.

tragem da ordem de 20 e um passo de 150 amostras (68.3 ms de um áudio original a 44.100 Hz, subamostrado com ordem 20). Nas Figuras 4 (c) e (d), são apresentados os espectograma do sinal chirp utilizando a CQT-FW para resolução de 12 e 96 frequências de de análise, respectivamente. Assim como no caso da CQT, as frequências de análise já estão expressas em MIDI. Observe que o problema da não-uniformidade do espalhamento no tempo foi resolvido.

Com a resolução de frequência mais baixa, como nas Figuras 4 (a) e (c), não vemos muita diferença entre a CQT e a CQT–FW. Mas, podemos observar que existe uma diferença significativa entre os gráficos da CQT e da CQT–FW quando aumentamos a resolução de frequência, como mostrado nas Figuras 4 (b) e (d). Isso se deve a dois fatores. Primeiro, a resolução temporal da técnica proposta não fica comprometida nas baixas frequências, como na CQT, uma vez que para todas as frequências temos o mesmo tamanho de janela. Segundo, o tamanho da janela na técnica proposta não depende das frequências de análise, nem da resolução (como é verdade para a CQT e a STFT). Assim, podemos manter uma boa resolução temporal, mesmo com uma alta resolução de frequência.

5.2 Reconhecimento de alturas musicais

Nesta seção, apresentamos o resultado da aplicação da técnica proposta para análise de um trecho da música 'While my guitar gently weeps' (Beatles, 1968), com um arranjo para ukulele[2]. Neste sinal temos polifonia com quatro vozes formando acordes com frequências variando de 246 Hz ("Si 3") até 740 Hz ("Fá sustenido 5"). Utilizamos a CQT-FW para encontrar o espectograma, que é apresentado na Figura 5(a). O *f0-grama* é o gráfico que contém as estimativas das alturas musicais no tempo. Em seguida, utilizando a técnica *Gathered log spectrum* revisada (ver Seção 4.2), obtemos o *f0-grama* do sinal, apresentado na Figura 5(b).

O espectrograma mostrou corretamente o conteúdo espectral na faixa de frequências escolhida. É interessante observar que as frequências mais baixas aparecem mais grossas do que as frequências mais altas. Mas,

[2]Disponível em http://www.youtube.com/watch?v=3mEqEe-PIC8, de 5:07 até 5:14

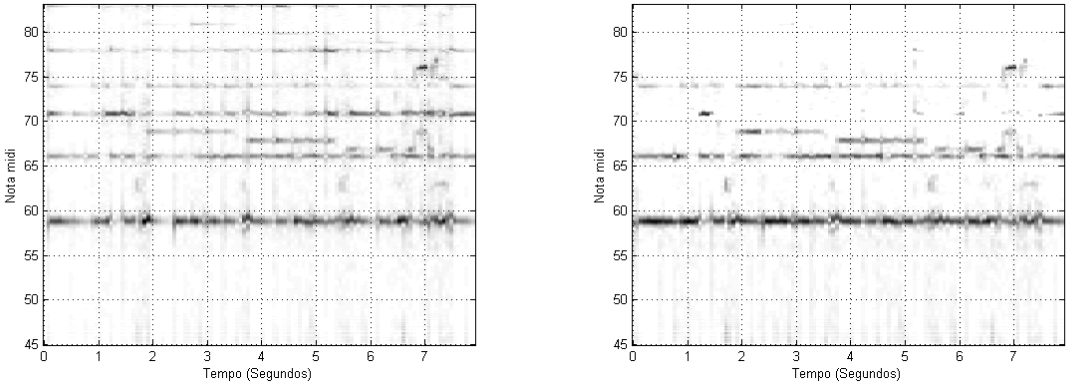


Figura 5: Espectrograma (esquerda) e f0-grama (direita) de um trecho da música 'While my guitar gently weeps' (Beatles, 1968), obtido utlizando a técnica proposta.

esta diferença está dentro de um limite aceitável. Além disso, podemos observar que conseguimos eliminar a energia de harmônicos de ordem mais alta, deixando apenas as frequências fundamentais. Para a nota *MIDI* 71 ("Si 4"), por exemplo, até o tempo 4000 a análise mostra apenas uma fundamental de baixa energia. A técnica identificou essa nota como segundo harmônico da nota *MIDI* 59 ("Si 3") e a eliminou. Ou seja, para o reconhecimento de alturas musicais a técnica apresenta bons resultados. As exceções correspondem aos casos em que fundamentais de baixa energia são interpretadas como harmônicos de uma frequência fundamental mais grave e de maior energia.

6 Conclusão

Apresentamos uma nova transformada para análise musical que, diferentemente da STFT, possui espaçamento entre as frequências de análise exponencial, como na música. A técnica proposta utiliza um banco de filtros, sendo possível utilizar uma resolução temporal fixa para todas as frequências, diferentemente do que acontece na CQT. Além disso, conseguimos uma independência entre a resolução de frequência e a resolução de tempo, o que permite obter simultaneamente uma boa resolução de frequência e tempo. Em relação à STFT e à CQT, as características da técnica proposta apresentam grandes vantagens para a análise de sinais de música. O espalhamento no domínio da frequência para as baixas frequências é aceitável e a técnica proposta apresenta bons resultados para o reconhecimento de alturas musicais, mesmo em polifonia.

Referências Bibliográficas

[1] J.C. Brown. Calculation of a constant q spectral transform. *Journal of Acoustical Society of America*, 89:425–434, 1991.

[2] J.C. Brown and M.S. Puckette. An efficient algorithm for the calculation of a constant q transform. *Journal of Acoustical Society of America*, 92:2698–2698, 1992.

[3] P. Cancela, E. López, and M. Rocamora. Fan chirp transform for music representation. In *Proc. of the 13th Int. Conf. on Digital Audio Effects – DAFx10*, pages 1–8, 2010.

[4] L.L. Henrique. *Acústica musical*. Fundação Calouste Gulbenkian, 2002.

[5] M. Kepesi and L. Weruaga. Adaptive chirp-based time-frequency analysis of speech signals. *Speech Communication*, 48(5):474–492, 2006.

[6] Sanjit K. Mitra. *Digital Signal Processing: A Computer Based Approach*. McGraw-Hill Higher Education, 2005.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

"Estúdio Virtual" para monitoração através de fones de ouvido de sistemas baseados em Ambisonics de 1ª ordem

Fábio Wanderley Janhan Sousa
University of York
York, North Yorkshire, YO10 5DD, UK
fabiojanhan@gmail.com

RESUMO

Esse artigo, parte adaptada da dissertação apresentada pelo presente autor para obtenção do título de mestre em música e tecnologia pela Universidade de York na Inglaterra, apresenta um sistema binaural baseado em tecnologia Ambisonics como solução imediata para monitoração de áudio multicanal através de fone de ouvidos. Desenvolvido para MAX e utilizando ferramentas livres, o "estúdio virtual" pretende ser de fundamental importância para a popularização de produções sonoras multicanais, assim como para a difusão do sistema Ambisonics.

0. INTRODUÇÃO

Basicamente o formato B Ambisonics consiste de um sinal de áudio composto de quatro canais, um omnidirecional – W – e três bidirecionais – X frente-trás, Y esquerda-direita e Z cima-baixo – que podem ser decodificados para diferentes arranjos multicanais de alto-falantes. O que aqui é descrito como um "estúdio virtual" consiste em um patch desenvolvido para MAX que utiliza as vantagens desse sinal que, depois de decodificado, passa por um processo de convolução através de filtros de HRTFs, correspondentes à posição de cada um desses alto-falantes virtuais, e é reproduzido finalmente por apenas dois canais de áudio sem perder informações referentes à espacialização tridimensional do sistema Ambisonics.

O conjunto de HRTFs pode ser escolhido entre as medições realizadas em diversos indivíduos e encontrados no banco de dados "LISTEN" do IRCAM [1], assim como entre as versões compensadas (que eliminam a influência do amplificador, microfones e canais auditivos) ou sem compensação. Um sistema de rastreamento de movimento da cabeça é implementado através do uso de uma webcam (normalmente embutidas nos computadores atuais). São implementados também uma compensação para a influência da utilização do fone de ouvido utilizado na monitoração (por enquanto somente para o modelo AKG 271mkII) e um filtro com ganho na região dos agudos como tentativa de melhorar a localização das fontes sonoras.

O principal objetivo dessa ferramenta de suporte é permitir ao engenheiro de som ou compositor trabalhar em um ambiente virtual baseado na reprodução binaural para fones de ouvido, antes de se deslocar para um estúdio para a etapa de finalização de sua produção, quando deve realizar a monitoração através de um conjunto de alto-falantes reais.

1. SISTEMAS AMBISONICS E BINAURAIS

O sistema Ambisonics se desenvolveu a partir das experiências de Alan Blumlein, na Inglaterra, nos anos 1920, com pares coincidentes de microfones direcionais. Posteriormente se tornou objeto de pesquisa de Peter Fellgett, Peter Craven, Michael Gerzon, entre outros, sendo

patenteado e documentado na década de 1970, época em que o áudio multicanal para fins comerciais começou a se desenvolver e surgiu a possibilidade de se gravar fitas magnéticas de quatro canais.

O princípio fundamental do sistema Ambisonics é que um campo sonoro tridimensional completo observado em um ponto poderia ser captado por um microfone com design específico (microfone *Soundfield*) composto por quatro cápsulas cardióides formando um tetraedro. Este sinal, após codificado para o formato B, pode ser decodificado para diferentes arranjos de alto falantes multicanais, independente de sua quantidade e posição, desde que estejam a uma mesma distância do centro e que apresente pares de alto-falantes diametralmente opostos, onde a soma do sinal de cada par seja sempre igual a dos outros pares. Uma outra característica interessante do sistema Ambisonics é que o sinal no formato B pode ser processado como um todo, realizando rotações, inclinações ou tombamentos. A adição de reverberação também pode ser aplicada ao sinal como uma entidade única, processo realizado preferencialmente por convolução com outro sinal no formato B obtido da resposta a impulso de uma sala gravada com um microfone *Soundfield*. Sua compatibilidade com os sistemas mono e estéreo também é um atributo que diferencia este sistema dos demais sistemas surround.

Os sistemas binaurais surgiram provavelmente antes dos sistemas estéreo, uma vez que já no inicio da história das gravações se observava que, como temos somente dois ouvidos, para se reproduzir um ambiente sonoro tridimensional, dois sinais de áudio deveriam bastar [2]. Alguns autores descrevem os sistemas binaurais como uma tentativa de se isolar o ouvinte do ambiente externo (quando se utiliza os fones de ouvido) e reproduzir exatamente o que se ouviria na melhor posição de uma sala de concerto, com todo o seu conteúdo espacial tridimensional. Sistemas estéreo por outro lado, apesar de também reproduzirem dois canais de áudio são incapazes de reproduzir todo o conteúdo espacial tridimensional.

Preocupações relativas à reprodução binaural estão, na maioria dos estudos, relacionadas a HRTFs individualizadas ou generalizadas, ou seja, alguns apóiam a idéia de que só se é possível uma boa reprodução binaural uma vez que o ouvinte utiliza suas próprias HRTFs na reprodução, outros não consideram HRTFs individualizadas tão importantes quanto outros fatores. Um desses fatores, que também é muito discutido pelos pesquisadores, seria a implementação de um sistema para se rastrear o movimento da cabeça do ouvinte e modificar o som da reprodução para se adaptar a nova posição do mesmo. Tais sistemas, mais conhecidos como *head tracking*, em sua grande maioria possuem a desvantagem da necessidade de hardware e software específico para sua medição e processamento, sendo que alguns autores já se referiram às vantagens que um sistema baseado em webcam traria, como é o caso de [3]. Outro fator a se levar em conta em qualquer reprodução binaural é a reprodução de uma reverberação realística, uma vez que dela se extrai grande parte da informação referente ao ambiente, seja ele real ou virtual.

2. ESTUDOS ANTERIORES

A primeira referência a um sistema que se assemelhe ao presente conceito de "estúdio virtual" foi encontrada em [4], onde o autor descreve aplicações para sistemas binaurais. Ele afirma que, entre outras, a aplicação do sistema para a "reprodução de alto-falantes simulados" é "especialmente atrativa para a simulação de salas de controle", e que essa abordagem "pode competir com as salas de controle ou ser um suplemento das mesmas" uma vez que são mais baratas e menores do que a estrutura normalmente utilizada para produções musicais ou cinematográficas multicanais.

Em [5], os autores mencionam a possibilidade de se utilizar técnicas binaurais para melhorar a qualidade da reprodução estéreo convencional através de fones de ouvido, objetivando evitar o efeito de reprodução "dentro da cabeça". Tal processo também é relatado por [6], que afirma que, entre outras aplicações, tais como simulação de salas, mixagem binaural, teleconferência, reprodução de áudio tridimensional e realidade virtual, o processamento binaural

> pode também ser usado para transformar o som *Dolby Surround* em reprodução para fones de ouvido, ou qualquer outro padrão multicanal, para o qual não haja uma forma simples de se obter uma representação em dois canais para fones de ouvido. Isso é válido para sistemas estéreo também, de fato, onde o sinal estéreo pretendido para alto-falantes dá ao ouvinte a sensação não usual do som sendo reproduzido dentro de sua cabeça.
>
> (Hammershoi, 1996: 3)

No trabalho de [7], do mesmo ano, os autores discutem a implementação de um sistema de rastreamento de movimento da cabeça e afirmam que o "novo decodificador binaural" deles, implementado utilizando o processador da Lake Huron, "soluciona os tradicionais problemas do som binaural e oferece uma forma de se ouvir gravações realizadas em formato B através de fones de ouvido". Na abordagem deles, um conjunto de filtros binaurais estáticos são utilizados e o usuário pode escolher entre diversos conjuntos de HRTFs disponíveis. Simulações de distância são obtidas através de reverberação processada pelo software CAAT Acoustic [8] em formato B Ambisonics. Eles assinalam a possibilidade de monitoração da reprodução de arranjos de alto-falantes através de fones de ouvido:

> Até agora, não havia forma alguma de se monitorar gravações espacializadas durante o processo de produção sem se recorrer a inconvenientes e espaçosos arranjos de alto-falantes. O decodificador binaural resolve este problema. Ele pode ser utilizado para monitorar sons espacializados em todos os estágios do processo de produção.
>
> (McGrath and McKeag, 1996: 5)

Os autores de [9] estabelecem o conceito de formato B binaural após discussão a respeito do custo de se implementar a interpolação entre HRTFs objetivando executar o rastreamento do movimento da cabeça do ouvinte ou o movimento de fontes sonoras e especialmente quando muitas fontes sonoras estão em jogo. De acordo com eles, esse conceito pode ser explorado de duas formas distintas: uma abordagem "pós-filtragem", onde o sinal original é espacializado e mixado em um formato intermediário e posteriormente transcodificado para um sinal binaural utilizando-se filtros não variáveis no tempo; e uma abordagem "pré-filtragem", onde as fontes sonoras são inicialmente processadas pelos filtros e posteriormente espacializadas em tempo real.

O conceito de "downmix" binaural é apresentado pelos autores mencionados acima e consiste na "exploração de uma técnica de espacialização conhecida para reprodução em alto-falantes, e na utilização de filtros binaurais para

gerar uma 'downmix' para a reprodução em fones de ouvido". É importante observar que em tal processo o resultado final é um arquivo de dois canais que não permite mudanças em tempo real que se adaptem ao posicionamento da cabeça do ouvinte. Para que tais mudanças sejam possíveis, o formato B Ambisonics é apresentado como formato intermediário de codificação. "O formato B oferece a possibilidade de se mixar com uma ambiência gravada com um microfone *Soundfield*, e compensar as rotações da cabeça do ouvinte após a mixagem ao se aplicar uma matriz de rotação antes da etapa de decodificação para o formato binaural". Os mesmos autores, um ano depois [10], acrescentam que "o paradigma do 'alto-falante virtual' é uma abordagem generalizada para desenvolver decodificadores que se adaptem a qualquer formato de codificação para reprodução em fones de ouvido ou qualquer arranjo de alto-falantes multicanal".

Os autores de [11], em suas experiências para avaliar a influência do processamento de rastreamento do movimento da cabeça, também apresentam um método de auralização baseado na reprodução de um conjunto de alto-falantes através de fones de ouvido, que eles mesmos denominam "binaural room scanning" (*escaneamento binaural de sala*), ou BRS. Este processo é descrito por eles como um sistema de monitoração de diferentes setups de múltiplos alto-falantes. Alguns outros trabalhos que confirmam que a confusão entre fontes sonoras localizadas posterior ou frontalmente pode ser solucionada com a implementação de um sistema de rastreamento do movimento da cabeça do ouvinte são citados, assim como experimentos com um motor de passos controlando uma *dummy head* de acordo com o movimento da cabeça de um ouvinte. Eles afirmam que:

> considerando o movimento da cabeça como um fator importante na localização, não apenas inversões frontais-traseiras são resolvidas, mas a performance da localização quase se equipara à audição natural. Até mesmo quando o típico conteúdo espectral do ouvido [HRTFs] está ausente, a localização no plano horizontal permanece boa enquanto o sistema de rastreamento da cabeça estiver habilitado, embora sensações de elevação sejam relatadas. Logo, um sistema de auralização tem que permitir que o movimento da cabeça possibilite uma localização sonora equiparada à audição normal.
>
> (Felderhoff et al., 1999: 2)

Eles apresentam um problema desse tipo de reprodução que deve ser considerado em qualquer processo de escuta e que está no fato de que se "utilizando somente fones de ouvido para a reprodução, freqüências graves não são transmitidas". Os autores mencionam também outros estudos que afirmam que a informação referente ao posicionamento de fontes sonoras na vertical pode ser melhor representada com o rastreamento do movimento da cabeça. Entre os principais objetivos de seu sistema eles apresentam:

> É sabido que a acústica de uma van é menos conveniente do que a de um estúdio de monitoração. Apesar disso, se a sala de controle do engenheiro de som for analisada de forma binaural e carregada no processador BRS, ele terá uma experiência através de fones de ouvido, comparável com a audição normal em um ambiente de estúdio familiar. Devido ao fato de que várias salas podem ser armazenadas no processador BRS, tanto uma sala de controle multicanal quanto um teatro de reprodução de filmes multicanais ou uma sala comum, podem ser escolhidas.
>
> (Felderhoff et al., 1999: 3)

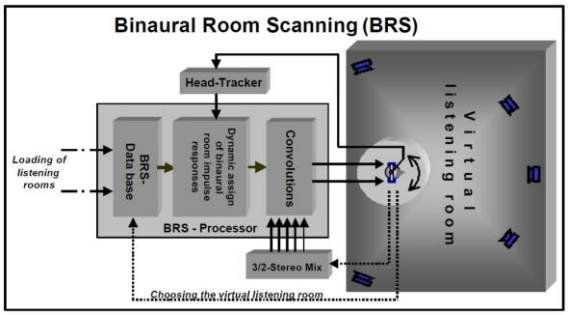


Figura 1. Esquema do sistema BRS - Binaural Room Scanning. Extraído de [10].

Nos trabalhos [12] [13] e [14], um outro sistema para reprodução virtual de arranjos de alto-falantes Ambisonics é descrito, assim como a implementação de um sistema de rastreamento da cabeça do ouvinte através de um giroscópio usando a plataforma PureData. De acordo com eles,

> para superar o problema da interpolação variável no tempo, de alta qualidade, entre as diferentes HRTFs na reprodução binaural em tempo real do sistema, uma abordagem do sistema Ambisonics virtual foi utilizada. Essa abordagem é baseada na idéia de se decodificar o sinal Ambisonics em alto-falantes virtuais. Depois disso, sinais binaurais são criados através da convolução dos sinais de cada um dos alto-falantes com a HRTF correspondente àquela posição no espaço. Agora, os sinais filtrados são superpostos para criar um sinal esquerdo e um sinal direito para os fones de ouvido.
>
> (Höldrich et al. 2003a: 2-3)

Na implementação dos autores de um sistema de rastreamento da cabeça, a abordagem de fazê-lo no domínio do Ambisonics foi adotada, onde matrizes simples de rotação ao redor do eixo Z mudam a orientação no plano horizontal. O aparelho para o rastreamento é montado no próprio fone de ouvido e, para a implementação da simulação da sala, atrasos simples são calculados para as primeiras reflexões de uma sala retangular. Os autores argumentam que a vantagem de se adotar o uso de Ambisonics é que o número de fontes sonoras a serem espacializadas não aumenta o processamento das HRTFs.

Eles também pontuam que, devido ao fato da interpolação de HRTFs variáveis no tempo para a reprodução de movimento das fontes sonoras assim como a implementação de sistemas de rastreamento da cabeça serem processamentos muito pesados para serem feitos em tempo real, a abordagem da utilização de Ambisonics é bastante plausível, resultando no uso de HRTFs fixas e conseqüente diminuição da necessidade de grande poder de processamento.

Entre as aplicações de seu sistema, os autores de [15] descrevem uma recriação em ambiente virtual do Pavilhão Philips de Varèse adicionado à visualização em tempo real da construção em questão. Nessa descrição, o ouvinte ou compositor pode caminhar pelo prédio assim como compor novas obras para serem executadas nesse ambiente. Em outro artigo [16], dos mesmos autores, a adição de um controle baseado em rastreamento visual de objetos, representando tanto o ouvinte quanto as fontes sonoras dispostas em uma mesa, é descrito.

Como pode ser observado das descrições acima, a forma como implementar o rastreamento da cabeça do ouvinte, se no domínio Ambisonics ou no binaural, tem sido questão de pesquisa de muitos autores. Sistemas de rastreamento da

cabeça sobre sinais binaurais podem ser implementados interpolando as HRTFs correspondentes às posições dos alto-falantes em tempo real a realizarem a convolução. No domínio Ambisonics isso pode ser feito se utilizando HRTFs fixas correspondentes a arranjos de alto-falantes pré-definidos e manipulando as matrizes de rotação, tombamento e inclinação do sinal no formato B como um todo. Parece que a abordagem de se manipular o formato B Ambisonics para simular o movimento da cabeça do ouvinte tem sido largamente aceito, principalmente devido ao fato de que dessa forma se reduz significamente o poder de processamento necessário.

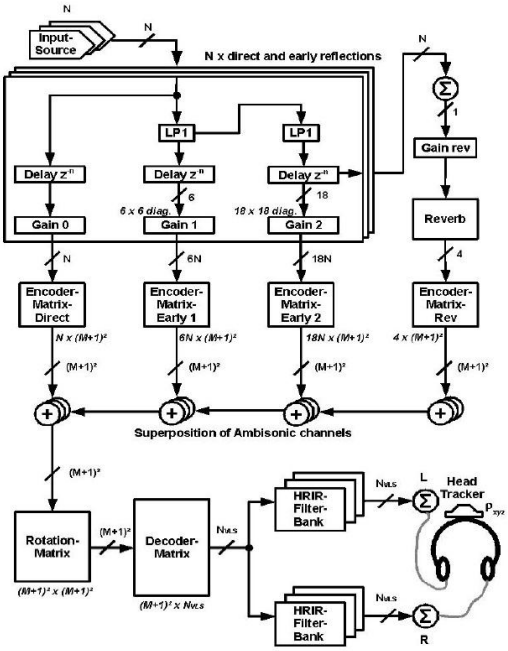


Figura 2. Diagrama de blocos do sistema binaural baseado em Ambisonics desenvolvido por Höldrich et al. Extraído de [12].

A primeira abordagem é descrita também em [17] enquanto se refere ao sistema Convolvotron que realiza o rastreamento no domínio binaural. A segunda abordagem é descrita em várias publicações, entre elas [18] e [19], onde o autor argumenta novamente que quando se reproduz um campo sonoro 3D completo através de HRTFs e fones de ouvido, muito poder de processamento é exigido e, se utilizando sinais Ambisonics no formato B, esse processamento pode ser reduzido e operações tais como as de rotação, tombamento e inclinação podem ser facilmente realizados.

> Ao se posicionar todas as fontes sonoras no campo sonoro em formato B incluindo, se necessário, campos sonoros de complexidade natural gravados com microfones *Soundfield*, o processamento que envolve a rotação, tombamento, inclinação, etc. do campo sonoro como um todo, como se mostra necessário em configurações de rastreamento da cabeça, é simplificado significamente se comparado ao que envolve processar HRTFs diretamente. O sinal em formato B pode posteriormente ser decodificado para sinais correspondentes a qualquer arranjo de alto-falantes e somente estes precisam ser passados pelos filtros HRTFs. Uma vez limitado a um conjunto fixo de HRTFs, é possível então se realizar qualquer operação em equipamentos comuns, até mesmo quando sistemas de rastreamento da cabeça estão atuando.
>
> (Malham, 1999b: 26)

Os autores de [5], ao observarem a necessidade de processamento digital eficiente para aplicações em tempo real de reproduções binaurais, também argumentam que o mesmo poderia ser obtido ao se realizar o rastreamento da cabeça em domínio Ambisonics uma vez que este consumiria menos poder de processamento. O autor de [20] também discute essa abordagem Ambisonics para o rastreamento e aponta o mesmo como uma escolha natural:

> A representação em Ambisonics possui uma característica bônus, que consiste em sua relativa facilidade de rotação do campo sonoro. Isso abre a possibilidade de se fazer filtros estáticos de HRTF e acomodação aos movimentos da cabeça do ouvinte na rotação dinâmica do conjunto. A comutação e interpolação dos filtros seriam então desnecessárias, possibilitando uma implementação mais simples e mais eficiente.
>
> (Travis, C. 1996a: 115)

Em [21], os autores sugerem que, para as várias aplicações do som tridimensional, a adoção de uma forma de renderização do áudio que seja completamente independente da configuração do sistema de reprodução se mostra necessária. Essa idéia se encaixa com a implementação de um "estúdio virtual" utilizando-se sinais no formato B Ambisonics. Os mencionados autores também citam [18] e [22], ao apresentar que a solução para diminuir o consumo de processamento por sinais binaurais poderia ser a utilização de pré-processamento no domínio Ambisonics e realizar a convolução dos sinais provenientes da decodificação para alto-falantes virtuais, teria-se então o processamento binaural estático e os processos dinâmicos realizados no domínio Ambisonics, que exige menos processamento.

O autor de [23] é outro que menciona a possibilidade de renderização de sinais Ambisonics no formato B para fones de ouvido através do "processamento de alto-falantes virtuais, que consiste na simulação binaural de cada um dos alto-falantes para uma posição central do ouvinte, por exemplo, através da filtragem desses sinais com a HRTF correspondente". O trabalho [24] também descreve a virtualização do áudio multicanal baseado em processamento por HRTFs. De acordo com eles a "reprodução virtual do áudio 3D, seja de gravações de dois ou mais canais, tradicionalmente objetiva a reprodução através de fones de ouvido da sensação de escuta da reprodução por alto-falantes" e consiste em aplicar filtros de HRTFs ao sinal dos alto-falantes e somá-los posteriormente. Os autores de [25] completam afirmando que

> atualmente, a maioria do conteúdo de áudio é produzida para reprodução através de alto-falantes. Para se renderizar tal material para reprodução em fones de ouvido, cada alto-falantes pode ser representado por uma fonte sonora virtual posicionada em um local determinado em relação ao ouvinte. Isso é obtido normalmente através de processamento binaural.
>
> (George et al., 2009: 2-3)

Um outro autor que discute essa abordagem é [26], que nos apresenta interessantes considerações a respeito da reprodução de alto-falantes virtuais através de fones de ouvido. Ela aponta a diferença entre a implementação da síntese binaural de dois canais e de múltiplos canais. No que ela denomina implementação de dois canais, existem tantos filtros de HRTF e convoluções quanto fontes sonoras; a implementação de múltiplos canais por outro lado se relaciona ao

> conceito de alto-falante virtual que consiste na simulação de uma configuração de alto-falantes para

> fones de ouvido. Cada alto-falante virtual é sintetizado através de síntese binaural com apropriados pares de filtros. Em outras palavras, essa é uma solução que permite se adaptar material multicanal (5.1, 6.1, 7.1, 10.2, 22.2,...) para reprodução através de fones de ouvido, o que permite qualquer um a ouvir áudio espacializado sem a necessidade de qualquer configuração de alto-falantes, seja ela simples ou complexa. Esse é motivo pelo qual isso é referido como *downmix binaural*. A espacialização da fonte sonora é controlada pelo formato multicanal primário (por exemplo, panorâmico de intensidade), que define as funções espaciais, enquanto os filtros são determinados pelas HRTFs correspondentes à posição de cada alto-falante.
>
> (Nicol, 2010: 41)

De acordo com a autora ainda:

> para sistemas caseiros, tais como home theaters ou até mesmo reprodução de conteúdo musical, a tecnologia binaural oferece uma melhora no que diz respeito à reprodução espacial através da abordagem de alto-falantes virtuais (ou mixdown binaural) que permite a qualquer um ouvir material originalmente mixado para sistemas multicanais (5.1, 6.1, 7.1, 10.2, ... ou Ambisonics) se utilizando fones de ouvido.
>
> (Nicol, 2010: 63)

Ela também aponta alguns desenvolvimentos distribuidos comercialmente que possuem semelhante abordagem aplicada a sistemas 5.1, tais como o Dolby®Headphone ou Dolby®Virtual Speaker e o Beyerdynamics®Headzone. Ela conclui que "uma grande vantagem da tecnologia binaural se deve a ser compacta, o que a faz apropriada para qualquer aparelho portátil".

Em [27] são citados outros trabalhos que têm demonstrado o surgimento de artefatos auditivos causados pela descontinuidade das ondas sonoras no sinal binaural obtido após a convolução quando da mudança das HRTFs que definem diretividade com o movimento da cabeça. Outro motivo para a criação de alto-falantes virtuais a partir das HRTFs, solução flexível onde o ouvinte estaria posicionado no centro imaginário do arranjo de alto-falantes.

A partir da análise de todos esses estudos descritos acima, a implementação do presente "estúdio virtual" teve como parte fundamental a busca de ferramentas que permitissem fácil acesso à maioria dos desenvolvimentos descritos, sendo o sistema de rastreamento de movimento da cabeça e um decodificador versátil, as principais preocupações do mesmo.

3. O DESENVOLVIMENTO PRESENTE

O presente desenvolvimento de um "estudio virtual" onde, através de fones de ouvido, se possa monitorar reproduções de mixagens multicanais, é baseado nos estudos descritos anteriormente e envolve o uso de cinco ferramentas principais, escolhidas devido a sua facilidade de acesso por compositores e engenheiros de som:

>Codificador Ambisonic (VST ou objeto MAX)

Pretende processar um sinal monofônico para o formato B Ambisonics. Como VST, pode ser insertado em qualquer track de uma DAW (*digital audio workstation*) e usado para enviar ao patch no MAX os quatro sinais correspondentes ao formato B de primeira ordem. Opcionalmente o usuário pode ter um reprodutor monofônico dentro do próprio MAX, que envie um sinal para um codificador (também dentro do MAX) antes do estágio de processamento. Existem vários codificadores livres, tanto no formato VST [28] [29] quanto em formato de objetos do MAX [30] [31]. Na presente implementação, plugins VST foram escolhidos devido à familiaridade que os engenheiros de som têm na manipulação de uma DAW e à simplicidade do processo de automação dos seus parâmetros, como se faz com qualquer outro plugin.

>Processador Ambisonics

O papel desse processador é executar as mudanças na orientação do ambiente tridimensional virtual, nos três eixos – X, Y e Z – de acordo com as saídas do sistema de rastreamento de movimentos da cabeça. O processador utilizado nessa implementação é o objeto ambiproc~, desenvolvido para MAX [32], adaptado por Matthew Paradis do plugin VST desenvolvido por Dave Malham. Essa escolha foi feita a partir da necessidade de integração com o objeto (também para MAX), que realiza a interpretação do rastreador de movimentos da cabeça do ouvinte.

>Decodificador Ambisonics

O decodificador gera o sinal de áudio dos alto-falantes para um determinado arranjo dos mesmos, que podem ser direcionados tanto para alto-falantes reais quanto para o estágio de processamento binaural. Para essa implementação foi escolhido o decodificador do conjunto de ferramentas para Ambisonics da ICST [30], uma vez que ele permite fácil mudança entre os diversos conjuntos de alto-falantes através do envio de mensagens simples escolhidas através de um menu pop-up.

>Processador Binaural

A função do processador binaural seria a de convoluir o sinal destinado a cada um dos alto-falantes com uma HRTFs correspondente à sua posição no espaço ao redor do ouvinte, posteriormente, após a soma de todos os sinais convoluídos, se obtém o sinal para ser endereçado ao fone de ouvido. Tal processo pode ser obtido com diversos plugins VST, tais como o SIR [33], que é livre, ou com o objeto spat~ para MAX [34], que, entre diversas outras funções, realiza essa convolução estática e ainda oferece a possibilidade de se realizar interpolações entre as HRTFs, realizando o processamento referente ao movimento da cabeça no domínio binaural. Algumas diferenças foram observadas nesse processo nos diferentes domínios, no entanto não foram profundamente investigadas devido ao fato de o objeto spat~ não ser de livre acesso, o que limitaria o uso do patch. Por outro lado, a escolha pelo uso do VST SIR limitou o patch à reprodução de configurações de até oito alto-falantes, uma vez que incluir mais unidades do referido plugin aumentaria consideravelmente a necessidade de maior poder de processamento do computador.

>Sistema de rastreamento do movimento da cabeça

A primeira tentativa de implementação de um sistema de rastreamento do movimento da cabeça foi baseado em marcas fiduciais [35], no entanto, devido à sua lenta resposta aos movimentos e baixa resolução, outras opções foram examinadas. O objeto headtracker [36], para MAX, infelizmente disponível somente para usuários windows, foi escolhido devido à sua facilidade de uso e não necessidade de hardware adicional. Ele permite que o conjunto de alto-falantes responda aos movimentos da cabeça do engenheiro de som ou compositor através do processamento do sinal de uma webcam. As saídas usadas na atual implementação do patch são os três sinais correspondentes à rotação nos três eixos (X, Y e Z). Estes são ajustados em sensibilidade e logo em seguida enviados às três entradas do processador ambiproc~ para manipulação, nesse caso executado no domínio

Ambisonics. Um chaveamento permite que o sistema de rastreamento do movimento seja ligado ou desligado para comparação pelo usuário.

Outra implementação feita no patch foi a compensação pelo uso do fone de ouvido (AKG 271 mkII), numa tentativa de se excluir a sua coloração tonal, realizando uma convolução do sinal de saída do sistema com uma resposta a impulso inversa do fone de ouvido em questão, obtida de um banco de dados de IRs [37]. Uma chave de liga e desliga dessa compensação assim como medidores de nível tanto dos sinais do formato B quanto dos sinais binaurais foram adicionados à interface do usuário. Um monitor de uso de CPU, extraído do pacote de ferramentas spat~, foi incluído assim como um filtro high shelf que pode ser configurado para um ganho na região aguda, na tentativa de se compensar a falta de resolução espacial nessa região do próprio sistema Ambisonics e melhorar a localização das fontes sonoras.

Apesar de alguns trabalhos descreverem a necessidade de simulação de reverberação para melhor reprodução binaural de ambientes virtuais, isto não foi necessário na presente implementação. Essa escolha se deve ao fato de que um dos principais objetivos aqui é a de se reproduzir gravações onde a reverberação original do local gravado já está presente ou é obtida a partir de convolução com sinais IR em formato B pré-gravados. Adicionar simulação de salas consistiria também em adicionar certa coloração à reprodução pelos fones de ouvido. Em outras palavras, esta implementação objetiva simular um ambiente ideal para mixagem, ausente de interferências acústicas senão aquelas próprias da gravação.

Na interface do usuário, o mesmo tem também, acesso a seis diferentes configurações de alto-falantes (stereo, quadrafonic, 5.1, hexagonal, 8horizontal, 8cubic) e a um conjunto de 51 HRTFs (compensadas e não-compensadas – *raw*) onde deve escolher uma que lhe sirva melhor ou simplesmente se acostumar a trabalhar com uma delas.

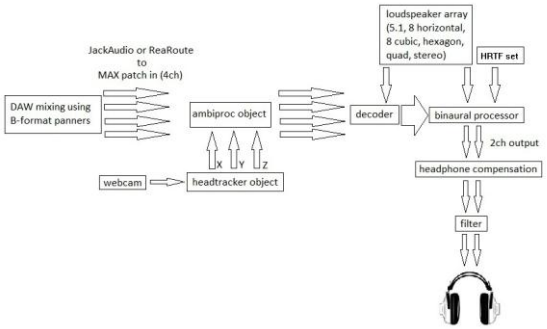


Figura 3. Diagrama de blocos do sistema desenvolvido.

O sinal em formato B Ambisonics pode ser enviado da DAW utilizada para o patch no MAX através da utilização do software JackAudio [38], driver virtual de endereçamento de sinais de áudio dentro do próprio computador, ou através do ReaRoute, driver virtual presente na DAW Reaper. Tal procedimento já foi adotado por diversos engenheiros e pesquisadores da área, por exemplo [39].

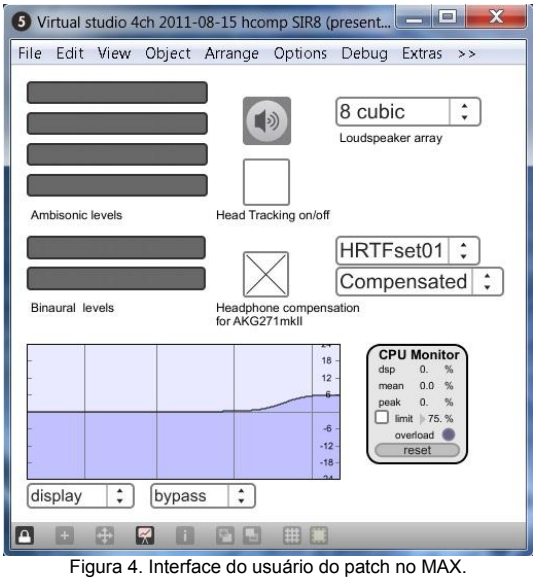


Figura 4. Interface do usuário do patch no MAX.

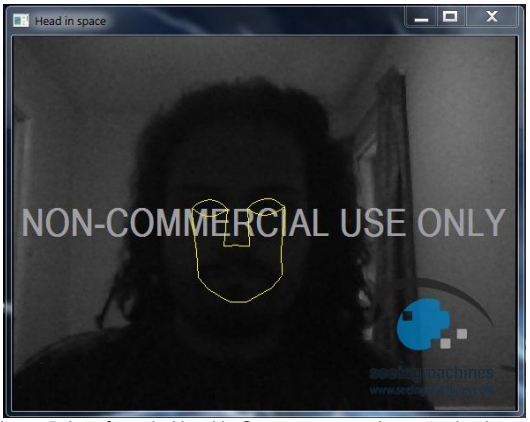


Figura 5. Interface do Head in Space para monitoração do sistema de rastreamento dos movimentos da cabeça.

4. AVALIAÇÃO DA FERRAMENTA DESENVOLVIDA

Durante o período de um ano foram realizadas diversas gravações no departamento de música da Universidade de York com o objetivo de se testar o presente trabalho. Tais gravações foram realizadas utilizando tanto microfones próximos das fontes sonoras, para serem posicionados no ambiente tridimensional virtual através de *panners* em Ambisonics, quanto com um microfone *Soundfield* de onde se pretendia extrair material referente à espacialização natural da execução, de forma semelhante com o que se faz em uma gravação estéreo ou 5.1. Composições também foram desenvolvidas pelo presente autor, onde a espacialização para um arranjo multicanal de alto-falantes foi um objetivo em comum. Tanto as gravações quanto as composições foram monitoradas até o estágio final da produção através do patch descrito. Na etapa final, quando da reprodução através de um arranjo real de alto-falantes somente pequenos ajustes foram necessários para se obter relativa satisfação, sendo a diferença entre um ambiente relativamente isento de coloração ou imperfeições (ambiente virtual através dos fones de ouvido) e um ambiente real (sala de controle Trevor Jones ou auditório Rymer) facilmente percebida, mas não incômoda ou impossibilitadora da apreciação das composições ou gravações.

Também para se avaliar o patch desenvolvido, vários estudantes do curso de música e tecnologia da Universidade de York foram convidados para uma sessão onde poderiam testar seu próprio material musical, no qual vinham trabalhando, através do sistema binaural, e comparar com a reprodução original através de alto-falantes reais no estúdio Trevor Jones no Centro de Pesquisa em Música da Universidade de York. Um questionário foi entregue onde eles poderiam expressar suas considerações com relação a vantagens e fidelidade do sistema.

Devido ao fato de um número de estudantes insuficiente para uma análise quantitativa terem se apresentado à sessão, considerações foram expressas verbalmente e de forma livre para o grupo presente. Foi concordado o fato de que o "estúdio virtual" se mostrou útil e de fácil utilização, assim como a possibilidade de utilizá-lo como ferramenta de suporte durante o processo de produção musical. A naturalidade com o sistema reproduz as diversas configurações de alto-falantes ainda precisa de mais dados estatísticos para análise, e foi concordado também que a mixagem e produção em tal sistema não deve ser tomada como definitiva e que todo material deveria ser ouvido em um arranjo de alto-falantes reais antes de finalizado, considerando principalmente a própria natureza da reprodução através de fones de ouvido e da função do corpo como um todo no processo de escuta.

Sistemas de rastreamento do movimento da cabeça definitivamente melhoram a reprodução de fontes sonoras virtuais, mesmo que as HRTFs não sejam individualizadas, no entanto a compensação da influência do fone de ouvido e o ganho na região aguda do sinal de áudio ainda precisam de maiores pesquisas e análises para serem consideradas importantes implementações. Um dos ouvintes, por exemplo, preferiu definitivamente a reprodução binaural sem a compensação para a influência do fone de ouvido, o que nos leva a refletir sobre as reais vantagens que isso trás ao sistema. De uma forma geral os ouvintes ficaram satisfeitos com os resultados e o patch se mostrou uma ferramenta muito útil para futuros trabalhos com áudio multicanal.

5. CONCLUSÕES E TRABALHOS FUTUROS

Uma ferramenta para se trabalhar em um ambiente Ambisonics virtual foi desenvolvida e colocada em prática através de gravações, composições e testes de escuta. A influência de diversos fatores tais como o material musical, a experiência e fisiologia individual, o conjunto de alto-falantes a ser reproduzido, entre outros, foram consideradas e devem ser analisadas em qualquer contexto de produção ou composição musical, assim como o processo de monitoração através de fones de ouvido. A hipótese inicial de que qualquer conjunto de alto-falantes pode ser reproduzido fielmente através de fones de ouvido não pôde ser confirmada, uma vez que reprodução binaural dos alto-falantes virtuais se mostrou muito sensível a diversos elementos que constituem o sistema em si. Apesar disso, monitorar arranjos de alto-falantes multicanais através de fones de ouvido se mostrou muito útil mesmo que não definitivo.

Apesar da falta de resolução espacial da gravação e do processamento em formato B Ambisonics de primeira ordem, este se mostrou uma poderosa ferramenta para produções musicais desde que o compositor ou engenheiro de som esteja ciente de suas limitações e usabilidade no contexto de objetivos finais de seu trabalho. Usar sinais Ambisonics no formato B como formato intermediário para outros formatos, tais como o binaural se mostrou uma escolha muito interessante, mas com suas próprias limitações, a serem adaptadas ou consideradas em cada caso particular.

Novas ferramentas para se trabalhar com sistemas Ambisonics, tais como delays e reverbers baseados em resposta a impulsos já estão em desenvolvimento e apresentam resultados interessantes e promissores. O próximo passo seria o encorajamento da criação de mais material musical para tal sistema, trazendo essas ferramentas ao conhecimento dos compositores e engenheiros de som.

Mais pesquisas precisam ser feitas no que diz respeito aos padrões direcionais tanto dos alto-falantes quanto das fontes sonoras na sua reprodução através de arranjos de alto-falantes reais ou em ambientes virtuais objetivando a melhoria dos mesmos. Sistemas Ambisonics de ordens elevadas (HOA) ainda precisam fazer disponíveis soluções práticas de microfones assim como melhorar a resolução espacial sem o aumento do número de canais necessários para sua transmissão ou armazenamento. Técnicas híbridas que podem manter o uso de apenas quatro canais Ambisonics como saída e melhorar a resolução espacial seriam muito bem vindas e mais pesquisas precisam ser feitas também no que diz respeito a psicoacústica e outros fenômenos que influenciam a percepção auditiva, tais como a relação dos sons com imagens, a influência da memória na escuta, a medição de HRTFs individuais, o design de fones de ouvido, entre outros.

6. REFERÊNCIAS BIBLIOGRÁFICAS

[1] *LISTEN HRTF Database.* Acessado em http://recherche.ircam.fr/equipes/salles/listen/ em 23 de junho de 2011.

[2] Malham, D. G. *Toward reality equivalence in spatial sound diffusion*. Computer Music Journal 25/4, 31-38.

[3] Bresson, J.; Schumacher, M. *Compositional control of periphonic sound spatialization.* Procedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics. Paris (France), 6-7 maio 2010.

[4] Moller, H. *Fundamentals of binaural technology.* Applied Acoustics v.36, p.171-218. 1992.

[5] Jot, J.-M.; Larcher, V.; Warusfel, O. *Digital signal processing issues in the context of binaural and transaural stereophony.* Convention paper presented at the 98th AES convention. Paris (France), 25-28 Mar 1995.

[6] Hammershoi, D. *Fundamental aspects of the binaural recording and synthesis techniques.* Convention paper presented at the 100th AES convention. Copenhagen (Denmark), 11-14 May 1996.

[7] McGrath, D.; McKeag, A. *Sound field format to binaural decoder with head tracking.* Convention paper presented at the 6th Australian regional AES convention. Melbourne (Australia), 10-12 Sep 1996.

[8] *CATT Acoustics.* Acessado em http://www.catt.se/ em 4 de julho de 2011.

[9] Jot, J.-M.; Larcher, V.; Wardle, S. *Approaches to binaural synthesis*. Convention paper presented at the 105th AES convention. San Francisco (USA), 26-29 Sep 1998.

[10] Jot, J.-M.; Larcher, V.; Pernaux, J.-M. *A comparative study of 3-D audio encoding and rendering*

techniques. Convention paper presented at the 16th AES international conference. Rovaniemi (Finland), 10-12 Apr 1999.

[11] Felderhoff, U.; Horbach, U.; Mackensen, P.; Pellegrini, R.; Theile, G. *Binaural room scanning –a new tool for acoustic and psychoacoustic research.* 1999. Acessado em http://www.hauptmikrofon.de/theile/BRS_DAGA_1999_Paper.PDF em 9 de janeiro de 2011.

[12] Höldrich, R.; Musil, T.; Noisternig, M.; Sontacchi, A. *A 3D ambisonic based binaural sound reproduction system.* Convention paper presented at the 24th AES international conference. Banff (Canada), 26-28 Jun 2003a.

[13] Höldrich, R.; Musil, T.; Noisternig, M.; Sontacchi, A. *A 3D real time rendering engine for binaural sound reproduction*. Proceedings of the 2003 International Conference on Auditory Display. Boston (USA), 6-9 Jul 2003b.

[14] Höldrich, R.; Musil, T.; Noisternig, M.; Sontacchi, A. *3D binaural sound reproduction using a virtual ambisonic approach.* Convention paper presented at the International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems. Lugano (Switzerland), 27-29 Jul 2003c.

[15] Höldrich, R.; Lorenz, R.; Musil, T.; Noisternig, M.; Sontacchi, A. *Hearing Varèse's Poème Électronique inside a virtual Philips pavilion.* Proceedings of ICAD 05-Eleventh meeting of the international conference in auditory display, Limerick (Ireland), 6-9 Jul 2005a.

[16] Höldrich, R.; Musil, T.; Zmölnig, J. M.; Zouhar, V. *Virtual audio reproduction engine for spatial environments*. International computer music conference proceedings, 2005b.

[17] Wenzel, E. M. *What perception implies about implementation of interactive virtual acoustic environments.* Convention paper presented at the 101st AES convention, Los Angeles (USA), 8-11 Nov 1996.

[18] Malham, D. G. *3D sound for virtual reality systems using Ambisonic techniques.* Apresentação na 4a Conferencia de realidade virtual, Londres (Inglaterra), Abril de 1993. Acessado em http://www.york.ac.uk/inst/mustech/3d_audio/vr93papr.htm em 9 de marco de 2011.

[19] Malham, D. G. *Homogeneous and non-homogeneous surround sound systems.* Proceedings of the 14th AES UK conference: Audio –the second century, London (UK), 7-8 Jun 1999.

[20] Travis, C. *A virtual reality perspective on headphone audio*. Convention paper presented at the 11th AES UK conference, London (UK), 25-26 Mar 1996a.

[21] Jot, J.-M.; Philip, A.; Walsh, M. *Binaural simulation of complex acoustic scenes for interactive audio*. Convention paper presented at the 121st AES convention, San Francisco (USA), 5-8 Oct 2006.

[22] Travis, C. *Virtual reality perspective on headphone audio*. Convention paper presented at the 101st AES convention, Los Angeles (USA), 8-11 Nov 1996b.

[23] Daniel, J. *Spatial sound encoding including near field effect: introducing distance coding filters and a viable, new Ambisonic format*. Convention paper presented at the 23rd AES international conference, Copenhagen (Denmark), 23-25 May 2003.

[24] Goodwin, M. M.; Jot, J.-M. *Binaural 3-D audio rendering based on spatial audio scene coding.* Convention paper presented at the 123rd AES convention, New York (USA), 5-8 Oct 2007.

[25] George, S.; Neugebauer, B.; Plogsties, J.; Silzle, A. *Binaural processing algorithms: importance of clustering analysis for preference tests*. Convention paper presented at the 126th AES convention, Munich (Germany), 7-10 May 2009.

[26] Nicol, Rozenn. *Binaural Technology*. AES Monograph, 2010.

[27] Boland, F.; Gorzel, M.; Kearney, G.; Rice, H. *Depth perception in interactive virtual acoustic environments using higher order ambisonic soundfields*. Procedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics, Pairs (France), 6-7 May 2010.

[28] *Swiss center for computer music VST plugins*. Acessado em http://www.dmalham.freeserve.co.uk/vst_ambisonics.html em 30 de janeiro de 2011.

[29] *Wigware VST plugins*. Acessado em http://www.brucewiggins.co.uk/?page_id=78 em 30 de janeiro de 2011.

[30] *ICST Ambisonics external for MAX website.* Acessado em http://www.icst.net/research/downloads/ambisonics-externals-for-maxmsp/ em 27 de abril de 2011.

[31] *Ambisonic external for MAX/MSP*. Acessado em http://www.grahamwakefield.net/soft/ambi~/index.htm em 20 de junho de 2011.

[32] *Ambisonic objects*. Acessado em http://www.york.ac.uk/music/mrc/software/objects/ em 20 de junho de 2011.

[33] *SIR VST plugin*. Acessado em http://www.knufinke.de/sir/sir1.html em 5 de julho de 2011.

[34] Jot, J.-M.; Warusfel, O. *Spat~: a spatial processor for musicians and sound engineers*. CIARM: International Conference on Acoustics and Music Research, Ferrara (Italy), 1995. Acessado em http://articles.ircam.fr/textes/Jot95a/ em 3 de maio de 2011.

[35] *Fiducial symbols.* Acessado em http://reactivision.sourceforge.net/data/fiducials.pdf em 26 de junho de 2011.

[36] *Head in space*. Acessado em http://www.lim.dico.unimi.it/HiS em 24 de junho de 2011.

[37] *Impulse response measurements.* Acessado em https://dev.qu.tu-berlin.de/projects/measurements/files em 23 de junho de 2011.

[38] Jack audio router. Acessado em http://jackaudio.org em 27 de fevereiro de 2011.

[39] Nettingsmeier, J. *General purpose ambisonic playback systems for electroacoustic concerts –a practical approach*. Procedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics. Paris (France), 6-7 May 2010.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Discussão teórica sobre a inclusão do escoamento e perfis de velocidade na modelagem física de instrumentos de sopro

Luís Carlos Oliveira,[1] Ricardo Goldemberg,[2] and Rafael Santos Mendes[1]

[1] Universidade de Campinas (UNICAMP), Faculdade de Engenharia Elétrica e de Computação (FEEC)
Departamento de Controle e Automação (DCA),
Campinas, SP, CEP:13083-970, Brasil

[2] Universidade de Campinas (UNICAMP), Instituto de Artes (IA), Departamento de Música (DM),
Campinas, SP, CEP:13083-970, Brasil

lucaoliv@fee.unicamp.br, rgoldem@iar.unicamp.br, rafael@fee.unicamp.br

ABSTRACT

A síntese de instrumentos de sopro por modelagem física não inclui o escoamento de ar no interior do tubo. Grosso modo, o algoritmo de síntese usando o guia de ondas digital (*Digital Waveguide - DWG*) corresponde a modelos de reverberação sem a inclusão do escoamento. Este trabalho inicia a discussão sobre a inclusão do escoamento em instrumentos de sopro da família das madeiras, em particular, para a clarineta. A possibilidade de incluir o escoamento em diferentes perfis de velocidade é analisado. Para esta finalidade, sugere-se que o tubo é dividido em camadas de mesma espessura ou em camadas que produzem a mesma área. Para ilustração, três perfis de velocidade são analisados – parabólico, constante e triangular. A inclusão do escoamento no modelo proporciona uma curva de pressão no interior do instrumento com melhor aproximação das condições reais.

0 INTRODUÇÃO

A síntese por modelagem física de instrumentos de sopro é ferramenta empregada desde a década de 80. O emprego desta técnica possibilita que os modelos de síntese utilizem parâmetros que são facilmente reconhecidos pelos usuários sendo estes ou não instrumentistas com prática. De modo geral, são utilizados algoritmos de reverberação formados por filtros digitais através do guia de ondas digital ("*Digital Waveguide*").

Nesta abordagem constata-se que o escoamento do ar no interior do instumento não está incluso, apesar de utilizar denominação como "*volume flow*" que se refere efetivamente ao movimento das moléculas de ar durante a produção sonora.

Para a inclusão do escoamento faz-se necessário modificar a velocidade de propagação sonora no interior do instrumento. Nas deduções para o equacionamento por modelagem física de instrumentos de sopro encontradas em diferentes autores [1, 2, 3, 4] parte-se do princípio que o meio de propagação sonora está em repouso. No entanto esta não corresponde à realidade dos instrumentos de sopro no qual o meio onde a onda sonora se propaga está em contínuo movimento. O tratamento teórico foi extraído do texto de Landau e Lifshitz [5] que consideram a possibilidade de propagação sonora com o meio ambiente em movimento.

Com esta nova abordagem torna-se possível incluir um perfil de velocidades no interior do tubo e verificar sua influência na sonoridade produzida pelo modelo de síntese. Possibilita ainda uma proposta de discussão sobre o desenvolvimento da sonoridade do instrumentista associada à evolução do escoamento no interior do instrumento.

A próxima seção apresenta alguns resultados já estabelecidos para a modelagem física. Em seguida a propagação sonora em um meio em movimento é discutida para destacar os resultados que serão utilizados em futuras simulações. Finalmente, exemplos envolvendo perfis de velocidade parabólica, constante e triangular juntamente com os resultados das curvas de pressão no interior do tubo são discutidos na última seção.

1 SÍNTESE POR MODELAGEM FÍSICA

O aprimoramento dos recursos computacionais teve influência sobre a velocidade de processamento dos modelos dos instrumentos musicais a ponto de possibilitar a execução em tempo real.

O ponto de partida foi lançado pelo trabalho de McIntyre *et al* [6] que ao resgatar os conceitos estabelecidos por d'Alembert, simula o comportamento dissociado da onda no instrumento. Desta forma, o fenômeno acústico foi descrito analiticamente por ondas que caminham em sentidos opostos.

Em seguida, destaca-se o trabalho de Julius O. Smith [3] que concatena os resultados de McIntyre *et al.* [6] para a simulação de instrumentos reais ao associar uma estrutura de filtros em cascata através do **guia de ondas** ("**waveguides**") digital. A equação da onda é resolvida em determinada geometria e condições de contorno pelo guia de ondas proporcionando rápido cálculo computacional do termo linear do instrumento. Estes resultados renderam a Smith a patente de um sintetizador.

Reunindo estes aspectos da modelagem física o instrumento musical é compreendido, genericamente, pelo conjunto de módulos descritos na Figura 1. As legendas sob os módulos indicam as partes constituintes da clarineta.

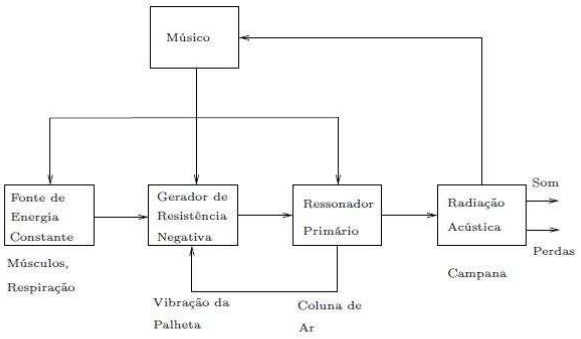


Figure 1: Diagrama Genérico da Clarineta.

Neste estudo a ação do músico foi desconsiderada. Os blocos indicados por "Fonte de Energia Constante" e "Gerador de Resistência Negativa" são agrupados e correspondem ao "*mecanismo de excitação*, termo *não linear* do modelo. Representam o conjunto formado pela boquilha e palheta sob a ação de uma pressão na entrada.

O tubo do instrumento, responsável pela radiação primária, corresponde ao modelo linear descrito pelo guia de ondas digital (DWG). O modelo pode incluir a variação da geometria do tubo ao longo do comprimento. Neste caso aproxima-se sua geometria por uma sucessão de cilindros [4]. Também pode indicar a presença de furos que também são responsáveis pela radiação sonora. Estes podem ser fechados por chave ou pelo dedo do instrumentista. A teoria para um único furo foi desenvolvida por Keefe [7] e a síntese por guia de ondas baseada nesta teoria por Smith [3].

A radiação sonora, quando todas as chaves estão fechadas, é totalmente proporcionada pela campana. Ao final da campana, parte da propagação sonora é transmitida e parte retorna ao tubo (reflexão), tanto para estabelecer a ressonância no interior do tubo como para a operação adequada do mecanismo de excitação [4]. O modelo corresponde a um filtro de primeira ou segunda ordem de acordo com Scavone [4].

A descrição da modelagem de suas partes estão descritas nas seções seguintes.

1.1 Modelo do Mecanismo de Excitação

A sonoridade resultante da síntese da clarineta terá resultados próximos ao instrumento real quanto melhor for a representação do mecanismo de excitação, correspondente ao conjunto boquilha/palheta. A Figura 2 representa a cavidade bucal (*cb*) em contato com a boquilha e palheta.

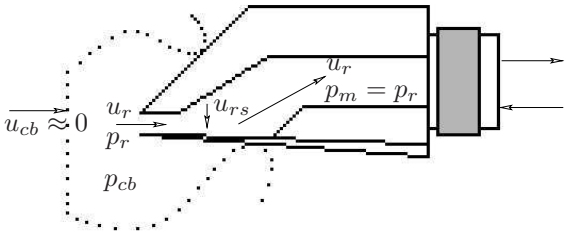


Figure 2: Cavidade Bucal, Boquilha e Palheta.

Usualmente, a cavidade bucal é considerada como sendo um tanque grande com pressão (p_{cb}) constante e velocidade (u_{cb}) nula. Para diferentes autores [3, 8, 4] a estrutura formada pela palheta simples e boquilha funciona como uma válvula controlada por pressão. Este conjunto possibilita a entrada de energia no instrumento para a inicialização e manutenção das oscilações no ressonador (tubo).

O movimento da palheta controla a vazão no canal formado pela palheta e boquilha. No regime permanente esta vazão é determinada pela equação de Bernoulli, [4, 2]:

$$u_r = S_0 \mathbf{sgn}\left(p_\Delta\right)\sqrt{\frac{2p_\Delta}{\rho}} \tag{1}$$

onde $S_0 = wy$ é a área disponível para escoamento do ar após a cavidade bucal, sendo w a largura do canal e y sua altura. A diferença de pressão entre a cavidade bucal (p_{cb}) e o tubo (p_b) é $p_\Delta = p_{cb} - p_b$. Finalmente, ρ é a densidade do ar e $\mathbf{sgn}$ é um operador que indica o sinal de (p_Δ).

Coeficiente de Reflexão Dependente da Pressão

No contexto de modelagem por guia de onda digital (DWG), Smith [3] propôs modelar a fronteira entre palheta e tubo com um coeficiente de reflexão sem memória ("*memoryless reflection coefficient*"). Este varia em resposta devido à diferença de pressão entre a cavidade bucal (p_{cb}) e o tubo (p_b). A vantagem deste tratamento consiste que o comportamento não-linear da palheta é obtido com um cálculo simples, podendo-se representar o comportamento da palheta por uma "look-up table", porém o comportamento dinâmico de "ordem superior" da palheta não é considerado.

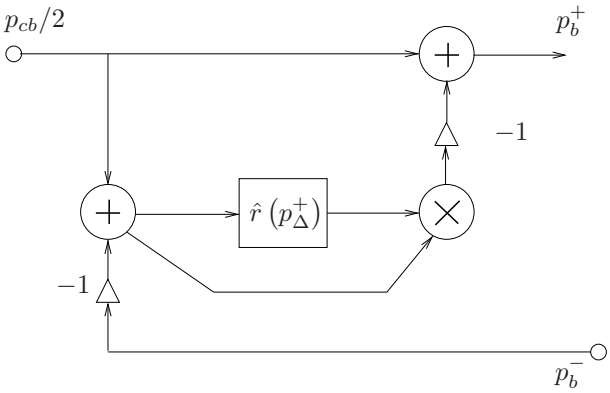


Figure 3: Implementação por Guia de Ondas do Coeficiente de Reflexão Polinomial como Função da Pressão.

O desenvolvimento das equações do modelo pode ser acompanhado em Scavone [4, págs.148-151]. Basicamente, o mecanismo de excitação é representado por uma *impedância acústica concentrada*, $Z_r(p_\Delta)$, sobre a qual p_Δ atua. Grosso modo, $Z_r(p_\Delta)$ corresponde à impedância da abertura da palheta, cujo tamanho depende da diferença de pressão p_Δ através da palheta.

A vazão volumétrica, $u_r(p_\Delta)$, através da abertura da palheta é designada por,

$$u_r\left(p_\Delta\right) = \frac{p_\Delta}{Z_r\left(p_\Delta\right)} \tag{2}$$

enquanto a vazão no tubo é obtida pela expressão,

$$u_b = \frac{p_b^+ - p_b^-}{Z_{0b}} \tag{3}$$

onde Z_{0b} representada a impedância característica da coluna de ar na interface entre a boquilha e tubo, p_b^+ é a componente da pressão da onda viajante que *sai* da boquilha para o tubo e p_b^- é a componente da pressão da onda viajante que *vem* do tubo para a boquilha. Ao estabelecer conservação de massa na junção boquilha/tubo chega-se a:

$$\begin{aligned} p_b^+ &= \left[\frac{Z_r\left(p_\Delta\right) - Z_{0b}}{Z_r\left(p_\Delta\right) + Z_{0b}}\right] p_b^- + \left[\frac{Z_{0b}}{Z_r\left(p_\Delta\right) + Z_{0b}}\right] p_{cb} \\ &= r\left(p_\Delta\right) p_b^- + \left[\frac{1 - r\left(p_\Delta\right)}{2}\right] p_{cb} \\ &= r\left(p_\Delta\right)\left[p_b^- - \frac{p_{cb}}{2}\right] + \frac{p_{cb}}{2} \end{aligned} \tag{4}$$

onde foi definido o *coeficiente de reflexão dependente de pressão* por,

$$r\left(p_\Delta\right) = \left[\frac{Z_r\left(p_\Delta\right) - Z_{0b}}{Z_r\left(p_\Delta\right) + Z_{0b}}\right] \tag{5}$$

O coeficiente de reflexão depende da dureza da palheta e do valor da abertura de equilíbrio entre a palheta e boquilha (y).

A solução da Equação 4 é recursiva pois $r\left(p_{\Delta}\right)$ depende de p_{Δ} que por sua vez é função de p_b^+. Smith [3] propõe solução através de uma *lookup table*. No entanto, para uma representação adequada isto exige grande quantidade de memória. O próprio Smith [3] obteve sons realísticos da clarineta através de uma expressão simples para o coeficiente de reflexão:

$$\hat{r}\left(p_{\Delta}^{+}\right)=\begin{cases} 1+m\left(p_{\Delta}^{+}-p_{C}\right), & p_{\Delta}^{+}<p_{C} \\ 1, & p_{\Delta}^{+}\geq p_{C} \end{cases} \quad (6)$$

onde, por definição, $p_{\Delta}^{+}=p_{cb}/2-p_b^-$, p_C é a pressão necessária para fechar o canal de entrada formado entre a palheta e boquilha e m é o coeficiente angular.

A implementação deste procedimento por guia de ondas está indicado na Fig. 3. A pressão no tubo (p_b) é composta pela soma das variáveis acústicas (p_b^+) e (p_b^-).

1.2 Modelo do Tubo

Os resultados estabelecidos por Smith [3] para a síntese sonora de instrumentos musicais (principalmente aerófonos e cordofones) podem ser apresentados a partir da solução geral da equação da onda, descrita por d'Alembert em 1747:

$$y(x,t)=y^{+}(x-ct)+y^{-}(x+ct) \quad (7)$$

onde y^+ e y^- são funções arbitrárias duas vezes diferenciáveis. O termo $y^+(x-ct)$ representa a função acústica caminhando na direção de x positivo enquanto $y^-(x+ct)$ no sentido oposto.

Quando as formas de onda são limitadas em banda é possível amostrar a solução por ondas viajantes e a Eq. 7 torna-se,

$$y(n,m)=y^{+}(n,m)+y^{-}(n,m) \quad (8)$$

Neste caso tanto a variável temporal t como a espacial x foram discretizadas através da relação,

$$\begin{aligned} t &= nT \quad \text{para} \quad n=0,1,2,3... \\ x &= mX \quad \text{para} \quad m=0,1,2,3... \end{aligned} \quad (9)$$

onde T é o intervalo de amostragem, isto é, o inverso da freqüência de amostragem e X é o intervalo espacial de amostragem o qual relaciona-se com T através da relação,

$$X=cT \quad (10)$$

A Figura 4 é a representação das ondas viajantes através do guia de ondas digital, cuja sigla em inglês é DWG (*Digital Waveguide*). Cada bloco corresponde a uma *delay line*. Computacionalmente, esta consiste em um processo de acesso, atualização e armazenamento em cada instante de amostragem.

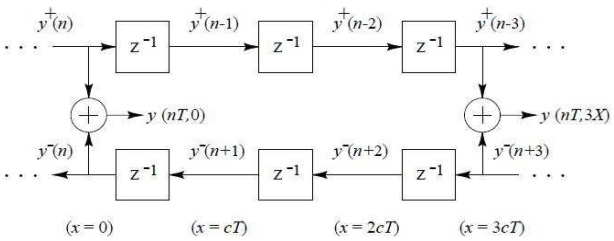


Figure 4: Guia de Ondas Digital.

Geralmente, os instrumentos de sopro da família das madeiras possuem tubos com geometria cilíndrica (como é o caso da clarineta) ou cônica (saxofone, por exemplo). Para estas geometrias, uma eficiente simulação por guia de ondas digital é obtida com ondas uniparamétricas.

A geometria real do instrumento é aproximada por seqüências de secções cilíndricas ou cônicas e será implementada em estudo posterior.

Neste trabalho, a síntese por modelagem física foi obtida para um tubo cilíndrico, sem furos e sem perdas causadas pelo atrito.

1.3 Modelo da Radiação Sonora

A radiação sonora da clarineta é feita através dos furos e parcialmente pela campana. Exceção ocorre quando **todos** os furos estão fechados e assim a radiação sonora é totalmente pela campana.

A configuração de furos abertos e fechados irá depender da nota emitida. No entanto, Scavone [4] adverte que a influência dos furos abertos e fechados sobre a propagação da onda no interior do tubo faz com que o modelo se distancie do comportamento ideal. Isto é, ondas planas em tubos de geometria cilíndricos e ondas esféricas em tubos cônicos. A aproximação mais simples, porém mais incorreta, é assumir que furos *fechados* não têm efeito sobre a propagação da onda e que o tubo é simplesmente *truncado* no primeiro furo aberto.

Desta forma, toda reflexão e transmissão da onda sonora ocorre no primeiro furo aberto. Porém esta hipótese está distante do comportamento real do instrumento.

Importante abordagem foi proposta por Keefe [7] ao analisar o efeito de um único furo no tubo. A configuração de furos desejada é obtida através de superposição linear. Ele caracteriza a influência de um furo através de uma junção de espalhamento simétrica representada por uma matriz de transmissão. Os elementos da matriz indicam a presença de furo aberto ou fechado ou ainda se o furo é tampado pelo dedo ou por

chave.

No caso da campana, da mesma forma que a energia de um sinal sonoro em um alto-falante é dividido entre o *woofer* e o *tweeter*, a campana da clarineta deixa passar as "altas" freqüências e reflete as "baixas" freqüências, como genericamente descreve Smith [9]. Onde freqüências "altas" e "baixas" estão relacionadas com o diâmetro da campana.

Neste estudo a geometria da campana e do tubo também foram consideradas. A radiação na saída do instrumento é calculada através de um filtro digital FIR ou IIR, indicado por $R(z)$ na Figura 5, cujos parâmetros estão baseados na solução de Levine e Schwinger [10] para a radiação sonora em tubos abertos sem flange, [4, pág.130].

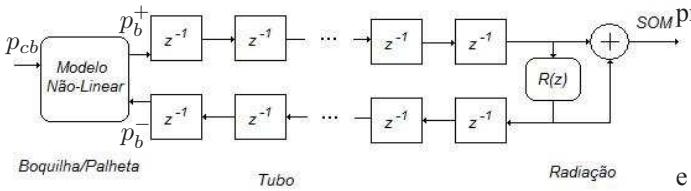


Figure 5: Modelo por Guia de Ondas Digital da Clarineta.

2 PROPAGAÇÃO SONORA COM ESCOAMENTO

Após análise das sínteses da clarineta por modelagem física disponíveis na literatuta verificou-se que estas não incluem a vazão, apesar do termo "*volume flow*" ser amplamente utilizado na elaboração dos modelos. Entende-se por esta denominação o movimento oscilatório das moléculas de ar e não propriamente o escoamento no interior do tubo destes instrumentos.

Efetivamente, o termo é utilizado para designar o fluxo de velocidade proveniente das ondas de compressão e rarefação e não para indicar a velocidade de escoamento do fluido. Estas ondas resultam na propagação da velocidade de perturbações infinitesimais de pressão, que define a velocidade do som.

Em resumo, todo o desenvolvimento da modelagem física analisado foi elaborado para um meio de propagação sonora em *repouso*. Isto, naturalmente, não corresponde à realidade de produção sonora dos instrumentos de sopro.

O resultado teórico de propagação sonora com o meio em movimento pode ser obtido no texto de Landau e Lifshitz [5] que é apresentado resumidamente neste trabalho. Genericamente, seja $\mathbf{n}$ o versor (ou seja, um vetor de módulo unitário) na direção de propagação sonora.

A notação aqui empregada utiliza grafia em **negrito** para vetores e normal para escalares. Assim, o vetor

$$\mathbf{k} = \left(\frac{\omega}{c}\right)\mathbf{n} \tag{11}$$

é denominado *vetor de onda* ("*wave vector*") e seu módulo k, *número de onda* ("*wave number*"). Aqui, ω é a freqüência e c a velocidade do som.

Quando a onda sonora propaga-se em um meio com velocidade de escoamento $\mathbf{u}$ a velocidade de propagação torna-se,

$$\tilde{\mathbf{c}} = c\frac{\mathbf{k}}{k} + \mathbf{u} \tag{12}$$

que é o vetor soma da velocidade c na direção de $\mathbf{k}$ e a velocidade $\mathbf{u}$ com a qual o som é "carregado" pelo fluido em movimento, conforme descrito em Laudau e Lifshitz [5].

2.1 Determinação da Propagação por Camadas

A aplicação dos resultados da secção anterior para a síntese da clarineta por modelagem física através do guia de ondas digital segue o seguinte procedimento.

O guia de ondas representa a propagação dos componentes da onda em direções opostas no tubo acústico. O número de elementos de *delay lines*, L, está relacionado com o comprimento efetivo do tubo, l_{ef}, através da relação,

$$L = \frac{l_{ef}F_S}{\tilde{\mathbf{c}}} = \frac{l_{ef}F_S}{c + \mathbf{u}} \tag{13}$$

admitindo que $\mathbf{u}$ está na mesma direção da propagação sonora. Assim, quando o escoamento está no mesmo sentido da propagação sonora (nas *delay lines* para a direita), tem-se $\tilde{\mathbf{c}} = c + u$, e quando o escoamento está no sentido oposto (nas *delay lines* para a esquerda), $\tilde{\mathbf{c}} = c - u$.

Pode-se reescrever a expressão anterior, Equação 13, em duas outras:

$$L_d = \frac{l_{ef}F_S}{c + u} \tag{14}$$

e

$$L_e = \frac{l_{ef}F_S}{c - u} \tag{15}$$

indicando a quantidade de delay lines para a direita (L_d) e esquerda (L_e), respectivamente. O resultado é sempre superior para a propagação à esquerda.

Igualam-se apenas quando a velocidade de escoamento é *nula* (na parede do instrumento).

A inclusão de um perfil de velocidades segue naturalmente. Determina-se a velocidade média de escoamento em um ponto do tubo através da solução do conjunto de equações 1 a 6 apresentadas na seção 1.1. Estes tratamentos podem ser observados em [11, 2, 4].

Em seguida pode-se dividir o tubo em diferentes camadas onde em cada camada instala-se um guia de ondas para a direita e outro para a esquerda. Deste modo tem-se,

$$L_{di} = \frac{l_{ef} F_S}{c + u_i} \qquad i = 1, \ldots, N_c \tag{16}$$

e

$$L_{ei} = \frac{l_{ef} F_S}{c - u_i} \qquad i = 1, \ldots, N_c \tag{17}$$

onde N_c indica o número total de camadas.

2.2 Duas Sugestões de Divisão do Tubo por Camadas

Para a divisão do tubo por camadas sugere-se aqui duas possibilidades. Na primeira, divide-se o raio do tubo em camadas de *mesma largura*. Na segunda, divide-se o diâmetro do tubo de modo que as camadas tenham *mesma área*. A Figura 6 representa a escolha de um perfil arbitrário, aparentemente triangular, ao longo do diâmetro. Neste caso foram utilizadas três camadas ($N_c = 3$), além de uma fina camada próxima à parede, que será referenciada por **camada limite**. Estas camadas foram obtidas dividindo-se a secção transversal do tudo, excluindo a camada limite, em três partes iguais, neste caso.

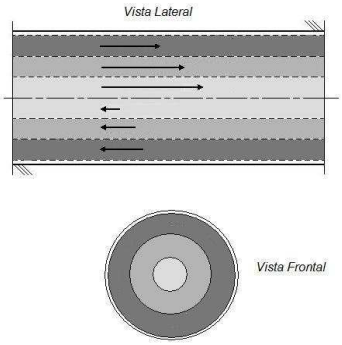


Figure 6: Divisão da Secção Transversal do Tubo por $N_c = 3$ Camadas de *mesma* **Largura**, além da Região da Parede.

Uma outra maneira de dividir a secção transversal em camadas está indicado na Figura 7. Neste caso as três camadas foram obtidas de modo a obter secções transversais com a mesma área. Excluindo a camada limite.

Observa-se ainda que no cálculo do filtro de radiação sonora, $R(z)$, o raio do tubo é um parâmetro

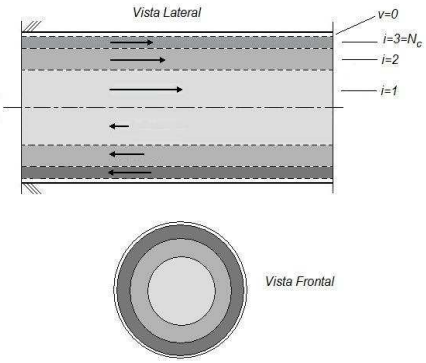


Figure 7: Divisão da Secção Transversal do Tubo por $N_c = 3$ Camadas de *mesma* **Área**, além da Região da Parede.

de entrada. Ao elaborar a divisão do tubo por camadas, os anéis das camadas serão transformados em um cilindro *equivalente* de modo a obter o *raio equivalente* como parâmetro de entrada do filtro. Com isso, ao elaborar a divisão do tubo por camadas de *mesma área* o filtro de radiação será idêntico para todas as camadas. Quando a divisão das camadas for de mesmo comprimento os filtros serão distintos.

Nos trabalhos de Scavone [4, págs.26-29] e Nackaerts [2, págs.44-46] estão apresentados expressões para o cálculo da espessura da camada limite viscosa e térmica. Nestas expressões não aparecem a velocidade de escoamento. É uma relação que envolve a viscosidade e densidade do fluido além freqüência para o caso viscoso e outra relação entre condutividade térmica, calor específico e densidade do ar como também a freqüência.

Segundo Perry [12, pág.40, Cap.6] "A espessura da camada limite é indefinida pois a velocidade de escoamento deve ser nula na parede (hipótese do não-deslizamento) e aproxima-se assintoticamente à velocidade do escoamento longe da parede".

Para efeito de síntese, neste trabalho será considerada a camada limite como sendo 1/10 do raio do tubo e nesta região a velocidade será nula.

3 Exemplos para Perfis Parabólico, Constante e Triangular

Para verificar o efeito do escoamento no tubo do instrumento propõe-se o estudo de uma clarineta com formato de campana catenoidal cujas dimensões estão apresentadas na Tabela 1. A aproximação da geometria do tubo é efetuada por uma sucessão de cilindros de diâmetros e comprimentos hipotéticos. A campana tem comprimento de 0.084 (m), diâmetro inicial de 0.0189 (m) e final de 0.053 (m). Ela foi subdividida em 10 partes iguais a partir do trapézio formado pelo diâmetro inicial, comprimento da campana e diâmetro final.

O perfil parabólico representa escoamento laminar.

O perfil constante aproxima o escoamento pistonado ("*plug-flow*"). O escoamento triangular para representar algo intermediário entre os dois primeiros.

Table 1: Aproximação da geometria do tubo por uma sucessão de cilindros hipotéticos.

L (10^{-3} m)	65	65	75	50	80	65	20
D (10^{-3} m)	18,9	20	18,9	20	18,9	20	18,9
L (10^{-3} m)	30	40	30	20	30	20	84
D (10^{-3} m)	20	18,9	20	18,9	20	18,9	53

A título de exemplo foi estabelecida a divisão do tubo em 4 camadas, além da camada limite que corresponde a $1/10$ do raio do tubo, que neste exemplo varia ao longo dos diferentes diâmetros.

A Figura 8 mostra a divisão do tubo por 4 camadas cujo critério é que as camadas apresentem a *mesma área*, exceção da camada limite.

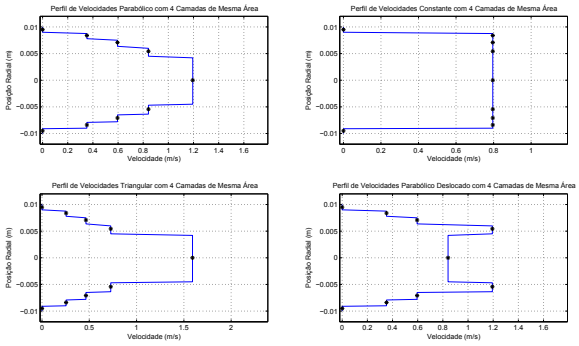


Figure 8: Divisão do tubo da clarineta em 4 camadas de **mesma área** para perfis parabólico, constante, triangular e parabólico deslocado.

Analogamente, a Figura 9 mostra a divisão do tubo por 4 camadas cujo critério é que as camadas apresentem a *mesma área*, exceção da camada limite.

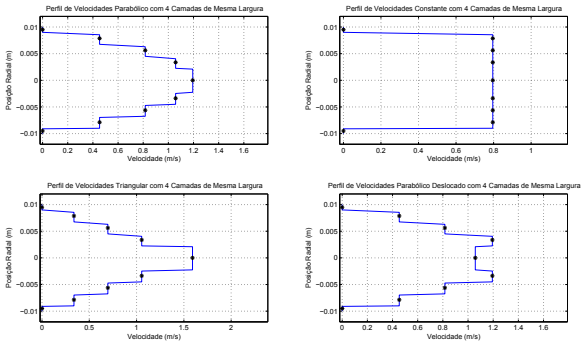


Figure 9: Divisão do tubo da clarineta em 4 camadas de **mesma largura** para perfis parabólico, constante, triangular e parabólico deslocado.

As amostras sonoras têm duração de 4s e em torno da metade do tempo são apresentados os perfis de pressão no interior do instrumento tanto para a divisão por mesma área, Figura 10, como por mesma largura, Figura 11.

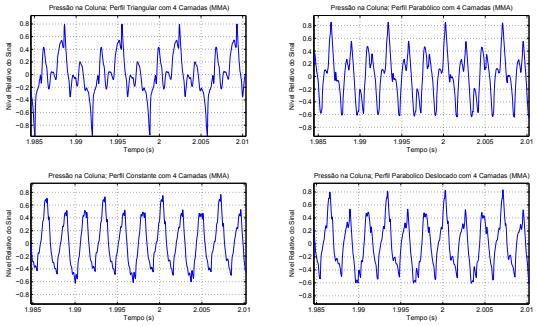


Figure 10: Pressão interna no tubo da clarineta dividido em 4 camadas de **mesma área** para perfis parabólico, constante, triangular e parabólico deslocado.

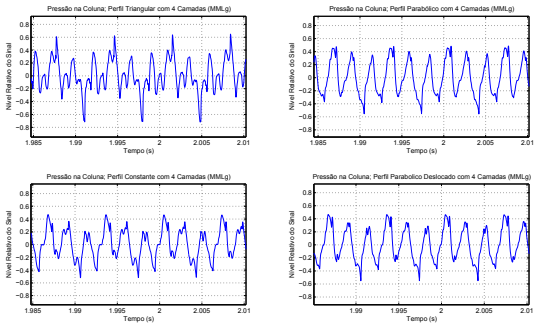


Figure 11: Pressão interna no tubo da clarineta dividido em 4 camadas de **mesma largura** para perfis parabólico, constante, triangular e parabólico deslocado.

4 Conclusão e Continuidade

A inclusão do escoamento oferece um resultado interessante pois proporcionou uma *distorção* da onda de pressão no interior do tubo que é característica dos instrumentos de sopro. Proporciona também uma sonoridade mais adequada aos intrumentos de sopro.

Associar os perfis de velocidade à sonoridade de instrumentistas, por exemplo iniciantes e experientes, faz parte da continuidade dos estudos.

Algumas dúvidas ainda são pertinentes e terão discussão futura, por exemplo, como distribuir as reflexões internas na geometria do tubo entre as diferentes camadas. Neste estudo foi proposta uma distribuição linear ao longo do raio, isto é, as camadas mais próximas à parede tem maior reflexão do que aquelas mais ao centro do tubo.

REFERENCES

[1] Julius Orion Smith III, "Bibliography: Physical modeling of musical instruments," Home Page, April 2008.

[2] Axel Nackaerts, *Sound Synthesis by Simulation of Physical Models of Musical Instruments*, Ph.D. thesis, KATHOLIEKE UNIVERSITEITE LEUVEN, Leuven, 2003.

[3] Julius Orion Smith III, "Efficient simulation of the reed-bore and bow-string mechanisms," in *Proc. 1986 Int. Computer Music Conf.(ICMC'86)*, The Hague,The Netherlands, 1986, pp. 275–280, Computer Music Association.

[4] Gary Paul Scavone, *An Acoustic Analysis of Single-Reed Woodwinds Instruments with an Emphasis on Design and Performance Issues and Digital Waveguide Modeling Techniques*, Ph.D. thesis, CCRMA - Stanford University, California, 1997.

[5] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, Butterworth-Heinemann, Oxford, 2nd edition, 1987.

[6] M.E. McIntyre; R.T. Schumacher; J. Woodhouse, "On the oscillations of musical instruments," *J. Acoust. Soc. Am.*, vol. 74, pp. 1325–1324, November 1983.

[7] D. H. Keefe, "Acoustic streaming, dimensional analysis of nonlinearities, and tone hole mutual interactions in woodwinds," *J. Acoust. Soc. Am.*, vol. 75, no. 5, pp. 1804–1820, 1983.

[8] Neville H. Fletcher e Thomas D. Rossing, *The Physics of Musical Instruments*, Springer, New York, 2nd edition, 2005, 5th corr. printing.

[9] Julius O. Smith III, "Physical modeling using waveguides," *Computer Music Journal*, vol. 16, no. 4, pp. 74–91, November 1992.

[10] H. Levine; J. Schwinger, "On the radiation of sound from an unflanged circular pipe," *Phys. Rev.*, vol. 73, no. 4, pp. 383–406, 1948.

[11] Luís Carlos de Oliveira, *Síntese por Modelagem Física da Clarineta: Modelo por Guia de Ondas com Escoamento*, Ph.D. thesis, Faculdade de Engenharia Elétrica e de Computação(UNICAMP), Campinas-SP, 2011.

[12] Robert H. Perry; Don W. Green, *Perry's Chemical Engineers' Handbook*, Mc Graw-Hill, USA, 7th edition, 1999, Prepared by a staff of specialists under the editorial direction of late editor Robert H. Perry: editor, Don W. Green.

OLIVEIRA, L. C.; GOLDEMBERG, R.; MENDES, R. S. Discussão teórica sobre a inclusão do escoamento e perfis de velocidade na modelagem física de instrumentos de sopro. In: CONGRESSO DE ENGENHARIA DE ÁUDIO AES BRASIL, 10., 2012, São Paulo, *Anais*... Rio de Janeiro: Sociedade de Engenharia de Áudio, 2012. p. 82-89.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Síntese por amostragem de piano baseada em codificação wavelet

Thiago Rossi Roque[1] e Rafael Santos Mendes[1]

[1] Universidade Estadual de Campinas, FEEC, DCA
Campinas, São Paulo, 13083-970, Brasil

ra072515@fee.unicamp.br, rafael@dca.fee.unicamp.br

RESUMO

Este trabalho propõe a criação de um sintetizador de sinais musicais utilizando a técnica de amostragem aliada a uma compressão de dados de alta eficiência pela transformada wavelet. Inicialmente é apresentada uma breve introdução a respeito das técnicas de síntese sonora, em especial da síntese por amostragem, seguido por uma descrição do som do piano, instrumento escolhido nesse trabalho, e da criação do banco de amostras. Em seguida é apresentada a transformada wavelet e a técnica de compressão utilizada, para então ser descrito o sistema completo de síntese a partir do banco de amostras comprimidas seguido pela demonstração dos resultados obtidos.

0 INTRODUÇÃO

Desde o surgimento da válvula de triodo e da eletrônica, o homem tem se esforçado para incorporar as novas invenções elétricas na música buscando novas formas de expressão [1]. A partir dessa busca surgiu uma nova classe de instrumentos que não mais produziam som a partir de cordas, membranas ou colunas de ar, mas sim a partir de oscilações na corrente elétrica de um sistema.

A essa nova classe de instrumentos musicas deu-se o nome de sintetizadores, pela capacidade que esses tinham de sintetizar sons. Com o avanço da tecnologia, diversas técnicas de síntese sonora surgiram. Merecem destaque a síntese subtrativa, síntese aditiva, síntese por modulação de frequência (FM), síntese por amostragem e por modelagem física, por seus sucessos tanto no meio musical quanto no meio comercial.

Cada técnica de síntese favorece diferentes aplicações e diferentes resultados sonoros. A síntese por amostragem em especial se demonstra muito eficiente em sintetizar instrumentos de difícil modelagem e com variações de dinâmica previsíveis, como é o caso do piano e de instrumentos de percussão. Pelo fato da síntese por amostragem exigir baixa capacidade computacional e possibilitar sintetizar sons já existentes com grande fidelidade, esta técnica se tornou a mais utilizada atual-

mente [2]. Porém, por se basear inteiramente em amostras de um som específico, a síntese por amostragem necessita de uma grande quantidade de informações a respeito do som a ser sintetizado para que o processo ocorra com qualidade.

Para contornar o problema da quantidade de informação necessária para este processo de síntese sonora, técnicas de compressão de sinais geralmente são utilizadas em conjunto com a síntese. Muitas vezes ao invés de armazenar longos trechos de amostras de um sinal de áudio, somente um período desse sinal é armazenado e no momento da síntese é executado em laço (*loop*), a partir da técnica chamada *wavetable* [3], ou mesmo usando diversos períodos de diferentes trechos da amostra do sinal que se combinam criando múltiplas *wavetables* [1].

Porém ao armazenar somente pequenos instantes de um sinal, muita informação a respeito da variação temporal do som é perdida causando uma degradação na qualidade da síntese. Para que as informações tanto temporais quanto frequenciais possam ser preservadas mesmo quando se dispõe de quantidades reduzidas de memória, técnicas mais complexas de compressão de sinais devem ser empregadas.

Apesar de atualmente o custo de memórias não voláteis ter caído consideravelmente, com o crescente mercado de *tablets* e *smartphones*, a questão de economia de memória em processos de síntese por amostragem voltou a se tornar um fator limitante em aplicações portáteis [4]. Com possíveis bancos de amostras de menos de 1 MB, novos aplicativos de síntese possam se tornar possíveis.

1 O SOM DO PIANO

Para tornar possível uma compressão de alta eficiência, foi necessário um estudo do som do piano de forma a propiciar um melhor aproveitamento das amostras em um esforço de concentrar a maior quantidade de informação possível em amostras mais curtas.

A mecânica de um piano consiste em um complexo conjunto de dispositivos e sistemas de alavancas que transmitem o movimento das teclas para os martelos que percutem as cordas e as fazem vibrar [5]. De acordo com Henrique [5], a dinâmica de uma nota do piano é composta por duas fases, a primeira é chamada de *som imediato* e corresponde ao instante de vibração da corda provocada diretamente pela percussão do martelo, onde a oscilação predominante da corda ocorre na mesma direção em que o martelo se movimenta e é perpendicular ao tampo harmônico, onde a corda se encontra fixada. A segunda fase é chamada de *som remanescente*, onde a vibração da corda é paralela ao tampo harmônico e por isso oferece maior sustentação ao som.

O tempo de duração de cada fase do som varia muito de acordo com a nota analisada. As notas mais graves de um piano chegam a ter um *som imediato* de quase 10 segundos, e um *som remanescente* de até 20 segundos, já as notas mais agudas podem ter um *som imediato* de apenas 1 segundo, enquanto que o *som remanescente* de pouco menos de 3 segundos, conforme observado por Meyer [6].

Ao analisar tanto a envoltória temporal quanto o espectro de uma amostra de uma nota de um piano pode-se notar claramente a distinção entre esses dois momentos do som do instrumento tanto com relação ao decaimento da amplitude do sinal, quanto pela composição harmônica mais rica no trecho de *som imediato*. Com o intuito de reduzir a quantidade de informação armazenada, decidiu-se não armazenar as amostras inteiras do piano, mas apenas trechos dos dois instantes do som para cada nota do instrumento. Notou-se no trecho de *som imediato* uma variação intensa tanto em conteúdo espectral quanto na amplitude do sinal, portanto optou-se por manter um instante mais longo para esse trecho de som, que compreende cerca de dois terços da duração do trecho de *som imediato*, e varia entre 250 milissegundos a 800 milissegundos de acordo com a frequência. Já para o trecho de *som remanescente* o sinal é mais constante nos domínios temporal e frequencial, portanto de maior previsibilidade, caracterizando-se por uma quantidade de informação [7], o que justifica utilizar um trecho mais curto a ser armazenado, sem grandes prejuízos na qualidade sonora a ser sintetizada. Após diversos testes empíricos chegou-se a trechos de cerca de 10 períodos do sinal, de forma a conciliar um tamanho razoável que facilitasse a execução em laço (como veremos mais a frente) sem que fossem notadas alterações de amplitude pelo decaimento natural do som do piano ao longo desse trecho.

1.1 O Banco de Amostras

Para criar o banco de amostras do sintetizador foram obtidas amostras de 85 notas de um piano *Steinway & Sons* tocadas em intensidade média (*mezzo forte*) a partir da biblioteca de sons de instrumentos musicais da universidade de Iowa [8], e a partir de cada amostra foram extraídos os dois trechos pertencentes aos instantes de *som imediato* e *som remanescente* que irão compor o banco de amostras. A partir somente dos dois trechos é possível sintetizar o som de uma nota de um piano com grande fidelidade. O processo de síntese a partir desses trechos será descrito na Seção 4.1.

2 A TEORIA WAVELET

Similar a análise de Fourier [9], a teoria wavelet consiste em decompor um sinal através de outras funções mais simples. Porém as funções utilizadas nas transformadas wavelet são funções de amplitude e suporte finitos e não possuem componentes de frequência nula, sendo assim quadrado integráveis. A decomposição de sinais pela transformada wavelet possui a característica de decomposição de sinais em níveis de resolução progressivos, isto é, permite a representação de sinais em grau de refinamento crescente [10]. É baseada na ideia de representar um sinal a partir de múltiplas escalas temporais (o que remete

ao domínio frequencial) e de múltiplos deslocamentos temporais (o que remete ao domínio propriamente temporal), característica que proporciona a multirresolução à transformada [11].

A partir da função escala $\varphi(t)$ é definido um conjunto de funções escala $\varphi(t-k)$ para qualquer k inteiro de forma que a Equação (1) seja satisfeita.

$$\varphi(t) = \sum_{n \in Z} h(n)\, \sqrt{2}\, \varphi(2t-n) \tag{1}$$

.

Definindo uma função wavelet $\psi(t)$, ortogonal a $\varphi(t)$, obtém-se a família de funções wavelet $\psi(2^j t - k)$ de forma que o espaço descrito por essa família de funções seja a diferença entre dois sub-espaços descritos por duas funções escala de níveis subsequentes. Esta condição é formalizada na Equação (2).

$$\psi(t) = \sum_{n \in Z} h_1(n)\, \sqrt{2}\, \varphi(2t-n) \tag{2}$$

Qualquer sinal $x(t)$ de energia finita pode ser decomposto pela transformada wavelet conforme apresentado na Equação (3).

$$\begin{aligned} x(t) &= \sum_{k=-\infty}^{\infty} c_r(k) \cdot \varphi(2^r t - k) + \\ &+ \sum_{k=-\infty}^{\infty} \sum_{j=r}^{\infty} d_j(k) \cdot \psi(2^j t - k), \end{aligned} \tag{3}$$

$c_r(k)$ e $d_j(k)$ são os coeficientes da transformada wavelet do sinal $x(t)$. Na Equação (3) o número inteiro r é arbitrário e representa "a escala mais grosseira" de representação do sinal $x(t)$ e cada valor do índice j está associado a um nível de detalhe do sinal.

Caso $x(t)$ seja um sinal discreto, sua banda é limitada e pode-se provar que existe um nível de resolução r, tal que a função $\varphi(2^r t)$ se aproxima da função impulso $\delta(t)$. Consequentemente os coeficientes $c_r(k)$ são convenientemente aproximados por $x(k)$.

A partir dessas aproximações e dos coeficientes $h(n)$ e $h_1(n)$, que são constantes definidas pela função wavelet adotada, pode se obter a Equação (4) onde é possível obter tanto os coeficientes $c_j(k)$ quanto $d_j(k)$ a partir somente dos coeficientes $h(n)$ e $h_1(n)$ e de $c_{j+1}(n)$ e consequentemente de $x(t)$.

$$\begin{aligned} c_j(k) &= \sum_n h(n-2k) c_{j+1}(n) \\ d_j(k) &= \sum_n h_1(n-2k) c_{j+1}(n) \end{aligned} \tag{4}$$

Sendo o número de coeficientes $h(n)$ e $h_1(n)$ finito, essa equação pode ser interpretada como um processo de filtragem de resposta a impulso finita (FIR) seguido de decimação, onde $h(n)$ funciona como um filtro passa-baixa e $h_1(n)$ funciona como um filtro passa-alta.

3 A TRANSFORMADA WAVELET NO PROCESSO DE SÍNTESE

Dentre os diversos métodos de realização das transformadas wavelet discreta, o método adotado foi a wavelet packet onde tanto os coeficientes $c_j(k)$ quanto $d_j(k)$ são sucessivamente filtrados. Esta transformação pode ser representada por uma árvore binária na qual cada nível decomposto dá origem a dois níveis subsequentes. O último nível obtido correspondente às folhas da árvore binária constituem a representação wavelet do sinal decomposto.

Desde o surgimento da teoria wavelet diversas funções wavelets foram criadas para as mais diversas aplicações, pois a realização da análise wavelet a partir de diferentes funções base leva a resultados diversos. Para representação de sinais sonoros Faria [12] chegou a conclusão de que a suavidade da função wavelet (seu número de momentos nulos) é o parâmetro de maior importância na escolha da base a ser usada, chegando a conclusão que a wavelet que possui a maior suavidade com menor custo computacional seria a Daubechie 16, com oito momentos nulos e 32 coeficientes nos filtros $h(n)$ e $h_1(n)$ da wavelet. Porém para reduzir ainda mais o custo computacional com o objetivo de criar um sistema orientado ao funcionamento em tempo real optou-se por usar a wavelet Daubechie 8 por ter metade do número de coeficientes em seus filtros.

3.1 A Compressão

Para realizar a compressão das amostras foi desenvolvida uma técnica a partir da pesquisa de Luvizotto [13] onde o sinal é decomposto pela transformada wavelet packet e os coeficientes de cada folha são ordenados ascendentemente e então aproximados por curvas mais simples. Assim pode-se recuperar o sinal original com grande proximidade a partir somente da descrição da curva e do posicionamento original dos coeficientes.

O método aqui utilizado segue os mesmos princípios utilizados por Luvizotto [13], porém refinados para utilização nas amostras de som do piano. Aqui cada um dos dois trechos de cada amostra é submetido à transformada wavelet packet de 4 níveis [1]. Para maximizar a taxa de compressão é dado um limiar mínimo para os coeficientes, abaixo do qual estes são descartados.

Cada folha da wavelet packet é avaliada com relação à quantidade de coeficientes que estão abaixo ou acima do limiar. Caso todos os coeficientes estejam abaixo do limiar, a folha inteira é descartada, caso contrário é analisado quais coeficientes estão acima do limiar e somente esses relevantes são armazenados.

Para proporcionar melhor resultado sonoro, foi aplicado um processo que varia o limiar em determinadas folhas. Dado o fato de que cada folha compete a uma

[1] A princípio desejou-se realizar a transformada com 8 níveis, porém para maiores níveis o tamanho da matriz da transformada inversa se torna muito grande aumentando o custo computacional

determinada faixa de frequência, utilizou-se a curva proposta por Fletcher-Munson [14] que determina a variação da sensibilidade do ouvido humano para diferentes faixas de frequência. Assim as folhas que contém frequências mais sensíveis à audição humana têm seu limiar reduzido, enquanto frequências menos sensíveis têm seu limiar aumentado, dado que possuem menos importância para o resultado sonoro.

Em seguida todos os coeficientes da folha são ordenados de forma ascendente para então serem aproximados por um polinômio de grau pré-definido. Essa aproximação é feita pelo método de mínimos quadrados [15], porém utilizando diferentes pesos para os coeficientes, garantindo aos coeficientes mais próximos do limiar maior importância na aproximação polinomial pois esses coeficientes possuem mais influência no timbre e na qualidade sonora. Após a realização do mínimos quadrados é obtido o polinômio que melhor se aproxima à curva, e assim somente os coeficientes do polinômio são armazenados, ao invés de cada elemento da curva original. Um exemplo gráfico dos coeficientes de uma folha pode ser visto na Figura 1 enquanto, na Figura 2, notam-se os mesmos coeficientes ordenados sobrepostos pelo polinômio encontrado que melhor se aproxima da curva obtida.

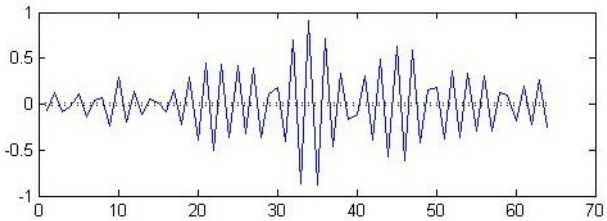


Figura 1: Coeficientes de uma folha.

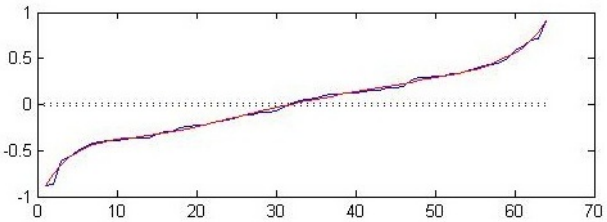


Figura 2: Coeficientes de uma folha ordenados e polinômio aproximado.

3.2 A Transformada Inversa

Nos métodos tradicionais da transformada wavelet packet inversa, a recuperação do sinal transformado é realizada de forma semelhante, porém no sentido oposto, ao método tradicional de decomposição descrito na Seção 3 pelo método dos filtros na criação da árvore binária. Porém essa técnica tem a característica de recuperar o sinal inteiro de uma única vez, o que não é desejável para este projeto. Para um processo de síntese sonora é preferível que a recuperação do sinal transformado ocorra amostra por amostra[2], de forma que o processamento de síntese possa ser feito paralelamente à execução do som favorecendo a execução em tempo real.

Para que isso fosse possível foi necessário implementar um método de realização de forma matricial da anti-transformada explicitada na Equação (5).

$$c_{j+1}(k) = \sum_n c_j(n)h(k-2n) + \sum_n d_j(n)h_1(k-2n) \tag{5}$$

A solução encontrada foi dispor os coeficientes dos filtros em uma matriz similar a uma matriz de *Toeplitz* e concatenar os coeficientes das folhas de um mesmo nível da transformada em um único vetor. A multiplicação dessa matriz de filtros pelo vetor de coeficientes gera o nível seguinte da árvore wavelet packet. A matriz de filtros está esquematizada na Figura 3, onde n é o comprimento de cada folha para o mesmo nível da matriz e nh é o comprimento dos filtros, que são dispostos verticalmente nas colunas e cada coluna de filtro é defasada em duas linhas garantindo o processo de decimação.

Nota-se que a matriz pode ser dividida verticalmente em duas, onde as duas metades são praticamente iguais, porém a primeira utiliza os coeficientes do filtro $h0$ e a segunda utiliza os coeficientes do filtro $h1$. Essa divisão ocorre, pois cada metade será multiplicada por cada folha binária de mesmo nível.

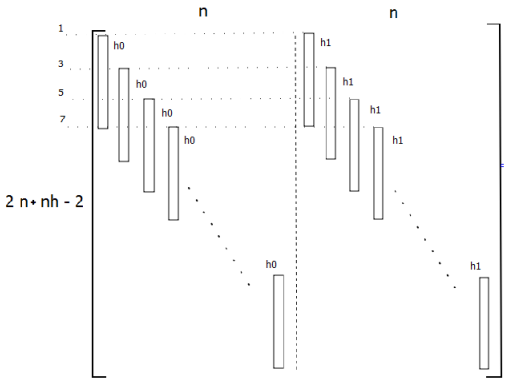


Figura 3: Matriz básica.

Porém essa matriz apresentada, que será denominada de matriz básica, só pode ser usada para o caso de duas folhas da árvore wavelet packet. Para níveis de mais de duas folhas é necessário compor uma nova matriz. Para o caso de quatro folhas, essa nova matriz

[2] Aqui se usa amostra com significado similar ao resultado da multiplicação de um sinal por uma função impulso, ao contrário do uso mais decorrente nesse trabalho que se refere a um trecho representativo de áudio.

pode ser obtida pela multiplicação de uma matriz diagonal composta pela matriz básica em sua diagonal com uma nova matriz similar à básica, porém de mesmas dimensões que a matriz diagonal.

Esse processo pode ser estendido para qualquer nível, onde a matriz de anti-transformada de nível k pode ser obtida a partir da matriz de nível $k-1$, e consequentemente da matriz básica. Esse processo está descrito na Figura 4.

Inicialmente é criada a matriz básica e então são feitas as sucessivas multiplicações até atingir a matriz para o nível necessário. Dada a característica recursiva desse cálculo, esse se apresenta muito eficiente, apesar das grandes dimensões envolvidas.

Nota-se que a cada multiplicação do vetor concatenado das folhas com cada linha da matriz se obtêm uma única amostra do sinal recuperado, como desejado inicialmente.

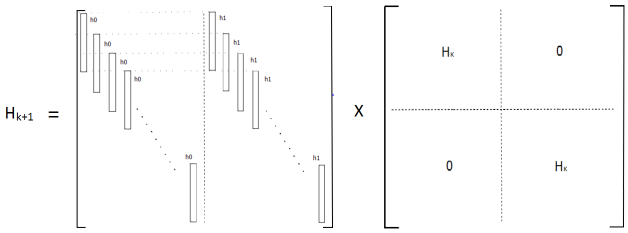


Figura 4: Construção da matriz de anti-transformada para nível genérico.

4 O SINTETIZADOR

O diagrama do sistema está apresentado na Figura 5. O processo começa com a recepção de uma mensagem MIDI [16], que é decodificada e inicia o processo de gerenciamento das notas a serem executadas. Esse processo de gerenciamento é feito por um método chamado de quadro negro, baseado nas estruturas de mesmo nome muito utilizadas em arquiteturas de softwares [17], onde a mensagem MIDI decodificada, antes de iniciar a síntese, é escrita juntamente com variáveis que representam o estado do processo de síntese. Esse método garante que mais de uma nota possa ser executada ao mesmo tempo, ou seja, que o sintetizador seja polifônico.

Após a escrita no quadro-negro da mensagem recebida, o sistema requisita as amostras correspondentes a nota do evento MIDI ao banco e, enfim se inicia o processo de síntese onde as amostras são descomprimidas e processadas.

Caso o sistema receba dois eventos de execução de notas simultaneamente, ou receba uma nova nota enquanto uma outra está sendo gerada, um gerenciador da síntese faz com que n notas sejam geradas ao mesmo tempo, ou seja, em cada período de amostragem é gerada uma amostra[2] de cada nota e a soma das n amostras resulta no som correspondente ao som simultâneo

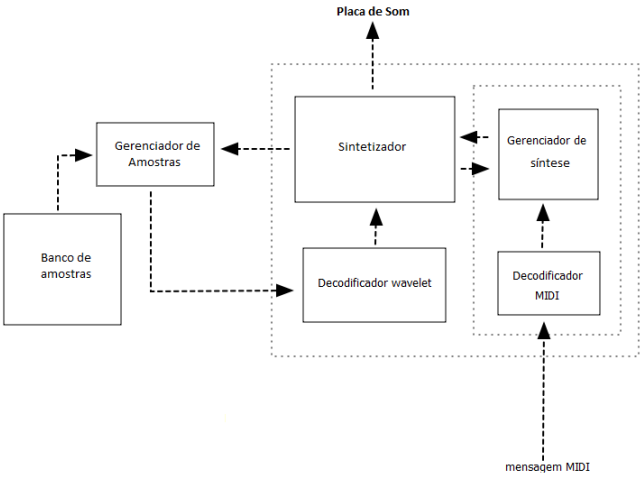


Figura 5: Estrutura do Sintetizador.

das n notas. Isto é possível pois para cada amostra gerada, o estado da nota que esta sendo sintetizada é atualizado, possibilitando ao sistema interromper o processo em qualquer etapa para sintetizar uma nova nota em qualquer momento.

Conforme as amostras são geradas, estas são submetidas a fatores de amplitude que determinam a envoltória da nota, além de simular a dinâmica de intensidade da nota de acordo com o valor da intensidade na mensagem MIDI. Caso sintetizador receba uma mensagem de fim de nota durante a execução da mesma, o sistema atualiza o evento dessa nota no quadro-negro e submete as amostras que estão sendo geradas a um novo decaimento mais acentuado, simulando a mecânica real de um piano.

4.1 A Síntese

O processo de síntese se inicia no momento em que o quadro negro é preenchido por uma mensagem MIDI de evento do tipo "NOTE-ON", assim o sistema analisa qual nota deve ser executada e requisita ao banco de amostras os trechos necessários. Porém, como os trechos se encontram comprimidos e codificados sob forma de polinômios, estes são previamente decodificados até que as folhas da wavelet packet sejam reconstruidas. A partir das folhas a recuperação do sinal ocorre através da matriz da transformada inversa apresentada na Seção 3.2 conforme cada amostra de cada trecho é requisitada em cada instante de período de amostragem.

O processo de síntese em si consiste em avaliar o estado do processo e decidir se a amostra que será gerada deve ser sintetizada a partir do trecho de *som imediato* ou do trecho de *som remanescente*, e qual envoltória deverá ser aplicada à amostra. Em um instante inicial, o processo de síntese é feito por um método próximo ao *PCM Playback* [18] onde as amostras são sintetizadas a partir do trecho de *som imediato*, porém conforme esse

trecho chega ao seu final, a síntese passa a ocorrer a partir do trecho de *som remanescente*. No momento em que o contador de amostras do segundo trecho chega ao final, este é zerado e o trecho começa a ser lido desde seu início, realizando a execução em laço.

A partir do momento em que a síntese passa a ocorrer a partir do trecho de sustentação, cada amostra gerada é submetida a um fator que gera um decaimento exponencial na amplitude do sinal, de forma a simular o decaimento natural do som do piano. Para cada nota do piano um decaimento diferente é aplicado, pois como visto na Seção 1, a duração do som varia muito de acordo com a nota analisada.

Aliado a esse processo um segundo fator é aplicado a todas as amostras geradas, inclusive do trecho de *som imediato*, que é proporcional à mensagem MIDI de *velocity* e determina a intensidade da nota tocada. Sendo que o valor máximo que a mensagem contém é 127, esse valor é associado a um fator unitário.

Porém um problema muito comum ocorre durante a tentativa de emendar trechos de sinais periódicos. Caso os dois trechos a serem emendados sejam muito diferentes ou estejam em fases distintas de seu desenvolvimento é muito comum o surgimento de ruído de alta frequência (chamado comumente de *click*) na emenda dos sinais. Para que isso não ocorra foi necessário utilizar algumas ferramentas.

Previamente, quando os dois trechos de cada nota foram selecionados, foi calculado o segmento de tamanho n do final do trecho de *som imediato* que possui melhor correlação com o segmento inicial de mesmo tamanho n do trecho sustentação, e então o segmento de melhor correlação do trecho de *som imediato* foi feito de segmento final simplesmente apagando qualquer segmento sucedente. Assim pode-se garantir que a sobreposição desses dois segmentos possuirá alta semelhança e a mesma fase. O mesmo processo foi feito também entre um segmento final do trecho de sustentação com o segmento inicial do mesmo trecho, garantindo uma sobreposição de alta semelhança e mesma fase entre o segmento final e inicial do trecho de sustentação.

Garantindo semelhança entre os segmentos a serem emendados, não mais se muda a leitura de cada trecho no momento em que o trecho chega ao final, mas quando a leitura de primeiro trecho chega ao último segmento de tamanho n, ocorre uma sobreposição entre esse segmento e o segmento inicial do próximo trecho, porém o segmento final sofre um processo de "fade out" e o segmento inicial recebe um "fade in". O processo de "fade" consiste em submenter um segmento a um fator crescente de 0 a 1 para o "fade in" e a um fator de 1 a 0 para o "fade out", porém em qualquer instante a soma dos dois fatores deve ser constante e de valor unitário, de forma que é mantida a linearidade da sobreposição. Esse processo é feito também tanto para a emenda do trecho de *som imediato* com o trecho de sustentação, como entre o final e o início do trecho de sustentação para criar o *loop*.

Aplicando esse processo de "fade" em trechos de alta correlação, as chances de ocorrer *clicks* diminuem consideravelmente, criando emendas praticamente imperceptíveis, o que é de grande importância para garantir um som contínuo como o som natural de um piano.

5 CONCLUSÃO E PRÓXIMOS PASSOS

Os primeiros resultados obtidos foram a respeito do método de compressão das amostras, onde os primeiros testes foram muito promissores. A princípio foram criados arquivos binários contendo todos os trechos de amostras de cada nota comprimidos e codificados juntamente com dados necessários para a síntese a partir dessas notas. O conjunto das 85 notas originais ocupam uma memória de 421MB. Somente pelo fato das amostras serem reduzidas aos trechos de *som imediato* e *som remanescente*, esse valor pôde ser reduzido a apenas 5MB. Após as amostras serem comprimidas pelo método descrito o banco de amostras criado pode ser concentrado em apenas 1,44MB contendo todas as informações necessárias para a síntese, mesmo garantindo um resultado sonoro de qualidade com grande fidelidade ao som de um piano acústico. Nota-se que esse resultado foi obtido com amostras de todas as 85 notas, para aplicações de alta exigência de redução de memória é possível reduzir o número de amostras armazenadas e sintetizar as a notas não armazenadas através de técnicas de *pitch shifting* [19].

Os resultados sonoros obtidos através da escuta dos sons sintetizados foram muito positivos, notou-se boa fidelidade do som obtido em relação ao som original. A execução em laço do trecho de *som remanescente* se mostrou bem satisfatória assim como a simulação da dinâmica temporal do som do instrumento. Porém ainda é notado um ruído durante a execução da nota sintetizada.

Para avaliar os resultados de forma objetiva, executou-se a síntese de uma nota A4, cuja forma de onde é representada na Figura 6, para que a mesma fosse comparada com a nota original de onde foi extraída suas amostras. Na Figura 7 temos um trecho de *som imediato* para o som original e para o som sintetizado. Enquanto que na Figura 8 temos um trecho de *som remanescente* extraído de ambos os sons.

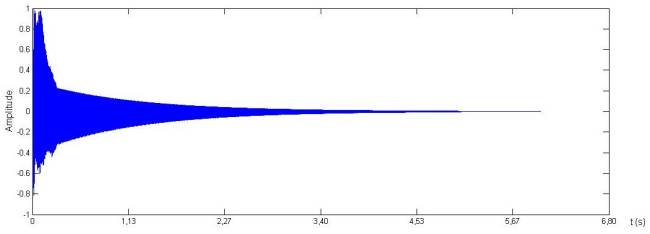


Figura 6: A4 sintetizado.

Pela análise dos formatos de onda pode-se notar grande semelhança entre os trechos de *som imediato*.

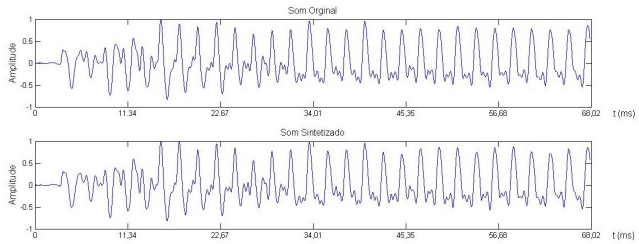


Figura 7: Comparação Entre Trechos de *Som Imediato*

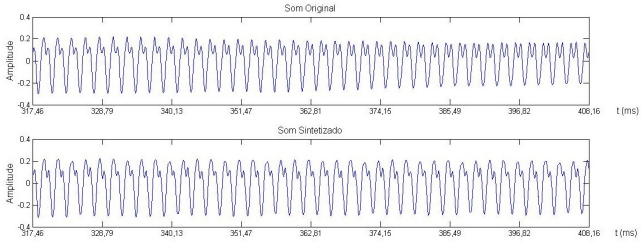


Figura 8: Comparação Entre Trechos de *Som Remanescente*

Para o trecho de *som remanescente* nota-se uma considerável diferença, inclusive é possível notar a repetição do trecho executado em laço, porém essa diferença entre o som sintetizado e o som original não é de grande distinção sonora.

Uma análise mais complexa pode ser feita a partir da análise de espectro dos dois sinais. Os espectros dos sons podem ser analisados através das Figuras 9 e 10. Como esperado pela análise anterior, o conteúdo espectral para os trechos iniciais de ambos os sons são semelhantes, porém nota-se mais uma vez a diferença entre os trechos de *som remanescente*. Pelo fato de ser utilizado somente um trecho curto executado em laço, o espectro desse trecho se torna constante, variando somente em amplitude pelo decaimento artificial inserido no processo de síntese, em oposição ao espectro do som original onde o decaimento é distinto para cada harmônica.

Um grande problema encontrado se dá no tempo de execução do programa, que é muito longo e inviabiliza uma execução em tempo real, como originalmente idealizado. O principal motivo encontrado para esse tempo de processamento está na plataforma utilizada, por usar uma linguagem de alto nível o programa Matlab não é a ferramenta ideal para projetos que necessitem de funcionamento em tempo real. Uma solução para esse problema seria a implementação desse projeto em linguagens de mais baixo nível como a linguagem "C", diminuindo o tempo de execução do programa consideravelmente. Seria interessante também estudar a esparsidade da matriz da transformada wavelet inversa de forma a desenvolver algoritmos mais eficientes para o cálculo de recuperação do sinal transformado.

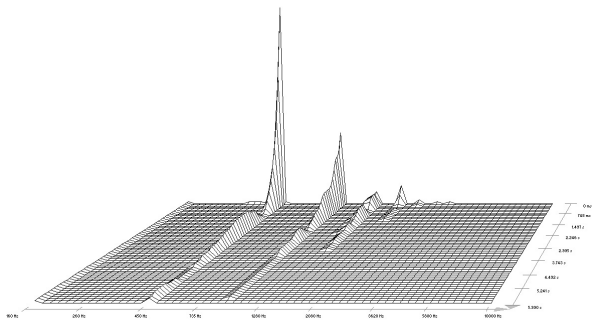


Figura 9: Espectro da Nota Original.

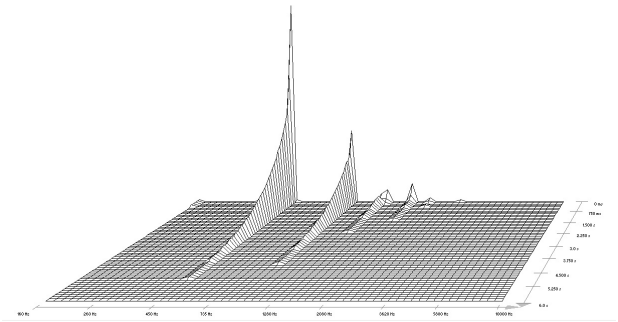


Figura 10: Espectro da Nota Sintetizada.

Outra limitação que a plataforma Matlab apresentou foi a falta de ferramentas para implementar uma comunicação externa com o programa para realizar o envio e recebimento de mensagens MIDI. Para isso foi feito um programa que simulasse o recebimento das mensagens nos instantes de ocorrência dos eventos mediante um temporizador. A utilização da linguagem "C" como plataforma de implementação também resolveria esse problema.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Curtis Roads, *The Computer Music Tutorial*, MIT Press, 1995.

[2] Ang Yaw Feng e Minni Ang, "Sampling synthesis technique applied for the digital generation of musical tones of malaysian folk instruments," Tech. Rep., Music Department, Faculty of Human Ecology, Universiti Putra Malaysia.

[3] Vesa Välimäki e Matti Karjalainen Tero Tolonem, "Evaluation of modern sound synthesis methods," Tech. Rep., Helsinki University of Technology, 1998.

[4] ROBERT C. MAHER, "Wavetable synthesis strategies for mobile devices," *JAES*, 2005.

[5] Luís L. Henrique, *Acústica Musical*, Fundação Calouste Gulbenkian, 2009.

[6] Jürgen Meyer, *Acoustics and the Performance of Music*, Springer, 2009.

[7] Bernardo N. B. Lima e Leandro Martins Cioletti e Marcelo de O. Terra Cunha e Gastão A. Braga, *Entropia: Introdução à Teoria Matemática da (des)Informação*, 2004.

[8] University of Iowa Electronic Music Studios, "http://theremin.music.uiowa.edu/mispiano.html," 2001.

[9] Alan V. Oppenheim e Alan S. Willsky e S. Hamid Nawab, *Signals & systems (2nd ed.)*, Prentice-Hall, Inc., 1996.

[10] C. Sidney Burrus e Ramesh A. Gopinath e Haitao Guo, *Introduction to Wavelets and Wavelet Transform: A Primer*, Prentice Hall, 1998.

[11] Paul S. Addison, *Illustrated Wavelet Transform Handbook*, Institute of Physics Publishing, 2002.

[12] Regis Rossi Alves Faria, "Aplicação de wavelets na análise de gestos musicais em timbres de instrumentos acústicos tradicionais," M.S. thesis, Escola Politécnica da Universidade de São Paulo, 2007.

[13] André Luiz Luvizotto, "Modelos de representação de sinais musicais via transformada wavelets," M.S. thesis, Faculdade de Engenharia Elétrica e de Computação - Unicamp, 2007.

[14] Harvey Fletcher e W. A. Munson, "Loudness, its definition, measurement and calculation," *Journal of the Acoustical Society of America*, 1933.

[15] Charles L. Lawson e Richard J. Hanson, *Solving least squares problems*, Prentice-Hall, 1974.

[16] MIDI Manufacturers Association, *MIDI 1.0 Detailed Specification*, MIDI Manufacturers Association, 1995.

[17] Antonio Mendes, *Arquitetura de Software*, Ed. Campus, 2002.

[18] Martin Russ, *Sound Synthesis and Sampling*, Elsevier, 2009.

[19] Keith Lent, "An efficient method for pitch shifting digitally sampled sounds," *Computer Music Journal*, 1989.

ROQUE, T. R.; MENDES, R. S. Síntese por amostragem de piano baseada em codificação wavelet.
In: CONGRESSO DE ENGENHARIA DE ÁUDIO AES BRASIL, 10., 2012, São Paulo, *Anais*... Rio de Janeiro: Sociedade de Engenharia de Áudio, 2012. p. 90-97.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

AUDIENCE para Pd: uma biblioteca para áudio espacial orientada a cena sonora

Regis Rossi Alves Faria1[1,2]
[1]Departamento de Música, FFCLRP, Universidade de São Paulo
Ribeirão Preto, SP, 14.040-900, Brasil
[2]Núcleo de Engenharia de Áudio e Codificação Sonora, Lab. Sistemas Integráveis da USP
São Paulo, SP, 05508-010, Brasil
regis@usp.br

RESUMO or *ABSTRACT*

AUDIENCE para Pd é uma biblioteca de funções para auralização, imersão e produção de cena sonora para a plataforma Pure Data (Pd). Com ela pode-se trabalhar de forma independente a composição da cena de som, a sua renderização acústica, a codificação de seu campo sonoro multicanal em um formato adequado e, finalmente, a sua audição em modos 2D/3D. A arquitetura do sistema é escalável, muitas cenas e fontes sonoras podem ser simultaneamente processadas, e é possível mover objetos nas cenas interativamente em tempo real. Este artigo apresenta a implementação da versão 2.0.3 do *software*, abordando a arquitetura do sistema e as premissas subjacentes. A estrutura da biblioteca é explicada, bem como a forma de usá-la, dando exemplos de aplicação. Desenvolvimentos em curso e as perspectivas futuras concluem o artigo.

0. INTRODUÇÃO

AUDIENCE para Pd (AUDIENCE4Pd) é uma biblioteca de funções para o programa Pure Data (Pd)[1] concebida na forma de objetos e abstrações que permitem construir aplicações genéricas de áudio espacial. Uma característica principal está na independência da concepção da cena sonora do modo de audição, uma vez que o usuário pode trabalhar uma experiência espacial desejada para a cena sonora sem se preocupar com qual será a configuração de alto-falantes usada para escutar. O campo sonoro final pode ser exportado e reproduzido em estéreo e em formatos de saída *surround* 2D e 3D.

Seu desenvolvimento iniciou-se em 2005 e, juntamente com um aparato de *hardware* construído para distribuição e amplificação de áudio multicanal, constitui o principal resultado do projeto homônimo AUDIENCE (*AUDio Immersion ExperieNce by Computer Emulation*).

Desde então resultados de vários trabalhos de desenvolvimento de aplicações, novas funções, conceitos e testes foram incorporados à biblioteca, chegando a atual versão 2.0.3, liberada em duas distintas distribuições: uma completa e restrita (AUDIENCE4Pd), e uma pública, básica e livre (OpenAUDIENCE).

Este artigo apresenta os conceitos de projeto e implementação da biblioteca, explicando como ela funciona, funções disponíveis, e dá exemplos de aplicações construídas com ela. A arquitetura subjacente ao sistema é brevemente revista e os componentes da versão atual da

[1] http://puredata.info

biblioteca são apresentados. Conclui-se com perspectivas acerca de novos desenvolvimentos em curso na direção de uma próxima geração do sistema.

1. ARQUITETURA DO SISTEMA

A arquitetura de processamento AUDIENCE é baseada no conceito de camadas funcionais desacopladas.

A primeira proposta da arquitetura data de 2005 [1] e desde então ela tem sido detalhada e consolidada em artigos posteriores [2-6]. Muitas aplicações práticas foram valiosas para se validar conceitos e distinguir características.

Esta seção cobre uma visão geral da arquitetura, explicando seu núcleo funcional e conceitos fundamentais, como o de camada funcional e interfaces.

1.1 Camadas funcionais

O núcleo da arquitetura está nas camadas funcionais, que podem ser vistas como aglomerados de funções relacionadas. Há quatro camadas funcionais principais nesta arquitetura, mandatórias em qualquer cadeia de produção-reprodução de áudio espacial. São elas:

- L1: camada de descrição e composição de cena, resolvendo a cena sonora (ou auditiva);
- L2: camada de simulação ou renderização do modelo acústico, usando algum método acústico;
- L3: camada de codificação de áudio espacial, empacotando dados de áudio (com ou sem metadados) em um formato específico para distribuição ou armazenamento;
- L4: camada de reprodução de áudio espacial, decodificando e gerando as saídas multicanal para alguma configuração de audição final (modo multicanal de alto-falantes).

Além das funções pertencentes a estas camadas nucleares da arquitetura há todavia funções adicionais que se encaixam numa categoria complementar, uma camada *auxiliar*. Estas implementam funcionalidades que, enquanto não diretamente conectadas às tarefas de processamento de áudio espacial, são requisitadas para ligar o núcleo a aplicações e dispositivos. Usualmente estas são dependentes da plataforma e não deveriam interferir na matéria processada pelas camadas principais. Estas funções incluem basicamente:

- *Interfaces de comunicação*, implementando mecanismos de comunicação e portas de entrada/saída (*I/O*) para a transmissão e recepção de dados de áudio e metadados de/para as camadas principais;
- *Interfaces e controle de usuário*, implementando os blocos de interfaces (ex: *graphical user interface* ou *GUIs*) que comandam o núcleo principal. Estes são responsáveis por admitir ações interativas e de controle do usuário/ouvinte e por enviar comandos às camadas principais. Entretanto, se um bloco de interface embute funcionalidades da camada L1, então ele obterá um *status* de L1.

A figura 1 mostra a hierarquia de camadas da arquitetura.

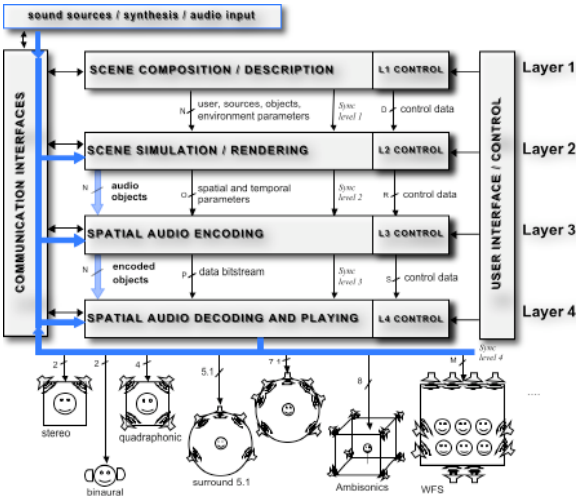


Figura 1 Arquitetura de referência do sistema AUDIENCE

Observe que a arquitetura propõe uma hierarquia de camadas onde L1 é o nível semântico mais alto e L4, o mais baixo. A taxa de dados de sincronização e de outros dados aumenta de L1 para L4. Observe também que há diferentes tipos de informação nas interfaces entre camadas, como metadados, dados de áudio, sinais de sincronização de dados e sinais de controle. A figura 1 sugere um fluxo de dados a partir da camada *i* para camada *i+1*. Em algumas circunstâncias uma cadeia de processamento pode alimentar outra cadeia e neste caso uma camada *i* pode alimentar uma camada *j*, em que $i, j \in \{1, 2, 3, 4\}$.

Uma seqüência prática engloba funções em diferentes estágios, e estes dados incluem naturalmente áudio, metadados, parâmetros de controle e sinais de sincronismo.

As camadas devem ser independentes uma da outra e ter um conjunto auto-contido e suficiente de regras de operação e algoritmos para controlar seu comportamento, por exemplo algoritmos para agir em caso de corte ou saturação de áudio, intervalos de dados inválidos, etc. Entradas provenientes de interfaces gráficas (*GUIs*) provêm controle para todas as camadas, fornecendo comandos específicos para controlar as funções e comportamentos em todas as camadas.

Funções em uma camada podem enviar mensagens, dados e instruções para as outras camadas, utilizando algum tipo de interface de comunicação, que é mostrado de uma forma generalizada na figura 1. Os dados (sinais) de áudio (indicados por setas cheias) podem fluir diretamente a partir de L2 para L3 e de L3 para L4, bem como de L4 para outras camadas (indicado pela realimentação de áudio de L4 para as interfaces de áudio).

A próxima seção apresenta a concepção e implementação da biblioteca de acordo com esta arquitetura.

2. PROJETO E IMPLEMENTAÇÃO DO SISTEMA

Esta seção apresenta as premissas de projeto de *software*, a história de desenvolvimento, a organização da biblioteca, suas características e objetos, e como ela foi concebida para trabalhar.

2.1 Premissas de projeto

Uma realização prática de um *software* para áudio espacial baseada na arquitetura AUDIENCE era vantajoso para ser desenvolvida sobre Pd devido a várias razões: *(a)* plataforma de programação flexível para prototipagem de funções de processamento de áudio sob a forma de blocos que podem ser ligados através de "cabos", *(b)* operação em tempo real, *(c)* portabilidade, e *(d)* acesso livre.

Um *software* de auralização executa várias funções que caberão adequadamente em uma das camadas da arquitetura. O requisito principal foi mapear todas as funções necessárias em cada camada para objetos individuais em Pd, tal que o fluxo de dados pudesse ser estabelecido ligando-os.

Para construir um objeto funcional do AUDIENCE em Pd alguns requisitos devem ser observados. Em primeiro lugar a definição da sua camada de destino. Cada objeto implementa uma funcionalidade em uma e apenas uma das camadas, isto é, cada objeto deve pertencer a uma camada válida entre as camadas principais 1, 2, 3 ou 4, ou à camada auxiliar. Para permitir uma identificação imediata da camada a que pertence todos os blocos são nomeados de acordo com a seguinte sintaxe:

audce + _ + *camada* + _ + *functionName*

Em segundo lugar, requer a definição do projeto de suas interfaces e de seu processo central: as funções que executa irão utilizar dados de entrada e/ou sinais de áudio para produzir dados de saída e/ou áudio. Os dados podem ser fornecidos de várias formas mas principalmente através de argumentos de criação ou passados em mensagens para o bloco. Os dados de saída podem ser canalizados para pontos de saída (*outlets*). A regra geral é que um bloco de camada *i* deve aceitar saídas de blocos de camada *i-1* e produzir sinais de entrada para a camada *i +1*.

O terceiro requisito é a sua conformidade com o conjunto de mensagens da camada: um objeto pode utilizar parâmetros definidos (por exemplo atributos de objetos da cena) que são utilizados ou endereçados por outros blocos no *patch*. Por exemplo, os blocos de L1 podem alterar a posição do usuário e precisam propagar esta nova posição usando uma sintaxe definida. AUDIENCE tem um mecanismo para atualizar os valores de dados usando *passagem de mensagens* que podem ser propagadas através de um mecanismo de envio/recebimento, o mais simples sendo uma conexão de dados da saída de um bloco para a entrada de outro bloco. Um bloco na camada *i* deve ser capaz de receber mensagens com dados para esta camada e interpretá-los. Normalmente as mensagens têm o nome dos parâmetros em si e são seguidas pelos seus novos valores.

A documentação que acompanha o *software* traz uma lista de mensagens implementadas que podem ser usadas para construir mecanismos de comunicação inter-objeto (por exemplo, usando conexões diretas com *inlets*/*outlets*, mensagens *send*/*receive* do Pd ou datagramas TCP/UDP).

Finalmente, a forma de implementação deve ser definida. Pode-se construir um bloco na forma de *patch* (abstração) utilizando objetos internos do Pd, ou construir um objeto (*external*), com uma função escrita em C/C++.

2.2 Histórico do desenvolvimento

O software AUDIENCE tem sido desenvolvido desde 2005 no Centro de Engenharia de Áudio e Codificação Sonora (NEAC) no Laboratório de Sistemas Integráveis da Escola Politécnica da Universidade de São Paulo (LSI-EPUSP) e foi usado em vários projetos científicos e artísticos desde então.

Os primeiros objetos foram projetados e construídos durante o projeto AUDIENCE[2]. Objetivando ferramentas de áudio imersivo flexíveis e escaláveis (*hardware* e *software*) para aplicações em realidade virtual completa (imersão sonora e visual), este projeto foi concluído em 2005 entregando um sistema de audição 3D octofônico para uma CAVE com uma máquina de auralização capaz de receber dados da cena a partir de aplicações de imersão visual e de renderizar cenas sonoras usando Ambisonics [7] para codificar campos sonoros 2D/3D e decodificá-los para um *rig* cúbico de alto-falantes. Diversos experimentos com cenas auditivas virtuais com instrumentos musicais e *rigs* planos para escuta também foram implementados, o que levou a ampliar o número de objetos de software e *patches* de aplicações específicas.

Esta evolução naturalmente abriu espaço para uma segunda rodada de investigações, uma vez que a arquitetura de auralização havia demonstrado potencial para resolver problemas em um âmbito mais geral relacionado à concepção de sistemas de som, e culminou com um segundo projeto de pesquisa: o OpenAUDIENCE. A partir de 2006 este projeto levou ao estabelecimento de diretrizes para estender a biblioteca através da criação de novos componentes abertos à comunidade, e levou à consolidação de uma versão livre do *software*, trazendo também recursos para a programação.

A biblioteca de *software* atual é um resultado direto desses projetos de desenvolvimento. Nas primeiras versões que trabalharam-se configurações de 2, 4, 6 e 8 alto-falantes, suporte de Ambisonics, e incorporou uma interface gráfica simples de L1 para a criação de cenas sonoras e seu controle. As impressões da percepção espacial foram bastante satisfatórias para mundos virtuais c/ campo-aberto e foi melhorado com a inclusão de processos de camada 2 (L2), tais como um algoritmo de simulação acústica fonte-imagem [3]. Progressivamente métodos e técnicas novas foram incorporadas e atualmente a versão completa trabalha com Ambisonics (até a 3ª ordem), PCM multicanal, MPEG-4 AAC (*Advanced Audio Coding*), MPEG *Surround*, e outros blocos estão desenvolvimento.

O OpenAUDIENCE (OA) é uma distribuição livre da biblioteca de *software*, baseado na versão completa. A principal diferença entre eles é a existência de objetos restritos (proprietários) como bibliotecas e códigos-fonte incluídos apenas na versão completa.

Enquanto a versão completa da biblioteca está disponível somente para a equipe de desenvolvedores e pesquisadores envolvidos em projetos de desenvolvimento, a versão OA pode ser baixada da internet gratuitamente a partir do *website* do projeto[3]. O desenvolvimento continuado de ambas as distribuições está aberto a colaboradores tanto de formações técnicas quanto artísticas dispostos a criar novas aplicações ou desenvolver novos objetos (funções).

[2] www.lsi.usp.br/interativos/neac/audience

[3] www.lsi.usp.br/audience/

2.3 Organização da biblioteca

O AUDIENCE4Pd é uma coleção de componentes funcionais na forma de objetos e abstrações (*patches*) para Pure Data (Pd). Um componente funcional nesta biblioteca é simplesmente um bloco de processamento construído como um objeto auto-contido de Pd ou um *patch*.

Os blocos são distribuídos em 4 camadas funcionais mais uma auxiliar funcionando no ambiente Pd. A figura 2 mostra a pilha das camadas de *software*.

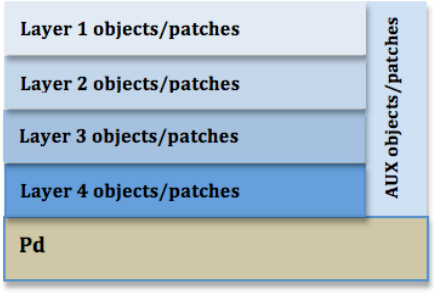


Figura 2 Pilha de camadas

A estrutura da biblioteca completa atual está organizada em 6 diretórios:

- *audce*: objetos e abstrações das camadas L1, L2, L3, L4, camada auxiliar, imagens usadas em interfaces, sons de teste, binários extras e bibliotecas de terceiros (necessários em algumas aplicaçoes)
- *app*: diretórios de aplicações e demonstrações
- *doc*: documentação
- *include*: arquivos para a compilação de objetos
- *src*: arquivos fontes
- *tst*: *patches* de testes

A distribuição completa AUDIENCE4Pd é destinada a desenvolvedores e licenciados, já que inclui todos os componentes. A distribuição OA exclui todos os componentes restritos (diretórios */include*, */src*, bibliotecas e binários proprietários).

Exemplos de aplicações básicas (inclusas no diretório */app*) são distribuídas nesta versão em pacote separado de maior tamanho, por conter arquivos de áudio. A documentação (em */doc*) também é disponibilizada separadamente do núcleo básico. Um núcleo executável mínimo precisa apenas dos componentes das camadas {L1, L2, L3, L4 e aux} para ser executado.

No diretório */tst* encontram-se *patches* de testes úteis para verificar o *status* de muitas funcionalidades. Por exemplo, o *patch test_audience_volume_patches.pd* resume todas as abstrações de volume disponíveis na versão atual, e o *patch test_audce_L4_objplay~.pd* testa o funcionamento do tocador *audce_L4_objplay~*.

Um arquivo *readme.txt* incluído resume em poucas palavras como funciona o *software* e informações sobre a estrutura de diretórios, as diretrizes de desenvolvimento, notas operacionais e de instalação, um guia de documentação, e uma lista de referência de versões (notas de versões). Todos os objetos externos são compilados para Win32 (.dll), Mac OSX (.Pd_darwin) e Linux (.Pd_linux).

2.4 Funcionalidades e componentes da biblioteca

A Tabela 1 apresenta os principais objetos funcionais e abstrações implementados na versão 2.0.3 do AUDIENCE4Pd. Para cada bloco é apresentado *(a)* a sua camada de grupo (L1, L2, L3, L4 ou auxiliar), *(b)* o nome do objeto ou abstração no Pd, e *(c)* uma breve descrição.

A distribuição completa versão 2.0.3 suporta uma interface gráfica (GUI) de cena sonora escalável e interativa, Ambisonics (até a 3ª ordem) e codecs MPEG-4 AAC, simulação acústica baseada no método de fonte-imagem [8], gravação e reprodução multicanal (incluindo 5.1) e pode decodificar para 11 configurações de alto-falantes diferentes.

Os objetos básicos descritos na tabela 1 processarão áudio e informações para uma fonte sonora em cena. Para facilitar o processamento de várias fontes ao mesmo tempo há abstrações que combinam vários objetos básicos e estendem a funcionalidade a um conjunto de N objetos sonoros (por exemplo 3, 5 ou 8 fontes). Por exemplo, para codificar e misturar 3 fontes em Ambisonics de 1ª ordem de uma só vez pode-se usar a abstração *audce_L3_3obj_amb1st~*. Abstrações de extensões estão listadas e disponíveis no diretório */audce* da distribuição.

Um número de objetos e abstrações são incluídos na camada auxiliar com recursos para as funções essenciais não pertencentes às 4 camadas principais, tais como a formatação de dados, interfaces gráficas de usuários, operações aritméticas e de processamento digital de sinais (DSP), de mistura, de entrada/saída, de temporização e escalonamento de dados, de sincronização e comutação de sinais. Eles são úteis para construir novas abstrações e *patches* aplicativos, e para definir e controlar parâmetros.

Objetos auxiliares incluem gravadores de arquivo de áudio multicanal, conversores de formato de dados (por exemplo, de tempo-para-amostras), conversores de formato de áudio (por exemplo, 6 sinais de áudio para arquivo 5.1), misturadores, operadores de convolução e abstrações para a construção de interfaces de controle para transporte, volume e no contagem de tempo.

Algumas abstrações alternativas implementam variações úteis no estilo de excursão da barra de um *slider* (por exemplo, logarítmica ou linear) e nas faixas de operação (por exemplo, min/max linear e valores de ganho dB). Por exemplo, o *audce_aux_objvol1~* é um controle de volume com VU com o ganho no intervalo [-100, +12] dB. As abstrações para controle de volume tem entradas para receber de fora comandos de liga/desliga e de configuração de nível de entrada.

Tabela 1 Principais objetos/abstrações disponíveis no AUDIENCE4Pd versão 2.0.3

Camada	Objeto/abstração	Descrição
L1	audce_L1_gui	GUI de cena para posicionamento de fontes sonoras (objetos sonoros em cena)
L2	audce_L2_allen	Simulador acústico (método fonte-imagem) para uma fonte, c/ saída de Respostas Impulsivas em B-Format (B-IR)

L2	audce_L2_RMgen	Gerador de Matriz de Renderização para *n* objetos sonoros com saída separada para cada um
L3	audce_L3_amb_3rd~	Codificador Ambisonics até 3a ordem
L3	audce_L3_aacenc	Codificador MPEG-4 AAC
L3	audce_L3_5.1render	Gerador de programa de áudio em modo 5.1 usando metadados produzidos por um gerador de Matriz de Renderização em L2
L3	audce_L3_ambirconv~	Operador de convolução de conjunto de Respostas Impulsivas Ambisonics (B-IR) p/ uma fonte sonora, produzindo sinais em B-Format
L4	audce_L4_amb_3rd~	Decodificador Ambisonics até 3a ordem
L4	audce_L4_aacdec	Decodificador MPEG-4 AAC
L4	audce_L4_amb_rot~	Objeto para rotacionar cena codificada em Ambisonics (campo B-Format) ao redor dos eixos de referência
L4	audce_L4_objplay~	Um tocador de arquivo de fonte sonora, interface c/ *file open* + transporte
L4	audce_L4_objplay1~	Um tocador de arquivo de fonte sonora, interface c/ *file open* + transporte + contador de tempo
L4	audce_L4_objplay2~	Um tocador de arquivo de fonte sonora, interface c/ *file open* + transporte + contador de tempo + controle de volume
Aux	audce_aux_convolver~	Objeto para convoluir fluxo de um sinal de áudio c/ Resposta Impulsiva em tempo real
Aux	audce_aux_counter	Abstração c/ contador de tempo em minutos e parciais
Aux	audce_aux_dbvol~	Pequeno *patch* para controle de volume em dB
Aux	audce_aux_mixer~	Misturador para n sinais (canais) c/ configuração de nível individual via mensagem
Aux	audce_aux_recordmc~	Abstração GUI para gravar um arquivo sonoro multicanal c/ configuração de no. de canais (até 6) e taxa de amostragem
Aux	audce_aux_playmcfile~	Abstração GUI para reproduzir um arquivo sonoro multicanal c/ configuração de no. de canais (até 6) e taxa de amostragem
Aux	audce_aux_move	Abstração experimental para automatizar o movimento de um objeto sonoro em cena
Aux	audce_aux_objtime	*Patch* para controlar o tempo de um objeto sonoro em cena no *patch* principal
Aux	audce_aux_vol~	*Patch* para controlar o volume of de uma linha de áudio
Aux	audce_aux_objvol~	*Patch* para controlar o volume de uma linha de áudio com medidor VU

Além disso, há um número de mensagens padrão usados para transmitir comandos e para definir os parâmetros dos blocos em um *patch*. A simplicidade é uma regra na definição do estilo de sintaxe, como neste exemplo: "*src_pos n x y z*" é uma mensagem para posicionar o objeto sonoro *n* nas coordenadas cartesianas *{x, y, z}* da cena, e da mesma forma "*rcv_pos x y z*" posiciona o ouvinte na cena.

A maioria dos objetos têm *patches de ajuda* (*help files*) que contêm informações sobre os argumentos de criação, os dados retidos pelo objeto (por exemplo, a posição de uma fonte sonora no espaço, a ordem do codificador Ambisonics, etc.), exemplos de uso, os sinais de interface de entrada/saída (ou seja, os sinais válidos que são aceitos nos *inlets* e sinais disponíveis nos *outlets*), *bugs* conhecidos e limitações. Eles também incluem um histórico de versão (notas) e ponteiros para objetos afins ("*see also*").

2.5 Licenças

AUDIENCE e OA são liberados sob uma série de licenças diferentes que se aplicam com base no uso dos objetos. A maioria dos arquivos objetos/abstrações .pd e/ou arquivos de ajuda incluem notas de versão e informações sobre licenças aplicáveis, já que individualmente usam diferentes recursos disponíveis na versão estendida *Pd-extended*, em objetos proprietários e/ou de terceiros.

Objetos nativos do AUDIENCE são licenciados sob BSD, e onde bibliotecas de terceiros ou *software* GPL são usados os seus respectivos termos são aplicados e reconhecidos. Objetos de terceiros podem ser licenciados sob GPL ou adotar outro esquema. Deve-se consultar as condições de licenciamento de arquivos individuais, bem como o manual que acompanha, com mais informações sobre este assunto. Alguns objetos – como por exemplo os operadores de convolução – usam a biblioteca GPL fftw3 de Matteo Frigo, e codecs MPEG-4 AAC usados em L3 e L4 AAC utilizam a biblioteca aacPlus© licenciada pela Coding Technologies (mais tarde Dolby Laboratories).

2.6 Criando novos objetos

Criando novos *externals* para a biblioteca vai exigir um ambiente de desenvolvimento C/C++ e acesso aos arquivos fontes, de cabeçalhos e *makefile*. Os arquivos fontes dos objetos e o *makefile* estão no diretório */src*, e os arquivos de cabeçalhos, dentro de */include*. O roteiro de *makefile* inclui instruções de compilação e construção de objetos para uso em todos os sistemas operacionais suportados.

A documentação inclusa na distribuição tem uma seção para instruir desenvolvedores interessados em construir objetos novos para o AUDIENCE e testá-los.

3. FUNCIONAMENTO DA BIBLIOTECA

Uma aplicação de áudio espacial prática é construída criando-se fluxos de dados instanciando e conectando objetos. A Figura 3 mostra o exemplo de um típico fluxo de dados de cima abaixo para ilustrar a lógica do processamento.

Ela mostra um *patch* usando objetos "genéricos" por camada gerando o campo de som combinado de 3 fontes de som irradiando em uma cena, e gerando uma audição em modo 5.1 *surround*. Na camada 1 ao alto há um analisador de dados de cena que filtra os dados para os objetos de processamento subsequentes na camada 2. Simuladores acústicos na camada 2 podem produzir, por exemplo, respostas impulsivas em B-Format e transmiti-las para os codificadores espaço-temporais na camada 3. Estes codificadores também recebem o fluxo de áudio de cada

fonte sonora (src1, src2 e src3) e geram os seus campos sonoros individuais como seriam ouvidos em uma posição determinada na cena (a posição do receptor ou ouvinte).

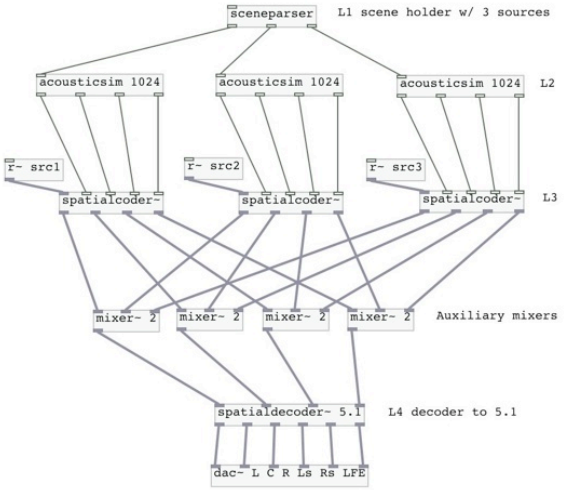


Figura 3 Fluxo de dados genérico de cima abaixo

Como estamos interessados no som da cena completa, os 3 conjuntos individuais de dados de campo sonoro são misturados (um misturador para cada canal do campo sonoro) e o resultado é o programa de saída de distribuição (por exemplo, em B-Format).

O último componente, o decodificador de camada 4 (L4), pode ser local ou remoto, e será capaz de descodificar a *mídia* de distribuição para um modo específico de saída do alto-falantes. Esta operação reconstrói o campo sonoro da cena, agnosticamente a partir do que foi feito nas outras camadas.

Muitos fluxos de processamento similares podem ser criados em paralelo e sua operação em tempo real pode ser estabelecida. Havendo um fluxo de dados ativo, processos em tempo real são executados continuamente, exatamente como processos acústicos ocorrem na natureza. Por exemplo, enquanto uma câmara de reverberação existe ela sempre processará todos os sons presentes. Da mesma forma, os processos podem ser terminados simplesmente desmontando a cadeia, desconectando os objetos ou destruindo-os. Este processo intuitivo de criação e destruição de aplicações é um recurso poderoso e uma característica típica do Pd e de outras plataformas gráficas de programação semelhantes orientadas a objeto.

4. INSTALAÇÃO E USO

Esta seção cobre a instalação e como usar a biblioteca, com exemplos de construção de aplicações.

4.1 Instalando a biblioteca

A partir dos arquivos de instalação, replique a estrutura de diretórios */audce* em seu sistema de arquivos, e insira esta direção no caminho de procura do Pd, para que seus componentes sejam localizados. Arquivos de ajuda individuais de objetos/abstrações e informações para solução de problemas no manual podem ser consultados em caso de problemas. A maioria dos objetos irá exibir mensagens de erro no console se não criados.

4.2 Construindo aplicações

AUDIENCE4Pd oferece recursos para construir uma ampla gama de aplicações de áudio espacial, como por exemplo gravadores e reprodutores de campo sonoro, simuladores acústicos, geradores, processadores e mixadores de som multicanal, codificadores de cena sonora 2D/3D (por exemplo, usando Ambisonics até a 3 ª ordem), e reprodutores de som espacial (auralizadores) com até 11 configurações de saída possíveis de alto-falantes. Há ainda recursos para convolução de respostas impulsivas com fontes sonoras secas, e para estabelecer fluxos de áudio e de metadados de/para outras aplicações (por exemplo, de computação gráfica ou de realidade virtual).

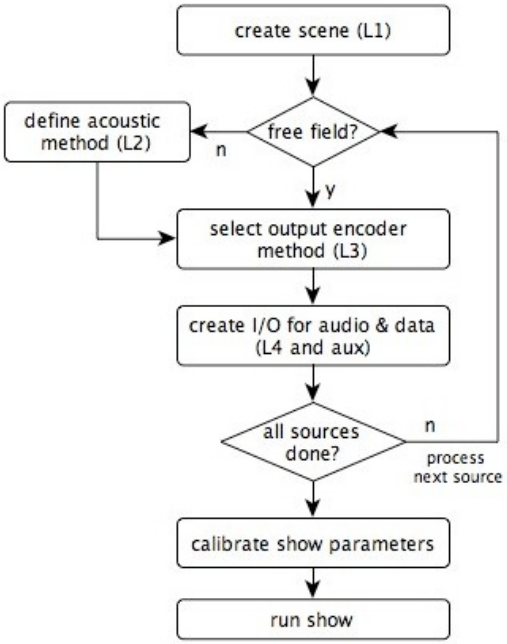


Figura 4 Fluxograma básico de uso

O fluxograma da figura 4 resume as tarefas básicas para criar aplicações usando o AUDIENCE4Pd. Um fluxo de típico de projeto envolve a criação de uma cena sonora (um objeto de L1), a configuração de um sistema de renderização acústico (um objeto de L2), a codificação do som da cena renderizada (um objeto de L3) e, em seguida, a decodificação (para um modo de saída específico) e reprodução da cena sonora através de uma plataforma de alto-falantes (um objeto de L4). Funções auxiliares, como os transportes, controles de mistura e volume estão disponíveis a partir de objetos na camada auxiliar.

Logo que uma novo *patch* de Pd é criado o primeiro passo é *instanciar um objeto da cena L1* (por exemplo, uma interface *GUI* L1) com a geometria e número de fontes sonoras desejadas. Alternativamente, pode-se ter um aplicativo externo para enviar dados de cena para uma abstração de L1 que irá analisar e distribuir as mensagens L1 daí por diante, como descrito em [2].

Se uma simulação acústica é desejada, *selecione um processador de L2* adequado (por exemplo, um simulador de fonte-imagem de Allen e Berkley [8]) que irá processar o cenário de uma fonte sonora irradiante e produzir um conjunto de parâmetros (por exemplo, uma resposta ao impulso) para dirigir um método de codificação de L3.

Para a espacialização em campo livre em vez disso parta diretamente para a seleção de um codificador de L3. No AUDIENCE a seleção padrão de codificador espaço-temporal usa Ambisonics [7], e para uma cena de *n* fontes serão precisos *n* pares de objetos L2-L3, um por fonte.

Os processos em L3 consomem áudio e metadados para gerar a mídia multicanal codificada, portanto a próxima etapa é estabelecer as ligações de entrada para áudio e dados (por exemplo, um objeto L4 reprodutor de arquivo de som ou um objeto *adc~*) e as ligações de saída para monitores de L4, utilizando adicionalmente quaisquer objetos auxiliares necessários para configurá-los (por exemplo, controladores de volume, contadores de tempo e comandos de transporte).

O processo acima deve ser repetido para todas as fontes de som na cena. Depois disso, os sons codificados em L3 de cada fonte devem ser misturados produzindo um sinal composto que corresponde a toda a cena processada (por exemplo, linearmente combinando os 4 canais em B-Format da fonte 1 com os da fonte 2).

A tarefa final básica é calibrar os parâmetros do *show*, isto é, configurar todos os parâmetros relevantes das fontes sonoras e os seus comportamentos (ex: posição de origem e seu volume calibrados a uma determinada distância a partir do ouvinte). Se os parâmetros do *show* mudam deterministicamente durante uma dada audição, talvez seja útil usar uma abstração de automação para guiá-los (ex: o *audce_aux_move*, que automatiza trajetórias simples para mover fontes de som no tempo).

Assim que todos os laços estejam conectados o *patch* inteiro se comportará como um ambiente acústico virtual que contém as fontes de som irradiando. A diferença é que o usuário pode definir o formato e o modo de saída de som, ou seja, o conjunto de sinais de saída para cada alto-falante de uma configuração de audição.

4.3 Exemplos de aplicações

AUDIENCE foi usado em 2007 para espacializar a peça "The Unanswered Question" de Charles Ives (1906) com 9 instrumentos (um quarteto de cordas, um quarteto de sopros e um trompete) [6]. No mesmo ano, sua arquitetura foi proposta como um quadro de referência para o novo codec MPEG SAOC (*Spatial Audio Object Coding*) [9].

Em 2009 a peça eletro-acústica "Kitchen Miniatures", de Fernando-Lopez Lezcano, com 8 fontes sonoras, foi espacializada durante o simpósio SBCM 2009 em Recife[4]. AUDIENCE está presentemente instalado para prover auralização em um sistema de realidade virtual imersivo do tipo CAVE na Universidade de São Paulo[5], e tem ainda sido utilizado como máquina de áudio de referência para aplicações de música digital reconfigurável, aplicações de projeto e avaliação de paisagens sonoras, e mixagem orientada a cena sonora.

A figura 5 mostra um exemplo de *patch* de aplicação com uma cena que contém 2 objetos de diferentes fontes de áudio: uma flauta (a partir de um arquivo de som) e uma voz (a partir de um porta *adc~* em tempo real). A figura ilustra a flexibilidade na renderização do campo sonoro independente para cada fonte, com a voz usando o simulador acústico de fonte-imagem L2 *audce_L2_allen* e a flauta sendo processada em campo aberto. Com base no método da fonte-imagem proposto por Allen (1979) [8], o objeto L2 *audce_L2_allen* calcula a propagação de raios a partir de uma fonte de som para o receptor e todas as reflexões obtidas numa sala em forma de caixa em um intervalo de tempo dado, produzindo um conjunto de respostas impulsivas em Ambisonics B-Format [3].

Em L3 ambas as fontes são codificadas em Ambisonics (1ª ordem) e misturadas para compor um único campo sonoro. No entanto, dois diferentes modos de saída são usados: um estéreo (por exemplo, para simples monitoramento) e um quadrifônico.

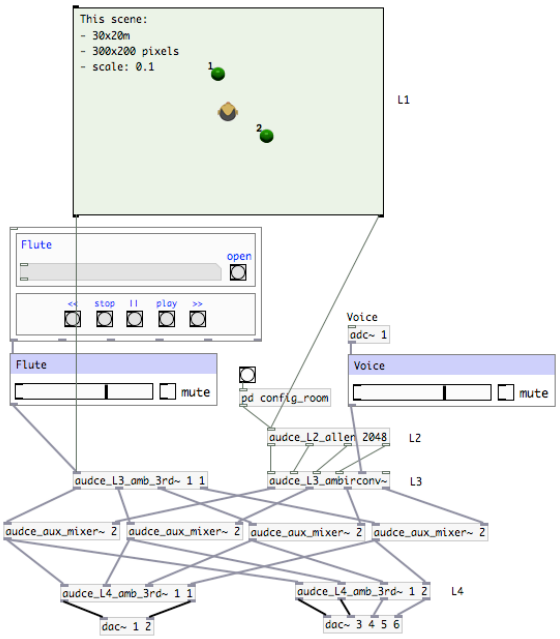


Figura 5 *Patch* processando uma cena com 2 fontes (uma com processamento L2 à direita), codificando em Ambisonics e ouvido através de 2 diferentes modos (2.0 e 4.0)

5. CONCLUSÕES

Este trabalho apresentou um exemplo de um *framework* para Pd voltado para a construção de aplicações de áudio espacial em geral empregando a arquitetura AUDIENCE, suas funções básicas e utilização. O sistema caminha para a sua 3ª geração, com melhorias na usabilidade e qualidade de som, ampliando o número de funções de processamento espacial, e objetivando expandir sua disponibilidade para novas aplicações e grupos de usuários considerando, por exemplo, a sua inclusão na distribuição *Pd-extended*.

Um novo desenvolvimento começou no âmbito do projeto MÓBILE (www.eca.usp.br/mobile/) para suportar a codificação e decodificação por *Wave Field Synthesis (WFS)*. Além disso, um simulador acústico L2 com suporte a traçado de raios e radiosidade, um gerenciador de volume (*loudness manager*) e uma interface de *timeline* orientada a fontes sonoras estão atualmente em desenvolvimento.

Um *modelo de objeto universal* para a arquitetura do sistema é um futuro objeto a ser desenvolvido, de modo que a criação de novos blocos fique mais fácil e a incorporação de um algoritmo externo ou a integração a um *plug-in* externo seja feita mais rápido, como num esquema *plug-and-play*.

[4] SBCM 2009, http://compmus.ime.usp.br/sbcm/2009/

[5] Caverna Digital, http://www.lsi.usp.br/interativos/nrv/

Há também um número de objetos e abstrações experimentais em desenvolvimento. Em codecs de áudio espacial, os objetos L3 *MPSenc* e L4 *MPSdec* são um par experimental de codificador e decodificador ISO/IEC 23003-1 (*MPS - MPEG Surround*) baseados no *software* de referência publicado em 2007. Estes, no entanto, são ainda desenvolvimentos experimentais parciais desta tecnologia.

Outra abstração experimental em desenvolvimento é o *audce_L4_sceneplay*, que visa estender os objetos reprodutores de L4 a todas as fontes sonoras em uma determinada cena.

Informações mais detalhadas sobre os recursos avançados da arquitetura e aplicações específicas estão previstas para artigos futuros.

Agradecimentos

Este software tem tido contribuições de muitos pesquisadores ao longo do caminho. Gostaríamos de agradecer especialmente aos pesquisadores Leandro F. Thomaz e Renan Vital, ao Laboratório de Sistema Integráveis da Escola Politécnica da Universidade de São Paulo (LSI-EP-USP) por sediar os projetos, e a empresa Coding Technologies (posteriormente Dolby Laboratories) pelo suporte com codecs MPEG-4 AAC. Agradecemos também o apoio da agência brasileira FAPESP através do processo 2008/08632-8.

6. REFERÊNCIAS BIBLIOGRÁFICAS

[1] R. R. A. Faria. *Auralização em ambientes audiovisuais imersivos*. Tese de doutorado, Engenharia de Sistemas Eletrônicos, Escola Politécnica da Universidade de São Paulo, 2005.

[2] R. R. A. Faria, L. F. Thomaz, L. Soares, B. T. Santos, M. K. Zuffo e J. A. Zuffo. *AUDIENCE – Audio Immersion Experiences in the Caverna Digital*. In: Anais do SBCM 2005 – 10º Simpósio Brasileiro de Computação Musical, 2005, Belo Horizonte, p.106-117.

[3] R. R. A. Faria e J. A. Zuffo. *An auralization engine adapting a 3D image source acoustic model to an Ambisonics coder for immersive virtual reality*. In: Proceedings of the AES 28th International Conference, 2006, Piteå, Suécia, p.157-166.

[4] L. F. Thomaz, R. R. A. Faria, M. K. Zuffo e J. A. Zuffo. *Orchestra spatialization using the AUDIENCE engine*. In: Proceedings of ICMC 2006 – International Computer Music Conference, 2006, New Orleans.

[5] R. R. A. Faria, M. K. Zuffo e J. A. Zuffo. *Improving spatial perception through sound field simulation in VR*. In: Proceedings of VECIMS 2005 – IEEE Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2005, Giardini Naxos, Itália.

[6] L. F. Thomaz. *Aplicação à música de um sistema de espacialização sonora baseado em Ambisonics*. Dissertação de mestrado, Engenharia de Sistemas Eletrônicos, Escola Politécnica da Universidade de São Paulo, 2007.

[7] M. Gerzon. *Periphony: with-height sound reproduction*. Journal of the Audio Engineering Society, v.21, n.1, 1973. p.2-10.

[8] J. B. Allen, D. A. Berkley. *Image method for efficiently simulating small-room acoustics*. Journal of the Acoustical Society of America, v.65, n.4, p.943-950, Abril, 1979.

[9] R. R. A. Faria. ISO/IEC JTC1/SC29/WG11 (MPEG). Document m14753. *New reference architecture for broad-scope spatial audio object coding and spatial sound frameworks*. Lausanne, Suíça, Jul. 2007. /MPEG, 81st MPEG & 42nd JPEG Meeting/

SESSÃO 3 / *SESSION 3 - PAPERS 3*

Foto: SXC

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Comparação entre métodos espectrais para melhoria de sinais de voz sob ruído branco

Irineu Antunes Jr.[1] e Claudio J. Bordin Jr.[1]

[1] Universidade Federal do ABC, Av. dos Estados, 5001, Santo André - SP, 09.210-580.

(irineu.antunes; claudio.bordin)@ufabc.edu.br

RESUMO

Analisam-se, neste trabalho, métodos para redução de ruído branco em sinais de voz. Os métodos analisados modificam componentes da transformada de Fourier de tempo curto do sinal, apenas em magnitude (SoftSoft) ou apenas em fase (PSC). Tais modificações visam atenuar componentes ruidosas, sem idealmente eliminar coeficientes importantes do sinal de voz. Verifica-se por simulações computacionais que os parâmetros ajustáveis desses métodos são robustos, identificando-se valores fixos que incrementam os valores de PESQ ("Perceptual Evaluation of Speech Quality") sob uma ampla faixa de níveis de ruído. Constata-se, também, que o desempenho médio do método SoftSoft é superior ao de outros métodos espectrais convencionais.

0 INTRODUÇÃO

Diversos métodos para redução de ruído em sinais de voz no domínio da transformada de Fourier de tempo curto (STFT, *Short-Time Fourier Transform*) foram desenvolvidos [1]. Neste artigo, consideram-se métodos monocanais aplicados a trechos de sinal de voz corrompidos por ruído aditivo gaussiano e branco (AWGN, *Additive White-Gaussian Noise*).

Neste cenário, o estimador ótimo linear no sentido de mínimos quadrados é descrito pelo filtro de Wiener [2], que pode ser implementado no domínio transformado atenuando-se independentemente cada componente da STFT segundo a sua razão sinal-ruído [1, Sec.11.1.3]. A técnica de subtração espectral [3], que consiste em subtrair do espectro corrompido uma versão aproximada do espectro do ruído, pode ser interpretada de maneira semelhante, também atenuando as componentes conforme a sua razão sinal-ruído.

Mais recentemente, surgiram métodos alternativos que efetuam a modificação não linear da magnitude das componentes espectrais. Muitos desses métodos modificam os coeficientes através de funções de limiar (*thresholding functions*), concebidas originalmente para suavização de imagens no domínio *wavelet*. A função de limiar Soft e suas generalizações (SoftSoft[4]) alcançam bons resultados para sinais de voz, com a vantagem de exibirem baixa complexidade computacional,

sendo assim adequados à implementação em dispositivos de baixo custo.

Em geral, métodos baseados na STFT alteram apenas a magnitude dos coeficientes, preservando a fase das componentes ruidosas na reconstrução da saída. Como sinais de voz são relativamente insensíveis a alterações de fase [5], esse procedimento não adiciona pertubações significativas.

Ainda assim, a informação de fase pode ser usada para se obter uma estimativa menos ruidosa. O recente método de Stark *et al.* [6], *Phase Spectrum Compensation* (PSC), propõe que se modifique apenas a fase das componentes da STFT do sinal ruidoso, conservando-se a suas magnitudes.

Neste trabalho, comparam-se duas abordagens recentes para redução de ruído em sinais de voz no domínio da STFT: o método SoftSoft, que altera apenas a magnitude dos coeficientes da STFT, e o método PSC, que modifica apenas a fase destes. Para essas comparações, realizaram-se simulações Monte Carlo nas quais se avaliou o desempenho de ambas abordagens por meio da metodologia PESQ [7], que fornece uma medida objetiva da qualidade percebida por um ouvinte.

Utilizando a metodologia PESQ, foi possível fixar valores para os parâmetros de ambos os métodos. Mais especificamente, através de simulações, pôde-se constatar que tais valores levam a resultados robustos, relativamente independentes das realizações particulares de sinal e ruído.

Oferece-se ainda uma comparação experimental dos métodos com abordagens tradicionais, como a subtração espectral, estimadores MMSE [8] e logMMSE [9]. Pôde-se constatar que os dois métodos analisados superam em desempenho esquemas tradicionais, com destaque para o SoftSoft.

O texto a seguir está organizado da seguinte forma: na Seção 1, descrevem-se os métodos PSC e SoftSoft; na Seção 2, por sua vez, apresentam-se os resultados dos experimentos realizados; ao final, na Seção 3, são oferecidas conclusões e comentários.

1 Descrição dos Métodos Empregados

Considerem-se amostras do sinal de voz ruidoso $x[n] = v[n] + d[n]$, em que $v[n]$ é o sinal de voz sem o ruído aditivo $d[n]$. Todo o sinal a ser processado $\{x[n]\}_{n=0}^{N-1}$ é dividido em quadros de comprimento L, com sobreposição de $S < L$ amostras. Inicialmente, cada quadro é multiplicado por uma janela de análise $w[n]$ e são calculados os coeficientes da STFT discreta

$$X_m[n,k] = \sum_{l=-\infty}^{\infty} x_m[l]\, w[n-l]\, e^{-j2\pi kl/N}, \quad (1)$$

em que x_m é o quadro de amostras, j é a unidade complexa, k é o índice da frequência da FFT (*Fast Fourier Transform*) de comprimento $N_{\text{FFT}} = 2^{\lceil \log_2(2L) \rceil}$, maior ou igual ao dobro de L, em que $\lceil \cdot \rceil$ denota a operação de arredondamento em direção a mais infinito. Emprega-se a janela de Hanning modificada segundo Griffin e Lim [10]

$$w[n] = \frac{2\sqrt{S/L}}{\sqrt{4a^2+2b^2}} \left[a + b\cos\left(\frac{2\pi n}{L} + \phi\right)\right], \quad (2)$$

para $n = 0, 1, \ldots, L-1$ e nula fora deste intervalo, sendo as constantes $a = 0{,}5$, $b = -0{,}5$ e $\phi = \pi/L$.

Em seguida, pela modificação de coeficientes em cada quadro $X_m[n,k]$, segundo o método SoftSoft ou PSC, obtém-se uma estimativa de coeficientes $Y_m[n,k]$ do sinal original.

No caso em que L é múltiplo de quatro vezes S, a janela de Hanning modificada obedece à seguinte propriedade:

$$\sum_{l=-\infty}^{\infty} w^2[lS-n] = 1, \quad (3)$$

permitindo, finalmente, que se empregue o procedimento padrão de sobreposição e soma dos quadros (*overlap-add*) para reconstruir a saída completa

$$y[n] = \frac{\sum_{m=-\infty}^{\infty} \text{Re}\{y_m[n]\}}{\sum_{m'=-\infty}^{\infty} w[m'S-n]}, \quad (4)$$

sendo o quadro reconstruído $y_m[n] = STFT^{-1}\{Y_m[n,k]\}$ e a STFT discreta inversa dada por $\frac{1}{N}\sum_{k=-\infty}^{\infty} Y_m[n,k]\, e^{j2\pi k/N}$.

Observe-se que a janela modificada é praticamente nula nas extremidades para valores de L elevados, contribuindo para reduzir as descontinuidades que possam surgir pelo processamento em blocos.

1.1 Modificação da Magnitude Espectral

A função SoftSoft [4] [11] pode ser usada para modificar individualmente a magnitude de cada um dos coeficientes ruidosos, $|\mathcal{X}|$, provendo a seguinte estimativa para o coeficiente de saída

$$|\mathcal{Y}| = g_{t_2}(|\mathcal{X}|) + (|\mathcal{X}| - g_{t_1}(|\mathcal{X}|)), \quad (5)$$

em que $g_t(|\mathcal{X}|) = (|\mathcal{X}| - t)\,\mathrm{u}(|\mathcal{X}| - t)$ é a função Soft [13] com limiar t, $\mathrm{u}(\cdot)$ a função degrau unitário (nula para argumentos negativos). Deve-se observar que a função SoftSoft depende dos parâmetros t_1 e t_2 denominados, respectivamente, de limiar inferior e limiar superior. Além disso, vale notar que a função Soft é um caso particular da SoftSoft com limiar inferior nulo.

Como é usual nos métodos que modificam a magnitude, para cada coeficiente ruidoso de magnitude $|\mathcal{X}| = |X_m[n,k]|$, forma-se um coeficiente estimado combinando-se a magnitude $|\mathcal{Y}|$ com a fase ruidosa $\angle X_m[n,k]$, ou seja, a estimativa $Y_m[n,k] = |\mathcal{Y}|\angle X_m[n,k]$ é usada para reconstruir a saída.

A estratégia dos métodos de limiar, como os que empregam a função Soft, consiste em anular os coeficientes de pequena magnitude que em grande parte são devidos ao ruído, especialmente no caso de sinais ditos "suaves", *i.e.*, que podem ser bem representados por poucos coeficientes. Como a voz não se enquadra nesta classe de sinais [4], foi proposta a função especializada SoftSoft.

Para coeficientes de grande magnitude, a função SoftSoft opera como a Soft, atenuando levemente esses coeficientes. Já para os coeficientes pequenos, a SoftSoft deixa passar intactos os coeficientes com magnitude abaixo do valor do limiar inferior. Deste modo, consegue-se preservar todos os coeficientes importantes para representar o sinal de voz, ao mesmo tempo em que a faixa de magnitudes que é mais afetada pelo ruído, notadamente entre os dois limiares, sofre maior atenuação.

1.2 Modificação de Fase das Componentes Espectrais

Os métodos apresentados anteriormente modificam a magnitude dos coeficientes espectrais. Em alguns desses métodos, como o de subtração espectral, a atenuação é função da razão sinal-ruído [1, Sec.11.1.3]. Em outros, como os de função de limiar, a atenuação é não-linear, pois depende da magnitude dos coeficientes.

Possivelmente inspirados pela estratégia desses métodos, os autores do PSC também observaram que as componentes espectrais de baixa energia contêm maior contribuição do ruído e, portanto, devem sofrer maior atenuação.

O método PSC fundamenta-se em obter um coeficiente estimado

$$Y_m\left[n,k\right] = \left|X_m\left[n,k\right]\right| \angle X_m^{\Lambda}\left[n,k\right], \qquad (6)$$

no qual se combina a magnitude ruidosa $\left|X_m\left[n,k\right]\right|$ com uma fase modificada $\angle X_m^{\Lambda}\left[n,k\right]$, sendo $X_m^{\Lambda}\left[n,k\right] = X_m\left[n,k\right] + \Lambda\left[k\right]$. Nesta última, emprega-se a função de modificação de fase

$$\Lambda\left[k\right] = \begin{cases} +\lambda\sigma, & 0 < k/N_{\text{FFT}} < 1/2 \\ -\lambda\sigma, & 1/2 < k/N_{\text{FFT}} < 1 \\ 0, & \text{caso contrário} \end{cases}, \qquad (7)$$

que depende do parâmetro λ multiplicado por uma estimativa do valor eficaz do ruído, σ.

Como o sinal processado é real, os coeficientes ocorrem em pares complexos conjugados[1]. Segundo a modificação da Equação (7), a fase de um deles é deslocada em direção de $-\pi$ rad e a fase do outro é deslocada na direção oposta (0 rad), destruindo a simetria hermitiana. Essas componentes tendem a se cancelar uma vez que, para se obter uma saída real, deve-se tomar a parte real da transformada inversa durante o processo de reconstrução da saída.

[1]Exceto, para possíveis coeficientes reais em $k = 0$ e $k = N/2$ no caso de N par.

Deve-se notar que a atenuação de cada componente assim reconstruída depende da magnitude e da fase do par complexo conjugado frente aos valores $\pm\lambda\sigma$. Também é oportuno notar que o coeficiente modificado, Equação (6), em geral, não apresenta simetria conjugada. Deste modo, ao se considerar apenas a parte real da transformada inversa, acaba-se realizando uma modificação de magnitude.

Disto tudo conclui-se que o PSC realiza de fato alteração de magnitude espectral, ainda que exclusivamente pela modificação de fase dos coeficientes de entrada. De maneira semelhante aos métodos que operam na modificação de magnitude espectral, é esperado que coeficientes de grande magnitude sofram pouca atenuação, enquanto que os de baixa magnitude sejam mais atenuados.

2 Avaliação Experimental dos Métodos

2.1 Base de dados de voz e medidas de desempenho usadas

Foram tomados dois conjuntos de sinais de voz do banco de dados Timit [12]: a) um conjunto de treinamento, com 5 locutores, usado para ajustar os parâmetros dos métodos (Seção 2.2); b) um conjunto de teste, com 4 locutores, posteriormente, usado para avaliar o desempenho (Seção 2.3).

Os sinais de voz escolhidos possuem a duração aproximada de 2,5 s e se encontram amostrados à frequência de 16 kHz. Tomou-se o cuidado de empregar frases diferentes, com número semelhante de vozes masculinas e femininas.

O sinal de entrada ($x\left[n\right]$) foi obtido pela adição de ruído ($d\left[n\right]$, AWGN) considerando-se uma ampla faixa de valores de razão sinal-ruído (SNR), definida como

$$\text{SNR} = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1}\left|v\left(n\right)\right|^2}{\sum_{n=0}^{N-1}\left|v\left(n\right)-x\left(n\right)\right|^2}\right). \qquad (8)$$

Para cada valor de SNR, adicionaram-se 25 realizações distintas de ruído ao sinal de voz correspondente a cada frase, perfazendo um total de 125 sinais de entrada ruidosos para treinamento e 100 para a etapa de testes.

O método PSC foi implementado como descrito pelos autores em [6], utilizando quadros de 32 ms com sobreposição de 4 ms, processados pela STFT discreta com a janela de Hanning modificada, Equação (2). O método SoftSoft foi implementado da mesma forma a fim de permitir a comparação de desempenho.

Para avaliar de maneira objetiva o desempenho perceptual dos métodos, adotou-se a medida PESQ, que prevê as notas que seriam atribuídas por ouvintes ao julgar a qualidade de sinais de voz num teste de opinião MOS (*Mean Opinion Score*). Os valores da medida PESQ variam desde $-0{,}5$ (muito ruim) até 4,5 (sem distorção). No caso dos métodos de redução de ruído, é importante também comparar a diferença da medida

PESQ de saída em relação a de entrada para avaliar a eficácia da redução de ruído.

2.2 Determinação dos valores dos parâmetros

Em geral, valores ótimos de parâmetros de estimadores dependem da distribuição dos coeficientes do sinal de voz. Supondo-se, simplificadamente, que a maior parte da energia do sinal se encontra representada em coeficientes de grande magnitude, torna-se possível obter uma estimativa de parâmetros "universais", ou seja, que se aplicam a toda a classe de sinais de voz e que, grosso modo, dependam apenas do ruído.

Para sinais contaminados por AWGN, o método Soft possui um limiar universal [13] frequentemente adotado, dado por $\bar{\sigma}\sqrt{2\ln N_{\mathrm{B}}}$, em que $\bar{\sigma}$ é o desvio-padrão do ruído e N_{B} o comprimento de uma transformada real, normalmente *wavelet*. Em trechos apenas com ruído, este valor de limiar garante que a probabilidade de algum coeficiente ultrapassar o limiar é pequena, proporcionando saída nula. Já nos trechos em que a energia do sinal é elevada, a magnitude dos coeficientes deve ultrapassar facilmente o limiar, garantindo sua participação na saída.

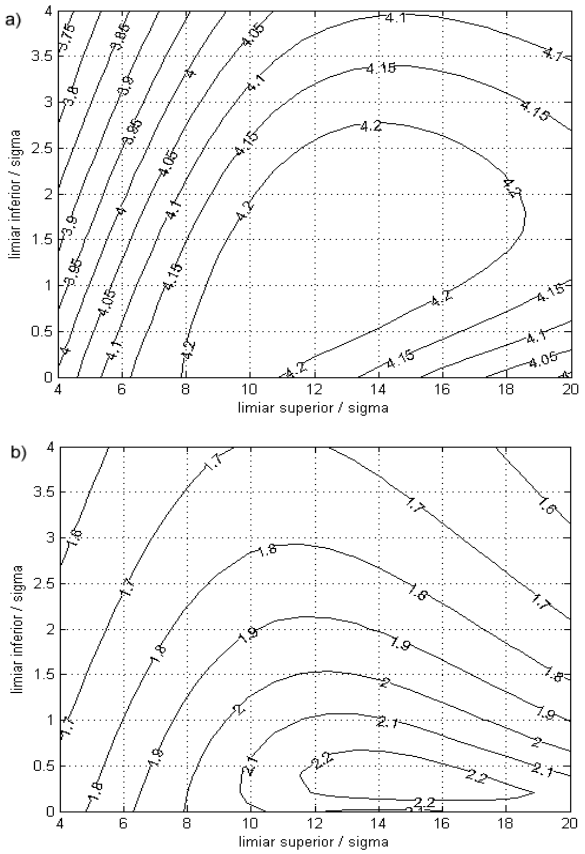


Figura 1: Curvas de nível médias da superfície PESQ em função dos limiares da função SoftSoft para entrada com a) SNR=30 dB e b) 0 dB, para o conjunto de sinais de treinamento.

De maneira análoga, para o método PSC é empregado um parâmetro universal $\lambda = 3{,}74$ sugerido por Stark *et. al.*, obtido empiricamente em [6]. Este parâmetro, multiplicado por σ, representa o quanto a fase deve ser modificada.

Tendo-se em consideração o sucesso de se empregar um valor universal de parâmetro para o PSC e para o Soft, procurou-se, neste trabalho, encontrar valores para os limiares SoftSoft como esses, dados por $t_1 = \bar{t}_1\sigma$ e $t_2 = \bar{t}_2\sigma$, sendo $\bar{t}_1$ e $\bar{t}_2$ constantes determinadas experimentalmente e que, na média de várias realizações de ruído e de sinais de voz, melhorem a qualidade da saída para entradas dentro de uma ampla faixa de valores de SNR.

Para estimar limiares SoftSoft universais utilizando o conjunto de sinais de entrada de treinamento, estimou-se a superfície PESQ média em função dos limiares $\bar{t}_1$ e $\bar{t}_2$ para valores extremos de SNR: 30 dB na Figura 1.a; 0 dB na Figura 1.b. Como se pode observar nestas figuras, a escolha $\bar{t}_1 = 0{,}5$ e $\bar{t}_2 = 12$ oferece melhoria de qualidade tanto para SNR baixa (situação de maior interesse prático) quanto para SNR elevada.

2.3 Avaliação de desempenho: método *SoftSoft versus PSC*

A Tabela 1.a apresenta os valores médios de PESQ sobre o conjunto de teste para os métodos SoftSoft e PSC. Para permitir comparações, a Tabela 1.b reproduz os valores obtidos em [6] para outros métodos sob as mesmas condições, como se pode constatar pela semelhança dos valores de PESQ para as entradas ruidosas em ambas as tabelas.

Tabela 1: Valores médios de PESQ $\times$ SNR: a) para a entrada ruidosa e métodos estudados neste trabalho: SoftSoft com limiares $\bar{t}_1 = 0{,}5$ e $\bar{t}_2 = 12$; PSC com parâmetro $\lambda = 3{,}74$; b) resultados obtidos por Stark *et al.* [6]: SSUB, Subtração Espectral [Schwartz e Makhoul *apud* 6]; MMSE [8]; logMMSE [9].

Métodos	Desempenho						
SNR:	0	5	10	15	20	25	30
a)	Valores de PESQ						
Entrada	1,48	1,83	2,24	2,66	3,03	3,38	3,73
SoftSoft	2,24	2,66	3,04	3,43	3,77	4,02	4,20
PSC	2,05	2,51	2,91	3,29	3,65	3,95	4,17
b)	Valores de PESQ [6, Tab.1]						
Entrada	1,55	1,90	2,26	2,62	2,97	3,32	3,65
PSC	2,06	2,47	2,87	3,23	3,56	3,88	4,14
SSUB	1,78	2,28	2,73	3,18	3,58	3,89	4,12
MMSE	2,03	2,40	2,74	3,05	3,33	3,61	3,87
logMMSE	2,12	2,54	2,88	3,18	3,45	3,71	3,96

Os resultados da Tabela 1 indicam que, em média, o método SoftSoft exibe um desempenho superior ao dos demais algoritmos. Esses resultados são evidenciados pelos gráficos da Figura 2.b, nos quais são mostradas as

médias dos incrementos de PESQ entre o sinal de entrada e o de saída. Tais resultados são semelhantes aos obtidos em [6], como se pode visualizar na Figura 2.a. Além disso, para os métodos SoftSoft e PSC, o incremento médio de PESQ apresentado na Figura 2.b permaneceu razoavelmente estável para uma ampla faixa de SNR, apontando a robustez dos parâmetros escolhidos.

A Figura 3 contém a representação no tempo e no domínio tempo-frequência de uma gravação de voz (a), dela corrompida por ruído com $\mathrm{SNR} = 10\,\mathrm{dB}$ (b), tratada pelo método PSC (c) e pelo SoftSoft (d). Pode-se observar que ambos os métodos preservam razoavelmente características do espectrama do sinal de voz (comparar a Figura 3.a, gravação original, com Figuras 3.c, PSC, e 3.d, SoftSoft). Além disso, o espectrograma obtido pelo método SoftSoft exibe menos ruído que o observado para o método PSC (ver Figuras 3.c e 3.d), como os valores de PESQ da saída atestam: 2,89 no SoftSoft contra 2,87 no PSC. Observe-se que isto ocorre a despeito do que indicam os valores de SNR da saída: 12,40 dB no SoftSoft contra 14,38 dB no PSC, mostrando que a SNR global não representa bem a qualidade perceptual de sinais de voz. Ademais, também deve-se ter em conta que os parâmetros não foram ajustados para otimizar a SNR.

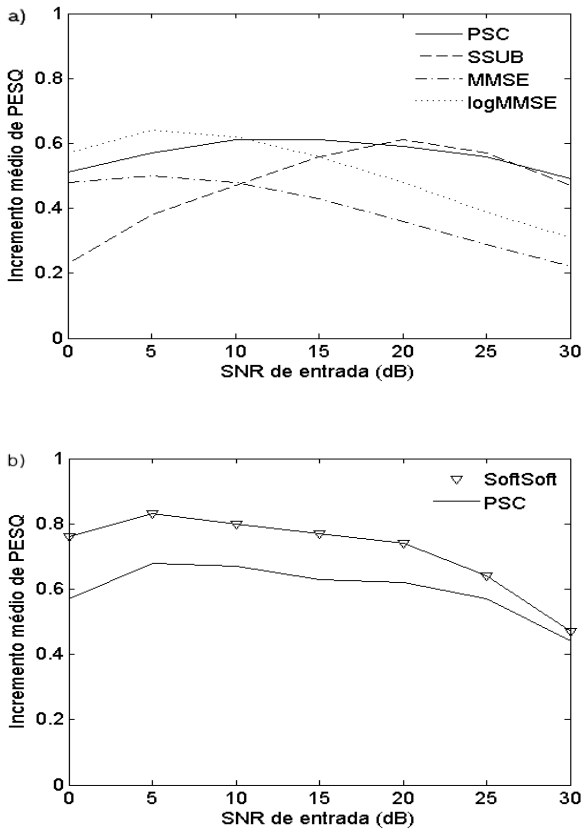


Figura 2: Incremento nos valores médios de PESQ: a) resultados obtidos por Stark *et al.* [6, Fig. 5]; b) métodos estudados neste trabalho.

Da mesma forma que observado por Stark *et al.* [6], o espectrograma da saída do PSC (Figura 3.c) revela um ruído de fundo parecido com o ruído branco e com menos artefatos característicos que os métodos de subtração espectral. O mesmo é verificado para o caso SoftSoft (Figura 3.c), porém com um ruído de fundo bem menor.

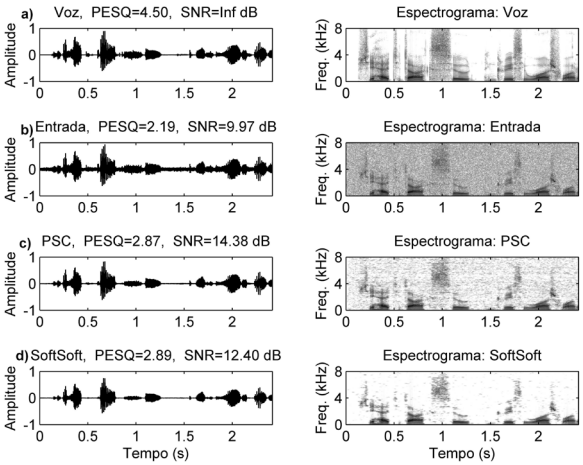


Figura 3: Representação no tempo e no domínio tempo-frequência dos sinais: a) Voz (frase "she had your dark suit in greasy wash water all year", locutor masculino); b) Entrada ruidosa; c) Resultado do PSC; d) Resultado do SoftSoft.

3 Conclusões e Perspectivas

Neste trabalho, compararam-se dois métodos de redução de ruído no domínio STFT, tendo-se alcançado resultados favoráveis na melhoria de sinais de voz contaminados por ruído branco. Nas simulações, o método PSC forneceu os mesmos resultados obtidos por seus autores em [6], tendo-se mostrado como uma abordagem promissora, mas que ainda pode ser aprimorada a partir de um estudo teórico mais rigoroso que ofereça uma justificativa formal para este método. Já o método SoftSoft, que pode ser justificado por argumentos estatísticos [4] ou variacionais [11], proporcionou uma qualidade perceptual superior à obtida pelos demais métodos aqui avaliados. Deve-se salientar que o método SoftSoft, na forma aqui apresentada, não é adequado para o caso de ruído não branco, necessitando adaptações semelhantes àquelas implementadas para o PSC em [6]. Por exemplo, pode-se substituir o valor constante σ por estimativas obtidas pela densidade do ruído no domínio da STFT. Deste modo, os limiares SoftSoft poderão se ajustar ao nível de ruído em cada porção do espectro.

Ou fato importante verificado nas simulações foi a possibilidade de se fixar o valor dos parâmetros nos métodos SoftSoft e PSC, evidenciando a robustez dos mesmos para uma ampla faixa de níveis de ruído. Apesar disto, é esperado que a adaptação dos

parâmetros forneça resultados superiores, o que incentiva a adoção de parâmetros variáveis. Numa possível implementação prática dos métodos, esses parâmetros poderiam ser determinados empiricamente, de modo a se ajustarem ao nível de ruído e às características dos sinais.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] S. V. Vasegui, "Advanced Signal Processing and Noise Reduction", *John Wiley & Sons*, 2nd edition, 2000.

[2] N. Wiener, The Extrapolation, Interpolation, and Smoothing of Stationary Time Series With Engineering Applications. New York:Wiley, 1949.

[3] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. Acoust. Speech, Signal Proc.*, vol. ASSP-27, pp. 113-120, Apr. 1979.

[4] I. Antunes Jr., and P. M. S. Burt, "Speech Denoising by SoftSoft Thresholding", *Proc. ISIE, IEEE International Symposium on Industrial Electronics*, v.1., pp. 532-536, Montreal, 2006.

[5] Wang, D. L., and J. S. Lim. "The unimportance of phase in speech enhancement," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 30, pp. 679-681, Aug. 1982.

[6] A.P. Stark, K.K. Wójcicki, J.G. Lyons, K.K. Paliwal, and K.K. Paliwal, "Noise driven short-time phase spectrum compensation procedure for speech enhancement", in Proc. Interspeech, 2008, pp.549-552.

[7] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evalua-tion of Speech Quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'01), Salt Lake City, UT, 2001, vol. 2, pp. 749–752.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process. , vol. ASSP-32(6), pp. 1109–1121, Dec 1984.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process. , vol. ASSP-33(2), pp. 443–445, 1985.

[10] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[11] I. Antunes Jr., "Redução de Ruído em Sinais de Voz Usando Funções de Limiar SoftSoft para Minimizar a Distorção Log-Espectral", Congresso de Engenharia de Áudio da AES-Brasil, 2011, São Paulo - SP.

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. L. Dahlgren, "The darpa timit acoustic-phonetic continuous speech corpus cdrom", Cdrom, NIST, Gaithersburg, MD, 1996.

[13] D. L. Donoho, I. M. Johnstone, "Threshold selection for wavelet shrinkage of noisy data", *Proc. of the 16th Annual International Conf. of the IEEE Eng. in Medicine and Biology Society*, pp. A24-A25, Nov. 1994.

ANTUNES JUNIOR, I.; BORDIN JUNIOR, C. J. Comparação entre métodos espectrais para melhoria de sinais de voz sob ruído branco.
In: CONGRESSO DE ENGENHARIA DE ÁUDIO AES BRASIL, 10., 2012, São Paulo, *Anais*... Rio de Janeiro: Sociedade de Engenharia de Áudio, 2012. p. 107-112.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Marca d'água em áudio: um estudo voltado ao direito de propriedade intelectual

Ana Lúcia Martins de Jesus[1], Dorirley Rodrigo Alves[1]
[1]Pontifícia Universidade Católica de Minas Gerais
Programa de Pós-graduação em Engenharia Elétrica
Belo Horizonte, Minas Gerais, 30170-131, Brasil

almjesus@sga.pucminas.br, dorirley@gmail.com

RESUMO
As facilidades de distribuição e manipulação de conteúdo proporcionadas pela Internet colocam em questão as dificuldades em preservar os direitos autorais de uma obra. A marca d'água digital é definida como potencial ferramenta a ser aplicada nesta área. Este trabalho apresenta a avaliação de uma técnica de inserção de marca d'água em áudio sob a perspectiva da preservação do direito de propriedade intelectual de arquivos de áudio no formato wave. Os resultados obtidos informam a viabilidade da técnica para este tipo de aplicação.

0. INTRODUÇÃO

Ao longo das últimas décadas presenciou-se um crescimento acelerado da Internet, sobretudo após a criação da WWW (World Wide Web) pelo físico Tim Bernes-Lee e integração da mesma ao domínio público e comercial [20]. Diversas vantagens foram acrescentadas com a popularização da grande rede e dentre elas, está a facilidade de copiar, modificar e distribuir conteúdos digitais [16]. Se por um lado existem vantagens, por outro, existem desvantagens. A maioria delas está relacionada aos direitos de propriedade intelectual e confiabilidade do conteúdo [25]. A possibilidade de fazer cópias ilimitadas do conteúdo sem perda de fidelidade e a facilidade de modificação envolvem discussões sobre como preservar os direitos do autor sobre sua obra, uma vez que o número de usuários que podem ter acesso a este conteúdo é grande. Neste contexto, torna-se necessária a busca de alternativas que possibilitem o reconhecimento da autoria, a fim de evitar que outras pessoas se apropriem de obras intelectuais indevidamente.

A Marca d'água digital é vista como uma tecnologia potencial para proteção de direitos autorais [7]. Constitui-se de um código inserido em um arquivo que contém informações a respeito da origem, estado, e/ou destino dos dados[16]. Bender et al [2] aponta que os dados inseridos devem ser codificados diretamente no corpo da mídia, ao invés de um cabeçalho ou invólucro para que possam permanecer intactos em diferentes formatos de arquivo. Além disso, o processo de inserção deve causar o mínimo de alteração possível [2] para que não comprometa a qualidade do arquivo. Especificamente para o áudio, esse é

um desafio, pois, embora seja imperfeito, o sistema auditivo humano ainda é o mais sensível, se comparado aos outros sentidos [11].

Este estudo trata da aplicação de marca d'água em arquivos de áudio no formato wave. Será selecionada a técnica de inserção de marca d'água em áudio mais relatada na literatura para a realização de um estudo voltado à preservação de direitos de propriedade intelectual de uma obra. O método adotado consiste em cinco etapas:

- Identificar a técnica de inserção de marca d'água mais referenciada na literatura
- Determinar os requisitos para utilização da marca em direitos autorais
- Estabelecer os critérios a serem adotados para realizar a avaliação
- Testar a técnica selecionada com base nos requisitos e critérios levantados
- Analisar e apresentar os resultados

O trabalho está organizado da seguinte forma: A seção1 faz uma abordagem das diversas publicações na área de marca d'água em áudio que foram de grande relevância para a realização do trabalho; a seção 2 apresenta os conceitos em direito autoral e marca d'água. A execução do método é relatada na seção 3 e na seção 4 foram descritos os testes realizados. Os resultados encontrados são apresentados na seção 5 e por fim, a seção 6 traz a conclusão do trabalho e sugestões de trabalhos futuros.

1. REVISÃO DA LITERATURA

Apesar de ser uma área ainda pouco relatada na literatura o número de publicações sobre marca d'água em áudio vem crescendo nos últimos 15 anos [23][6]. No ano de 1998 foram publicados 103 trabalhos internacionais relacionados à área, ao passo que o ano de 1993 rendeu apenas duas publicações [6]. Esse crescimento se deve principalmente à realização do Primeiro Workshop internacional sobre o assunto que aconteceu em Berlim, na Alemanha no ano de 1996[1]. A partir deste ano, a ocultação de dados tem se tornado área de grande interesse no meio acadêmico.

Diversos esquemas de marca d'água em áudio têm sido propostos tais como Espalhamento de Espectro (Spread Spectrum), Ocultação de Eco (Echo Hiding), Modificação de Amplitude (Amplitude Modification), Modificação de Escala de Tempo (Time-scale Modification), LSB (least-significant bit), dentre outros. Trabalhos como aqueles apresentados em [15], [2] e [11] fizeram uma abordagem geral destas técnicas, no formato de estado da arte. Rodriguez, Miyatake e Meana [19] e Marnani, Karami e Molavianjazi [14] focaram na avaliação de desempenho dos esquemas. Outras publicações como, por exemplo, a tese de mestrado de Nunes [16] propuseram novas abordagens de técnicas já existentes. O que se tem percebido é que a maior parte das obras encontradas têm considerado apenas o desempenho individual dos algoritmos, sob diversas perspectivas. Logo, torna-se necessário um estudo voltado especificamente para comprovação de direitos autorais.

Dos trabalhos comparativos entre diferentes esquemas de marca d'água destaca-se [19]. Os autores avaliaram três esquemas de marca d'água em áudio: Ocultação de Eco (Echo Hiding), Miscelânia no Domínio da freqüência (Patchwork) e Domínio do Cepstro (Cepstrum Domain). Com o objetivo de verificar a resistência destas técnicas ao processamento comum de sinais, este trabalho está entre as poucas pesquisas comparativas em marca d'água em áudio [14]. Por meio deste estudo, os autores concluíram que os esquemas Miscelânea e Domínio do Cepstro resistem melhor a ataques de processamento de sinal que a técnica de Ocultação de Eco. Como foi dito em [11], esta técnica é mais vulnerável a ataques como processamento de sinal, embora possua uma boa imperceptibilidade, passando despercebida ao ouvido humano mais facilmente.

O estudo apresentado em [19] foi tratado em [14]. Neste último, os autores avaliaram os mesmos esquemas abordados em [19], adotando outra forma de avaliação, levando em conta o fenômeno de mascaramento no modelo psicoacústico. Deste modo, concluiu-se que a técnica de Ocultação de eco além de possuir imperceptibilidade aceitável, possui resistência a modificação de sinal, ou seja, robustez contra ataques. As outras técnicas mostraram resultados semelhantes aos apresentados em [19].

2. REFERENCIAL TEÓRICO

2.1 Direito autoral

Segundo Sola [22] o direito autoral é definido como "[...] a proteção dos interesses do autor e seus sucessores, em relação às suas obras criadas, sejam elas ou não literária, artística e científica". A proteção de propriedade intelectual tem por objetivo proteger as criações originais de reproduções ou alterações que venham a ferir os direitos de seus autores [8]. Portanto, constitui a forma de garantir que seus créditos serão preservados, além de garantir ao autor o direito de autorizar ou proibir o uso de sua obra e obter a devida retribuição.

A legislação assegura proteção de direitos autorais por meio da lei nº 9.610 de 19 de Fevereiro de 1998, e pelo Código Civil instituído em 2002. Segundo a lei nº 9.610 a proteção dos direitos autorais independe se o proprietário registrou ou não trabalho. Ainda que não houvesse registro em órgão competente, a legislação oferece plena garantia de que os direitos sobre a criação serão preservados.

Sob a ótica jurídica, a propriedade intelectual é considerada um direito duplo: moral e patrimonial. O primeiro garante ao autor a possibilidade de reivindicar a autoria da obra, garantindo que ao ser utilizada, seu nome será referenciado. Garante também o ineditismo e a integridade da obra, além de poder retirá-la de circulação quando isso implicar em afronta à sua reputação ou honra (BRASIL, 1998 – art. 24). O direito moral é inalienável e irrenunciável, ao contrário do direito patrimonial. Este último caracteriza-se por garantir ao autor a possibilidade de explorar economicamente a obra, exercendo atividades de promoção e arrecadando proventos que compensem seu esforço. Isso também pode ser feito por um terceiro, desde seja autorizado pelo titular, sendo esse direito estendido também à possibilidade de autorizar ou não a exploração de suas obras publicamente [8].

[1] *Information Hiding: First international Workshop.* Este workshop fazia parte de um programa de pesquisa sobre Segurança na Informática, Criptografia e Teoria da Codificação e foi realizado no Instituto Isaac Newton em 1996 (ANDERSON; PETICOLAS, 1999).

2.2 Marca d'água digital

A marca d'água digital, juntamente com a Esteganografia e a Criptografia faz parte de uma área de pesquisa conhecida como Ocultação de Informação ou Information Hiding. Como exposto anteriormente, a marca d'água constitui um código que é inserido em uma mídia [16]. Esse código pode ser utilizado para inserir informações sobre o autor e sua origem diretamente no corpo do arquivo, criando uma ligação entre o proprietário da obra e sua criação [6]. Havendo referência do autor na obra, as disputas por propriedade podem ser resolvidas mais facilmente, uma vez que o mesmo deixará seus dados registrados a fim de facilitar a comprovação da origem do arquivo.

2.2.1. Classificação de esquemas de marca d'água

Existem diversos modos de inserir marca d'água em áudio. Para todo processo de inserção, existe uma forma de recuperar a informação inserida no sinal. Para que isso ocorra, o detector pode ou não exigir a presença do arquivo original para realizar a extração da informação [16]. Esquemas de detecção não cega definem sistemas de marca d'água que necessitam da presença do sinal original para que ocorra a detecção da informação inserida. Por outro lado, esquemas de detecção cega não requerem a presença do áudio original para detectar a marca d'água [6]. Neste contexto, esquemas de detecção não cega também são interessantes para aplicações relacionadas a direitos autorais [11]. De posse do arquivo original, a verificação do áudio marcado se torna mais eficiente em relação aos métodos de detecção cega. Isso ocorre pelo fato de que o arquivo marcado pode sofrer alterações que podem ser causadas inclusive pelo próprio método de marcação, podendo comprometer a fidelidade da marca d'água [17]. Isso poderia impossibilitar a detecção da informação. Por outro lado, esquemas de detecção não cega requerem dupla capacidade de armazenamento e dupla banda de comunicação, ao contrário dos esquemas de detecção cega [11].

3. DESENVOLVIMENTO

Esta seção aborda o método descrito na Introdução. Descreve o levantamento da técnica de marcação mais referenciada na literatura, a busca dos requisitos para utilização de marca d'água na proteção de direitos autorais e a forma de avaliação do esquema de marca d'água.

3.1 Levantamento da técnica de marca d'água mais referenciada

Dada a variedade de técnicas de inserção de marca d'água propostas desde o 1º Workshop internacional em Berlim em 1996, foi realizado um estudo a fim de verificar sua popularidade na literatura. Para isso, utilizou-se a ferramenta de busca Engineering Village onde foram pesquisados termos ligados às técnicas de marca d'água em áudio digital. A Tabela 1 mostra os resultados da busca realizada. Observe que a técnica Espalhamento de Espectro (Spread Spectrum) obteve o maior número de documentos retornados, mostrando-se como a técnica mais referenciada, segundo esta ferramenta.

Com base neste estudo adotou-se a técnica "Espalhamento de Espectro" neste trabalho. Não somente pelas buscas realizadas no site *Engineering Village*, mas também por sua popularidade, uma vez que foi referenciada por um grande número de autores como a técnica mais promissora encontrada na literatura [3], [12], [11], [23].

Técnica	Documentos retornados
Spread-Spectrum Method	73
Synchronization Code	61
Echo Hiding	37
Watermark Shaping	28
Time-Scale Modification	26
Amplitude Modification	22
Patchwork	20
Pseudo-Random Sequence	12
Outras técnicas	12

Tabela 1 - Número de publicações relacionadas às principais técnicas de inserção de marca d'água em áudio.
Fonte: dados da pesquisa

3.1.1. Espalhamento de espectro

A Segunda Guerra Mundial trouxe consigo um forte desenvolvimento tecnológico quando a necessidade de privacidade na troca de informações possibilitou o desenvolvimento de tecnologias cada vez sofisticadas. O Espalhamento de espectro é uma técnica de modulação de sinal utilizada de maneira geral em sistemas de comunicação, radares e navegação, dentre outras aplicações [20]. Neste tipo de sistema, a banda do espectro utilizada na comunicação é bem maior que a necessária para transmitir a informação. Essa característica faz com que o sinal que contém a informação sofra uma redução na intensidade ao ser incluído no sinal portador, dificultando sua detecção por terceiros. A segurança do sistema é maximizada com a utilização de códigos pseudo-aleatórios durante a incorporação do sinal sendo que, para recuperar a mensagem é necessário que se tenha conhecimento desse código de espalhamento. Durante a Segunda Guerra Mundial comprovou-se que as características desta técnica a tornam aplicável à utilização em processos esteganográficos e marcação de sinais [21].

A técnica de Espalhamento de Espectro está entre as mais promissoras técnicas de inserção de marca d'água encontradas na literatura [3], [12], [11], [23]. Consiste basicamente em disseminar dados através do áudio [9], onde o fluxo de informação é espalhado pela maior quantidade do espectro de frequência possível [2]. Diferentemente da técnica de modificação do bit menos significativo (LSB) (para mais detalhes consulte [2]) que insere a informação no áudio por meio da modificação dos bits de menor importância, esta técnica incorpora informação nos bits mais significativos do espectro de frequência; fato que lhe confere robustez. Uma característica interessante desta técnica, é que ela utiliza o próprio áudio como regra para inserir a informação, uma vez que as características do próprio arquivo influenciam diretamente nos locais de ocultação da mensagem. Isso permite que a detecção por elementos não-autorizados se torne cada vez mais difícil [1]. A ideia é que a alteração no áudio não afete a marca, a menos que haja comprometimento significativo do sinal [5].

A Figura 1 apresenta o modelo geral para a técnica de Espalhamento de Espectro. A sequência de ruído com a informação a ser inserida é espalhada no sinal no domínio do tempo ou no domínio da transformada, não importa qual

transformada é utilizada. Normalmente são utilizadas a DCT (Discrete Cosine trasnform), DFT (Discrete Fourrier transform) e a DWT (Discrete wavelet transform) (KIM, 2004). Em (α) tem-se uma mensagem binária a ser incorporada ou b = {+1, -1} a variável bipolar equivalente a esta mensagem. Esta última assume valores +1 ou -1 de acordo com o bit a ser transmitido pela marca d'água *w*. Se o bit da mensagem a ser transmitida tiver valor 1, a variável bipolar *b* assume valor 1, se for 0 *b* assume valor -1. A mensagem ou sua variável bipolar são moduladas por uma seqüência pseudo-aleatória (PRN) *u* que é gerada por meio de uma chave secreta *K,* na figura visualizada em (β). Esta modulação dá origem ao sinal de marca d'água *w* pela seguinte fórmula: $w(n) = bu(n)$ [11]. Em (γ) o sinal hospedeiro representado na figura por *x,* logo depois de passar para o domínio da transformada utilizada, é acrescido da marca d'água modulada de modo que a sequência *u* é adicionada ou subtraída de *x* de acordo com o valor da variável *b* dependendo da informação a ser incorporada [5]. O sinal *s* constitui o sinal acrescido da marca *w* e segundo [11] é dado pela seguinte função: $s(n) = x(n) + \alpha w(n)$. O valor de *α* na equação representa a amplitude da marca e controla a relação entre robustez e inaudibilidade. Em (δ), o vetor *y* é o sinal original no domínio da transformada depois de passar pelo canal de distorção.

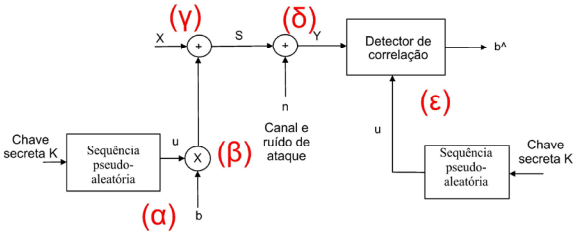


Figura 1 - modelo geral de um sistema de espalhamento de espectro
Fonte: adaptado de [5]

(ε)A detecção da marca é feita por meio do cálculo da correlação entre o sinal marcado *s* e a seqüência pseudo-aleatória *u*, podendo a mesma ser gerada e recuperada por meio de uma chave secreta [11]. A equação 1 mostra o cálculo da correlação, onde N representa o tamanho do sinal e *α* a amplitude da marca d'água.

$$c = \frac{1}{N}\sum_{i=1}^{N} s(i)u(i) + \frac{1}{N}\sum_{i=1}^{N} \alpha b u^2(i) \qquad (1)$$

3.2 Requisitos de marca d'água para proteção de direitos autorais

Conforme apontado por [16], a eficiência de um sistema de marca d'água é medida por meio de três parâmetros: robustez, taxa de inserção e fidelidade. A robustez refere-se à capacidade de sobrevivência da marca d'água a diversos tipos de ataques, intencionais ou não. Um método é dito robusto se após passar pelos diversos ataques, ainda é possível recuperar a informação contida na marca [20]. A taxa de inserção representa a razão entre o número de bits inseridos e o número de amostras, caracterizando a carga de dados no arquivo. Esta quantidade pode variar de acordo com o método de inserção devido às particularidades de tratamento do arquivo. Um fator de diferença é a utilização ou não de redundância [2], que tem relação direta com a robustez da técnica de inserção. Quanto mais redundante, mais resistente a modificações se torna a marca, o que evita que sua remoção ocorra ilegalmente. E por fim, a fidelidade representa o quanto o arquivo original foi alterado em virtude da inserção da marca, o que pode ser verificado pela quantidade de ruído acrescentada após o processo. De acordo com Nunes [16], a magnitude do sinal de marca d'água deve ser menor que a do sinal hospedeiro, a fim de evitar a introdução de distorção.

3.3 Critérios de avaliação dos requisitos de marca d'água

Os requisitos citados possuem formas particulares de avaliação. A *robustez* é avaliada por meio da promoção de ataques ao sinal de áudio marcado [16]. Após esses ataques, verifica-se a possibilidade de detecção da marca ou o quanto restou dela.

A imperceptibilidade é avaliada através da verificação da quantidade de ruído que foi inserida no arquivo após a incorporação da marca. Para isso, foi utilizada a ferramenta PEAQ que utiliza as variáveis de um modelo psico-acústico para gerar uma única nota denominada ODG (Objetctive difference grade) ou grau de diferença objetiva. Por meio desta nota é possível classificar a quantidade de ruído inserida em uma escala que vai de "muito irritante" a "Imperceptível" com notas variando de -4 a 0 respectivamente [1]. Para a realização da avaliação foi estabelecida uma escala de valores variando de "Ruim" a "Ótimo" de acordo com o resultado obtido em cada arquivo marcado. Com essa classificação, é possível verificar o nível de atendimento ao requisito. Os valores ODG e a classificação juntamente com os respectivos pesos são apresentados na Tabela 2

Valor	Significado	Critérios detecção	Faixa de valores (%)	Peso
-4	Muito irritante	Ruim	0 – 20	1
-3	Irritante	Regular	21 – 40	2
-2	Ligeiramente irritante	Bom	41 - 60	3
-1	Perceptível mas não irritante	Muito bom	61 - 80	4
0	Imperceptível	Ótimo	81 - 100	5

Tabela 2 – Faixas de valores ODG e classificação utilizada no trabalho.
Fonte: adaptado de [10]

Por fim, para a avaliação da carga de dados foi adotada a especificação da RIAA[2], que estabelece uma carga útil de dados na marca d'água que deve ser de pelo menos 108 bits para que possa utilizada na proteção de direitos autorais.

3.4 Definição dos níveis de classificação

Para verificar a conformidade da técnica com os requisitos levantados utilizou-se uma adaptação do esquema de avaliação proposto em [13]. Neste estudo as autoras descrevem a auditoria da qualidade de um software de contabilidade considerando as características e subcaracterísticas para softwares preconizadas na NBR ISO/IEC 9126 e na NBR ISO/IEC 13596. Por meio da

[2] Recording Industry Association of America

atribuição de pesos às características e subcaracterísticas é possível chegar a uma nota única, que reflete o nível de qualidade final do sistema.

Para a realização dos testes de robustez os arquivos de áudio foram submetidos aos ataques e posteriormente, procedeu-se ao cálculo do percentual de detecção da marca d'água em cada arquivo analisado. Com base no número de caracteres inseridos verificou-se o percentual de caracteres recuperados corretamente após cada ataque. Para avaliar o nível de resistência da marca a cada ataque, os valores percentuais foram classificados em uma escala, onde cada nível de classificação recebeu um peso de acordo com a faixa de valores que representa. Com isso, é possível classificar os resultados para verificar o nível de resistência da marca a cada ataque. Os conceitos adotados estão numa escala que vai de "Ruim" a "Ótimo" de acordo com o percentual de caracteres recuperados corretamente após os ataques ao áudio marcado. Veja detalhes apresentados na tabela 3.

Critérios detecção	Faixa de valores (%)	Peso
Ótimo	81 - 100	5
Muito bom	61 - 80	4
Bom	41 - 60	3
Regular	21 - 40	2
Ruim	0 - 20	1

Tabela 3 – Níveis de classificação para o número de caracteres recuperados após os ataques ao áudio marcado
Fonte: dados da pesquisa

Determinados os critérios de classificação para os requisitos, é necessário que se estabeleça uma forma de compactar todos os resultados obtidos em um resultado que defina o nível de atendimento da técnica de marca d'água em relação aos requisitos estabelecidos. Os cálculos dos níveis de atendimento a cada requisito, bem como da técnica de marca d'água como um todo são descritos na seção 3.5. A escala de julgamento estabelecida para a avaliação destes resultados é apresentada na tabela Tabela 4.

Classificação	Nível de Classificação (%)
Excelente	90,1 a 100
Bom	75,1 a 90
Satisfatório	60,1 a 75
Regular	50,1 a 60
Insatisfatório	0 a 50

Tabela 4 - níveis de classificação para avaliação final da técnica de marca d'água
Fonte: adaptado de [13].

3.5 Cálculo do nível de atendimento da marca aos requisitos estabelecidos

Para chegar a um resultado único no final da avaliação, foi realizado o cálculo do nível de atendimento de cada requisito em cada um dos arquivos analisados. A avaliação desses requisitos é estabelecida através da avaliação dos respectivos parâmetros. No caso da robustez, os parâmetros avaliados constituem cada um dos ataques aplicados ao áudio marcado. Logo, a resistência da marca à compressão dinâmica, à remoção de ruído, filtragem, adição de eco e demais ataques são adotados como parâmetros para avaliação desse requisito.

Para avaliar imperceptibilidade o parâmetro adotado é o valor ODG. A ocorrência dos conceitos "ruim", "regular", "bom", "muito bom" e "ótimo" especificados na Tabela 2 associados aos respectivos pesos é utilizada para calcular o nível de atendimento a esse requisito.

Por fim, para carga de dados tem-se o a quantidade de informação inserida no arquivo.

Para cada parâmetro de avaliação de cada requisito é verificado o número de ocorrências em cada nível da escala estabelecida de acordo com o desempenho dos arquivos nos testes. A avaliação de cada parâmetro é feita com base em uma nota que é calculada da seguinte forma:

$$NPa = \frac{\sum(Pa*pe)}{\sum(Mpa*Mpe)} * 100 \quad (2)$$

Onde:

NPa = nota do parâmetro de avaliação

Pa = número de ocorrências de arquivos de áudio em determinado nível de classificação na avaliação sob determinado parâmetro.

Pe = peso do nível de classificação

Mpa = número máximo de ocorrências.

Mpe = peso máximo.

A nota de cada parâmetro de avaliação em cada requisito é obtida através da soma do número de ocorrências em cada nível de classificação, multiplicada pelos respectivos pesos e dividida pelo número máximo de ocorrências multiplicado pelo peso máximo.

Após o cálculo da nota dos parâmetros é possível proceder ao cálculo da nota dos requisitos. Isto é feito por meio do somatório das notas dos parâmetros de avaliação dividido pelo somatório das notas máximas possíveis. A fórmula 3 apresenta o cálculo da nota dos requisitos:

$$NR = \frac{\sum NPa}{\sum MPa} * 100 \qquad (3)$$

Onde:

NR = nota do requisito

MPa = nota máxima do parâmetro de avaliação

Para o cálculo da nota final da técnica de marca d'água são utilizadas as notas de todos os requisitos. É feito o somatório dessas notas multiplicado pelos respectivos pesos e dividido pela soma dos mesmos, conforme apresentado a seguir:

$$NFT = \frac{\sum(NR*peR)}{\sum peR} \qquad (4)$$

Onde:

NFT = Nota final da técnica de inserção

[3]peR = peso do requisito

3 Os requisitos avaliados neste trabalho foram considerados como sendo de igual importância para aplicação em direitos autorais, sendo atribuído peso 10 aos mesmos.

4. EXECUÇÃO DOS TESTES

Os testes foram realizados em 20 arquivos de áudio no formato wave com resolução de 16 bits e taxa de amostragem de 44.100Hz. Como mencionado anteriormente, a ferramenta utilizada para avaliar a imperceptibilidade da marca d'água nos arquivos foi o PEAQ. Durante os testes para avaliação da imperceptibilidade da marca, os arquivos foram reamostrados para 48.000Hz em virtude de o PEAQ trabalhar somente nesta taxa de amostragem. As músicas possuem diferentes características tais como presença acentuada de percussão, níveis diferentes de energia, gravações ao vivo e de estúdio além da presença de pequenos períodos de silêncio ao longo do tempo, com comprimento do áudio variando de 3m06s a 5m59s. A ferramenta utilizada para inserir e detectar a marca d'água foi desenvolvia por Kirovski e Malvar em 2001 e foi descrita em [12]. O sistema insere uma string hexadecimal no arquivo de áudio em janelas de tempo de 11segundos num total de dois caracteres por janela.

Para a execução dos testes de robustez, foi inserida a string hexadecimal "AAAABBBBCCCCDDDDEEEEFFFF11112222" em cada arquivo analisado. A escolha destes caracteres foi dada em virtude da facilidade na conferência da marca recuperada após os ataques. Poderiam ser utilizadas outras strings, a serem aplicadas em diversas situações, como identificação do artista por exemplo. O tamanho desta string depende diretamente do tamanho do áudio. Pelo fato de serem analisados áudios com diferentes características e diferentes tamanhos, o arquivo com menor comprimento foi adotado parâmetro para determinar o tamanho da string utilizada nos testes.

Após o processo de inserção, o áudio foi submetido aos ataques, sendo gravado um log para cada processo de detecção. Os valores para os ataques executados foram baseados nos valores default apresentados no sistema Stirmark Benchmarck e no trabalho de Peticolas e outros [18]. Este valores são descritos no Quadro 1. Na análise dos logs, levou-se em conta a recuperação da marca como um todo. A presença de símbolos que não foram incluídos na mensagem original e a sequência incorreta dos caracteres foram consideradas. Foi avaliado o percentual de informação detectada e não detectada sendo que, para a informação detectada foi verificado o percentual de erro e acerto na recuperação da string.

Grupo	Ataque	Valores
Ataques de dinâmica	Compressão	Ataque: 1ms / Release: 500ms / Output: 0dB / Thresrhold: -6,123 / Proporção: 2:1
	Remoção de ruído	Thresrhold: -60dB
Filtragem	Filtro passa-baixa	Freqüência de corte: 15.000kHz
	Filtro passa-alta	Freqüência de corte: 50Hz
Ambiente	Eco	Atraso: 100ms
	Delay	Atraso: 400ms / Volume: 10% do audio original / Número de ecos: 2
Conversão	Reamostragem	Amostragem: 8.000Hz
	Alteração da taxa de bits	Resolução: 8 bits
Compressão com perda	Compressão MP3	Qualidade: 56 kbps
Adição de ruído	Adição de ruído branco	Amplitude: 0,5
Modulação	Chorus	Numero de vozes: 5 /Atraso: 5ms / rate: 1,2Hz / Feedback: 10% / DryOut: 100% / Wet out: 5%
	Flanger	Depth: 10% / Rate: 0,300Hz / Dry out: 50% / Wet out: 50%
Alteração de pitch e tempo	Shift pitch	Incremento de 5%
	Alteração de tempo	Incremento de 5%
Permutação de amostras	Corte	Corte aleatório de amostras no áudio

Quadro 1 – Ataques aplicados ao áudio marcado
Fonte: dados da pesquisa

Na avaliação da imperceptibilidade o áudio marcado foi submetido à ferramenta PEAQ juntamente com o arquivo original para o cálculo do ODG.

Por fim, o cálculo do número máximo de caracteres a serem inseridos no arquivo foi feito através do número de janelas de 11seg possíveis no áudio, de acordo com sua duração.

5. RESULTADOS

Na análise da robustez, a marca d'água se mostrou resistente à maioria dos ataques sendo que, a remoção de ruído, a utilização de filtro passa-baixa, reamostragem e alteração do tempo do áudio não causaram nenhum dano à mensagem incorporada no arquivo. A nota alcançada para o requisito foi de 92,67, ou seja, a técnica alcançou um percentual de 92,67% do nível máximo de atendimento ao requisito para os arquivos analisados. Através dos dados mostrados na Tabela 5 é possível perceber que os ataques mais agressivos foram *Pitch Shift* e inserção de ruído, com média de detecção de 0,16% e 63,44%. O percentual de detecção para os demais ataques variou de 99,00% a 100,00% confirmando o excelente nível no requisito robustez da marca.

Ataque	Acerto %	Erro %	Não detecção %
Remoção de ruído	100,00	0,00	0,00
Filtro passa-baixa	100,00	0,00	0,00
Reamostragem	100,00	0,00	0,00
Alteração de tempo	100,00	0,00	0,00
Compressão	99,84	0,16	0,00
Filtro passa-alta	99,84	0,16	0,00
Quantização	99,84	0,16	0,00
Chorus	99,84	0,16	0,00
Flanger	99,84	0,16	0,00
Corte	99,84	0,16	0,00
Eco	99,69	0,31	0,00
Delay	99,69	0,31	0,00
Compressão mp3	99,22	0,78	0,00
Inserção de ruído	63,44	0,00	36,56
Pitch	0,16	0,00	99,84
Média geral	90,09	0,17	9,74

Tabela 5 – Percentual médio de recuperação da marca por ataque
Fonte: dados da pesquisa

A nota atingida no requisito imperceptibilidade foi de 96,00%. Os valores ODG encontrados nos arquivos analisados estão entre -1,457 e -0,521, compreendendo valores de "imperceptível" a "perceptível, mas não irritante", o que justifica a nota encontrada para este requisito.

Arquivo	ODG	Arquivo	ODG
Áudio 1	-0,646	Áudio 11	-0,614
Áudio 2	-0,623	Áudio 12	-0,577
3 Áudio	-0,714	Áudio 13	-0,695
Áudio 4	-1,457	Áudio 14	-0,630
Áudio 5	-0,638	Áudio 15	-0,573
Áudio 6	-0,629	Áudio 16	-0,959
Áudio 7	-0,678	Áudio 17	-0,710
Áudio 8	-0,656	Áudio 18	-0,521
Áudio 9	-0,638	Áudio 19	-0,632
Áudio 10	-0,626	Áudio 20	-0,605

Tabela 6 - Resultados do teste de imperceptibilidade
Fonte: dados da pesquisa

Com relação à carga de dados, verificou-se que a quantidade de informação a ser inserida varia de acordo com o tamanho do arquivo. O sistema de marca d'água insere dados numa proporção de 2 caracteres por cada janela de 11 segundos. Os arquivos analisados neste trabalho possuem uma duração média de 4m14s, o que confere uma capacidade média de incorporação de 46 caracteres. De acordo com a especificação da RIAA esta carga de dados é suficiente para proteger os direitos autorais de um arquivo de áudio, uma vez que estabelece um mínimo de 108 bits de informação, equivalendo a aproximadamente 13 caracteres. De acordo com o estudo realizado, verificou-se que a duração do áudio para que a técnica possa ser aplicada na proteção de direitos autorais deve ser superior a 1m17s, devido à particularidade da mesma de inserir 2 caracteres de informação a cada 11s de janela.

6. CONCLUSÃO E PERSPECTIVAS

Este trabalho propôs a avaliação da técnica Spread Spectrum de inserção de marca d'água em áudio por ser a mais referenciada na literatura e por ser considerada como a técnica mais promissora literatura [3], [12], [11], [23]. Os cálculos das notas dos requisitos e da técnica de marca d'água podem ser visualizados no Quadro 2. A técnica obteve um percentual de 94,33% de atendimento aos requisitos; nível considerado como excelente. Embora os três requisitos sejam conflitantes entre si, a marca se mostrou robusta à maioria dos ataques com um nível de imperceptibilidade alto. Além disso, a carga de dados atende à especificação da RIAA sobre a carga de dados para marca d'água nos áudios analisados bastando para isso, que tenham duração mínima de 1m17s.

Com base nos resultados encontrados neste estudo, verificou-se que alguns questionamentos podem ser abordados em trabalhos futuros. Sabendo-se que os três requisitos estudados são conflitantes entre si, é importante entender o quanto estão ligados no sentido de estabelecer uma relação entre eles, o que possibilitaria um entendimento maior sobre o comportamento da marca nos diversos tipos de mídia e aplicações. Além disso, foi verificado que os ataques baseados em alterações no domínio da freqüência se mostraram mais agressivos e removeram a maior parte da marca. Portanto, uma investigação do impacto desse tipo de ataque sobre marcas d'água baseadas em Espalhamento de espectro pode contribuir para o desenvolvimento de sistemas cada vez mais robustos. Percebeu-se também, que as características do áudio marcado podem interferir na eficiência da marca d'água. Seria interessante avaliar a sobrevivência da marca em músicas de diferentes estilos como forma de verificar a relação entre o estilo e o esquema de marca d'água mais adequado à aquele arquivo. Por fim, este trabalho apresentou a avaliação de uma técnica de marca d'água na preservação de direitos autorais. Um estudo comparativo com outras técnicas de inserção pode contribuir e facilitar o direcionamento desta tecnologia aos diversos tipos de aplicação a que pode ser destinada.

Mensur	Mensuração do nível de atendimento da técnica de inserção								Nota dos parâmetros			Nota dos requisitos			Nota Final da técnica de inserção			nal da técnica inserção	
Requisito	Requisito	Parâmetro de avaliação	Peso (peR)	Nível de classificação					∑(pa*pe)	∑(Mpa*pe)	NP	∑ NPa	∑ MPa	NR	∑NR* PeR	∑peR	NFT	∑peR	NFT
				Ruim	Regular	Bom	Muito Bom	Ótimo											
				1	2	3	4	5											
Robuste:	Robustez	Resistência a compressão dinâmica	10	0	0	0	0	20	100	100	100	1390,00	1500,00	92,67	926,67	20	94,33	20	94,33
		Remoção de ruído		0	0	0	0	20	100	100	100								
		Filtro passa-baixa		0	0	0	0	20	100	100	100								
		Filtro passa-alta		0	0	0	0	20	100	100	100								
		Adição de eco		0	0	0	0	20	100	100	100								
		Delay		0	0	0	0	20	100	100	100								
		Reamostragem		0	0	0	0	20	100	100	100								
		Quantização		0	0	0	0	20	100	100	100								
		Compressão MP3		0	0	0	0	20	100	100	100								
		Inserção de ruído		6	2	0	0	12	70	100	70								
		Chorus		0	0	0	0	20	100	100	100								
		Flanger		0	0	0	0	20	100	100	100								
		Shift Pitch		20	0	0	0	0	20	100	20								
		Alteração de tempo		0	0	0	0	20	100	100	100								
		Corte		0	0	0	0	20	100	100	100								
	Impercepti bilidade	Valor *ODG*	10	0	0	0	4	16	96	100	96	96	100	96	960				
		Chorus		0	0	0	0	20	100	100	100								

Quadro 2 – Cálculo do percentual de atendimento aos requisitos
Fonte: dados da pesquisa

7. REFERÊNCIAS BIBLIOGRÁFICAS

[1] AQUINO, Luiz Diego. 2011. *Sistema de ocultação de dados em áudio através de técnicas de compactação e espalhamento espectral*. Monografia (Bacharelado em Ciência da Computação) - Universidade Regional de Blumenau.

[2] BENDER, W.; GRUHL, D.; MORIMOTO, N.; LU, A. *Technics for data hiding*. 1996. IBM Systems Journal. MIT Media Laboratory, 20 Ames Street, Cambridge, Massachusetts 02139-4307, USA

[3] BONEY, Laurence et al. *Robust audio watermarking using perceptual masking*. 1998. Department of Electrical Engineering. University of Minnesota.

[4] BRASIL. MINISTÉRIO DA CULTURA. *Direito autoral.* Disponivel em: < http://www.cultura.gov.br/site/wp-content/uploads/2008/02/caderno-politicas-culturais-direitos-autorais.pdf> Acesso em: 06 jul 2011.

[5] CVEJIC, Nedeljko; SEPPÄNEN, Tapio. Audio watermarking: requeriments, algorithms and benchmarking. In: SEITZ, Juergen. *Digital watermarking for digital media*. 2005. Cap 6, p. 135-181. Information Science Publishing - Hershey. United States of America.

[6] DEFIGUEIREDO, Dimitri do Brasil. *Método de inserção de marcas d'água robustas em sinais de áudio*. 2001. 66f. Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP).

[7] DITTMANN, Jana et al. *Advanced audio watermarking benchmarking*. 2004. Proceedings of SPIE - The International Society for Optical Engineering, v 5306, p 224-235.

[8] FUNAKI, Ricardo. *A proteção dos direitos autorais na Internet.* 2010. UniBrasil - Faculdades Integradas do Brasil.Curutiba - Paraná.

[9] HANJALIC, A., LANGELAAR, G.C., VAN Roosmalen, P.M.G., Biemond, J., & LANGENDIJK, R.L. 2000. *Image and video databases: Restauration, watermarking and retrieval*. Amsterdam: Elsevier.

[10] KABAL, Peter. *PQevalAudio*. 2003. Disponível em: <http://www-mmsp.ece.mcgill.ca/documents/Software/Packages/AFsp/PQevalAudio.html>. Acesso em: 04 nov. 2011.

[11] KIM, Hyoung Joong. *Audio Watermarking Techniques. Intelligent Watermarking Technique.* Ed. J.S. Pan, H.C. Huang, and L.C. Jain, World Scientific Publishing Company, 2004, pp. 185-217.

[12] KIROVSKI, Darko; MALVAR, Henrique. *Robust Spread Spectrum audio watermarking*. 2001. IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, pp. 1345-1348

[13] MARÇAL Elizabeth K., BEUREN, Ilse M., 2005. *Auditoria da qualidade de um software de contabilidade*, I Seminário de Ciências Contábeis. Blumenau/S.

[14] MARNANI, Elham Karimi; KARAMI, Zahra; MOLAVIANJAZI Ebrahim. *A Comparison of Some Audio Watermarking Methods*. Dept. of Electr. & Comput. Eng., Isfahan Univ. of Technol., Isfahan, Iran. 2009.

[15] MOULIN, Pierre; KOETTER, Ralf. *Data hiding*. 2005. Proceedings of the IEEE, v 93, n 12, p 2083-2125.

[16] NUNES, Karoline Matias de Oliveira. *Inserção de marca d'água em sinais de áudio.* 2007. 97f. Tese (mestrado em Ciências) - Instituto Tecnológico de Aeronáutica.

[17] OLIVEIRA, L. P. M. *Marca d'água frágil e semi-frágil para autenticação de vídeo no padrão H.264*. 2009. 97f. Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal do Rio de Janeiro.

[18] PETICOLAS et al. 2001. *A public automated web-based evaluation service for watermarking schemes: StirMark Benchmark*. Disponível em: <http://www.petitcolas.net/fabien/publications/ei01-automated.pdf> Acesso em: 03/11/2011.

[19] RODRIGUEZ, G. Cano; MIYATAKE, M. Nakano; MEANA, H. M. Pérez. *Analysis of audio watermarking schemes*. 2005. 2nd International Conference on Electrical and Electronics Engineering (ICEEE) and XI Conference on Electrical Engineering (CIE 2005). Mexico City – Mexico.

[20] SEITZ, Juergen; JAHNKE, Tino. Digital Watermarking: An Introduction. In: SEITZ, Juergen. *Digital watermarking for digital media.* 2005. Cap 1, p. 1-29.

[21] SKLAR, Bernard. *Digital communications: fundamentals and applications*. 2nd ed. Upper Saddle River: Prentice-Hall, 2001. Apud AQUINO, Luiz Diego. 2011. *Sistema de ocultação de dados em áudio através de técnicas de compactação e espalhamento espectral*. Monografia (Bacharelado em Ciência da Computação) - Universidade Regional de Blumenau.

[22] SOLA, José Eduardo Martins. *A proteção dos direitos autorais a partir da realidade Internet: a perspectiva brasileira*. 2002. 172f. Dissertação (mestrado em Ciências da Informação) – Faculdade de Filosofia e Ciências – Universidade Estadual Paulista UNESP.

[23] SWANSON, Mitchell D. et al. *Robust audio watermarking using perceptual masking.* Signal processing, v. 66, p. 337 – 355, 1998.

[24] TANENBAUM, Andrew S. *Computer networks*. 4 ed. Upper Saddle River : Prentice-Hall, c2003. 889p.

[25] WANG, Xiang-yang et al. *A New Content-Based Digital Audio Watermarking Algorithm for Copyright Protection*. 2004. Proceeding InfoSecu '04 Proceedings of the 3rd international conference on Information security. New York – USA.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Interactions between audio and 3D stereoscopic image in the perception of distance and suitable recording distance

Luiz Fernando Kruszielski[1], Toru Kamekawa[1], Atsushi Marui[1],
[1]Tokyo University of the Arts
Tokyo, 120-0034, Japan
kruszielski@gmail.com, kamekawa@ms.geidai.ac.jp, marui@ms.geidai.jp

ABSTRACT

In this paper, a primary study about the interaction between a three-dimensional stereoscopic (3D) image and sound in the perception of distance was made. The most suitable sound recording distance for 3D and 2D image was also investigated. One 2D and three 3D images with three different eye-convergence points were used for the visual stimuli and three different recording distances were used for the sound stimuli. The result shows that the 3D images presented a stronger sense of distance compared to 2D the image. However, influence caused by the sound recording distance had a stronger effect in this factor. Although there was a difference in the perception of distance between 2D and 3D, it did not have a significant effect on the suitability of the sound recording distance.

0. INTRODUCTION

The visual sense is considered to be predominant and have the main role in the perception of space. However, this perception is not only done by one sense alone, but by complex interactions between senses. Those interactions are still not very well understood [1].

In the past few years, an exponential growth in the number of media using stereoscopic contents can be observed [2]. The use of three-dimensional (3D) stereoscopic content significantly increases the perception of space [3,4,5]. The slight difference in the image caused by the distance between two eyes when processed by the brain is one of the main factors in the perception of deepness. The production methods for this kind of media are still under development, and the possible interactions between the 3D image and audio perception/production is still unclear.

The audio content can also be used as an important factor in creating the sense of space [5]. The increase in the spatial impression caused by the 3D image suggests a need to use a more detailed spatial sound delivery to the spectator. Kamekawa, *et al*. have shown that a seven-channel surround system used with a 3D image increases the perceived depth [7]. Other researches also suggest a sound/depth rendering algorithm [8] and the use of speakers mounted above the listener for improved spatial delivery of sound [9] for 3D-image media.

In the currently available commercial stereoscopic 3D content, two variables are mainly responsible for the depth effect in the image: the stereo base and the image convergence. The stereo base is the distance between the two cameras and controls the strength of the stereo effect. The other parameter is the convergence, which is the aiming of both cameras towards a single point. This parameter changes the perceived projected depth of an object in relation to others. [2,3,10]

While it is clear that media using 3D is both a strong developing trend in cinema, television and games, researches concerning interactions in the sound perception and therefore improve sound for use in this media are very

small. This study has the objective to be an introductory research about how the convergence in the 3D image recording could interact with sound in the perceived sense of distance, and also to compare a 3D image to a 2D image in the perception of distance, and how it could influence the sound recording distance suitability.

1. EXPERIMENTAL METHODOLOGY

In this experiment, two aspects were evaluated: the perceived sense of distance and the image and sound suitability. The test consisted of a set of images of a saxophonist recorded in 2D and 3D with different convergences points combined with a convoluted sound simulating a recording distance.

1.1 Stimuli

The stimuli used in this test consisted of four different image patterns and 3 different sound patterns, resulting in a total of 12 stimuli.

2.1.1 Image Stimuli

Four different image patterns were used for this test. The images contained a saxophonist, recorded at Studio A in the Senju Campus of Tokyo University of the Arts (Figure 1). This Studio A has a space of 1260 m^3 and RT_{60} of 0.96s at 500 Hz.

In this test, a fixed value of 6.5cm was used for the stereo base. This value is considered to be an usual distance between the human eyes, although they may vary from 5.5cm to 7.0 according to the human physiology[3]. In this test, three different convergence points where used for the 3D images (Figure 2): one 1m behind the saxophone player's position (Convergence Point C), one at the saxophone player's actual position (Convergence Point B) and one at 1m in front of the saxophone player (Convergence Point A). For the 2D image, the left eye image of convergence point B was used. The cameras used in this test were two Sony Handycam DCR-SX45 placed at 2.5m from the saxophonist with a focal length of 38mm. The distance between the camera and the rear wall was 6m. The camera focus for all images was fixed at the saxophone player. Since the sound was not recorded together with the image, to avoid sound and image sync problems the saxophonist used a metronome. The images had a resolution of 1080p and 30 frames per second.

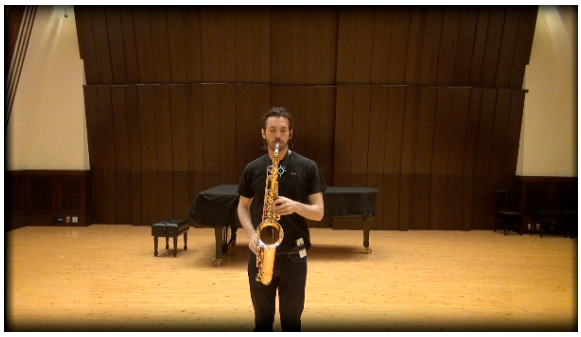


Figure 1 Image of the saxophonist in Studio A

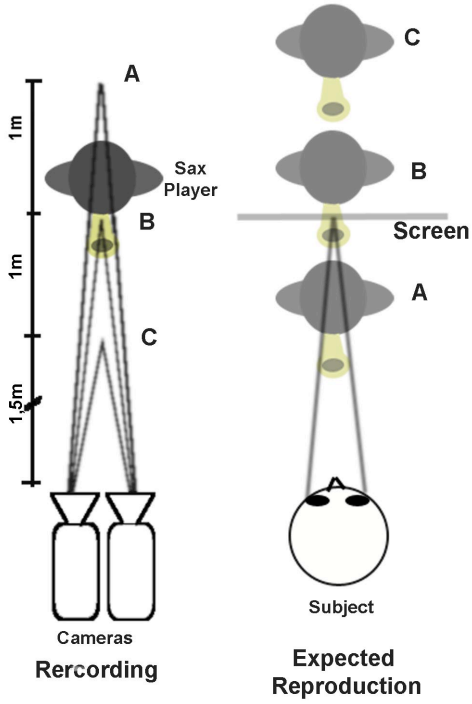


Figure 2 Top view of the convergence points A, B and C and respective expected projected images.

2.1.1 Sound Stimuli

The sound consisted in a musical phrase of 10s from the music "Ornithology" from the composer Charlie Parker. To have the same sound performance in all the stimuli, the sound was not recorded together with the image. It was recorded in a different room that has a short reverb time (RT_{60}<0.4s at 500Hz) with a lavaliere microphone DPA 4060 clamped to the saxophone bell. This recording method was used to make the sound as close as possible to sound recorded in an anechoic chamber.

The saxophone sound was convolved with three impulse responses recorded in Studio A, the same room used for the image, resulting in a location-simulated sound. The impulse sound was a time stretched pulse recorded using a stereo pair of omni-directional microphones, DPA 4007, spaced 42cm apart of each other as in an AB recording setup. The swept sine was processed with a program created in matlab, resulting in a two second impulse response (96000 samples) that was convolved with the sax sound using the Apple Space Designer reverb plug-in. The loudspeaker used as the impulse sound source was a Genelec 8040, and was placed at the same position where the saxophonist recorded the image. The distances between the sound source and the microphones were 150cm, 300cm and 600cm (Figure 3).

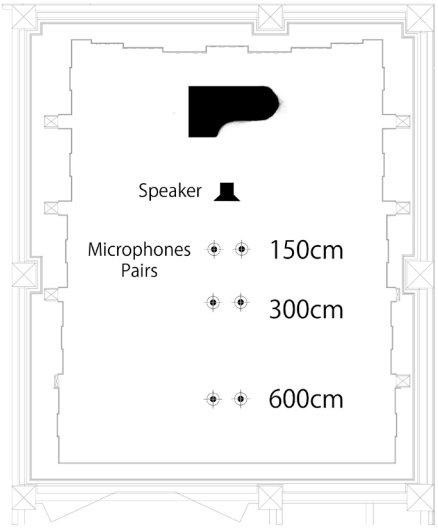


Figure 3 Studio A top view: microphone recording distances

1.2 Test place and subjects

The test was done in Studio B of the Tokyo University of the Arts. The room has an area of 50m^2 and has controlled acoustics with short reverberation time. Two Genelec 8260 speakers were used for the audio reproduction. The image display device was a Panasonic television, TH-P54VT, 54 inches, using active polarized shutter glasses for both the 2D and 3D images. The screen and the stereo speaker were set at 2.4m from the subject.

The subjects consisted of eight people, six students and two professors from Tokyo University of the Arts. All the subjects had previous experience with sound design and psychoacoustic evaluations.

1.3 Evaluation method

In this experiment, the subjects were asked to perform two tasks. The first one was to rate the subjective sense of distance perceived between the saxophonist presented in the image and the subject themselves. The second task was to rate how suitable the presented sound relative to the presented image was.

For the first task, sense of distance, a method inspired by S.S. Steven's Magnitude Estimation (ME)[11] was used. In this method, the subject has the freedom to rate the perceived sense of distance instead of using a fixed scale. A subject was asked to rate the perceived distance in one's own numeric scale with regard to their internal reference. They were instructed to use larger values if the egocentric distance to the presented saxophonist was larger. The results of each subject were standardized to have the same range of values. Three repetitions of 12 stimuli were presented in a random order, resulting in a total of 36 ratings per subject. There was a training session before the test, using the same 12 stimuli. The data from the training session was discarded before analysis.

In the second task, about sound and image suitability, the subjects used a seven-point bipolar scale to rate if the sound was suitable or not relative to the presented image.

All the stimuli were randomized. Each test took about 25 minutes to be completed.

2. RESULTS AND DISCUSSION

The results of the sense of distance task can be seen in figure 4. Analyzing the data with a two-way analyze of variance (ANOVA), it was possible to see that both the sound and the image had a significant value (image: F=8.099 p<.01; sound F=254.95 p<.001) (See Table 1 for more detail). The sound recording distance was clearly the strongest factor regarding the sense of distance. As the recording distance increased, the perceived distance also increased.

The convergence in the 3D images did not have a significant effect in the perception of the distances. The result shows that between the 3D images, it was unclear to the subject if the saxophonist was more or less distant regarding the three convergence points. According to Medinburu, a stereoscopic window violation occurs when an object is wrongly placed relative to the stereoscopic window. Therefore, one of the possible reasons for this result is the fact that since the saxophone player's body was not entirely inside the frame; it was cropped at the height of the knees as can be seen in Figure 2. The television screen external frame may have been used as a reference point to the subject, therefore creating a mismatch between the images presented on the screen and the real television frame.

The 2D image was perceived to be closer than the all the 3D images. Using the Tukey Honest Significant Difference Comparisons analysis method [12], it was found that there was a significant value comparing the 2D image and all 3D images (p>.001). However, between the 3D stimuli A, B and C, the significance was not perceived (p=.99). The deepness caused by 3D image may lead the perception of the whole picture to be further away than the 2D image. This difference was smaller than the one caused by the sound recording distance.

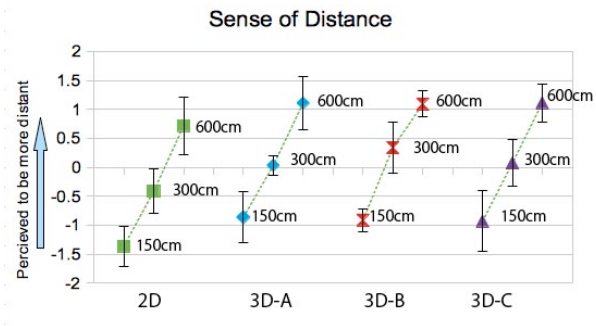


Figure 4 Relative sense of distance between the stimuli. The abscissa value represents a subjective scale of the perceived distance

	df	SS	MS	F	p
Sound	2	174.49	87.247	254.95	<.0001
Image	2	8.3156	2.7719	8.099	<.001
Interaction	2	0.2923	0.0487	0.142	0.99
Error	268	91.715	0.342		

Table 1 Anova table for the Sense of Distance

In the second task, the image and sound suitability, both 3D and 2D images presented a similar preference for the sound recorded at 300cm (Figure 5). An analysis made with two-way ANOVA shows that there was no

significant difference in the suitability between the images (p=.85 see Table 2 for more detail).

Comparing both tasks, it can be noticed that even with a noticeable difference in the sense of distance between the 3D images and the 2D images, there was no noticeable effect in the overall suitability of the recorded distance. A possible reason is that, as stated by the subjects, the differences between the sound recording distances had a much stronger and more evident effect than the differences in the images. Smaller differences between the sound recording distances could have been more compatible with the different levels in the perception caused by the image.

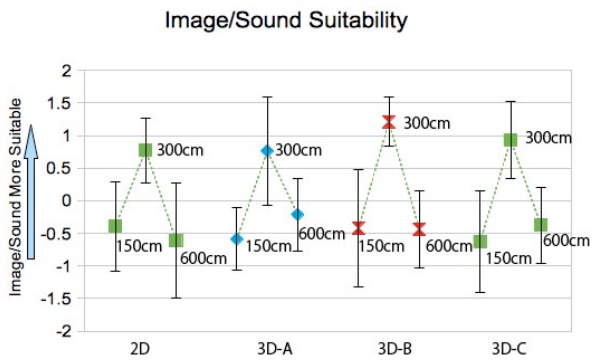


Figure 5 Relative Image and sound Suitability. Higher values mean a more suitable sound for the presented image.

	df	SS	MS	F	p
Sound	40.73	2	20.36	32.13	<.0001
Image	0.49	3	0.16	0.26	0.854
Interaction	1.54	6	0.26	0.4	0.8771
Error	53.24	84	0.63		

Table 2 Sound and Image suitability ANOVA table

It is interesting to notice that, in a similar previous 2D experiment about the camera lenses focal length, the focal length did change the suitability of the sound [13]. Although a direct comparison cannot be made, this suggests that the lenses focal length have a stronger effect in the suitability of the recording distance compared to the fact that the image is 3D or 2D. Also, that 2D cues in the perception of space had a stronger influence in the perception of sound compared to 3D cues.

3. CONCLUSION

In this study, a primary study on the possible effects of 3D image interaction with sound in the perception of distance and in the sound recorded distance suitability was made. The camera convergence did not have a significant effect in the sense of distance or in the sound suitability. Comparing 2D and 3D results, the 2D was perceived to be closer than the 3D image. However, it did not affect the sound suitability comparing to a 3D image. This suggests that a recording distance in a 2D content can be effective for a 3D content. Further studies need to be conducted in this field to confirm the results using other variables such as different sound sources, camera focal lengths, recording distances, stereo bases, and multi-channel surround or other sound rendering techniques.

4. REFERENCES

[1] A Stein, B.; and Meredith, M.: "The merging of the senses" MIT press, Cambridge, USA,1993

[2] Mendiburu, B.: "3D Movie Making" Focal press, Burlington, USA, 2009

[3] Zakia, R; "Perception and Imaging: Photography – A Way of Seeing, third edition", Focal press, Burlington, USA, 2007

[4] Mendiguru B.; Pupupin, Y.; Schklair, S. "3D TV and 3D Cinema" MIT press, Focal press, Burlington, USA , 2012

[5] Yewdall, D. "Pratical Guide of Motion Picture Sound, 3rd Edition" Focal press, Burlington, USA, 2007

[6] Fehn, C. "3D TV Broadcasting, in 3D Videocommunication: Algorithms, Concepts and Real-Time Systems in Human Centred Communication" John Wiley & Sons, Ltd, Chichester, UK, 2006

[7] Kamekawa, T.; Marui, A.; *et al*: "Evaluation of spatial impression comparing 2ch stereo, 5ch surround, and 7ch surround with height channels for 3D imagery" AES 130th Convention London, UK Proceedings, 2010

[8] Kim, Sunmin; Lee, Y.; and Kim, Y.: "3D Audio Depth Rendering for Enhancing Immersion of 3DTV" AES 131th Convention New York, USA Proceedings, 2011

[9] Kim, Sungyoung; Okamura, H.; *et al*: "Auditory depth control: a new approach utilizing a plane wave loudspeaker radiating from above a listener" AES 131th Convention New York, USA Proceedings, 2011

[10] Abe, K.: "Stereoscopic 3D Video Production Guidebook" (available only in Japanese language) Picture Industry Press, Tokyo, Japan, 2010

[11] Zwicker, E. and Fastl H.: "Psycho-acoustics Facts and Models, 2nd Updated Edition" Springer, Germany, 1999

[12] Ian E. Holliday, 2012, Two-Way ANOVA (v1.0.3) in Free Statistics Software (v1.1.23-r7), Office for Research Development and Education, URL http://www.wessa.net/Ian.Holliday/rwasp_Two%20Factor%20ANOVA.wasp/

[13] Kruszielski, L.; Kamekawa, T.; and Marui, A.: "The Influence of a Camera Focal Length in the Direct To Reverb Ratio Suitability and Its Effects in the Perception of Distance for a Motion Picture" AES 131th Convention New York, USA Proceedings, 2011

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 10º Congresso de Engenharia de Áudio
16ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2012, São Paulo, SP

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

Tempo estimation: evaluation of different spectral flux computation methods

Leonardo O. Nunes,[1] and Luiz W. P. Biscainho[1]

[1] Universidade Federal do Rio de Janeiro, DEL/Poli & PEE/COPPE, LPS
Caixa Postal 68504 – Rio de Janeiro, RJ, Brasil

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

ABSTRACT

In this paper, different calculation methods of features for tempo estimation are compared. In particular, several features based on the spectral flux found in different algorithms from the literature are evaluated. In order to compare the features, two well-known datasets are employed along with two figures of merit. By comparing the performance of different calculation configurations, it was possible to assess their impact on the performance of a tempo estimation algorithm. These results may serve as guidelines for building features for tempo estimation.

0 INTRODUCTION

Computational rhythmic analysis consists of estimating the overall time characteristics of a given music performance from either its symbolic representation or directly from its signal. Among rhythmic analysis targets, tempo estimation consists in finding the common period between consecutive events in the music signal. The tempo information can be employed in several applications of music information retrieval. For example, it can be used to assist the classification of music genres, as input information for the so-called beat tracking systems, as well as be employed in audio transformation tools.

The concept of tempo is widespread in music theory [1] and cognition where it is usually used to define the pace of the music: whether it is fast or slow. In order to estimate the tempo from a recording, one has to overcome several difficulties. For example, the tempo can vary along time. Also, the perceived tempo, defined as the beat[1] rate or period, varies greatly from person to person [2]. Usually, the tempo can be calculated [3] from a symbolic representation or directly from a time-varying signal. In this work, only the latter is considered. Also, the perceived tempo, i.e. the beat period, will be estimated.

Usually, when estimating the tempo directly from

[1] The perceived beat is informally determined by the taps of a listener's foot along with the music.

the signal, a set of features are first extracted from the signal. Ideally, these features capture all the relevant tempo-related information of the signal, facilitating the estimation. Several investigations about which features are better suited to tempo estimation have been made. In [4, 5], the ability of a large set of features to predict the beat information of a signal is studied. In [6], several onset detection functions that can be employed as features for tempo extraction are reviewed. More recently, in [7], several tempo estimation algorithms (each one employing different features) are compared.

In this paper, a family of features for tempo estimation based on the spectral flux onset detection function [6] are studied. In order to arrive at this family, a careful bibliography review of recent papers on tempo estimation is performed and the similarities and differences between feature sets are categorized. The envisaged study of the spectral flux features consists of the performance analysis of the different ways to calculate this feature, including varying parameter values, through the use of a simple tempo estimation algorithm and two different signal databases. The objective of this comparison is to serve as a guideline to future researchers when employing spectral flux-like features, indicating the good practices for choosing parameters. Also, considering that the choice of adequate parameters is dependent on the tempo estimation algorithm employed, this paper tries to provide a good starting point from which to search the parameter space.

After this introduction, a brief overview of tempo estimation algorithms is made in Section 1. Afterwards, in Section 2, a bibliography review of features for tempo estimation is performed. In Section 3, the algorithm employed for feature extraction and tempo estimation is described. The evaluation, the databases, figures of merit, and results are detailed in Section 4. Further evaluations are presented in Section 5. Finally, in Section 6 the conclusions are drawn.

1 TEMPO ESTIMATION OVERVIEW

In this section, a brief overview of signal-based tempo estimation algorithms is performed, conceptualizing the tasks involved without detailing specific algorithms. Tempo estimation algorithms can usually be divided into three stages [7] as depicted in Figure 1. In this figure, the tempo of signal $x[n]$ is estimated using three processing stages: feature extraction, similarity computation, and tempo estimation. In order to illustrate the stages, an example signal will be employed. This signal consists of a cowbell being hit at a rate of 120 beats per minute (BPM), sampled at 44.1 kHz. Figure 2 displays the signal in the time-domain, where the onsets of each note can be clearly observed.

The feature estimation has the objective of extracting a feature list from the audio signal that, hopefully, reduces the amount of data input to the next stage while facilitating its task. Hence, the features should be highly correlated with temporal events occurring in

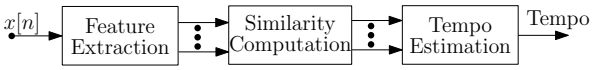


Figure 1: Three processing stages commonly employed in tempo estimation algorithms.

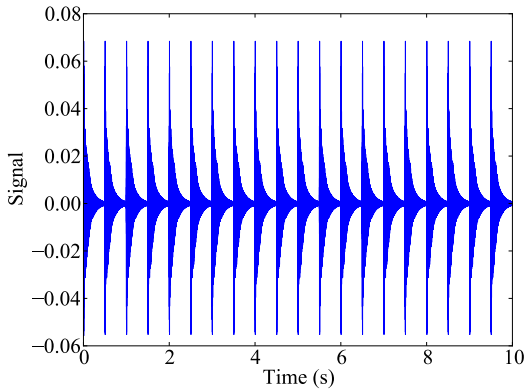


Figure 2: Example signal in the time-domain.

the music signal, such as onsets. In general terms, the features are time-varying signals that should preserve the time-structure of the signal. For example, in Figure 3, a feature signal extracted from the example signal is shown. It can be observed that the feature captures the occurrences of the onsets, while removing information that is not relevant for tempo estimation, such as the duration of each "note". Moreover, the sampling rate is reduced from 44.1 kHz to 100 Hz in the example. It should be noted that the output of this block can be a single feature (as in the example) or a set of such signals (e.g. each feature obtained for a different frequency range). In Section 2, a detailed discussion of which features are usually employed for tempo estimation can be found.

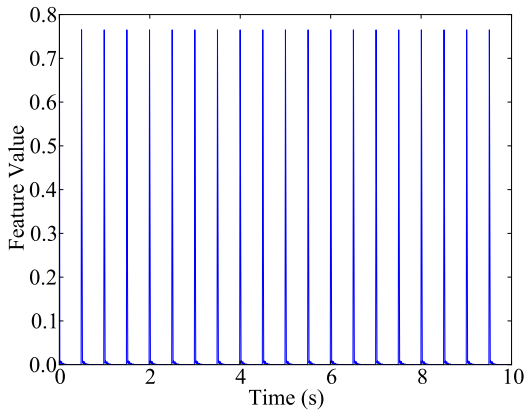


Figure 3: Features obtained from the signal in Figure 2.

The next stage consists of estimating the period of the extracted features. The output of this stage, the periodicity function, is usually a measure of how likely different periods (usually expressed in BPMs) are. Figure 4 displays one such function for the example signal. As can be observed, prominent peaks can be seen at the 60 and 120 BPM marks, indicating that these two periods are the most likely for the example signal.

The periodicity function can integrate the information of different features or can be computed on a feature-by-feature basis. Accordingly, one or more periodicity functions can be obtained. A discussion on periodicity functions can be found in [8].

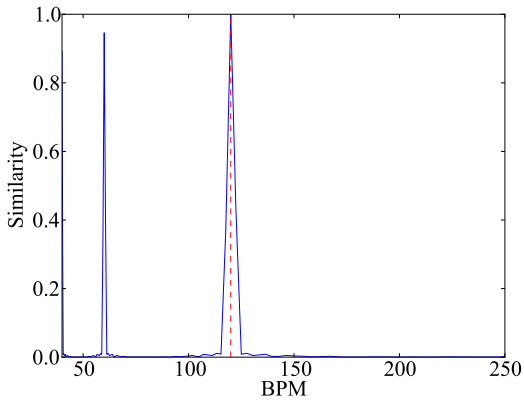


Figure 4: Periodicity function computed from the feature in Figure 3. The vertical dashed line denotes the tempo of the signal.

Once the periodicity function is obtained, the next stage consists of estimating the tempo or tempo candidates of the input signal. The strategies employed for this task vary greatly; from a simple peak picking algorithm to more sophisticated methods that try incorporate music theory-related information [9]. For the example signal, a simple peak picking algorithm would suffice: most prominent peak in the similarity function corresponds to the correct tempo, i.e. 120 BPM.

2 BIBLIOGRAPHY REVIEW

This section provides a bibliography review of features commonly used in tempo estimation algorithms. Emphasis is given to papers published from 2007 onwards.

Usually, features for tempo estimation try to capture sudden changes in the signal: a sudden energy surge or a noticeable change in the melodic line of the piece. The underlying assumption is that these changes occur "in time", i.e. they obey the overall tempo of the piece. Usually more than one type of feature is employed in the same work.

Features that capture melodic/harmonic change are employed in [10]: in these papers tonal features are extracted using a Constant-Q Transform. In [11] a similar idea is employed, but using a different method for acquiring of the tonal features. Another approach consists of detecting timbre changes in the signal [12] through the use of Mel-Frequency Cepstral Coefficients.

Features based on energy changes, however, were found to be the most popular. These features capture how the energy envelope of the signal (or of one of its sub-bands) evolves along time. Usually these energy variations are related to note onsets, which are closely related to the perception of the temporal structure of music [4]. In the literature these features receive different names (depending on their implementations): novelty functions, onset detection functions, or accent functions. One of the first papers to use these features and make them popular [13] first divide the signal into 6 different sub-bands, then estimates the energy envelope of each sub-band signal, and obtains the features as the time-differentiated and half-wave rectified envelopes. These series of steps (sub-band division, time-differentiation, and half-wave rectification) are found in almost all papers that employ energy-related features. In [14] and [15] modifications of the feature extraction algorithm described in [13] are detailed.

Among the energy-based features, the spectral flux [6] has been particularly popular for tempo estimation [16, 17, 18, 19, 20, 9, 21, 22, 8, 23]. As defined in [6], the spectral flux is calculated by first taking the Short-Time Fourier Transform (STFT) of the signal, time-differentiating the magnitude of the STFT, and then summing the result across all frequency bins of the STFT. Several researchers have proposed modifications on top of this algorithm. In [16] the signal is first down-sampled and a reassigned STFT is employed, in order to obtain a more precise time-frequency representation of the signal. In [18, 24, 25, 26, 27] the STFT output is mapped to the Mel scale. A sub-band division is performed in advance in [17, 22], leading to the calculation of one spectral flux function per sub-band. The logarithm of the STFT is taken before the differentiation in [16, 17, 18, 21, 22, 24, 26, 27]. Other modifications include the use of smoothing filters before the time-differentiation [8, 25] and after the summation [8, 9, 18]. Other works [25, 26, 27] do not perform the summation step, calculating one spectral flux function per frequency bin of the STFT.

Among the different algorithms using the spectral flux, there is also a great variation on the values used for the processing parameters of each of the steps. For example, the STFT window length can vary from 20 [26] to 92.8 ms [16] and the hop size between adjacent windows can vary from 4 [18] to 28 ms [23]. Among the methods that perform a mapping to Mel scale, the number of associated Mel filters also vary from 9 [25] to 40 [18].

Considering the popularity of the spectral flux and the plethora of modifications proposed to the method, one would wonder which configurations works best for tempo estimation. The remainder of this paper is dedicated to answering this question by comparing how the performance of a tempo estimation algorithm is influenced by different forms of calculating the spectral flux and the choice of its parameters.

3 METHODOLOGY

In this section, the algorithm employed to assess how different forms of calculating the spectral flux can

influence the performance of a general tempo estimation algorithm is described. The main stages of the prototype algorithm are: spectral flux calculation, autocorrelation computation, and peak-picking.

The spectral flux stage can be broken down in the steps shown in Figure 5. In the figure, the steps named within dashed blocks are optional. Overall, the non-optional steps consist of the basic steps commonly employed in the acquisition of the spectral flux. The signal can be down-sampled by a positive integer D prior to any processing. The STFT is performed using a Hann window with length N and hop size H, the number of frequency bins is determined by the window length. The frequency bins can be mapped to the Mel scale with an arbitrary number of filters F; if $F = 0$ the mapping is not performed. Afterwards, the difference between adjacent frames is calculated using either the magnitude spectrum or its logarithm in base 10 (if L=True, the logarithm is used). The resulting signal is then half-wave rectified. Finally, the signal can be summed across all frequency bins (S=True), in which case a single spectral function is obtained. If S=False, the summation is not performed and several functions are obtained, one for each frequency bin.

The periodicity function employed is the autocorrelation, defined by [4]

$$p[l] = \sum_{m=0}^{M-l-1} f[m]f[m+l], \tag{1}$$

where M denotes the total number of frames of the spectral flux signal and l are the lags for which the periodicity is computed (the lags were chosen as to encompass the interval [40, 250] BPM). In the case where several spectral flux functions are obtained (when S=False), one periodicity function is computed for each spectral flux function and the periodicity functions are summed. The tempo is estimated as the BPM rate corresponding to the largest peak of the periodicity function.

The proposed algorithm emphasizes the feature extraction stage, making the calculation configurable and able to mimic the strategies employed in different algorithms found in the literature. Moreover, the periodicity computation and tempo estimation stages were purposely chosen as simple and well-known methods. Hopefully, this selection will make the impact of the feature extraction stage more prominent. A more elaborate tempo estimation algorithm, for example, could mask problems of the feature extraction stage while making its impact on performance less interpretable.

4 PARAMETER EVALUATION

In this section, the performance of the algorithm described in the previous section with different parameter configurations is evaluated. The objective of the evaluation is to measure how the different calculation methods of the spectral flux influence the performance of the algorithm. In the following sections, the databases employed in the test, the figures of merit used to measure the performance, the different test configurations, and the evaluation results are presented.

4.1 Databases

Two databases commonly employed in the evaluation of tempo estimation methods were used: the Ballroom Dancer and Song Excerpts, both originally used in the Music Information Retrieval eXchange (MIREX) of 2004 and detailed in [4]. Table 1 contains relevant information about each database. Both databases have annotated tempo for every signal, to be used as ground truth for the figure of merit calculations.

Table 1: Information about the databases used in the evaluation.

	Ballroom	Excerpts
Number of signals	698	465
Signal duration (s)	30	20
Total duration (s)	20940	9300
Tempo range (BPM)	[60, 224]	[24, 242]

4.2 Figures of Merit

Two figures of merit are employed [4]:

- Accuracy 1: percentage of tempo estimates within 4 % of the ground truth;
- Accuracy 2: percentage of tempo estimates within 4 % of the ground truth or half, twice, three times, or one third of the ground truth.

Accuracy 1 is a more stringent figure of merit, measuring if the actual annotated tempo was correctly estimated. Accuracy 2, on the other hand, is more lenient, considering estimates that are multiples or divisors of the actual tempo as correct. If one wants to know the tempo of a piece, clearly Accuracy 1 is the more adequate measure. If one wants to use the estimated tempo as input to some other algorithm (e.g. beat tracking), estimated tempos that are multiples or divisors of the ground truth are not so problematic, specially if one considers that even among human listeners such errors are common [2].

4.3 Test Setup

The objective of this first evaluation is to verify the impact of each optional step on the computation of the spectral flux. Moreover, two window lengths and hop sizes are tested. The chosen parameters can be seen in Table 2.

The full combinations of all values chosen for each parameter were employed, totaling 64 different configurations. For each configuration, the tempo was estimated for each signal in one of the chosen databases. Then, the two figure of merits were computed.

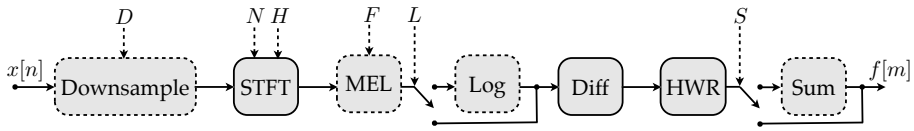


Figure 5: Steps employed in the calculation of the spectral flux. The steps in dashed boxes are optional (they can be omitted). The dashed vertical lines denote parameters that can be chosen in each step.

Table 2: Parameters chosen for the evaluation.

Parameter	Values
D (Down-sample)	$\{1, 4\}$
N (STFT)	$\{20, 40\}$ ms
H (STFT)	$\{5, 10\}$ ms
F (Mel)	$\{0, 20\}$
L (Log)	{True, False}
S (Sum)	{True, False}

4.4 Test Results and Discussion

The results will be separately presented for each figure of merit. Since the objective of this section is to measure how each feature influences the figure of merit, the average result of the figure of merit with a given parameter value, when varying every other parameter, was calculated. For example, 32 tests with $D = 1$ were performed being the result for this parameter value considered as the average result of these 32 tests.

4.4.1 Accuracy 1

Table 3 shows the average results (in %) and their standard deviation for both databases and for each fixed parameter value. The last row (named 'Joint') on the table displays the results when they are jointly calculated for both databases. From the table, the following conclusions can be drawn:

- The down-sample factor does not seem to have an overall impact on the results;
- Larger window sizes appear to yield better results, even though the effect is small face to the standard deviation of the results;
- Inconclusive results were observed for the hop;
- The use of Mel filters were found to produce the greatest impact on performance, regardless of the database;
- There is a clear indication that not taking the logarithm can improve the results;
- As regards S, it does not seem to affect the result.

The best result among the 64 configurations was found to be:

- Excerpts: 32 % for $\{D = 4, N = 40\,\text{ms}, H = 5\,\text{ms}, F = 20, L = \text{False}, S = \text{False}\}$;
- Ballroom: 41 % for $\{D = 4, N = 40\,\text{ms}, H = 5\,\text{ms}, F = 20, L = \text{False}, S = \text{True}\}$;
- Joint: 37 % for $\{D = 4, N = 40\,\text{ms}, H = 5\,\text{ms}, F = 20, L = \text{False}, S = \text{False}\}$.

As can be observed, the results for the Ballroom tend to be slightly better than the ones obtained for the Excerpts database; this fact has also been observed in other studies [28]. From the parameters configuration that yielded the best results, it was found that for all databases the smaller hop, the longer window, and $D = 1$ were employed. Moreover, the Mel-frequency mapping was used, while the logarithm was not.

When one selects the results obtained when all the better-performance parameters (per Table 3) for each database are jointly chosen, the following is found:

- Excerpts: 31 % for $\{D = 1, N = 40\,\text{ms}, H = 10\,\text{ms}, F = 20, L = \text{False}, S = \text{False}\}$;
- Ballroom: 41 % for $\{D = 4, N = 40\,\text{ms}, H = 5\,\text{ms}, F = 20, L = \text{False}, S = \text{True}\}$;
- Joint: 28 % for $\{D = 1, N = 40\,\text{ms}, H = 5\,\text{ms}, F = 20, L = \text{False}, S = \text{False}\}$.

As can be observed, the impact of choosing the best parameters relative to the best results is small for both databases. For the case of the Excerpts, the only difference in the chosen parameters are the down-sample factor and hop, with a difference of 1 percentage point relative to the best result. This corroborates the fact that the impact of these parameters on the overall result can be small. For the Ballroom database, the chosen parameters actually yielded the best overall result. When considering both databases jointly, a significant drop in performance is observed relative to the best overall result, even though only a single parameter was changed: the down-sample factor. This points out to the fact that, at least for D, the results that seem adequate for one database might not be adequate for another. The case of the down-sample will become clearer in the next section, when the results for Accuracy 2 are analyzed.

4.4.2 Accuracy 2

Table 4 shows the average results (in %) and their standard deviation for both databases and their joint results. The following can be gathered by observing the tabulated results:

- Differently from what was observed for the Accuracy 1, the use of down-sample had a great impact in performance.
- Again, larger window sizes seem to slightly improve the result;

Table 3: Accuracy 1 average results (in %) for each database and each parameter value. The standard deviation values are exhibited inside parentheses.

Database	D		N		H		F		L		S	
	1	4	20 ms	40 ms	5 ms	10 ms	0	20	True	False	True	False
Excerpts	**26** (4)	20 (7)	22 (6)	**24** (7)	22 (7)	**23** (6)	18 (6)	**27** (4)	**24** (5)	21 (8)	**25** (6)	21 (7)
Ballroom	28 (4)	**31** (5)	29 (5)	**30** (5)	**30** (6)	29 (5)	28 (5)	**31** (5)	**33** (5)	26 (3)	28 (5)	**30** (6)
Joint	**27** (3)	27 (5)	26 (4)	**27** (5)	27 (5)	**27** (4)	24 (4)	**29** (3)	**29** (4)	24 (4)	**27** (4)	26 (5)

Table 4: Accuracy 2 average results (in %) for each database and each parameter value. The standard deviation values are exhibited inside parentheses.

Database	D		N		H		F		L		S	
	1	4	20 ms	40 ms	5 ms	10 ms	0	20	True	False	True	False
Excerpts	**68** (5)	56 (7)	61 (9)	**63** (9)	62 (9)	**63** (9)	57 (9)	**67** (7)	**63** (7)	61 (11)	**64** (7)	60 (11)
Ballroom	**70** (11)	64 (5)	65 (9)	**69** (8)	66 (9)	**68** (9)	66 (10)	**68** (9)	**74** (7)	60 (4)	**68** (9)	66 (10)
Joint	**69** (8)	61 (5)	63 (8)	**67** (6)	64 (8)	**66** (7)	63 (8)	**67** (6)	**70** (6)	60 (6)	**66** (8)	64 (7)

- Larger hops also seem to be favored, but the difference is too small relative the standard deviation for the result to be conclusive;
- The use of Mel filters seems to improve performance, as was observed for Accuracy 1;
- As with Accuracy 1, the results show that using the logarithm leads to worse results;
- S=True yielded better result, indicating that combining the periodicity functions might facilitate the detection of multiples or divisors of the actual tempo estimate.

Overall, the results obtained for the Accuracy 2 test tended to corroborate what was observed for the Accuracy 1 test. The most interesting new finding was the influence of the down-sampling factor, which might indicate that the high-frequency region (discarded when D=4) carries relevant information relative to multiples or divisors of the actual tempo. The test employed was not able to make clear the effect of the window size. Moreover, even though the results indicated that mapping the features to the Mel scale might yield better results, the effect of the number of filters was not assessed.

5 FURTHER EVALUATIONS

In this section, two tests will be described: one to evaluate the influence of the window length and the hop size employed on the tempo estimation performance, and the other to investigate the influence of the number of Mel filters. Each test will be described separately.

5.1 Window Length and Hop Size

The objective of this test is to verify how different values of window length and hop size affect the performance of tempo estimation, considering a wider range of values than in the previous test. As found in the previous section, larger window lengths seem to improve the performance, but the improvements were too small to be conclusive.

In this test, longer windows and hops were employed, in order to try to make the differences in performance more noticeable: $N \in \{40, 80, 120\}$ ms and $H \in \{10, 20, 40\}$ ms. Apart from the window length and the hop size, all other parameters were kept fixed at values that exhibited a good performance according to the previous test: $D = 1$, $F = 20$, $L =$ False. Since the impact of S on the performance was not clear, its value was also varied: $S \in \{\text{True, False}\}$.

Table 5 contains the results for all combinations of window lengths and hop sizes averaged for the two possible values of S. For Accuracy 1, the results indicate that a hop size of 20 ms might lead to better results (specially for the Ballroom database). The results for larger window sizes are less clear, apparently being dependent on the database and the hop size. As regards Accuracy 2, smaller hops tend to produce better results, and the gains produced by smaller hops are greater than those yielded by longer windows. In particular, the performance increases significantly when the hop is reduced from 40 ms to 20 ms. As to the window size, the improvement is more clear for the Excerpts database than for the Ballroom, indicating that the effect of the window size might vary greatly with signal characteristics.

5.2 Number of MEL Filters

In the test described in Section 4 the results indicated that employing a mapping to the Mel scale is beneficial to performance. In this section, the effect of the number of MEL filters employed is studied. For this, the number of filters was chosen as $\{10, 20, 40, 80\}$, while the other parameters were chosen as $D = 1$, $F = 20$, $L =$ False, and S= $\{\text{True, False}\}$, following the same reasoning as for the previous test.

The results for both figures of merit can be seen in Table 6. From the results, it can be gathered that

Table 5: Average results for longer window length and hop size. All values in %.

	Accuracy 1									Accuracy 2								
	Excerpts			Ballroom			Joint			Excerpts			Ballroom			Joint		
$N \setminus H$ (ms)	10	20	40	10	20	40	10	20	40	10	20	40	10	20	40	10	20	40
40	30	30	30	30	35	28	30	33	29	73	72	67	84	83	74	79	79	71
80	32	32	32	28	34	28	30	33	30	76	75	70	84	83	76	80	80	74
120	34	32	33	30	31	28	32	32	30	78	76	71	83	82	76	81	80	74

Table 6: Results for different numbers of MEL filters. All values in %.

	Accuracy 1						Accuracy 2					
	Excerpts		Ballroom		Joint		Excerpts		Ballroom		Joint	
$F \setminus S$	True	False	True	False	True	False	True	False	True	False	True	False
10	31	32	33	28	32	29	76	76	85	83	81	80
20	31	32	32	26	31	28	77	75	84	84	81	80
40	32	31	33	26	32	28	77	76	83	83	81	80
80	31	29	32	26	32	27	77	75	82	82	80	79

the number of filters has a minor impact on the performance. Apparently, similarly to what was reported for the division of signals into sub-bands [13, 4], the number of Mel filters is not important, as long as they are employed. As regards the value of S, it only had some impact in the Accuracy 1 for the Ballroom; further investigations are needed to understand why this happens.

6 CONCLUSION

Considering the problem of tempo estimation, this paper briefly reviewed several different features employed in the literature, identifying that recently the spectral flux has gained popularity. Considering this feature, different variations and configurations were elicited. Using a common framework that encompasses different ways to compute the spectral flux, the performance of a prototype tempo estimation system when varying this feature calculation method was evaluated.

The main results found were that the use of Mel-frequency mapping improved the result, taking the logarithm before the acquisition of the features made the result worse, and that the use of down-sampling did not affect the Accuracy 1 figure of merit but did have a negative impact on Accuracy 2. The size of the observation window and the hop between adjacent windows was also studied: as expected, longer windows and smaller hops lead to better results (specially when considering Accuracy 2), but the improvement in performance is small face to the increase in computational complexity. The effect of summing the features across different frequency bins prior to tempo estimation was also evaluated. The results have shown that not summing can lead to slightly better results. It should be noted, however, that the choice of summing or not might be highly dependent on the particular tempo estimation algorithm. The effect of the number of Mel-frequency filters was also investigated, with no significant effect found on the system performance (even though it is lowered by not using the filters). Overall, the objective of the tests was to serve as guidelines to researchers employing spectral flux-like features. Moreover, the findings serve as a starting point when investigating which setup works better for specific tempo estimation algorithms.

Of course, the different parameters tested in this paper do not encompass all the variations observed in the literature for the computation of spectral flux-like features. For example, the use of smoothing functions has been reported in more than one paper, as well as the use of time-frequency reassignment. The investigation of the impact of these modifications on performance are left as future work.

ACKNOWLEDGEMENTS

The authors would like to thank the funding agencies CNPq and FAPERJ for financial support of their research projects.

REFERENCES

[1] G. Cooper and L. B. Meyer, *The Rhythmic Structure of Music*, The University of Chicago Press, Chicago, USA, 1960.

[2] M. Levy, "Improving perceptual tempo estimation with crowd-sourced annotations," in *Proc. of the 12th Int. Soc. for Music Information Retrieval Conf.*, Miami, USA, Oct. 2011, pp. 317–322.

[3] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *J. of New Music Research*, vol. 30, no. 1, pp. 39–59, 2001.

[4] F. Gouyon, *A Computational Approach to Rhythm Description: Audio Features for the Computation of Rhythm Periodicity Functions and Their Use in Tempo Induction and Music Content Processing*, PhD Thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2005.

[5] F. Gouyon, S. Dixon, and G. Widmer, "Evaluating low-level features for beat classification and tracking," in *IEEE Int. Conf. on Acoust., Speech and Sig. Proc.*, Apr. 2007, vol. 4, pp. IV 1309–1312.

[6] S. Dixon, "Onset detection revisited," in *Proc. of the 9th Int. Conf. on Digital Audio Effects*, Montreal, Canada, Sep. 2006, pp. 133–137.

[7] J. R. Zapata and E. Gómez, "Comparative evaluation and combination of audio tempo estimation approaches," in *Proc. of the 42nd Int. Conf. of the Audio Eng. Soc.*, Ilmenau, Germany, July 2011.

[8] G. Peeters, "Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal," *IEEE Tran. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 5, pp. 1242–1252, July 2011.

[9] J. L. Oliveira, F. Gouyon, L. G. Martins, and L. P. Reis, "IBT: A real-time tempo and beat tracking system," in *Proc. of the 11th Int. Soc. for Music Information Retrieval*, Utrecht, Netherlands, Oct. 2010.

[10] A. Gkiokas, V. Katsouros, and G. Carayannis, "Ilsp audio tempo estimation algorithm for mirex 2011," in *Proc. of the MIREX*, Miami, USA, Oct. 2011.

[11] A. Gkiokas, V. Katsouros, and G. Carayannis, "Tempo induction using filterbank analysis and tonal features," in *Proc. of the 11th Int. Soc. for Music Information Retrieval Conf.*, Utrecht, Netherlands, Oct. 2010, pp. 555–558.

[12] I. Antonopoulos, A. Pikrakis, and S. Theodoridis, "Self-similarity analysis applied on tempo induction from music recordings," *J. of New Music Research*, vol. 36, no. 1, pp. 27–38, Mar. 2007.

[13] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. of the Acoust. Soc. of America*, vol. 103, no. 1, pp. 588–601, Jan. 1998.

[14] J. Seppänen, A. Eronen, and J. Hiipakka, "Joint beat & tatum tracking from music signals," in *Proc. of the 7th Int. Conf. on Music Information Retrieval*, Victoria, Canada, Oct. 2006.

[15] O. Lartillot, "MIRTEMPO: Tempo estimation through advanced frame-by-frame peaks tracking," in *Proc. of the MIREX*, Utrecht, Netherlands, Oct. 2010.

[16] G. Peeters, "Template-based estimation of time-varying tempo," *EURASIP J. on Advances in Sig. Proc.*, vol. 2007, no. 67215, pp. 1–14, 2007.

[17] M. Alonso, G. Richard, and B. David, "Tempo estimation for audio recordings," *J. of New Music Research*, vol. 36, no. 1, pp. 17–25, Mar. 2007.

[18] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. of New Music Research*, vol. 36, no. 1, pp. 51–60, Mar. 2007.

[19] M. Davies, A. Robertson, and M. Plumbley, "Mirex 2009 audio beat tracking evaluation: Davies, robertson and plumbley," in *Proc. of the MIREX*, Kobe, Japan, Oct. 2009.

[20] G. Tzanetakis, "Marsyas submission to MIREX 2010," in *Proc. of the MIREX*, Utrecht, Netherlands, Aug. 2010.

[21] E. Aylon and N. Wack, "Beat detection using PLP," in *Proc. of the MIREX*, Utrecht, Netherlands, Oct. 2010.

[22] P. Grosche and M. Müller, "A mid-level representation for capturing dominant tempo and pulse information in music recordings," in *Proc. of the 10th Int. Soc. for Music Information Retrieval Conf.*, Kobe, Japan, Oct. 2009, pp. 189–194.

[23] F.-H. F. Wu, "A statistic learning approach to tempo estimation for audio music," in *Proc. of the MIREX*, Miami, USA, October 2011.

[24] B. Schuller, F. Eyben, and G. Rigoll, "Fast and robust meter and tempo recognition for the automatic discrimination of ballroom dance styles," in *IEEE Int. Conf. on Acoust., Speech and Sig. Proc.*, Honolulu, USA, Apr. 2007, vol. 1, pp. 217–220.

[25] F. Eyben and B. Schuller, "Tempo estimation from tatum and meter vectors (MIREX 2010 submission)," in *Proc. of the MIREX*, Utrecht, Netherlands, Aug. 2010.

[26] A. Gkiokas, V. Katsouros, and G. Carayannis, "Audio tempo extraction algorithm for MIREX 2010," in *Proc. of the MIREX*, Utrecht, Netherlands, Oct. 2010.

[27] S. Böck and M. Schedl, "Enhanced beat tracking with context-aware neural networks," in *Proc. of the 14th Int. Conf. on Digital Audio Effects*, Paris, France, Sep. 2011, pp. 135–139.

[28] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *IEEE Tran. on Audio, Speech, and Lang. Proc.*, vol. 14, no. 5, pp. 1832 –1844, Sep. 2006.

NUNES, L. O.; BISCAINHO, L. W. P. Tempo estimation: evaluation of different spectral flux computation methods.
In: CONGRESSO DE ENGENHARIA DE ÁUDIO AES BRASIL, 10., 2012, São Paulo, *Anais*... Rio de Janeiro: Sociedade de Engenharia de Áudio, 2012, p. 125-132.

ÍNDICE DE AUTORES / *AUTHORS INDEX*

Alves, D. R.: 113
Antunes Júnior, I.: 107
Avila, A. R.: 61

Basso, G. J.: 13
Biscainho, L. W.: 28, 125
Bordin Jr., C. J.: 107
Brandao, M. C. P.: 66

Dalcastagnê, A. L.: 55

Faria, R. R. A.: 98
Farias, M. C. Q.: 66
Fornari, J.: 41
Fraga, F. J.: 61

Goldemberg, R.: 82

Jesus, A. L. M.: 113

Kamekawa, T.: 121
Kruszielski, L. F.: 121

Lazzarini, V. : 3

Lima, J. A. S.: 66

Marui, A.: 121
Mendes, R. S.: 82, 90
Moraes, E. : 19

Noceti Filho, S.: 55
Nunes, L. O.: 28, 125

Oliveira, L. J.: 47,
Oliveira, L. C.: 82

Portela, M. S. : 1

Roque, T. R.: 90

Santos, R. S.: 47
Silva, G. K.: 55
Sousa, F. W. J.: 74

AUDIO ENGINEERING SOCIETY

BRAZIL SECTION