Anais do 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil

Proceedings of th 5th AES Brazil Conference
11th AES Brazil National Convention

INOVAÇÃO EM ÁUDIO
AUDIO INNOVATION

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

Coordenador da Convenção | *Convention Chair*
Joel Brito

Editado por | *Edited by*
Regis Rossi Alves Faria

AUDIO AES

Audio Engineering Society - Seção Brasil

Comitê Organizador

Coordenador Geral da Convenção:
Joel Brito (Presidente AES Brasil)

Coordenador Geral do Congresso:
Regis Rossi A. Faria (LSI-EPUSP)

Coordenador do Comitê de Programa:
Luiz W. P. Biscainho (Poli & COPPE-UFRJ)

Coordenador de Artigos e Pôsteres:
Fábio P. Freeland (COPPE-UFRJ)

Programação CD-ROM:
Tadeu Magalhães

Programação LaTeX:
Sálvio Marcelo Soares

Apoio logístico:
Leandro Ferrari Thomaz
Rosilene Louro

Editoração e arte:
Lídia Brito
Totum Marketing e Comunicação

Comitê de Programa Técnico:
Adolfo Maia Jr. (Imecc-UNICAMP)
Aníbal J. de S. Ferreira (Univ. do Porto, Portugal)
Eduardo R. Miranda (Univ. Plymouth, UK)
Fábio P. Freeland (COPPE-UFRJ)
Fernando Iazzetta (ECA-USP)
Francisco J. Fraga da Silva (UFABC)
Hugo B. de Paula (PUCMinas)
Jayme G. A. Barbedo (FEEC-UNICAMP)
João A. Zuffo (EP-USP)
Jônatas Manzolli (IA-UNICAMP)
Julio C. B. Torres (Poli-UFRJ)
Luiz W. P. Biscainho (Poli & COPPE-UFRJ)
Marcelo G. de Queiroz (IME-USP)
Márcio H. de Avelar Gomes (FEC-UNICAMP)
Mariane R. Petraglia (Poli & COPPE-UFRJ)
Miguel A. Ramírez (EP-USP)
Paulo A. A. Esquef (INdT/AM)
Pedro F. Donoso-Garcia (EE-UFMG)
Phillip M. S. Burt (EP-USP)
Regis Rossi A. Faria (LSI-EPUSP)
Rodrigo C. Velloso (EM-UFRJ)
Rubem D. R. Fagundes (FENG-PUCRS)
Sergio L. Netto (Poli & COPPE-UFRJ)
Sidnei Noceti Filho (CTC-UFSC)

Agradecimentos:
AES Board of Governors
Andres Mayo (VP AES LatinoAmerica)
Silvia Regina Saran Della Torre

Copyright © 2007
Audio Engineering Society – Brazil Section

Congresso de Engenharia de Áudio (5.: São Paulo: 2007); Convenção Nacional AES Brasil (11.: São Paulo: 2007)
Anais 5. Congresso de Engenharia de Áudio; 11. Convenção Nacional AES Brasil / ed. R.R.A. Faria - São Paulo: AES Brasil, 2007.
1 CD-Rom.

ISBN 978-85-99997-02-4

1. Engenharia de áudio (Congressos) 2. Computação musical (Congressos) 3. Processamento de sinais (Congressos)
I. Convenção Nacional AES Brasil (11.: São Paulo, 2007) II. Áudio Engineering Society. Seção Brasil III. Faria, Regis Rossi Alves IV. t

CDD621.3828

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

Impresso no Brasil.
Printed in Brazil.

Realização:

Audio Engineering Society - Seção Brasil

Organização:

Laboratório de Sistemas Integráveis da Escola Politécnica da USP

Apoio:

Sociedade de Engenharia de Áudio

AES – Audio Engineering Society – Brazil Section

Rua Carlos Machado 164, sala 305, Pólo Rio de Cine e Vídeo – Barra da Tijuca

Rio de Janeiro, Brasil – Cep. 22775-042 | e-mail: aesbrasil@aes.org | www.aesbrasil.org

telefone: +55(21) 2421-0112 | fax: +55(21)2421-0112

Administração

Presidente/Chairman: Joel Brito
Vice-Presidente/Vice-Chairman: Franklin G. Leite
Secretário/Secretary: Carlos Ronconi
Tesoureiro/Treasurer: Guilherme Figueira
Comição/Committemen: Luiz Wagner Biscainho
Luiz Campos Reis
João Américo Bezerra
José Pereira Jr.
Homero Sette Silva

Audio Engineering Society, Inc.

International headquarters

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

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

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

Sumário

Contents

Prefácio dos Organizadores / Organization Greetings 7

Revisores / Reviewers 9

Sessões de Artigos / Papers Sessions

Sessão 1 - Processamento de áudio, voz, transformações e efeitos
(Audio and voice processing, transforms and effects)

1. A sequential system for voice pitch modification
Rafael C. D. Paiva, Luiz W. P. Biscainho e Sergio L. Netto 11

2. Wavelet-based audio watermarking
Fabrício Lopes Sanches, Lucimar Sasso Vieira,
Sylvio Barbon Júnior e Rodrigo Capobianco Guido 17

3. Técnicas para separação cega de fontes aplicadas na melhoria de desempenho de um classificador de palavras isoladas
Diego Barreto Haddad, Mariane Rembold Petraglia
e Paulo Bulkool Batalheiro 20

4. Wavelets aplicadas à Dynamic Time Warping
Sylvio Barbon Júnior, Rodrigo Capobianco Guido,
Lucimar Sasso Vieira e Fabrício Lopes Sanchez 28

5. On the use of wavelets for voice conversion
Lucimar Sasso Vieira, Fabrício Lopes Sanchez,
Sylvio Barbon Junior e Rodrigo Capobianco Guido 33

6. Estudos teórico e prático sobre o emprego da transformada wavelet contínua na detecção passiva em águas rasas
José Roberto Motta da Silva, Mariane Rembold Petraglia
e Carlos Eduardo Parente Ribeiro 39

Sessão 2 - Análise, síntese, percepção auditiva e computação musical
(Analysis, synthesis, auditory perception and computer music)

7. Transcrição automática de sinais de áudio monofônico baseada em quadro do Artigo
Tiago Fernandes Tavares, Jayme Garcia Arnal Barbedo e Amauri Lopes 47

8. Modelo empírico da sonoridade da clarineta: análise e síntese sonora
Luís Carlos de Oliveira, Fábio Furlanete, Ricardo Goldemberg e Jônatas Manzolli 51

9. Avaliação do Bounded-Q Fast Filter Bank na transcrição automática de música
Filipe Castello da Costa Beltrão Diniz, Luiz W. P. Biscainho e Sergio L. Netto 58

10. Evaluation of threshold-based algorithms for detection of spectral peaks in audio
Leonardo de Oliveira Nunes, Paulo Antônio A. Esquef e Luiz W. P. Biscainho 66

11. Avaliação de uma representação compacta por “Matching Pursuit” em um esquema de análise/síntese de áudio
Michel P. Tcheou, Lisandro Lovisolo, Luiz W. P. Biscainho e Eduardo Antônio Barros da Silva 74

Sessão 3 - Áudio espacial, sonorização multicanal e acústica de ambientes
(Spatial audio, multichannel sound and environmental acoustics)

12. Modelagem eficiente de HRTFs por parametrização espacial
Arnaldo S. Gunzi, Fábio P. Freeland, Luiz W. P. Biscainho e Paulo S. R. Diniz 82

13. Desenvolvimento experimental de amplificadores multicanal para auralização
Renan Santos Vital e Regis Rossi A. Faria 89

14. Implementação do Método do Traçado de Raios e das Fontes Virtuais para a simulação acústica de salas
Márcio H. de Avelar Gomes, João G. Dedecca e Rhenan G. do Nascimento 97

Sessão 4 - Tratamento, restauração e avaliação de qualidade de áudio
(Audio treatment, restauration and quality assessment)

15. Plugin VST para remoção de ruído de fundo
Alan Freihof Tygel e Luiz Wagner Pereira Biscainho *105*

16. O PEAQ aplicado a sinais de áudio contaminados por ruído
Rafael A. de Jesus e Luiz W. P. Biscainho *113*

17. Sistemas de áudio no controle ativo de ruído
Bruno S. Masiero e Vítor H. Nascimento *119*

18. Fast Distortion vs. Level Measurements with smoothly increasing Signal Amplitudes
Swen Müller e Nelson Melo do Espírito Santo *127*

19. Correção de clicks através de métodos Bayesianos computacionalmente eficientes
Flávio R. Ávila e Luiz W. P. Biscainho *132*

Sessão 5 - Equalização, gerenciamento e mixagem
(Audio equalization, management and mixing)

20. Harmonic Equalizer
Andre Luiz Luvizotto, César Rennó Costa, Jonatas Manzolli, Rafael Santos Mendes e Fernando Von Zuben *138*

21. Gerenciador digital com alta faixa dinâmica
Swen Müller *143*

22. Designing the delay out of digital mixing systems
Robin Clark *151*

Índice de Autores I *Author Index* *158*

Prefácio dos Organizadores

Nesta edição do congresso e convenção da AES Brasil abordamos o tema Inovação nas áreas de engenharia de áudio, música, acústica, na indústria de produtos e serviços, na cadeia de produção de conteúdo audiovisual, nas telecomunicações e em várias outras áreas afins que trabalham ou utilizam áudio.

Mas afinal o que seria inovação em áudio? Como podemos estimular a pesquisa, o desenvolvimento de novos produtos e negócios nas áreas de áudio? Como estão alinhadas as linhas de produtos e de pesquisas no Brasil com as tendências tecnológicas e empresariais? Como os resultados de pesquisas e os desenvolvimentos inéditos são incorporados na cadeia produtiva? Quais ações são necessárias para se fazer inovação e revertê-la em bens para a sociedade?

Organizamos este ano um painel sobre Inovação em Áudio, que conta com a participação de especialistas no tema, que irão traçar um panorama focalizado no segmento de áudio. Nossa intenção é estimular os empreendedores e a indústria nacional a descobrirem como podem investir em inovação e não simplesmente trazerem soluções importadas para o país.

Os 22 artigos publicados nesta edição mostram a qualidade e a diversidade da nossa produção científica em tantas tecnologias e aplicações voltadas para o áudio, a música e a voz. Neste ano observamos uma ênfase dos trabalhos em processamento de sinais de áudio, voz, análise e síntese musical, áudio espacial, multicanal e codificação sonora, seguindo tendências semelhantes observadas nos grandes eventos e congressos internacionais.

Desejamos a todos os congressistas os melhores momentos para elaborarem novos contatos, absorverem novos conhecimentos, e que aproveitem a oportunidade para incorporarem a inovação em seu trabalho.

Comissão Organizadora
AES Brasil 2007

Organization Committee Greetings

In this AES Brazil convention edition we approach the theme Innovation in the areas of audio engineering, music, acoustics, in the software/hardware and service industries, in the audiovisual production chain, in telecommunications and many other related areas which work or use audio.

But, after all, what would be innovation in audio? How can we stimulate the research, the development of new products and business in the areas of audio? How are the research and product lines in Brazil and other countries aligned with technological and market trends? How are research results and novel developments incorporated in the production chain? Which actions are necessary to produced innovation and revert it to social assets?

We organized this year a panel about Innovation in Audio, which counts on the contribution of specialists, who will sketch a panorama focused on the audio segments. Our intention is to stimulate entrepreneurs and our domestic industry to discover how to invest in innovation and not simply bring imported solutions to the country.

The 22 papers published in this edition show the quality and diversity of our scientific production in many technologies and applications devoted to audio, music and voice. This year we observe an emphasis of works in audio signal processing, musical analysis and synthesis, spatial audio, multichannel and sound coding, following similar trends observed in large international conferences and events around the globe.

We wish all congress attendants the best moments to elaborate new contacts, to absorb new knowledge, and that take the opportunity to incorporate the innovation matter in their jobs.

Organization Committee
AES Brazil 2007

Revisores

Reviewers

Adolfo Maia Jr. (UNICAMP)
Anibal Ferreira (Universidade do Porto)
Eduardo Miranda (University of Plymouth)
Fábio Freeland (UFRJ)
Fernando Iazzetta (USP)
Francisco Fraga da Silva (UFABC)
Hugo de Paula (PUC Minas)
Jayme Barbedo (UNICAMP)
Juan R. V. Lopez (UFSC)
João Antonio Zuffo (USP)
Jônatas Manzolli (UNICAMP)
Julio Torres (UFRJ)
Luiz Biscainho (UFRJ)
Marcelo Queiroz (USP)
Mariane Petraglia (UFRJ)
Márcio Gomes (UNICAMP)
Mário Minami (USP)
Miguel Ramírez (USP)
Paulo Antonio Esquef (Inst. Nokia de Tecnologia)
Pedro Donoso-Garcia (UFMG)
Phillip Burt (USP)
Regis Rossi A. Faria (USP)
Rodrigo Velloso (UFRJ)
Rubem Fagundes (PUCRS)
Sergio Netto (UFRJ)
Sidnei Noceti Filho (UFSC)
Walter Antônio Gontijo (UNIVALI)

Sessões de Artigos

Papers Sessions

Sessão 1

Processamento de áudio, voz, transformações e efeitos

(Audio and voice processing, transforms and effects)

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

A Sequential System for Voice Pitch Modification

Rafael C. D. Paiva,[1] Luiz W. P. Biscainho,[1] and Sergio L. Netto[1]

[1] LPS PEE-COPPE, DEL-Poli,UFRJ Caixa Postal 68504 - Rio de Janeiro, RJ, 21941-972, Brasil

rcdpaiva@lps.ufrj.br, wagner@lps.ufrj.br, sergioln@lps.ufrj.br

ABSTRACT

This paper presents a system for pitch modification of voice signals. In the proposed structure, a recursive least-squares (RLS) adaptive algorithm is employed to determine the linear prediction (LP) model for the vocal tract during the analysis stage. In the subsequent synthesis, an excitation signal with the desired pitch contour is processed by the LP model obtained previously. The method avoids the frame delay inherent to block-processing techniques as well as discontinuities in the LP model. Listening tests evaluate positively the quality of the synthesized signal.

0 INTRODUCTION

Pitch modification and voice transformation have been major subjects for speech processing research in the last years. Perhaps the main reason for this continuous interest is the fact that the human perception of voice is very accurate. Therefore, practical processing of voice signals may require quite intricate techniques for a proper pitch shifting procedure.

Applications of pitch modification systems include voice editing for movies, automatic tuning of musical signals, concatenative synthesis of voice [1], voice morphing [2, 3] or even esophageal voice enhancement [4].

Voice production is often modeled as a source-filter system [5]. In this model, the *excitation* or source signal corresponds to the vocal folds (vibratory or not) output, and appears in guise of either a pseudo-periodic or a random-like waveform. Pseudo-periodic excitation, which can be characterized by a fundamental frequency value, leads to voiced phonemes, whereas random-like excitation leads to unvoiced phonemes. Some expressive information can be also conveyed by the excitation [6, 7]. The *filter*, which represents the vocal tract characteristics (mouth and tongue positions, lips opening, articulatory elements, *etc.*) processes the excitation information to ultimately generate the voice signal. The vocal tract is responsible for the discrimination of different phonemes, which present frequency patterns, often called spectral envelope, that do not change much among different speakers [5], even for distinct excitation characteristics.

The first experiences in pitch modification of speech signals were implemented by speed changes in recorded

signals. Of course, this approach could not be performed without changing the duration of the original signal. The main drawback of this technique was to change the speaker's timbre in a very annoying way, since it distorted the spectrum envelope of the original signal. Those experiences showed that the spectral envelope should be preserved for high quality time and pitch scaling of speech signals.

A more flexible approach is based on the direct spectrum envelope modeling using the *discrete Fourier transform* (DFT), which may lead to the so-called *phasiness* distortion when synthesis blocks are simply put together without proper care. The *phase vocoder* [8] belongs to this class of solution. Another idea developed from speech coding techniques: the pitch shifting can be performed over a *linear prediction* (LP) model [9].

While exhibiting great advantages when compared to primitive speed changing approaches, all the solutions listed before operate on a block-by-block basis: the signal is segmented by an adequate window, the blocks are processed and combined by an *overlap-and-add* (OLA) method. Improvements like the use of variable-length blocks and pitch-synchronization led to very successful methods, including the popular *pitch synchronous OLA* (PSOLA) and its variants [10].

This work proposes applying a sample-oriented solution, *viz.* the *recursive least-squares* (RLS) adaptive algorithm, to determine the LP filter in a pitch shifting system. Section 1 reviews and compares one classic block solution and the sequential RLS solution for LP modeling. Section 2 describes the overall system proposed, and Section 3 illustrates pitch modification using the new technique through practical examples. Section 4 concludes the paper by discussing some performance issues of the system.

1 Linear Prediction Modeling

Given a speech signal $s[n]$, suppose one wishes to obtain a set of P filter coefficients a_p such that

$$\hat{s}[n] = \sum_{p=1}^{P} a_p s[n-p] \tag{1}$$

is a good estimate of $s[n]$. The estimation (prediction) error is defined as

$$e[n] = s[n] - \hat{s}[n], \tag{2}$$

which can be seen as the result of submitting $s[n]$ to a filter with transfer function

$$G(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_P z^{-P}. \tag{3}$$

Conversely, Equations (1) and (2) can describe the so-called *autoregressive* (AR) model, in which $s[n]$ is interpreted as the result of filtering white noise by a filter with transfer function $H(z) = 1/G(z)$. In that sense, the model order P must be high enough to allow the excitation signal (formerly prediction error) $e[n]$ to be considered white. In speech processing techniques, this assumption is not particularly useful, since during voiced portions of speech the voice production model assigns to the excitation the shape of a pseudo-periodic pulse train. In fact, using a reduced-order predictor for voiced speech yields the expected pulsed $e[n]$, with almost equal-amplitude harmonics. Usually this kind of modeling is referred to as *linear prediction coding* (LPC). It leads to an economic way of speech storage and transmission.

There are several practical approaches to obtain the LPC coefficients of a given signal. The most popular and widely used in speech processing work in blocks and are often called just by LPC methods. Classic block solutions include the autocorrelation and the covariance methods [11]. Estimation of LPC coefficients can also be made through adaptive filtering techniques.

The speech spectral envelope is estimated by the LPC model, but it can alternatively be obtained by frequency or cepstrum transforms, as well as other techniques.

The classic autocorrelation block solution and the adaptive RLS solution are revisited in the next subsections.

1.1 Classic Block Solution

The block solution for LPC modeling is performed on a time frame of N samples, considering the signal to be modeled is ergodic and wide-sense stationary (WSS) within this frame.

The error function for this solution is given by

$$e_m[n] = s_m[n] - \widehat{s}_m[n] = s_m[n] - \mathbf{a}_{m,P}^T \mathbf{s}_m[n-1], \tag{4}$$

where

$$\mathbf{a}_{m,P} = (a_{m,1}\ a_{m,2}\ \ldots\ a_{m,P})^T, \tag{5}$$

$$\mathbf{s}_m[n-1] = (s_m[n-1]\ s_m[n-2]\ \ldots\ s_m[n-P])^T, \tag{6}$$

with the superscript T denoting matrix transposition.

Considering $n = (P+1), (P+2), \ldots, N$, Equation (4) can be written in matrix form as

$$\mathbf{e}_m = \mathbf{d}_m - \mathbf{S}_m \mathbf{a}_{m,P}, \tag{7}$$

where

$$\mathbf{e}_m = (e_m[N]\ e_m[N-1]\ \ldots\ e_m[P])^T, \tag{8}$$

$$\mathbf{d}_m = (s_m[N]\ s_m[N-1]\ \ldots\ s_m[P])^T, \tag{9}$$

and $\mathbf{S}_m$ is an $(N-P) \times P$ matrix defined as

$$\mathbf{S}_m = \begin{pmatrix} s_m[N-1] & s_m[N-2] & \ldots & s_m[N-P] \\ s_m[N-2] & s_m[N-3] & \ldots & s_m[N-P-1] \\ \vdots & \vdots & \ddots & \vdots \\ s_m[P] & s_m[P-1] & \ldots & s_m[0] \end{pmatrix} \tag{10}$$

Minimizing the mean squared-error $\mathbf{e}_m^T\mathbf{e}_m$, the optimal coefficient vector is given by

$$\mathbf{a}_{m,P} = \mathbf{R}_{D,m}^{-1}\,\mathbf{p}_{D,m}, \tag{11}$$

where $\mathbf{R}_{D,m} = \mathbf{S}_m^T\mathbf{S}_m$ is the so-called deterministic correlation matrix of the signal s and $\mathbf{p}_{D,m} = \mathbf{S}_m^T\mathbf{d}_m$ is the deterministic cross-correlation vector. For an ergodic WSS signal, the matrix $\mathbf{R}_{D,m}$ is Toeplitz, and efficient methods for calculating its inverse are available.

1.2 Sequential RLS Solution

The LP modeling using RLS adaptive filtering is made on a sample-by-sample basis. The objective function for the *weighted least-squares* (WLS) formulation is given by

$$\xi_{RLS}[n] = \sum_{i=0}^{n} \lambda^{n-i} e^2[i], \tag{12}$$

where $0 << \lambda < 1$ is the so-called forgetting factor. Equation (12) can be rewritten as

$$\xi_{RLS}[n] = \mathbf{e}^T[n]\Lambda^2[n]\mathbf{e}[n], \tag{13}$$

where

$$\mathbf{e} = (e[n]\; e[n-1]\; \ldots\; e[0])^T, \tag{14}$$
$$\Lambda^2[n] = \text{diag}\left(1, \lambda, \lambda^2, \ldots, \lambda^{n-1}\right). \tag{15}$$

Minimizing $\xi_{RLS}[n]$, the optimal coefficient vector is given by

$$\mathbf{a}_P[n] = \mathbf{R}_D^{-1}[n-1]\,\mathbf{p}_D[n], \tag{16}$$

with

$$\mathbf{R}_D[n-1] = \mathbf{S}^T[n-1]\Lambda^2[n]\mathbf{S}[n-1], \tag{17}$$
$$\mathbf{p}_D[n-1] = \mathbf{S}^T[n-1]\Lambda^2[n]\mathbf{d}[n], \tag{18}$$

where

$$\mathbf{d} = (s[n]\; s[n-1]\; \ldots\; s[0])^T. \tag{19}$$

For the RLS algorithm, the inverse matrix in Equation (16) is determined recursively as

$$\mathbf{R}_D^{-1}[n-1] = \frac{1}{\lambda}\left[\mathbf{R}_D^{-1}[n-2] - \frac{\Psi[n]\Psi^T[n]}{\lambda + \Psi^T[n]\mathbf{s}[n-1]}\right], \tag{20}$$

with

$$\Psi[n] = \mathbf{R}_D^{-1}[n-2]\mathbf{s}[n-1], \tag{21}$$
$$\mathbf{p}_D[n] = \lambda\mathbf{p}_D[n-1] + s[n]\mathbf{s}[n-1]. \tag{22}$$

In practice, one often uses $\mathbf{R}_D^{-1}[-1] = \delta\mathbf{I}$, where δ is a small positive constant, and $\mathbf{p}_D[0]$ is a null vector. For further details on the implementation of the RLS algorithm the reader may refer to [12].

1.3 Comparison of Block and Sequential Solutions

The classic autocorrelation and the sequential RLS solutions are closely related. For instance, one can rewrite the autocorrelation solution as

$$\mathbf{a}_{m,P} = (\mathbf{S}^T[n-1]\Lambda'_m[n]\mathbf{S}[n-1])^{-1}\mathbf{S}^T[n-1]\Lambda'_m[n]\mathbf{d}[n], \tag{23}$$

which closely resembles the RLS solution given in Equation (16) with

$$\Lambda'_m[n] = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{N\times N} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{24}$$

This indicates that both solutions are built around deterministic correlation matrices and vectors including weighting terms: $\Lambda'_m[n]$, in the LS solution of the block method; $\Lambda[n]$, in the WLS solution of the sequential method. Furthermore, one can see that both methods consider only part of the signal at a given time instant: the classic autocorrelation method uses a square window, whereas the RLS algorithm applies an exponential window over past samples, resulting in a smaller contribution of older observations of $\mathbf{s}[n]$.

Advantages of the RLS algorithm include avoiding the inversion of the correlation matrix, by employing the recursive estimate given by Equation (20), and allowing an estimate of the optimal coefficient vector for each time instant n, instead of working in blocks as in the classic method.

In order to provide a quick comparison between the classic block and RLS methods, we consider a low-order LP modeling of a segment of nonstationary speech. Figure 1 shows the pole tracking in the z-plane for each method: (a) Classic block model using 20-ms frames with no time overlap between consecutive frames; (b) The same as (a), but with a 5-ms linear interpolation on the LP models; (c) RLS model. In these figures, the filled circles indicate initial positions of the LP model and the rhombus marks correspond to the final positions. From these plots, one can observe a similar path being followed by all three schemes, with the RLS yielding the trajectory with better continuity.

2 Proposed System

The proposed pitch modification algorithm uses the LP model $\mathbf{a}_P$ obtained with by RLS algorithm to synthesize the voice signal $s'[n]$ with the desired pitch information, as indicated in Figure 2, by

$$s'[n] = e'[n] - \mathbf{a}_P^T[n]\mathbf{s}'[n-1], \tag{25}$$

where

$$\mathbf{s}'[n-1] = (s'[n-1]\; s'[n-2]\; \ldots\; s'[n-P])^T. \tag{26}$$

In practice, for low sampling frequency (around 8 kHz) one must use a model order $10 \leq P \leq 15$, and

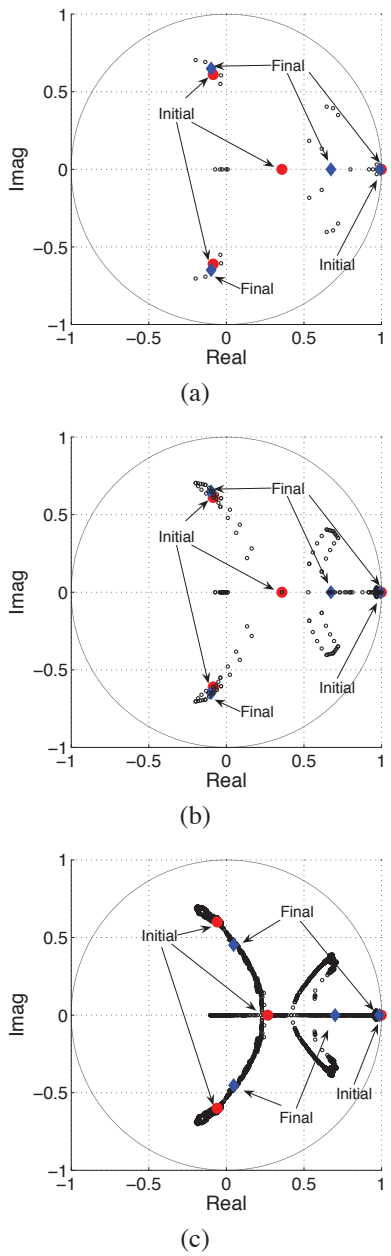


Figure 1: Pole evolutions: (a) Block algorithm with 20-ms frame; (b) Block algorithm with 20-ms frames, with 1:4 coefficient interpolation; (c) RLS algorithm.

for high quality audio (sampling frequencies between 30 and 44.1 kHz) a larger P is required.

A desired pitch-period contour $p'[n]$ can be determined from a scaling transformation $\beta[n]$ on the original pitch period $p[n]$. To do that, $p[n]$ and the corresponding pitch marks $p_m[n]$, associated to the vocal folds closures, as illustrated in Figure 3, must be known.

For that purpose, the pitch period $p[n]$ can be calculated using classic autocorrelation method [5], cepstrum-based method [13], the YIN estimator [14], and event-based techniques. The latter methods look

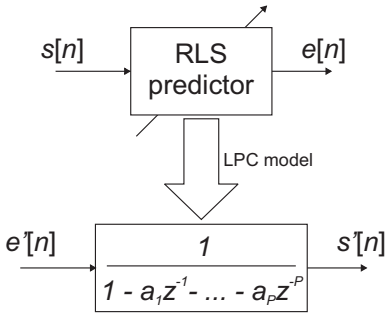


Figure 2: Analysis and synthesis scheme using the RLS algorithm.

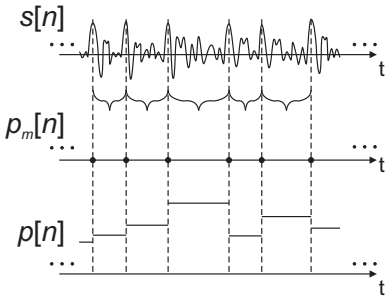


Figure 3: Example of pitch marking and associated pitch detection.

for periodicity in the amplitude envelope, and try to find pitch marks that would correspond to the instants of significant glottal excitation. Such methods identify each pitch period individually, requiring precise heuristic information on the signal at hand. Implementation of event-based techniques for speech includes wavelets [15], characteristics of autocorrelation matrix [16], and simpler methods like amplitude envelope calculation.

New pitch marks $p'_m[n]$ are then obtained incorporating the new pitch-period information $p'[n]$, as represented in Figure 4, where c is an auxiliary sample

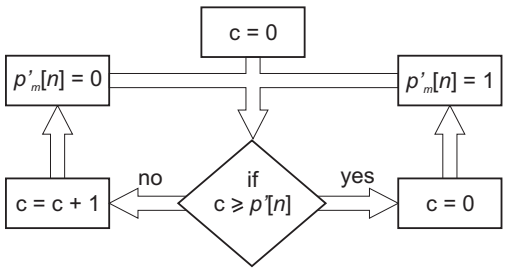


Figure 4: Search algorithm for modified pitch marks $p'_m[n]$ according to modified pitch period $p'[n]$.

counter. If $p'[n]$ is not an integer number, it may be rounded to its closest integer.

The excitation signal $e'[n]$ can then be formed by

concatenating several copies of a standard pitch interval of the RLS residual signal $e[n]$ according to the new pitch marks $p'_m[n]$. Alternative approaches include using a typical model of the glotal pulse [6, 7] or combining a pitch period of the LP residual with the PSOLA [10] algorithm. The overall procedure for composing $e'[n]$ is depicted in Figure 5, where $\beta[n]$ can be made variable for a wide range of pitch-modification scenarios.

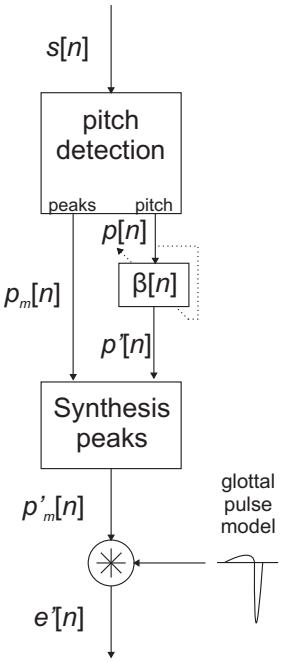


Figure 5: Synthesis of the modified excitation signal.

3 Experimental Results

The proposed system was implemented with $P = 30$ RLS coefficients and $\lambda = 0.999$ for a signal sampled at 44.1 kHz. Figure 6 shows the results for a voiced segment of a spoken phrase and Figure 7 shows the correspoding spectrogram. From these figures, one can clearly see that the spectral envelope is preserved in the generated signals, whereas the pitch information is modified as desired. Listening tests have validated the quality of the overall result for pitch scaling factors in the 2-octave range $0.5 \leq \beta \leq 2$.

4 Conclusions

This paper presented a pitch shifting scheme using adaptive filtering techniques for spectral envelope estimation. The adaptive RLS filter has proved to be a good solution when compared to conventional block solutions for estimating of the LPC model of speech signals, since it guarantees model smoothness through time.

The work has illustrated the applicability of this system on a pitch scaling procedure, which is the basis of automatic tuning, solo to unison transformation

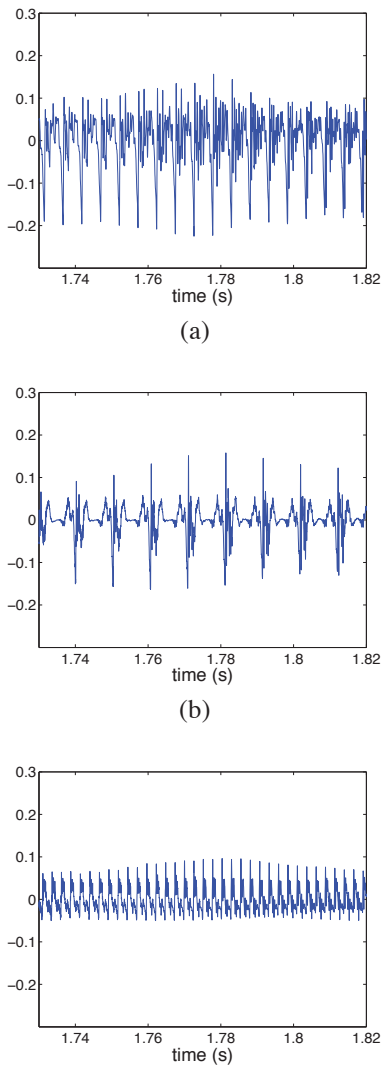


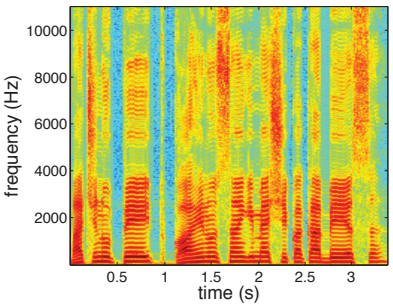
Figure 6: Time plot of: (a) Original signal; (b) Modified signal with lower pitch, $\beta = 2$; (c) Modified signal with higher pitch, $\beta = 0.5$.

or prosody transposition. Listening tests have verified the overall quality of the algorithm for a wide range of pitch transformation.

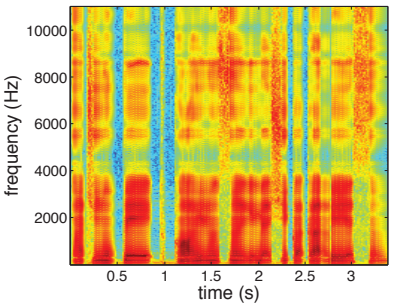
The next step of this work is to perform the systematic comparison of the proposed technique with others available in the literature. For this purpose, a wider voice corpus, including speakers and singers of different gender and age as well as different phonation types, will be employed.

5 Acknowledgments

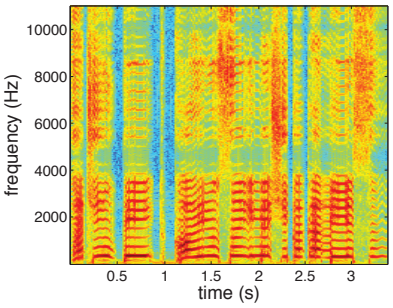
The authors would like to thank the Brazilian sponsors of this work *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) and *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*(FAPERJ) for their support.



(a)



(b)



(c)

Figure 7: Spectrogram of: (a) Original signal; (b) Modified signal with lower pitch, $\beta = 2$; (c) Modified signal with higher pitch, $\beta = 0.5$.

REFERENCES

[1] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 67–79, March 2007.

[2] L. Fabig and J. Janer, "Transforming singing voice expression - the sweetness effect," in *Proc. of the DAFx04 - 7th International Conference on Digital Audio Effects*, Naples, Italy, October 2004.

[3] A. Loscos and J. Bonada, "Emulating rough and growl voice in spectral domain," in *Proc. of the DAFx04 - 7th International Conference on Digital Audio Effects*, Naples, Italy, October 2006.

[4] J. Bonada and A. Loscos, "Esophageal voice enhancement by modeling radiated pulses in frequency domain," *121st Audio Engineering Society Convention*, October 2006, Preprint 6952.

[5] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, Wiley-IEEE, 1999.

[6] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 492–501, March 2006.

[7] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, February 1990.

[8] J. Laroche and M. Dolson, "Phase-vocoder: about this phasiness business," in *Proc. of the WASPAA'97 - Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, October 1997, IEEE.

[9] J. A. Moorer, "The use of linear prediction of speech in computer music applications," *Journal of Audio Engineering Society JAES*, vol. 27, no. 3, pp. 134–140, March 1979.

[10] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, no. 2, pp. 175–205, February 1995.

[11] J. Makhoul, "Linear prediction: a tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[12] P. S. R. Diniz, *Adaptive Filtering: Algorithms and Practical Implementations*, Kluwer, 2 edition, 2002.

[13] A. M. Noll, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293 – 309, February 1967.

[14] A. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, April 2002.

[15] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 917–924, March 1992.

[16] C. Ma, Y. Kamp, and L. F. Willems, "A frobenius norm approach to glottal closure detection from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 258–265, April 1994.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Wavelet-based Audio Watermarking

Sanchez, F.L.[1], Vieira, L.S.[2], Barbon, S.[3], Guido, R. C.[4]
[1]School of Engineering of São Carlos
São Carlos, São Paulo, 13566-590, Brazil
[2,3,4]Institute of Physics of São Carlos
São Carlos, São Paulo, 13566-590, Brazil
fabriciosanchez@ieee.org, guido@ifsc.usp.br, sassovieira@ursa.ifsc.usp.br, barbon@ursa.ifsc.usp.br

ABSTRACT

In this work, we present a novel algorithm for watermarking embedding in digital audio files. The proposed algorithm, that is based on the discrete wavelet transform (DWT) and artificial neural networks (ANN), presents satisfactory results. Implemented in C++ language, it works under Linux, Windows XP Professional SP2 and Windows Vista.

0 INTRODUCTION

The rapid development of network communications and new Internet technologies, such as ADSL, wireless, among others, has brought the possibility to generate and copy digital media files in a significantly easy way. Based on this, digital watermarking [1] can be effectively used to protect against piracy, achieving the authenticity of media content.

This work it presents a novel algorithm for digital audio watermarking that is based on the discrete wavelet transform (DWT) and artificial neural networks (ANNs). The embedded signature preserves the content of the file and is not perceivable by the listener [2].

1 BIBLIOGRAPHICAL REVIEW

1.1 DWT – Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT), which has as its main idea the process of multi-resolution analysis (MRA) [3][4][5][6], is one of the most appropriate techniques to make a joint time-frequency analysis of a discrete-time signal. considering S as the discrete-time signal under analysis, it may be decomposed in j levels of resolution, as follows:

$$\vec{S}=\vec{A}\,j+\sum_{i=1}^{j}\vec{D}\,i\,, \qquad (1)$$

where each vector A_s (reconstructed signal of approximation – low-frequencies) is orthogonal to the vector D_s (reconstructed signal of detail – high-frequencies). More specifically, it is possible to write each point i of the signal S, containing N samples, as follows:

$$S(i)=\sum_{k=0}^{\frac{N}{2^j}-1}H_{j,k}(i)\phi_{j,k}(i)+\sum_{l=1}^{j}\sum_{k=0}^{\frac{N}{2^j}-1}G_{l,k}(i)\psi_{l,k}(i), \qquad (2)$$

where,

$$\varphi(i)=\sum_{k}h_i\varphi(2i-k) \qquad , (3)$$

$$\psi(i) = \sum_k g_i \phi(2i-k) \quad , (4)$$

$$H_{j,k}(i) = < S, \phi_{j,k}(i) > \quad , (5)$$

and

$$G_{j,k}(i) = < S, \psi_{j,k}(i) > \quad , (6)$$

In this case, $\phi(i)$, defined by dilatations and translations of itself, is called scaling function, and $\psi(i)$ is called wavelet function, which is orthogonal to the scaling function. The h_i coefficients corresponds to a low-pass filter and the g_i coefficients to a high-pass filter, forming a Quadrature Mirror Filter bank (QMF) [7]. From a signal processing point of view, the DWT can be considered as the output of band pass filters whose bandwidth and the center frequency are proportional to $\frac{1}{2^j}$ for each scale *j*, and, for time information analysis, it is possible to show that the temporal resolution increases while *j* decreases. After each level of decomposition, the filtered signal is downsampled by 2, which allows only half band of the original signal to pass. In order to reconstruct the original signal *S* without loss of information, an upsampling by 2 is used and *S* can be determined as in equation (1). The most common family of wavelets, created by Ingrid C. Daubechies, well described in references [8][9], has its low-pass filter coefficients, h_i, determined by solving the following system of equations:

$$\begin{cases} \sum_{k=0}^{N-1} (-1)^k h_k k^b = 0 \\ \sum_{k=0}^{N-1} h_k = \sqrt{2} \\ \sum_{k=0}^{N-1} h_k h_{k+2l} = 2\delta_{0,l} \end{cases} \quad (7)$$

where *N* is the filter's support (length), $b = 0, 1, ..., (N/2) - 1$, $\delta_{0,l}$ is the Kronecker's delta and $l \in Z$. The first equation is responsible for producing *N/2* vanishing moments, the second one comes from the fact that the area under the scaling function is unitary: $\int \varphi(x)\,dx = 1$ and the third one assures the orthogonality among each pair of scaling function vectors. Daubechies' discrete wavelet transform (DWT-db), as other families of wavelets like Symmlets and Coiflets [10], are used in the tests.

To example, figure 1 shows a third-level decomposition tree of a signal, using the DWT.

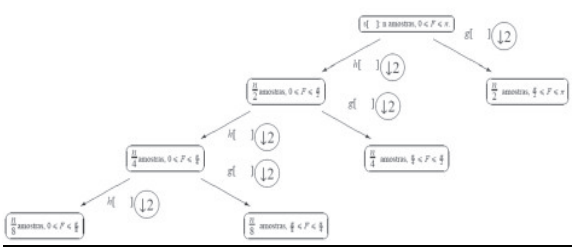


Figure 1: 3-level decomposition tree of a signal.

1.2 ANN – Artificial Neural Network

ANNs have seen an explosion of interest over the last few years [11], and are being successfully applied across an extraordinary range of problem domains, in areas as diverse as finance, medicine, engineering, geology and physics. Indeed, anywhere that there are problems of prediction, classification or control, neural networks are being introduced. ANNs are applicable in virtually every situation in which a relationship between the predictor variables (independents, inputs) and predicted variables (dependents, outputs) exists, even when that relationship is very complex and not easy to articulate in the usual terms of "correlations" or "differences between groups."

The proposed approach uses a particular type of ANN, called Radial Basis Function (RBF). It is based on an intuitive approach that is is to divide up space using circles or hyperspheres. A hypersphere is characterized by its center and radius. A RBF network responds non-linearly to the distance of points from the center represented by the radial unit. The response surface of a single radial unit is a Gaussian function, peaked at the center, and descending outwards. RBFs have one hidden layer of radial units, each modeling a Gaussian response. Since these functions are nonlinear, it is not actually necessary to have more than one hidden layer to model any shape of function: sufficient radial units will always be enough to model any function. A weighted sum of the Gaussians is used to produce the output.

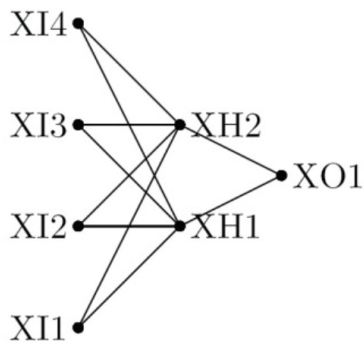


Figure 2: A RBF with 4 neurons in the input layer, two in the hidden layer, and one output.

1.3 Human Auditory System (HAS)

An important concept derived from working of the human auditory system, illustrated in figure 3, is the set of critical bands [12]. According to this concept, if we have two signals of different frequencies inside the same critical band, occuring at the same time, the stronger will mask the weaker. In this work, we take advantage of this concept to separate audio files into critical bands and then insert the watermark into each band of the input signal.

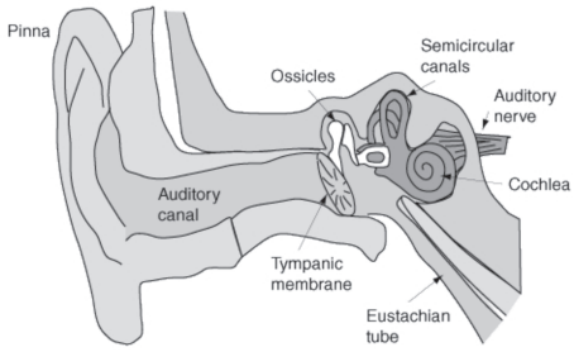


Figura 3: Physiological composition of human ear

1.4 Audio Watermarking

A wide bibliographical review on the literature shows that most of the recent algorithms for audio watermarking are based on a combination of the DWT and the HAS, where the watermark is embedded in different sub-bands of frequencies, observing the HAS so that the mark is inaudible [13].

2 THE PROPOSED ALGORITHM

The proposed algorithm to embed the watermark, that is divided in some steps, follows.

- STEP 1: extract the raw data from the input audio file, and segment it in frames of 2048 samples each;

- STEP 2: apply the 11-th level DWT to the previous signal. The family of wavelets and the corresponding support size will be discussed later;

- STEP 3: span the coefficients of the DWT of the input signal into 25 sets so that each set corresponds to a critical band of the human auditory system, according to [12];

- STEP 4: for each one of the previous set of coefficients, apply the transformation described in [14].

- STEP 5: apply the inverse DWT.

To detect the watermark, a RBF was trained with one second of data, receiving as input frames of length equals to 2048, being the corresponding output 0 or 1, for non-watermarked and watermarked signal, respectively. After training, the network was used to detect the watermarks.

3 RESULTS AND CONCLUSIONS

We proposed an algorithm to embed watermarks in audio files, base don DWT and HAS. The watermarks can be detected by using a RBF ANN, previously trained for this purpose. For a file that is one minute long, 50% of data were used for training and 50% for testing, achieving more than 90% of correct results for various segments of audio. The Daubechies' family of wavelets has shown good performance, that increases as the support-size of the filters increase too, i.e., a good frequency response is important to the algorithm.

4 REFERENCES

[1] Kim O.H., Lee B.K., Lee N.Y., *"Wavelet-Based Audio Watermark Techniques: Robustness and Fast Synchronization",* Division of Applied Mathematics, KAIST, November 2001.

[2] Cvejic N., *"Algorithms for Audio Watermarking and Steganography"*, Academic Dissertation, University of Oulu, 2004.

[3] Strang G., Nguyen T., *"Wavelets and filter banks"*, Wellesley-Cambridge Academic Press, Wellesley, 1997.

[4] Addison P.S.. *"The Illustrated wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance"*. Edinburg-UK: Institute of Physics Publishing, 2002.

[5] Walker J.S.. *"A Primer on Wavelets and Their Scientific Applications"*. Whashington-USA: Chapman and Hall/CRC, 1999.

[6] Jensen A., Cour-Harbo. A.A.. *"Ripples in Mathematics: The DiscreteWavelet Transform"*. New York-USA: Springer-Verlag, 2000.

[7] Oppenhein A.V., Schafer R.W.. *"Discrete Time Signal Processing"*. 2. ed. New York: Prentice Hall,1999.

[8] Cohen A, Daubechies I., and Feauveau J.-C., *"Biorthogonal Bases of Compactly Supported Wavelets"*, Comm. Pure Appl. Math., Vol 45, 1992, pp. 485-560.

[9] Daubechies I. *"Orthonormal bases of compactly supported wavelets"*, Comm. Purê and Appl. Math. 1988 v.41, pp. 909–996.

[10] Vieira L.S., *"Conversão de Voz Baseada em Transformada Wavelet"*. IFSC, USP. Dissertação de Mestrado, 2007.

[11] Guido R.C., Vieira, L.S., Fonseca E.S. (Status: to appear) *"A Neural network and wavelet-based approach for voice conversion. Neurocomputing"*. Elsevier, v. -, p. in–press, 2007.

[12] Guido R.C., Vieira L.S., Sanchez F.L., Slaets J.F.W, Almeida L.O., Gonzaga A., Bianchi M., *"A Matched FIR Filter Bank Bank for Áudio Coding"*, IEEE Computer Society, 2005.

[13] Deguillaume F., Voloshynovskiy S., Pun T. *"Secure hybrid robust watermarking resistant against tampering and copy attack"*. Signal Processing, 2003.

[14] Cvejic N., Seppanem T. *"Robust Audio Watermarking in Wavelet Domain Using Frequency Hopping and Patchwork Method"*, Proceedings of the 3rd International Synposium on Image and Signal Processing and Analisys, 2003.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, São Paulo, SP

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

Técnicas para Separação Cega de Fontes aplicadas na Melhoria de Desempenho de um Classificador de Palavras Isoladas

Diego Barreto Haddad,[1] Mariane Rembold Petraglia[1] e Paulo Bulkool Batalheiro[2]

[1] COPPE/UFRJ,
Rio de Janeiro, RJ, 21941-947, Brasil

[2] UERJ/FEN/DETEL
Rio de Janeiro, RJ, 20550-900, Brasil

diego@pads.ufrj.br, mariane@pads.ufrj.br, bulkool@pads.ufrj.br

RESUMO

Métodos de separação cega de fontes têm sido propostos para melhoria de desempenho em sistemas automáticos de reconhecimento de fala, dado que estes são muito sensíveis a ruído. Neste artigo apresentamos o projeto de um classificador de palavras isoladas e analisamos sua degradação de desempenho em ambientes ruidosos. Em seguida, são empregadas técnicas recentes de separação de sinais, baseadas na análise de componentes independentes, para reavaliarmos a acurácia do reconhecedor, sendo considerados tanto casos de mistura instantânea quanto os de mistura convolutiva.

0 INTRODUÇÃO

Durante os últimos vinte anos houve intensa pesquisa em técnicas de separação cega de fontes, comumente denominadas BSS (do inglês *Blind Source Separation*), baseadas em Análise de Componentes Independentes (ICA, *Independent Component Analysis*), comumente empregadas na resolução do célebre problema conhecido como *cocktail party* (em referência à habilidade humana de facilmente separar os sons de interesse - vozes de ruídos - numa festa). A separação é classificada como cega pois não utiliza quaisquer informações *a priori* sobre o conteúdo espectral ou posições das fontes, ou mesmo acerca do processo de mistura. Era comum a admissão do modelo de misturas lineares e instantâneas, porém mais recentemente surgiu um grande interesse em se contemplar modelos mais realistas, em especial para o caso de misturas de fontes sonoras. Estes modelos, embora ainda lineares, passaram a ser convolutivos, permitindo assim modelar-se a reverberação sempre presente em ambientes ecóicos. Uma das principais motivações destes métodos reside na significativa queda de desempenho

de programas de reconhecimento automático de palavra quando o extrato sonoro encontra-se contaminado por ruído (entendendo-se ruído como qualquer tipo de interferência indesejada). Um pré-processamento baseado na separação do ruído da parte que contém informação relevante traria a possibilidade de ao menos reduzir essa degradação de desempenho (essa melhora pode ser também empregada para avaliação de métodos de separação de fontes). Este artigo, a partir de um banco de dados com 10 palavras (gravações dos números de zero a nove) e três diferentes locutores (num total de 2100 gravações), implementa um reconhecedor de palavras isoladas, verifica a degradação de desempenho em ambientes ruidosos (tanto no caso de mistura instantânea quanto no de mistura convolutiva) e avalia a melhora que ocorre após a aplicação de alguns dos mais precisos métodos de separação de fontes.

1 RECONHECEDOR DE PALAVRAS ISOLADAS

1.1 Extração de características

O sistema de reconhecimento automático de palavras empregado pode ser caracterizado como multilocutor para palavras preparadas (gravadas especificamente para a tarefa de reconhecimento), isoladas e gravadas com diferentes microfones. O método de extração de características do reconhecedor consiste nos seguintes passos: (a) divisão da palavra a se reconhecer em um número fixo de janelas, com sobreposição; (b) extração dos coeficientes LP (do inglês *Linear Prediction*) de cada janela, bem como do erro quadrático médio do modelo LP e (c) concatenação dos resultados de (b), formando por fim um único vetor de características. O modelo LP de ordem N de uma seqüência $x(n)$ consiste num conjunto de coeficientes $a(i)$, $i = 1, ...N$, que minimiza o erro quadrático médio do seguinte preditor *forward*:

$$\hat{x}(n) = \sum_{i=1}^{N} a(i)x(n-i). \tag{1}$$

Os coeficientes LP são encontrados por meio da recursão de Levinson-Durbin, a qual ocorre no contexto de minimização do erro quadrático médio [1]. Assim, cada gravação possui um vetor de características associado, o qual será utilizado para teste, validação (caso esta seja implementada) ou treinamento do sistema de reconhecimento. Idealmente, o conjunto de vetores de características deveria ser branqueado, de forma a auxiliar a generalização dos métodos de classificação. Porém este pré-processamento não se traduziu numa melhora de desempenho, sendo portanto descartado no intuito de redução da carga computacional.

1.2 Métodos de Classificação

Dois métodos de classificação foram utilizados: redes neurais e a métrica de Minkowski de ordem s, uma generalização da distância euclidiana.

1.2.1 Métrica de Minkowski

A métrica de Minkowski de ordem s (equivalente à distância euclidiana quando $s = 2$) é definida, entre dois vetores $\mathbf{x}$ e $\mathbf{y}$ de dimensão N e k-ésimos componentes x_k e y_k, por [1]:

$$d_s(\mathbf{x}, \mathbf{y}) \cong \sqrt[s]{\sum_{k=1}^{N} |x_k - y_k|^s}. \tag{2}$$

Para utilizar a métrica supracitada, inicialmente extraímos o vetor de características das primeiras 40 amostras (de um total de 70) de cada palavra e de cada locutor, gerando o conjunto de treinamento. Para todas as outras amostras (conjunto de teste), calculamos a distância (de ordem s) entre os vetores características destas e de cada um dos vetores do conjunto de treinamento. Por meio do vizinho mais próximo (menor distância), efetuamos a classificação.

1.2.2 Redes Neurais

As redes neurais implementadas são do tipo *feed-forward*, com todos os neurônios dotados da função característica tangente hiperbólica, apresentando uma camada escondida (maiores detalhes em [2]), 10 neurônios de saída (um para cada palavra) e treinadas com o algoritmo rápido *resilient backpropagation* [3]. Das 70 gravações de cada palavra, efetuadas por cada um dos locutores, 40 foram destinadas para treinamento, 15 para validação e 15 para teste. No treinamento supervisionado, o neurônio da camada de saída correspondente à classe da palavra em questão apresenta 1 como resposta desejada e os neurônios restantes foram induzidos a ter -1 como saída desejada. No tocante ao teste, considera-se como estimativa da rede a classe correspondente ao neurônio da camada de saída que apresentar maior resposta.

2 MÉTODOS DE SEPARAÇÃO

A interferência será modelada por um ruído branco e gaussiano. Mostraremos a seguir dois métodos de separação para o caso de misturas instantâneas e dois para o caso de misturas convolutivas. Após a separação, o problema da permutação é resolvido mediante a mera medida da *kurtosis*, pois, sendo o ruído gaussiano, sua *kurtosis* (de certa forma, uma medida de não-gaussianidade) é próxima de zero. Para decidir qual saída seria considerada ruído, utilizamos a seguinte definição:

$$\textit{kurtosis}[x(n)] = \frac{E[(x - E[x])^4]}{\{E[(x - E[x])^2]\}^2}, \tag{3}$$

onde o operador $E[\cdot]$ significa média estatística. Assim, a saída sobre a qual se efetuará o processo de reconhecimento será a que possuir maior *kurtosis*. Todos os métodos supõem, sem perda de generalidade, 2 fontes e 2 microfones.

2.1 Misturas Instantâneas

Sejam $s_1(n)$ e $s_2(n)$ duas seqüências que representam sinais mutuamente independentes (em geral não gaussianos). Podemos formar com estas uma matriz **S**, cuja n-ésima coluna é dada por:

$$\mathbf{s}(n) = \begin{bmatrix} s_1(n) \\ s_2(n) \end{bmatrix}. \tag{4}$$

A mistura instantânea das seqüências independentes $s_k(n),\ k = 1, 2$, pode ser descrita mediante um produto por uma matriz de mistura **H** quadrada e de ordem 2, gerando as seqüências $x_k(n)$ (linhas da matriz **X**), as quais passam a ser dependentes entre si, da forma a seguir:

$$\mathbf{x}(n) = \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} = \mathbf{H} \begin{bmatrix} s_1(n) \\ s_2(n) \end{bmatrix}, \tag{5}$$

sendo $x_k(n)$ o sinal capturado pelo k-ésimo microfone.

A separação é obtida através do produto de uma matriz de separação **W** (de mesmas dimensões que **H**) pela matriz **X**, gerando a matriz **Y** das fontes estimadas, como mostrado abaixo:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \mathbf{WX} = \mathbf{WHS}, \tag{6}$$

sendo $\mathbf{y}_i$ o vetor que contém todas as amostras da i-ésima seqüência de saída $y_i(n)$ e **H** uma matriz não-singular.

Idealmente, para recuperarmos exatamente as fontes originais, **W** deveria ser a inversa de **H**, tornando o produto **WH** a matriz identidade. Porém, mesmo sendo bem sucedido o método de separação, isto não costuma ocorrer por duas razões. Primeiramente, há uma ambigüidade de escalamento, pois caso multiplicássemos uma fonte $s_k(n)$ por uma constante a, isso seria o equivalente a multiplicar uma coluna da matriz **H** pelo mesmo valor, e, desconhecida qualquer informação tanto sobre a fonte quanto sobre a matriz de mistura, $s_k(n)$ e $as_k(n)$ são, ambas, soluções igualmente aceitáveis. Assim, podemos restringir o objetivo a obter uma matriz **WH** diagonal. Porém, além dessa ambigüidade, há o problema da permutação, que implica não podermos sempre recuperar (a menos da ambigüidade de escalamento já abordada) a k-ésima linha de **X** na k-ésima linha de **Y**. Em última análise, este problema também ocorre devido à nossa falta de informação sobre as fontes e sobre a matriz de mistura, pois inverter as fontes é uma operação que possui os mesmos efeitos do que trocar as colunas da matriz **H**. Logo, somente devemos esperar que o produto **WH**, ao invés de ser a matriz identidade, seja uma matriz diagonal ou anti-diagonal.

2.1.1 Método I: Gradiente Natural da Informação Mútua

Almejando tornar nula a informação mútua entre $\mathbf{y}_1$ e $\mathbf{y}_2$, os métodos de separação cega de fontes não raro têm por objetivo anular a diagonal secundária da matriz $\mathbf{R}_Y$ definida como:

$$\mathbf{R}_Y = \begin{bmatrix} E[\Phi(\mathbf{y}_1)\mathbf{y}_1] & E[\Phi(\mathbf{y}_1)\mathbf{y}_2] \\ E[\Phi(\mathbf{y}_2)\mathbf{y}_1] & E[\Phi(\mathbf{y}_2)\mathbf{y}_2] \end{bmatrix}, \tag{7}$$

sendo

$$\Phi(y_i) = -\frac{\partial}{\partial |y_i|}[p(y_i)], \tag{8}$$

onde $p(y_i)$ é a função densidade de probabilidade do sinal y_i ($\Phi(y_i) = y_i$ implica um método que emprega apenas estatísticas de segunda ordem, não recorrendo às mais altas), com o operador $E[\cdot]$ não necessariamente contemplando todo o sinal; efetivamente, implementamo-lo apenas sobre blocos do mesmo, gerando uma componente estocástica que torna o algoritmo menos sensível a mínimos locais (uma opção muito utilizada seria branquear o sinal, antes de se aplicar o ICA, por meio de uma matriz de branqueamento). Por meio do gradiente natural (mais rápido que o gradiente padrão normalmente utilizado, além de mais robusto ao mau condicionamento da matriz de misturas) proposto em [4], a matriz de separação é obtida iterativamente da forma a seguir:

$$\mathbf{W}_{i+1} = \mathbf{W}_i + \eta\{\mathbf{I}_{2x2} - E[\Phi(\mathbf{Y})\mathbf{Y}^T]\}\mathbf{W}_i, \tag{9}$$

onde η é um coeficiente de aprendizagem, $\mathbf{I}_{2x2}$ é a matriz identidade de ordem 2 e $(\cdot)^T$ indica transposição.

2.1.2 Método II: ICA Rápido

A família de métodos denominados *Fast ICA* (ICA Rápido) [5] é muito popular. Antes de aplicar o algoritmo *Fast ICA*, cumpre branquear as misturas. Comumente, isto é feito mediante decomposição em autovalores da matriz **R** definida por:

$$\mathbf{R} = E[\mathbf{x}(n)\mathbf{x}^T(n)]. \tag{10}$$

Sejam λ_1, λ_2 os autovalores de **R** e $\{\mathbf{e}_1, \mathbf{e}_2\}$ seus autovetores correspondentes. A matriz branqueada **Z** da mistura **X** é calculada como:

$$\mathbf{Z} = \mathbf{VX}, \tag{11}$$

sendo **V** explicitada a seguir:

$$\mathbf{V} = \mathrm{diag}[\lambda_1, \lambda_2]^{-\frac{1}{2}}[\mathbf{e}_1, \mathbf{e}_2]^T, \tag{12}$$

onde o operador $\mathrm{diag}[a_1, \ldots, a_M]$ gera uma matriz diagonal com os valores $a_1, \ldots, a_M$ em sua diagonal principal. A versão do *Fast ICA* por nós implementado tem como objetivo maximizar a *kurtosis* de sinais de média nula. Delfosse e Loubaton provaram em [6] que o máximo local deste critério corresponde ao ponto de separação. Como já explicado, a *kurtosis* está relacionada à não-gaussianidade e o fato de se esperar que a soma de dois sinais seja mais gaussiana que cada um separadamente fornece a motivação para se empregar esta metodologia . Em poucas iterações, cada qual de baixo custo computacional, o algoritmo converge para

uma boa solução. A iteração fundamental deste algoritmo, obtida por meio do gradiente, é:

$$\mathbf{W}_{i+1} = E[\mathbf{Z}(\mathbf{Z}^T\mathbf{W}_i^T)^3] - 3\mathbf{W}_i. \quad (13)$$

Após cada iteração, é necessário manter a norma unitária da matriz $\mathbf{W}$, o que é obtido mediante a transformação:

$$\tilde{\mathbf{W}} = T\{\mathbf{W}\} = \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-\frac{1}{2}}. \quad (14)$$

Ao término das iterações, encontramos a matriz final de separação por meio da fórmula:

$$\mathbf{W}_{final} = \mathbf{W}^T\mathbf{V}. \quad (15)$$

2.2 Misturas Convolutivas

Existem diversos algoritmos que se aplicam à separação de fontes misturadas de forma convolutiva. Uma das mais freqüentes abordagens consiste em transformar as misturas para o domínio da freqüência, admitir (a partir do fato de que a convolução no tempo se torna produto na freqüência) que em cada freqüência a mistura é instantânea, e resolver o problema de separação via ICA instantâneo. Essa abordagem possui o mérito de reduzir enormemente a carga computacional, à custa de tornar não trivial o problema da permutação e do escalamento em cada freqüência, criando a possibilidade de se trocarem as componentes na freqüência entre as fontes na saída do sistema. O escalamento comumente é resolvido de forma razoável mediante o recurso ao Princípio da Mínima Distorção, apresentado em [7]. Já o problema da permutação, mais complexo, costuma ser abordado via correlação (ou uma versão modificada da mesma, inclusive utilizando critérios que contemplam simultaneamente diversas freqüências) de envelopes [8],[9], direção de chegada de cada fonte [10],[11] ou uma conjunção de ambas as técnicas [12], opção esta que tenta atenuar as desvantagens de uma abordagem com as vantagens de outra. Seguiremos a nomenclatura adotada por Aichner et al. [13], denominando este tipo de método (transformada na freqüência associada a ICA instantâneo) de *narrowband* (banda estreita), em contraposição aos métodos que não analisam cada freqüência isoladamente, chamados *broadband* (banda larga). Ambos os métodos aqui implementados utilizam apenas estatísticas de segunda ordem (basicamente correlações), são de banda larga, sendo um no domínio do tempo (onde não há transposição para a freqüência) e outro no domínio da freqüência, baseados em [14]. O método no domínio da freqüência contorna o problema da permutação em cada freqüência restringindo o tamanho dos filtros no tempo, o que equivale a tornar suaves na freqüência as respostas do sistema, de maneira semelhante à efetuada em [15],[16]. Assim, este artigo não precisa abordar os clássicos problemas de permutação passíveis de ocorrer em cada freqüência.

A separação de fontes misturadas de forma convolutiva, além de ser uma tarefa mais difícil, apresenta uma peculiaridade que contribui para uma queda maior do desempenho do reconhecedor de palavras. Reconhecendo que: (a) a separação é efetuada por filtros que idealmente geram saídas independentes e (b) se um sinal $\mathbf{s}_1$ é independente de $\mathbf{s}_2$, uma versão filtrada de $\mathbf{s}_1$ também é independente de $\mathbf{s}_2$, verificamos que, mesmo sendo bem sucedido o processo de separação, as saídas do sistema não são, em geral, versões meramente escaladas e permutadas das fontes, porém versões *filtradas* das mesmas. Essa filtragem costuma não ser perceptível auditivamente; porém, à medida em que altera o conteúdo espectral das fontes, prejudica o processo de extração de características e, indiretamente, o reconhecimento. A p-ésima mistura $x_p(n)$ (relativa ao p-ésimo microfone) do modelo convolutivo com filtros FIR de comprimento M (vide Figura 1) é escrita como:

$$x_p(n) = \sum_{q=1}^{2}\sum_{k=0}^{M-1} h_{qp}(k)s_q(n-k), \quad (16)$$

sendo $h_{qp}(n)$ a resposta ao impulso do filtro que modela o caminho acústico entre a q-ésima fonte e o p-ésimo sensor.

Em ambos os métodos a seguir, a atualização nos filtros de separação w_{pq} é implementada nas matrizes de separação $\mathbf{W}$ (ou $\underline{\mathbf{W}}$, ambas definidas mais à frente) por meio do gradiente (natural) descendente, da forma a seguir:

$$\mathbf{W}_{i+1} = \mathbf{W}_i - \mu \cdot \alpha^i \cdot \nabla_{W_i}^{GN}, \quad (17)$$

onde μ é o coeficiente de aprendizagem e $\alpha \leq 1$ um fator que poderá reduzir o ganho de adaptação a cada atualização.

2.2.1 Método I: Domínio do Tempo

Os filtros de separação também são do tipo FIR, de comprimento L, gerando a q-ésima saída $y_q(n)$ dada por:

$$y_q(n) = \sum_{p=1}^{2} \mathbf{x}_p^T(n)\mathbf{w}_{pq}, \quad (18)$$

onde

$$\mathbf{x}_p = [x_p(n), x_p(n-1), \ldots, x_p(n-L+1)]^T, \quad (19)$$

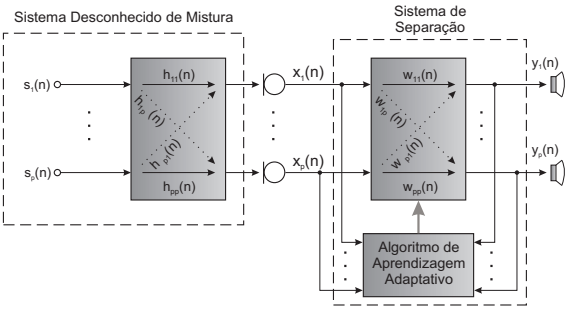


Figura 1: Ilustração do processo convolutivo de mistura e de separação.

e

$$\mathbf{w}_{pq} = [w_{pq,0}, w_{pq,1}, \ldots, w_{pq,L-1}]^T. \quad (20)$$

Dois novos parâmetros devem ser introduzidos para generalizar a formulação, a saber: o número D de intervalos no tempo considerados para as correlações na função custo ($1 \leq D \leq L$) e o comprimento N dos blocos de sinais de saída utilizados para estimar as correlações de intervalos de curta duração ($N \geq 2D$). O m-ésimo bloco correspondente à q-ésima saída é dado por:

$$\mathbf{Y}_q(m) = \sum_{p=1}^{2} \mathbf{X}_p(m)\mathbf{W}_{pq}, \quad (21)$$

com

$$\mathbf{X}_p(m) = \left[\mathbf{U}_p^T(m), \mathbf{U}_p^T(m-1)\right] \quad (22)$$

e

$$\mathbf{U}_p(m) = \begin{bmatrix} x_p(mL) & \cdots & x_p(mL-L+1) \\ x_p(mL+1) & \cdots & x_p(mL-L+2) \\ \vdots & \ddots & \vdots \\ x_p(mL+N-1) & \cdots & x_p((m-1)L+N) \end{bmatrix}. \quad (23)$$

Por fim, a matriz Sylvester $\mathbf{W}_{pq}$ de dimensões $2L \times D$ é definida como:

$$\mathbf{W}_{pq} = \begin{bmatrix} w_{pq,0} & 0 & \cdots & 0 \\ w_{pq,1} & w_{pq,0} & \ddots & \vdots \\ \vdots & w_{pq,1} & \ddots & 0 \\ w_{pq,L-1} & \vdots & \ddots & w_{pq,0} \\ 0 & w_{pq,L-1} & \ddots & w_{pq,1} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & w_{pq,L-1} \\ 0 & \ldots & 0 & 0 \\ \vdots & \ldots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}. \quad (24)$$

A redundância desta matriz significa que conhecer uma de suas colunas implica conhecê-la completamente. Há razões de convergência [14] que requerem o procedimento de, a cada atualização de $\mathbf{W}_{pq}$, utilizar somente a primeira coluna para preencher a matriz, desprezando todos os valores contidos nas outras colunas. Esse procedimento será adotado em nossa implementação. Combinando todos os canais, podemos escrever concisamente:

$$\mathbf{Y}(m) = \mathbf{X}(m)\mathbf{W}, \quad (25)$$

com as matrizes:

$$\mathbf{Y}(m) = [\mathbf{Y}_1(m), \mathbf{Y}_2(m)], \quad (26)$$

$$\mathbf{X}(m) = [\mathbf{X}_1(m), \mathbf{X}_2(m)], \quad (27)$$

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{bmatrix}. \quad (28)$$

Definida a formulação, sendo b o número de blocos, a implementação *off-line* do algoritmo tem por função custo:

$$\begin{aligned} \Im &= \sum_{i=1}^{b} \frac{1}{b}\{\log(\det(\mathrm{bdiag}(\mathbf{Y}^T(i)\mathbf{Y}(i)))) \\ &\quad -\log(\det(\mathbf{Y}^T(i)\mathbf{Y}(i)))\}, \end{aligned} \quad (29)$$

onde o operador bdiag$(\cdot)$ opera sobre uma matriz de submatrizes, anulando todas as submatrizes que não estejam situadas na diagonal principal da matriz (neste artigo, todas as matrizes sobre as quais atua este operador são constituídas por 4 submatrizes de idênticas dimensões). O gradiente natural da função custo acima é fornecido por:

$$\begin{aligned} \nabla_{\mathbf{W}}^{GN}\Im(m) &= \frac{2}{b}\sum_{i=1}^{b} \mathbf{W}\{\mathbf{R}_{yy} - \mathrm{bdiag}(\mathbf{R}_{yy})\} \\ &\quad \{\mathrm{bdiag}\,(\mathbf{R}_{yy})\}^{-1}. \end{aligned} \quad (30)$$

2.2.2 Método II: Domínio da Freqüência

Seja $\mathbf{F}_{4L\times 4L}$ a matriz DFT de dimensão $4L \times 4L$. Assumiremos, sem perda de generalidade, $N = L$, exigindo que os blocos possuam comprimento não mais igual a $2N$, mas a $4N$, pois o tamanho do bloco deve dobrar após a convolução e novamente duplicar após a estimativa de correlação. Definimos:

$$\begin{aligned} \underline{\mathbf{X}}_p(m) = \mathrm{diag}[&\mathbf{F}_{4L\times 4L}[x_p(mL-3L), \ldots, x_p(mL-1), \\ &x_p(mL), x_p(mL+1), \ldots, x_p(mL+L-1)]^T], \end{aligned} \quad (31)$$

$$\underline{\mathbf{X}}(m) = [\underline{\mathbf{X}}_1(m), \underline{\mathbf{X}}_2 m], \quad (32)$$

$$\underline{\mathbf{W}}_{pq} = \mathrm{diag}\left[\mathbf{F}_{4L\times 4L}[w_{pq,0}, \ldots, w_{pq,L-1}, 0, \ldots, 0]^T\right]. \quad (33)$$

Deve-se ressaltar que, a cada iteração na freqüência, deve-se voltar com os filtros na freqüência para o domínio do tempo de forma a zerar todos os componentes que ultrapassam o tamanho do filtro, res-tringindo o comprimento deste para L. Esta restrição pode ser representada como um produto pela inversa de uma matriz DFT, em seguida por uma matriz de restrição (que zera os componentes indesejáveis) e então pela matriz DFT. Definimos ainda:

$$\underline{\mathbf{W}} = \begin{bmatrix} \underline{\mathbf{W}}_{11} & \underline{\mathbf{W}}_{12} \\ \underline{\mathbf{W}}_{21} & \underline{\mathbf{W}}_{22} \end{bmatrix}, \quad (34)$$

$$\mathbf{S}_{xx} = \underline{\mathbf{X}}^H \mathbf{F}_{4L\times 4L} \begin{bmatrix} \mathbf{0}_{3L\times 3L} & \mathbf{0}_{3L\times 3L} \\ \mathbf{0}_{L\times 3L} & \mathbf{I}_{L\times L} \end{bmatrix} \mathbf{F}_{4L\times 4L}^{-1} \underline{\mathbf{X}}, \quad (35)$$

$$\mathbf{G}_{4L} = \mathbf{F}_{4L\times 4L} \begin{bmatrix} \mathbf{I}_{2L\times 2L} & \mathbf{0}_{2L\times 2L} \\ \mathbf{0}_{2L\times 2L} & \mathbf{0}_{2L\times 2L} \end{bmatrix} \mathbf{F}_{4L\times 4L}^{-1}, \quad (36)$$

$$\mathbf{G}_{8L} = \begin{bmatrix} \mathbf{G}_{4L} & \mathbf{0}_{4L\times 4L} \\ \mathbf{0}_{4L\times 4L} & \mathbf{G}_{4L} \end{bmatrix}, \quad (37)$$

$$\mathbf{S}_{yy} = \underline{\mathbf{W}}^H \mathbf{G}_{8L} \mathbf{S}_{xx} \mathbf{G}_{8L} \underline{\mathbf{W}}, \tag{38}$$

$$\mathbf{L} = \begin{bmatrix} \mathbf{F}_{4L\times 4L} \begin{bmatrix} \mathbf{I}_{D\times D} \\ \mathbf{0}_{4L-D\times D} \end{bmatrix} & \mathbf{0}_{4L\times D} \\ \mathbf{0}_{4L\times D} & \mathbf{F}_{4L\times 4L} \begin{bmatrix} \mathbf{I}_{D\times D} \\ \mathbf{0}_{4L-D\times D} \end{bmatrix} \end{bmatrix}, \tag{39}$$

$$\mathbf{L_I} = \begin{bmatrix} \mathbf{1}_{4L\times 1} & \mathbf{0}_{4L\times 1} \\ \mathbf{0}_{4L\times 1} & \mathbf{1}_{4L\times 1} \end{bmatrix}, \tag{40}$$

onde $(\cdot)^H$ signica hermitiano e $\mathbf{0}_{m\times n}$ é a matriz nula de dimensões m e n. O gradiente natural da função custo nesta abordagem é fornecido por:

$$\nabla_{\underline{\mathbf{W}}}^{GN} \Im = \sum_{i=1}^{b} \frac{2}{b} \mathbf{G}_{8L} \underline{\mathbf{W}} \mathbf{L} \mathbf{L}^H \left\{ \mathbf{S}_{yy} - \text{bdiag}(\mathbf{S}_{yy}) \right\} \mathbf{L} \left\{ \text{bdiag}(\mathbf{L}^H \mathbf{S}_{yy} \mathbf{L}) \right\}^{-1} \mathbf{L}^H. \tag{41}$$

A matriz $\underline{\mathbf{W}}$, à maneira de $\mathbf{W}$, também possui redundâncias, as quais são retiradas após cada iteração, por meio da transformada:

$$\underline{\mathbf{W}}' = \underline{\mathbf{W}} \mathbf{L}_I, \tag{42}$$

com a matriz $\underline{\mathbf{W}}'$ possuindo duas colunas, cada qual com $2L$ linhas (as primeiras L linhas da q-ésima coluna contendo as informações das primeiras L linhas da primeira coluna de $\mathbf{W}_{1q}$ e as restantes armazenando as primeiras L linhas da primeira coluna de $\mathbf{W}_{2q}$). O gradiente natural de $\underline{\mathbf{W}}'$ é:

$$\nabla \underline{\mathbf{W}}\text{'} = \mathbf{G}'_{8L} \nabla_{\underline{\mathbf{W}}}^{GN} \Im \mathbf{L}_I, \tag{43}$$

com

$$\mathbf{G}'_{4L} = \mathbf{F}_{4L\times 4L} \begin{bmatrix} \mathbf{I}_{L\times L} & \mathbf{0}_{L\times 3L} \\ \mathbf{0}_{3L\times L} & \mathbf{0}_{3L\times 3L} \end{bmatrix} \cdot \mathbf{F}_{4L\times 4L}^{-1}, \tag{44}$$

e

$$\mathbf{G}'_{8L} = \begin{bmatrix} \mathbf{G}'_{4L} & \mathbf{0}_{4L\times 4L} \\ \mathbf{0}_{4L\times 4L} & \mathbf{G}'_{4L} \end{bmatrix}. \tag{45}$$

3 RESULTADOS

3.1 Classificador de Palavras

Em todos os resultados aqui apresentados, cada gravação foi particionada em 12 janelas, com uma sobreposição de 15% das amostras entre as mesmas, pois, após testes, esta configuração sempre apresentou bons resultados. Todas as gravações tiveram variância normalizada antes da extração de características. Na Figura 2, vê-se o resultado da taxa de acerto do sistema de reconhecimento em função da ordem do modelo LP e do parâmetro s da métrica de Minkowski. A taxa de acerto máxima encontrada foi de 96,67% para uma ordem LP igual a 11, com $s = 1,5$. No tocante às redes neurais, variamos o número de neurônios na camada escondida, assim como a ordem LP e exibimos na Figura 3 a taxa de acerto correspondente (apresentamos o melhor resultado oriundo de 10 diferentes inicializações). A rede neural utilizada, que alia à pouca complexidade um bom desempenho, apresenta 49 neurônios na camada intermediária, ordem LP igual a 13 e taxa de acerto estimada também igual a 96,67%. Os resultados via métrica de Minkowski e redes neurais em geral assemelham-se muito; assim, apresentamos somente os devidos às redes neurais, as quais, após treinadas, efetuam a estimativa de forma muito mais rápida.

3.2 Mistura Instantânea

Para simular a interferência, empregamos um ruído branco de desvio padrão 0,12. A matriz de misturas não-singular arbitrariamente escolhida foi:

$$\mathbf{H} = \begin{bmatrix} 1 & 0,8 \\ 0,7 & -2 \end{bmatrix}. \tag{46}$$

A abordagem ICA via gradiente natural da informação mútua adotou os parâmetros $\eta = 10^{-6}$, 5 iterações, tamanho de bloco igual a 20 amostras e $\Phi(y_i) = \tanh(y_i)$, parâmetros estes que apresentaram os melhores resultados. As Figuras 4 e 5 exibem os resultados finais destas abordagens, com a iteração zero correspondendo a taxa de acerto anterior à inicialização dos métodos de separação. Os dois procedimentos evidenciaram melhoria na taxa de acerto do classificador, porém, com um número de iterações ligeiramente maior, o algoritmo *Fast* ICA apresentou um desempenho superior.

3.3 Mistura convolutiva

O desvio padrão do ruído branco utilizado na mistura convolutiva foi 0,3. Os filtros de mistura arbitrariamente escolhidos foram:

$$\begin{aligned} h_{11} &= \begin{bmatrix} 1 & 0,5 & -0,2 & 0,1 \end{bmatrix} \\ h_{12} &= \begin{bmatrix} 0,9 & 0,01 & 0,1 & -0,05 \end{bmatrix} \\ h_{21} &= \begin{bmatrix} 1,1 & 0,2 & 0 & -0,06 \end{bmatrix} \\ h_{22} &= \begin{bmatrix} 0,8 & -0,1 & 0,3 & 0,03 \end{bmatrix}. \end{aligned}$$

Ambos os métodos de separação utilizaram filtros de comprimento 4 para separação. O método de separação implementado no domínio do tempo teve escolhidos os parâmetros $\mu = 0,02$ e $\alpha = 0,9$. Na freqüência, os parâmetros adotados foram $\mu = 0,05$ e $\alpha = 0,7$. A taxa de acerto média do classificador, sem separação de fontes, foi de 33,33%. O parâmetro μ escolhido em geral foi relativamente alto, de forma a, em poucas iterações, fornecer melhoras. Devido à dificuldade de se obter um passo ótimo que atendesse todas as palavras, foram selecionadas algumas para verificação da melhora na taxa de acerto. Para o caso de separação no domínio do tempo, os resultados são mostrados na Figura 6; no caso de separação no domínio da freqüência, são exibidos os resultados na Figura 7. Os dois procedimentos melhoraram o desempenho do classificador

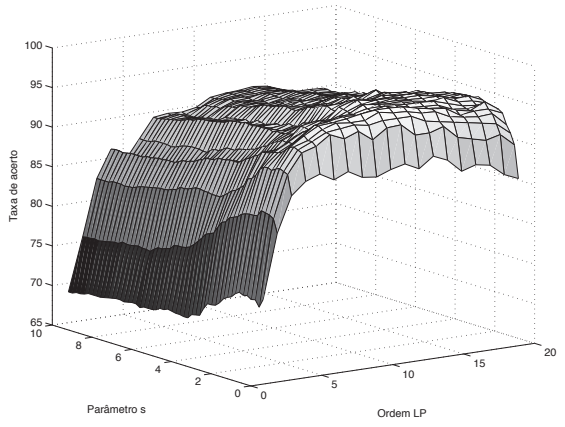


Figura 2: Evolução da taxa de acerto em função da ordem LP e do parâmetro s da métrica de Minkowski.

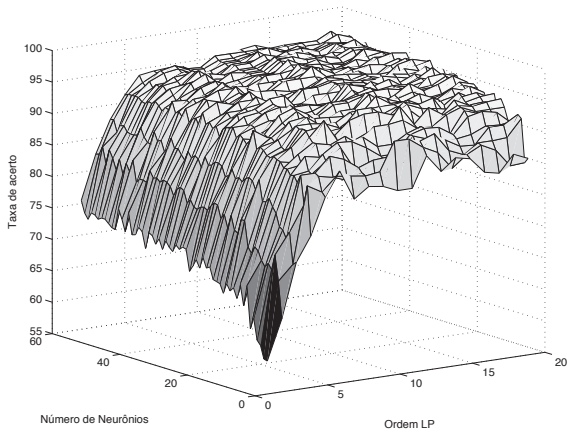


Figura 3: Evolução da taxa de acerto em função do número de neurônios na camada escondida e da ordem LP.

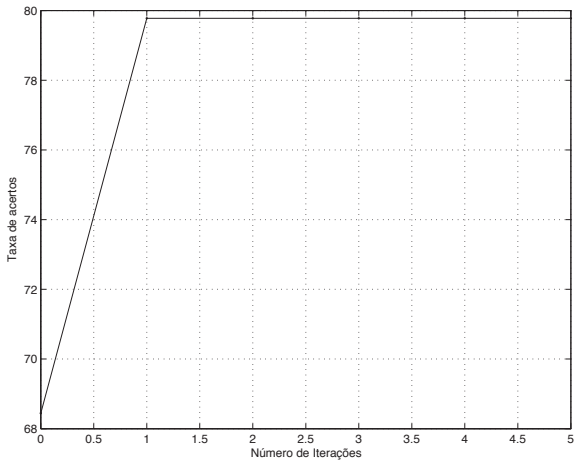


Figura 4: Evolução da taxa de acerto do ICA via gradiente natural para mistura instantânea.

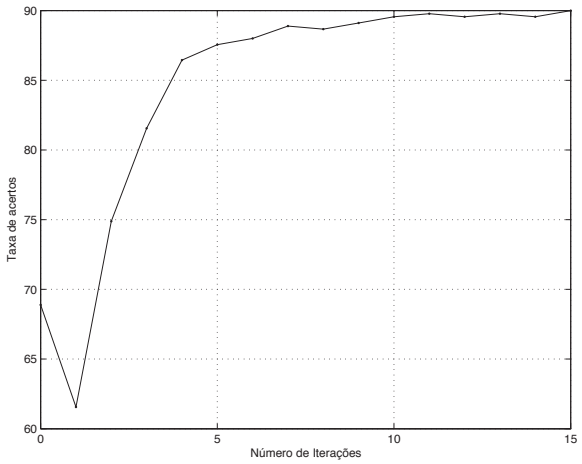


Figura 5: Evolução da taxa de acerto do *Fast ICA* para mistura instantânea.

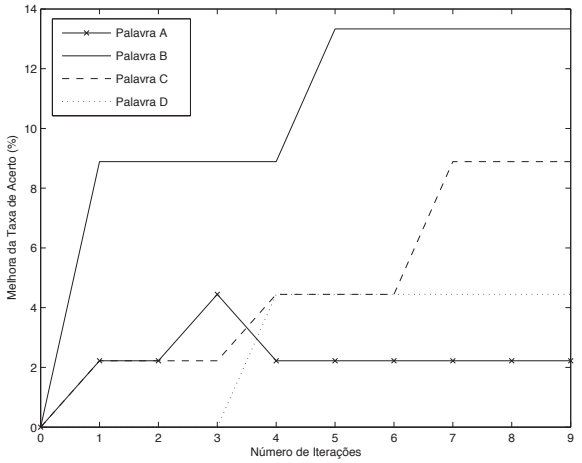


Figura 6: Evolução da melhora na taxa de acerto das palavras A, B, C e D para abordagem no tempo de mistura convolutiva.

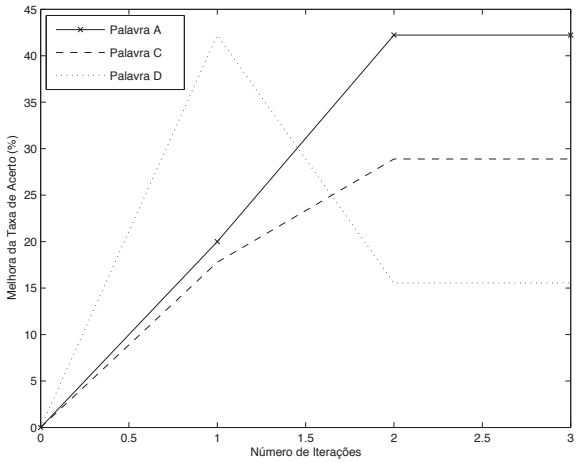


Figura 7: Evolução da melhora na taxa de acerto das palavras A, C e D para abordagem na freqüência de mistura convolutiva.

usando um passo único para todas as palavras. Em alguns casos a utilização de um μ não apropriado para uma determinada palavra resultou num descréscimo de desempenho para um número maior de iterações (palavra A no domínio do tempo e palavra D no domínio da freqüência). Na Figura 7, a evolução da palavra B foi omitida porque não houve melhoria na taxa de acerto.

4 CONCLUSÕES

A curta duração das palavras complica o processo de separação, seja este instantâneo ou convolutivo, já que não fornece muita informação (auditivamente, verifica-se que ainda há ruído presente na separação, o que não ocorre com gravações de duração mais longa). Porém, observa-se que, mesmo assim, ambos os métodos de separação de fontes misturadas instantaneamente se mostraram bem-sucedidos ao reduzirem a degradação da tarefa de reconhecimento, com larga vantagem para o *Fast ICA*, o qual, ao fim, apresentou uma redução da taxa de acerto (em relação às gravações sem ruído) em torno de 6%.

Sendo a informação relativamente pouca para o caso instantâneo, para o caso convolutivo a situação é mais crítica. Embora ambas as abordagens no caso convolutivo tenham apresentado melhoras, o método de separação no domínio da freqüência mostrou um desempenho significativamente superior. Uma explicação para este resultado seria o fato de que as seqüências no domínio da freqüência, possuindo uma distribuição menos gaussiana [17], se revelem mais propensas à separação (em especial quando há pouca informação). Uma explicação complementar poderia consistir no fato de a abordagem no domínio da freqüência efetuar uma filtragem das fontes menos severa (o que reduziria a distorção das saídas). Porém, cabe efetuar testes para verificar a validade desta afirmativa.

5 AGRADECIMENTOS

Os autores são especialmente gratos a Raphael Melo Guedes e a Felipe da Rocha Henriques por gentilmente fornecerem suas gravações para os testes efetuados neste artigo.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] John R. Deller, John H. L. Hansen, John G. Proakis, "Discrete-Time Processing of Speech Signals", IEEE Press, 2000.

[2] Simon Haykin, *Neural Networks - A Comprehensive Foundation* Prentice Hall, Segunda Edição, 1999

[3] Riedmiller, M., and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm", Proceedings of the IEEE International Conference on Neural Networks, 1993.

[4] S. Amari, "Natural gradient works efficiently in learning", Proc. ICASSP, 2002.

[5] A. Hyvärinen, "Fast and robust fixed-point algorithm for independent component analysis", IEEE Trans. Neural Networks, 1999.

[6] N. Delfosse e P. Loubaton, "Adaptive blind separation of independent sources: A deflation approach", Signal Process., vol. 45, 1995.

[7] K. Matusoka, "Minimal Distortion Principle for Blind Source Separation", SICE, pp. 2138-2143, 2002.

[8] J. Anemüller e B. Kollmeier, "Amplitude Modulation decorrelation for convolutive blind source separation", Proc. ICA, pp. 215-220, 2000.

[9] N. Murata, S. Ikeda e A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals", Neurocomputing, vol. 41, pp. 1-24, 2001.

[10] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda e F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions", Proc. ICASSP, pp. 3140-3143, 2002.

[11] M. Z. Ikram e D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation", Proc. ICASSP, pp. 881-884, 2002.

[12] H. Sawada, R. Mukai, S. Araki e S. Makino, "A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation", IEEE Transactions On Speech And Audio Processing, vol. 12, pp. 530-538, 2004.

[13] R. Aichner, H. Buchner, e W. Kellerman, "Exploiting Narrowband Efficiency for Broadband Convolutive Blind Source Separation", EURASIP Journal on Advances in Signal Processing, 2007.

[14] H. Buchner, R. Aichner, e W. Kellerman, "A Generalization of Blind Source Separation Algorithms for Convolutive Mixtures Based on Second-Order Statistics", IEEE Transactions on Speech and Audio Processing, vol. 13, no. 1, 2005.

[15] L. C. Parra e Clay Spence, "Convolutive blind separation of non-stationary sources", IEEE Transactions on Speech and Audio Processing, vol.8, pp. 320-327, 2000.

[16] M. Z. Ikram e D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment", Proc. of the IEEE International Conference on Acoustics, Speech ans Signal Processing, vol. 2, pp. 1041-1044, 2000.

[17] N. Mitianoudis e M. E. Davies, "Audio Source Separation of Convolutive Mixtures", IEEE Transactions on Speech and Audio Processing, vol. 11. no.5, 2003.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Wavelets aplicadas à Dynamic Time Warping

Sylvio Barbon Júnior[1], Rodrigo Capobianco Guido[1], Lucimar Sasso Vieira[1] e Fabrício Lopes Sanchez[1]
[1]Universidade de São Paulo
São Carlos, São Paulo, 13566-590, Brasil
barbon@ursa.ifsc.usp.br, guido@ifsc.usp.br, {sassovieria, sanchez}@ursa.ifsc.usp.br

RESUMO

Dynamic Time Warping (DTW) é uma técnica do tipo *pattern matching* para reconhecimento de padrões de voz, sendo baseada no alinhamento temporal de um sinal com os diversos modelos de referência. Uma desvantagem da DTW é o seu alto custo computacional. Este trabalho apresenta uma versão da DTW que, utilizando a Transformada Wavelet Discreta (DWT), reduz a sua complexidade. O desempenho obtido com a proposta foi muito promissor, ganhando em termos de velocidade de reconhecimento e recursos de memória consumidos, enquanto a precisão da DTW não é afetada.

INTRODUÇÃO

O reconhecimento automático de fala já tem sua importância e espaço conhecidos, assim diversas técnicas foram desenvolvidas e estas vem sendo aprimoradas para a obtenção de melhores resultados.

Existem duas abordagens básicas para reconhecimento de voz [1]: técnicas do tipo *pattern matching* e técnicas do tipo *knowledge-based*. Dois algorítmos que implementam tais técnicas são, respectivamente, *Dynamic Time Warping* (DTW) e *Hidden Markov Moddels* (HMM).

Neste artigo será utilizada a técnica DTW. Essa técnica compreende a identificação de uma palavra ou fonema baseada em uma biblioteca de modelos [2]. Assim a proposta do artigo é justamente melhorar o desempenho da comparação entre entrada e modelo, fazendo isso com a aplicação da Transformada Wavelet Discreta *(Discrete Wavelet Transform* – DWT).

A DWT consiste na versão discreta da Transformada *Wavelet*, que tem sido utilizada para análise de sinais diversos, pois é ótima para descrever tais sinais examinando-os tanto no domínio do tempo quanto no da freqüência [3].

Os experimentos realizados utilizaram a base de análises de fala DARPA TIMIT [5], a mesma é discutida em uma seção destinada a esse assunto, cujo nome é ”Base TIMIT”.

O trabalho mostra o ganho de desempenho na aplicação do algoritmo da DTW, uma vez que ao invés de utilizar o sinal original, o sinal filtrado (considerando apenas as baixas freqüências) é passado ao algoritmo, isto porque ele tem um comprimento menor e praticamente a mesma representatividade das características do sinal original.

Este artigo apresenta uma introdução dos assuntos, seguido por um aprofundamento a respeito da DTW e o reconhecimento de voz, citando e a técnica da DTW. Na seqüência é apresentada uma revisão breve sobre *Wavelets* e o processo de *downsampling*. Subsequentemente é apresentada a metodologia de trabalho, incluindo menção ao *software* de processamento e a base de dados TIMIT e como a DWT pode diminuir a complexidade do algoritmo da DTW. Os resultados são apresentados na sequência, sendo que o artigo é finalizado com uma conclusão sobre o trabalho.

DYNAMIC TIME WARPING E O RECONHECIMENTO DE VOZ

DTW pode ser usada em aplicações de reconhecimento de voz com um pequeno vocabulário de "protótipos" de pronúncias armazenadas [4].

Um exemplo de sua utilização pode ser visto na tarefa de reconhecer qual dos números (0-9) foi dito. A sinal de entrada (número ditado) pode ser comparado na biblioteca com as dez amostras armazenadas, a que mais se igualar ao sinal de entrada será considerada o número pronunciado.

Umas das características da DTW é que o sinal de entrada pode ter um comprimento diferente do sinal original. Isso faz com que o mesmo sinal pode ser reconhecido mesmo sendo falado em um intervalo de tempo diferente, o que possibilita que diferentes locutores possam ter pronunciado uma mesma palavra, cada um com suas características e tempo de fala diferente.

Técnica DTW

A técnica DTW tenta encontrar o melhor caminho *w* através da tabela de associação de índices (*best path*) construída pela delimitação do fonema [6]. O algoritmo da DTW tem sistematicamente variado para determinar os efeitos no desempenho do reconhecimento [2].

Assumindo que as amostras de entrada e modelo (referência) são conhecidas, a primeira matriz é montada, neste caso a matriz Solução. Os elementos desta matriz são formados pelo módulo da diferença entre cada ponto do sinal de entrada com o sinal de referência.

Assumindo como sinal de entrada x[ ] = {1, 1, 2, 3, 2, 0} e o sinal y[ ] = {0, 1, 1, 2, 3, 2, 1} como referência, temos a seguinte matriz solução:

1	0	0	1	2	1	1
2	1	1	0	1	0	2
3	2	2	1	0	1	3
2	1	1	0	1	0	2
1	0	0	1	2	1	1
1	0	0	1	2	1	1
0	1	1	2	3	2	0
	1	1	2	3	2	0

Figura 1 - Matriz Solução

A próxima etapa é justamente montar a matriz de Distância Acumulada, esta é formada pela iteração na soma dos valores de cada elemento da matriz Distância Acumulada (DA) com o elemento superior do mesmo na Matriz Solução (S) , como visualizado pela Equação 1.

$$DAi,j = DAi,j + Si\text{-}1,j \quad , i > 1, j > 1; \qquad (1)$$

Após preencher a primeira coluna assume-se:

$$DA_{i,j} = DA_{i,j} + S_{i,j+1} \quad , j < \text{quantidade de colunas}; \qquad (2)$$

A matriz DA deve ser inicializada com o valor 1 no último elemento da primeira coluna, como na Figura 2.

5+0	5					
4+1	5					
2+2	4					
1+1	2					
1+0	1					
1+0	1					
	1	2	4	7	9	9
		1+1	2+2	4+3	7+2	9+0

Figura 2 - Primeira iteração para o cálculo da DTW na matriz de Distância Acumulada.

Primeiramente devem ser preenchidos os valores das colunas, do último elemento para o primeiro. Depois devem ser calculados os elementos da esquerda para a direita, como nas setas indicativas na Figura 2.

Quando a matriz DA estiver finalizada, o próximo passo é criar a matriz de Movimento. A matriz de Movimento deve ter o último item da primeira coluna igual a 0. Deve ser realizada uma iteração no sentido baixo-cima analisando qual o valor é menor (na matriz DA). Caso o menor valor seja o elemento abaixo, a matriz Movimento deve ser preenchida com o valor 1, caso na matriz DA o menor elemento seja o elemento imediatamente à esquerda, deve-se preencher a matriz movimento com o valor 3. Caso o menor valor seja a diagonal inferior a esquerda ou os valores forem iguais, deve-se colocar o valor 2 na matriz Movimento, como exemplificado na Figura 3.

	5	5	3	4	2	2
	5	5	2	2	1	3
	4	4	2	1	2	5
	2	2	1	2	2	4
	1	1	2	4	5	6
3	1	1	2	4	5	6
2	1	2	4	7	9	9
		1				

Figura 3 - Matriz Distância Acumulada completa.

1					
1					
1					
1					
1					
1	2				
0	3	3	3	3	3

Figura 4 - Matriz Movimento.

Finalmente após completada a matriz Movimento, deve-se desenvolver a matriz *Best Path (w)*, melhor caminho. Para isso, observando a matriz Movimento, partindo do último elemento da primeira linha, escolhe-se o próximo elemento com a menor distância (m_d) valores dos elementos [4], como demonstrado na Equação 3.

$$m_d = \min \{ | w_{i,j} - w_{i-1,j} | , | w_{i,j} - w_{i,j-1} |, | w_{i,j} - w_{i-1,j-1} | \}, i > 1, j > 1 \qquad (3)$$

Desta forma, a cada elemento escolhido, cria-se uma marca na matriz *Best Path*, até que se alcance o último elemento da primeira coluna. O resultado é algo como demonstrado na Figura 5.

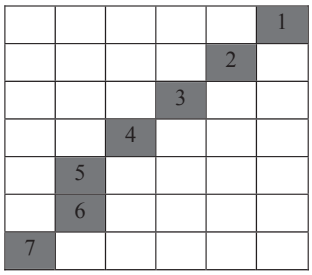


Figura 5 - Matriz Best Path

Como demonstrado pela Figura 5, a matriz *Best Path* para o sinal x[ ] quando comparado ao modelo y[ ], tem distância igual a 7.

Baseado nesta técnica, a aplicação da DTW a um sinal de voz armazenado em um arquivo de áudio consistiria, como na Figura 6, em:

1- Capturar o(s) arquivo(s);
2- Extrair as amostras relativas ao fonema a ser modelo e o fonema de entrada;
3- As amostras dos dois fonemas seria analisada pela DTW formando a matriz *Solution;*
4- As demais etapas seriam aplicadas até encontrar a matriz *Best Path*.

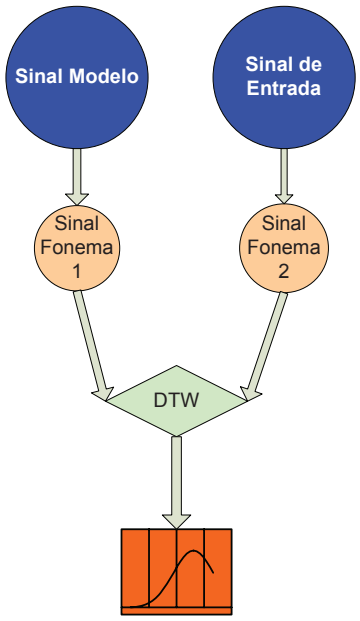


Figura 6 - Algoritmo da análise sem filtragem.

TRANSFORMADA DISCRETA WAVELET

A DWT de um sinal discreto f[ ], contendo n amostras (quantidade de pontos do sinal), é um outro sinal discreto y[ ], também de n pontos. O sinal y[ ] contém informações sobre as freqüências que compõe o sinal f[ ] (original), e também onde tais freqüências se localizam em f[ ].

Downsampling

Até então, o sinal de entrada continua com o mesmo comprimento e a complexidade computacional similar. Dessa maneira é empregada a técnica conhecida com *downsampling* por 2 (ou *subsampling* por 2), onde esta consiste em descartar uma amostra de y [ ] a cada duas [3].

Este descarte é realizado a cada nível da DWT, reduzindo a cada passo o número de amostras do novo sinal. É importante salientar que este descarte não afeta efetivamente o sinal, mantendo nele as propriedades do sinal f [ ] (o original), isso é uma das características da DWT.

Desta forma, é necessária a avaliação de um limiar para os níveis a serem alcançados, de forma a preservar uma quantidade mínima de pontos que preserve o sinal.

METODOLOGIA

Os resultados apresentados por este artigo foram conquistados através da utilização de um software para aplicação da DWT e posteriormente utilizar a DTW para verificar os sinais resultantes da filtragem com *wavelets*.

Todo o software de filtragem e análise foi desenvolvido em linguagem C++, este assunto será contemplado de forma mais ampla no tópico "Software de Processamento". A análise posterior foi realizada através da visualização do resultado gráfico da DTW por um software desenvolvido em linguagem Java.

Os sinais analisados foram fonemas, assim um fonema *template (*modelo) era comparado a diversos fonemas de entrada (*input*) outros.

A comparação mantinha ambos os sinais (modelo e entrada) sob as mesmas condições, por exemplo, se o modelo sofria uma filtragem com a wavelet da família *Daubechies* [3], de suporte quatorze até o nível 8, o mesmo deveria ser feito com o sinal de entrada.

Os sinais analisados eram todos da base de análises de fala DARPA TIMIT [5], está será mais comentada no tópico intitulado "Base TIMIT".

O software de analise recebia como análise a listagem de fonemas a serem analisados, entre estes era destacado o fonema *template*. As análises eram realizadas em batch, onde como saída o sistema exibia a matriz *Best Path*[4] de cada comparação.

Posteriormente para uma análise mais clara, as matrizes *Best Path* obtidas eram importadas pelo software desenvolvido para exibição dos resultados em um gráfico cartesiano.

Baseado nas análises sobre o gráfico cartesiano e comprimento (caminho) demonstrado pela matriz *Best Path* eram observados os resultados do reconhecimento.

Software de Processamento

O software de processamento dos dados relativos ao sinal foi desenvolvido com a função de automaticamente aplicar a filtragem (DWT) e comparação (DTW). A Figura 7 apresenta o diagrama dos passos realizados.

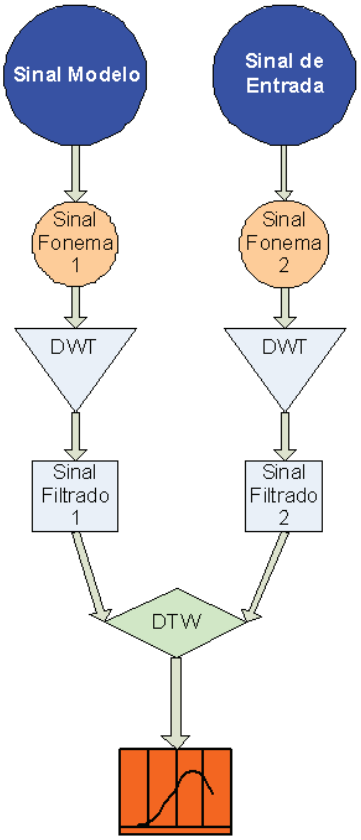


Figura 7 - Algoritmo de filtragem e análise.

Primeiramente um modelo padrão em formato de texto puro era passado ao software. Neste modelo estão as informações de qual o sinal modelo, qual o sinal de entrada, qual fonema deve ser analisado e se é necessária a aplicação do filtro e as demais especificações da filtragem.

A informação de qual sinal (modelo ou entrada) era para ser utilizado foi passado como localização física do arquivo em disco, arquivo este, em formato *wav ()*.

O *software* carrega o arquivo do sinal modelo, extraindo apenas os dados brutos do fonema em questão, parte que contém o valor das amostras efetivamente, aplica a DWT sobre os dados do fonema extraído. Em seguida, faz o mesmo para o arquivo do sinal de entrada.

Após extraídos e filtrados (caso requisitado) as novas amostras são passadas para o algoritmo DTW, que realiza os cálculos como comentado no tópico "Algoritmo DTW", dentro de *Dynamic Time Warping* e o Reconhecimento de Voz.

A saída do algoritmo da DTW é persistido em um arquivo, o conteúdo é relativo as informações da filtragem aplicada seguido da matriz *Best Path* da análise.

Diminuindo a Complexidade

Sendo a complexidade computacional de um algoritmo o termo referente à quantificação dos cálculos necessários durante a execução do código para resolver um problema, diminuindo a quantidade de dados de entrada para os cálculos a serem realizados, diminuímos também a carga de operações realizadas.

O sinal y [ ] (filtrado), que corresponde a DWT do sinal original f[ ], é formado pela concatenação seqüencial dos sub-sinais y1[ ] e y2[ ], como representado na Figura 8. O sub-sinal y1[ ], conhecido como sinal de aproximação, contém as freqüências de zero até a metade da máxima freqüência de f[ ]. Por sua vez, y2[ ], conhecido como sinal de detalhamento, contém as freqüências acima da metade da máxima freqüência presente em f[ ] . Muitas vezes, y1[ ] contém mais de 95% da energia total de f[ ] [7].

Consequentemente, ao invés de utilizar todos os pontos de f [ ] para aplicar a DTW, podemos utilizar somente uma parte do sinal filtrado, no caso, y1 [ ]. Em resumo, ao invés de utilizar f[ ], de comprimento *n* pontos, utilizamos y1[ ], com *n/2* pontos. A quantidade de pontos é diretamente proporcional ao número de iterações e cálculos que o algoritmo DTW irá calcular, desta maneira minimizando os pontos, teremos menos iterações, e assim uma complexidade menor e um desempenho consideravelmente superior ao modelo original.

Por sua vez, pode-se aplicar a DWT em nível superior, e assim, tentar minimizar ainda mais a complexidade. A complexidade pode ser diminuída a cada nível, isto ocorre pois conforme a *wavelet* avança um nível, a sua quantidade de amostras é reduzida pela metade, como observado na Figura 8.

Esta redução pela metade se deve ao método *downsampling* por 2. Porém existe um limite de níveis a ser atingido pela *wavelet*. Este limite varia de acordo com a preservação de informação no último nível em questão. Por exemplo, um trecho de 1233 amostras. A mesma deve sofrer uma adaptação para que seja aplicada a DWT, esta adaptação implica na aproximação da quantidade de amostras para um número potência de dois. Desta forma o número de amostras a serem utilizadas seriam 1024 (2^{10}), onde 209 amostras deveriam ser descartadas.

A quantidade de pontos a ser aproximada à potência de dois, normalmente, é um número de valor inferior a quantidade de amostras original, além de diminuir sensivelmente a quantidade de cálculos, o sinal não necessitaria ser coberto por amostras arbitrárias para alcançar o número requisitado.

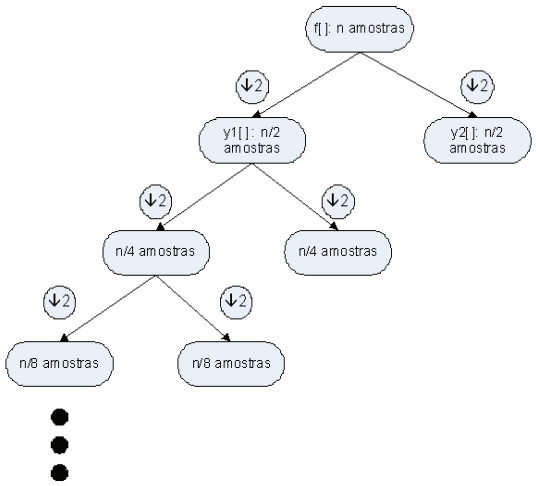


Figura 8 - Funcionamento da DTW para um sinal f[ ] com n amostras.

Base TIMIT

A base de fala TIMIT é um agrupamento de narrações de uma grande quantidade de frases em inglês narradas por nativos Americanos [5].

O TIMIT foi desenvolvido pela Texas Instruments and MIT, sendo distribuído por U. S. National Institute of Standards and Technology em um CD_ROM. Na mídia citada, são encontradas 2342 diferentes frases, lidas por cerca de 630 narradores [5].

Todas as frases narradas são catalogadas, neste catálogo é possível observar a divisão da sentença em palavras. Cada palavra tem em sua descrição a posição inicial e final das amostras que representam as informações da palavra.

Outra informação cadastrada é a posição de cada fonema na frase, onde este, como as palavras, apresentam a posição inicial e final. Esta informação propiciou a extração e filtragem de apenas do fonema requisitado. Na Figura 7 para observar esta extração basta atentar-se para a transição entre o circulo maior (p. ex. "Sinal Modelo") e o circulo menor, intitulado "Sinal Fonema".

Desta forma a base TIMIT é uma ferramenta muito útil, uma vez que a mesma apresenta uma coleção abrangente de arquivos para análise de voz e as informações relativas ao posicionamento de cada fonema no sinal.

RESULTADOS

Os resultados apresentados foram satisfatórios, uma vez que a DWT efetivamente aumentou o desempenho no cálculo da DTW.

Foi observado que a família da *wavelet* não afeta profundamente o resultado da análise. O suporte do filtro também não foi de eficiência relevante.

A principal contribuição da *wavelet* empregada foi o seu nível alcançado, onde este a cada iteração destacava um novo sinal pertinente com a metade da quantidade de amostras do anterior.

Era necessário controlar o nível alcançado, uma vez que caso as iterações alcançassem o nível máximo da *wavelet*, o sinal seria composto por apenas uma amostra, o que tornaria impossível a aplicação da DTW.

Assim é recomendado aprofundar-se a um nível que permita o sinal ser composto por uma quantidade de 32 ou 16 amostras.

Caso seja utilizadas menos amostras do que o proposto no parágrafo anterior, o reconhecimento poderá ficar comprometido.

CONCLUSÃO

Neste artigo foi proposta a utilização da filtragem digital de um sinal com a DWT de forma a reduzir a complexidade computacional do mesmo. Esta redução se faz necessária pois a aplicação da DTW é computacionalmente complexa, variando de acordo com o tamanho (quantidade de pontos) do sinal de entrada.

O tipo de família *wavelet* não é um fator de destaque entre as análises e os resultados. O mais significante para a mudança no desempenho é o nível máximo alcançado pelo sinal.

O que se deve considerar realmente é o nível da *wavelet,* pois este influencia diretamente a quantidade das amostras.

Os resultados obtidos foram plenamente satisfatórios, foi possível reduzir consideravelmente a quantidade de pontos do sinal original e o mesmo continuou a ser reconhecido.

Estão sendo realizados mais testes para verificar de forma mais aprofundada a influência das diferentes famílias de *wavelets* na formação da matriz *Best Path*, estes resultados futuramente serão publicados.

REFERÊNCIAS

[1] Peter H. W. Wong, Oscar C. Au, Justy W.C. Wong, William H. B. Lau, *Reducing Computational Complexity of Dynamic Time Warping Based Isolated Word ecognition with Time Scale Modification*, ICSP 98.

[2] Cory Myers, Lawrence Rabiner, *Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition*, Ieee Transactions On Acoustics, Speech, And Signal Processing, Vol. Assp-28, No. 6, Dezembro 1980.

[3] Paul S. Addison, *The Illustrated Wavelet Transform Handbook*, , Institute of Physics Publishing, 2002, Edinburg, United Kingdom.

[4] Coleman John, *Introducing Speech and Language Processing,* Cambridge University Press, 2005, Cambridge, United Kingdom.

[5] Keating A. P., Byrd D., Flemming E., Yuichi T., *Phonetic analyses of word and segment variation using the TIMIT corpus of American English*, 1993, Department of Linguistics, University of California, 405 Hilgard Avenue, Los Angeles, USA.

[6] Robert W. P., Damper I. Luk., *Interference of Letter-Phoneme Correspondeces by Delimiting and Dynamic Time Warping Techniques*, Departament of Eletronics and Computer Science, University of Southampton, Southampton, United Kingdom.

[7] Walker, J. S. *A Primer on Wavelets and Their Scientific Applications.* Chapman & Hall/CRC, 1999, Washington, USA.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

On the Use of Wavelets for Voice Conversion

Lucimar Sasso Vieira,[1] Fabrício Lopes Sanchez,[1] Sylvio Barbon Junior,[1] and Rodrigo Capobianco Guido[1]

[1] University of São Paulo at São Carlos, Institute of Physics of São Carlos, Department of Physics and Informatics, United Group for Audio and Speech Processing
400 Trabalhador São Carlense Avenue ,São Carlos , São Paulo, 13560-970, Brazil

sassovieira@ursa.ifsc.usp.br, fabricio.sanchez@apo.org.br, barbon@ursa.ifsc.usp.br, guido@ifsc.usp.br

ABSTRACT

Nowadays, among many usual techniques for voice morphing, the ones based on wavelet filter banks, associated with artificial neural networks, have been emphasized. This work concentrates on such techniques, presenting a study on the best wavelet for converting each particular voice pattern, with a detailed analysis of the characteristics that take to such outcomes. Tests are performed with speech data collected from TIMIT corpus provided by Linguistic Data Consortium (LDC).

0 INTRODUCTION

Voice conversion[1][2], also kown as voice morphing or voice transformation, is the process by which a sentence uttered by a source speaker is transformed as if it were uttered by a target speaker. Nowdays, the main applications of voice conversion are: reproduction of the voice of a famous person for applications in mobile telephone, Internet voice applications with e-mail readers and screen readers for the blinds, as well as computer and video game applications with game heroes speaking with desired voices, plus other applications.

The algorithm we describe in this paper concentrates on a wavelet-based [3-6] procedure for voice conversion. This particular class of algorithms takes advantage of a special property of the Discrete Wavelet Transform (DWT), that is the sub-band separation. Then, an Artificial Neural Network (ANN) is used to convert each band.

This paper is organized as follows. First, we present a brief review on DWT, ANN, perceptual auditory masking, and voice conversion systems. Then, we show the proposed algorithm for voice conversion. Next, we describe the tests and conclusions. Lastly, the bibliografical references and acknowledgments are presented.

1 BRIEF REVIEW ON THE LITERATURE

In this section we present a brief explanation on wavelet analysis and neural networks, as well as some techniques for voice conversion.

1.1 The Discrete Wavelet Transform

The DWT [3][4][5][6], which has as its main idea the process of multi-resolution analysis (MRA), is one of the most appropriate techniques to make a joint time-frequency analysis of a discrete-time signal. Let $\vec{f}$ be the discrete-time signal under analysis, containing N samples. Then, it may be decomposed in j levels of resolutions, according to figure 1, as follows:

$$f[n] = \sum_{k=0}^{\frac{n}{2^j}-1} H_{j,k}[n]\phi_{j,k}[n] + \sum_{t=1}^{j}\sum_{k=0}^{\frac{n}{2^j}-1} G_{t,k}[n]\psi_{t,k}[n] \quad (1)$$

In this case:

- $\phi[n]$ and $\psi[n]$ form a Riesz basis to write $\vec{f}$;
- $\phi[n] = \sum_k h_n\phi[2n-k]$, defined recursively by dilations and translations of itself, is called scaling function;
- $\psi[n] = \sum_k g_n\phi[2n-k]$, is called wavelet function, which is orthogonal to the scaling function;
- $H_{j,k}[n] =< f, \phi_{j,k}[n] >$;
- $G_{t,k}[n] =< f, \psi_{t,k}[n] >$;
- the h_k coefficients corresponds to a low-pass filter and the g_k to a high-pass one, formming a Quadrature Mirror Filter (QMF) [13] system. Then, $g_k = (-1)^k h_{T-k-1}$. Furthermore, a filter with T coefficients is called filter of support T and order $T-1$.

In pratice, the multiresolution decomposition is computed using a simple process of the discrete convolutions of the signal under analysis with the filter coefficients h_k, and g_k,

$$y_{low}[n] = f * h = \sum_{k=0}^{T-1} h[k]f[2n-k], 0 \leq n \leq N \quad (2)$$

$$y_{high}[n] = f * g = \sum_{k=0}^{T-1} g[k]f[2n-k], 0 \leq n \leq N \quad (3)$$

where T corresponds to the support of the filters $h[\ \ ]$ and $g[\ \ ]$, and $*$ denotes the discrete convolution. After each level of decomposition there is a downsampling by 2 of the filtered signal, due to filtering that allows only half-band of the original signal to pass. A much more detailed description of wavelet transform can be found in the references, specially[3] and [7].

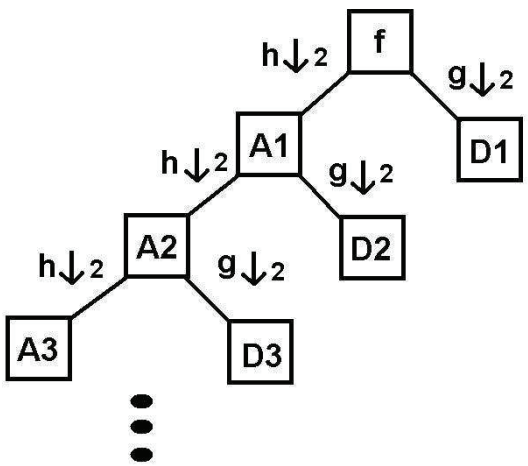


Figure 1: Wavelet decomposition tree, where f is the original signal split into an aproximation, $A1$ and a detail $D1$. Parting from $A1$, the process continues recursively.

1.2 Principles of Artificial Neural Networks

Humans and others animals process information with neural networks [8]. They are formed from trillions of neurons, i.e., nerve cells, exchanging brief electrical pulses called actions potentials. Computer algorithms that mimic these biological structures are formally called artificial neural networks (ANNs). In this work we use an artificial neural network to transform one person's speech pattern into another person's pattern, giving it a new identity while preserving the original content of the spoken message. Many different neural network structures have been tried, some based on imitating what a biologist sees under the microscope, some based on a more mathematical analysis of the problem. The most commonly used structure is shown in figure 2 and is called multi-layer perceptron.

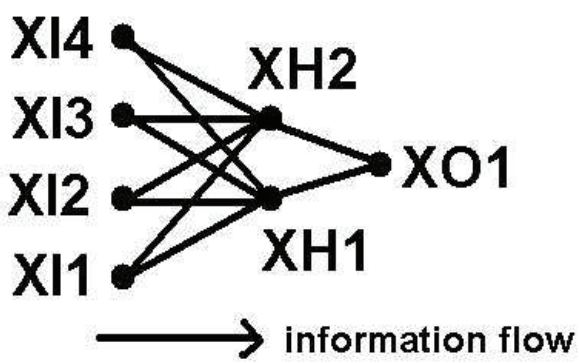


Figure 2: Example of a tipical structure of an ANN. In this case, there are 3 layers with 4 neurons in the input layer(XI), 2 in the hidden layer (XH), and finally 1 in output layer (XO).

This neural network is formed by three layers, called the input layer, hidden layer, and output layer. Each layer consists of one or more nodes (neurons). The lines between the nodes indicate the flow of information from one node to the next. In this particular type of neural network, the information flows only from the input to the output, that is, from left-to-right. The nodes of the input layer are passive, meaning that they do not modify the data. They receive a single value on their input, and duplicate the value to their multiple outputs. In comparison, the nodes of the hidden and output layers are active. This means that they modify the data, as shown in figure 3.

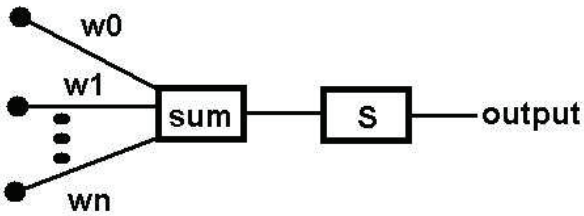


Figure 3: How an active neuron works in an ANN. Each input is multiplied by a weight, and the sum passes through an activation function to produce the output.

The nodes XI hold the data to be evaluated (see figure 2). For example, XI1 may be some pixel values from an image, samples from an audio signal, and so on. Each value from the input layer is duplicated and sent to all of the hidden nodes. This is called a fully interconnected structure. As shown in figure 3, the values entering a hidden node are multiplied by weights, a set of pre-determined numbers stored in the program. The weighted inputs are then added to produce a single number. Before leaving the node, this number is passed through a non-linear mathematical function called a sigmoid, that is given by:

$$s(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

This is an "S"-shaped curve that limits the node's output. That is, the input to the sigmoid is a value between - $\propto$ e + $\propto$, while its output can only be between 0 and 1. The outputs from the hidden layer are represented in the flow diagram (figure 2) by the variables XH . Just as before, each of these values XH1, XH2, and so on, is duplicated and applied to the next layer. The active nodes of the output layer combine and modify the data to produce the two output values of this network, XO1. Neural networks can have any number of layers, and any number of nodes per layer. Most applications use the three layer structure with a maximum of a few hundred input nodes. The hidden layer is usually about 10% the size of the input layer. In the case of target detection, the output layer only needs a single node. The ability of the neural network to provide useful data manipulation lies in the proper selection of the weights [8]. In this work we used a supervised training algorithm, the traditional back-propagation, for the training. This algorithm works as follows:

- i) an input example, and the corresponding desired output, are presented to the ANN;
- ii) the difference between the output and the desired output is found;
- iii) the error vector's gradient is found and the output layer's weights are updated;
- iii) the desired values are back-propagated to update the weights of the others layers.

1.3 The Perceptual Model of the Human Auditory System

The hearing threshold describes the lowest sound level at which a signal can be heard at a given frequency and given's listener age. A fundamental concept is the frequency auditory masking that occurs according to the time-frequency resolution of the auditory system, i.e., the critical bands [9]. Based on this concept, if two signals of different frequencies, that belong to the same critical band, occur at the same time, or very close one to the other, the weaker will not be perceived. This property is mostly explored in this work. Table 1 shows the critical bands in frequency scale (Hz) and the Bark scale (z). The conversion between Hertz and Bark is given by:

$$z = \frac{26.81f}{1960 + f} - 0.53 \quad , \tag{5}$$

where f is the frequency in Hertz.

1.4 Algorithms for Voice Conversion

At this moment we will make a brief review on the literature about voice conversion methods that use the

Bark (z)	*Hertz (Hz)*	*Bark (z)*	*Hertz (Hz)*
0	0 - 100	12	1720 - 2000
1	100 - 200	13	2000 - 2320
2	200 - 300	14	2320 - 2700
3	300 - 400	15	2700 - 3150
4	400 - 510	16	3150 - 3700
5	510 - 630	17	3700 - 4400
6	630 - 770	18	4400 - 5300
7	770 - 920	19	5300 - 6400
8	920 - 1080	20	6400 - 7700
9	1080 - 1270	21	7700 - 9500
10	1270 - 1480	22	9500 - 12000
11	1480 - 1720	23	12000 - 15500
		24	15500 - 22050

Table 1: The 25 critical bands of the human auditory system.

techniques presented above. In the past, elementary systems for voice conversion were based in two stages: in the first one, some technique for speech recognition (speech-to-text) was used, and then a morphed speech was synthesed according with the target speaker's pattern. The following items show some of the current approaches for voice conversion:

- Arslan and Talkin[10] proposed a system where both source and target's voices were segmented into phonemes and converted using a dictionary. Therefore, the final quality depends on the dictionary's size;
- Stylianou et al[11] proposed the use of Gaussian Mixture Models (GMM) to improve the above system. Particularly, source and target's voices are time-aligned, then the GMM is trainned using the expectation maximization algorithm (EM) and lastly the mapping is performed;
- Kain[12] presented a mapping algorithm of espectral envelope like Arslan and Talkin did. Additionaly, he included the error prediction to try best results.
- Orphanidou et al[13] used Generative Topographic Mapping to transform the source's spectra parameters into target's spectra ones, according the speakers' vocal tractal models.
- Orphanidou et al[14] proposed a better approach using time-frequency wavelet analysis and a ANN. The wavelet decomposition separates the speech signal into sub-bands and then, for each sub-band, the ANN makes the mapping from the source speaker's patter to the target speaker's patter.

2 THE PROPOSED ALGORITHM

In this section we describe the proposed algorithm for voice conversion with details, including the system's architecture.

2.1 System's Architecture

The proposed system is divided in two stages: one for training and another for testing, according to the figure 4.

2.2 The Algorithm for Voice Conversion

The detailed description of the algorithm follows.

- **BEGIN**
- **STEP 1**: Source and target voice files are chosen for two individuals uttering the same sentence;
- **STEP 2**: Extract the raw data from each wav file;
- **STEP 3**: The raw source and target samples are time-aligned so that both signals will be of the same length;

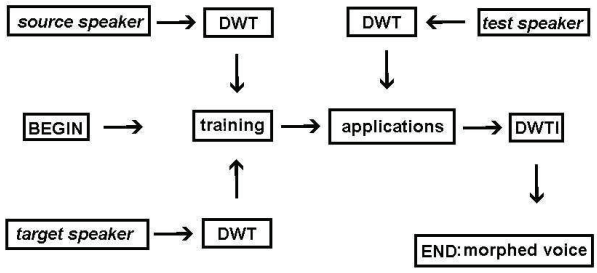


Figure 4: The architecture of the system: a stage for training and other for testing.

- **STEP 4**: Make zero-padding in the source and target signals in order to have multiple of 2048 lenght ;
- **STEP 5**: The source and target sample are normalized in order to have 0 mean and standard deviation;
- **STEP 6**: Divide source sample into n sequencial frames of 2048 samples each. For each frame s_i:
  - **STEP 6.1**: Apply the 11th-level Discrete Wavelet-packet Transform (DWPT), that has also length 2048;
  - **STEP 6.2**: Normalize DWTP11(s_i) in order to have a $\{0-1\}$ amplitude;
  - **STEP 6.3**: Separate DWTP11(s_i) in 25 sub-bands according to the perceptual model of the human auditory system [26] described in table 1.
- **STEP 7**: Divide target signal into n sequencial frames of 2048 samples each. For each frame t_i:
  - **STEP 7.1**: Repeat step 6.1 for DWTP11(t_i);
  - **STEP 7.2**: Repeat step 6.2 for DWTP11(t_i);
  - **STEP 7.3**: Repeat step 6.3 for DWTP11(t_i);
- **STEP 8**: For each frame (DWTP11(s_i)) and (DWTP11(t_i)):
  - **STEP 8.1**: Create a multi-layer perceptron to convert $DWTP11(s_i)_j$ into $DWTP11(t_i)_j$. The number of neurons in the input, hidden, and output layers is the same;
  - **STEP 8.2**: Train each network using the back-propagation algorithm so that the error is less than 10^{-3};

2.3 The Algorithm for ANN Training

The procedure used to train the perceptrons, i.e., the back-propagation, follows [15]:

- to find the news weigths (wh) of hidden layer, repeat:
  - for each neuron j in the hidden layer:
    - for each neuron i in the output layer:
      - $slope_o \leftarrow XO * (1 - XO)$
      - $slope_h \leftarrow XH_{,j} * (1 - XH_{,j})$
      - $\frac{\Delta XO}{\Delta w} \leftarrow X_{1,i} slope_h w_j slope_o$
      - $wh_{i,j} \leftarrow wh_{i,j} + \frac{\Delta XO}{\Delta wh} tm$
- while the error is more than 10^{-3};
- to find the news weigths (wo) of output layer, repeat:
  - for each neuron j in the hidden layer:
    - $slope_o \leftarrow XO * (1 - XO)$
    - $\frac{\Delta XO}{\Delta w} \leftarrow XH_{,i} slope_o$
    - $wo_{,j} \leftarrow wo_{,j} + \frac{\Delta XO}{\Delta wo} tm$
- while the error is more than 10^{-3}

  In the algorithms, t is the current input error, m is a constant set in 10^{-1}, XO are the neurons of the output layer and XH are neurons of the hidden layer. The implementation code of this work is written using C++ programming language under Mandriva Linux.

3 TESTS AND RESULTS

In this section we show the tests and results found according to the support-size of the wavelet filters used and the wavelet family. The system was tested using sentences from the TIMIT database [16]. The tests of convergence are in table 2. A perceptual ABX-like preference test was also carried out, where ten listeners, men and women, of ages between ten and seventy years old were asked to rate the morphed speech. The table 2 also shows these results.

4 COMMENTS AND CONCLUSIONS

In this work, we presented a study on voice conversion based on neural networks and wavelets. The tests have demonstrated that Daubechies' wavelet with support-size 30 presented the best results for the perceptual criterion. This result indicates that a maximally-flat frequency response close to the ideal response is important. Future work includes the construction of a matched wavelet filter-bank to improve the results.

5 ACKNOWLEDGMENT

We would like to thank FAPESP (The State of Sêo Paulo Research Foundation at http://www.fapesp.br) for the grants and support given to this work.

DWT	support	M	S
Haar	2	5.2	0.75
Daubechies	4	5.8	1.0
Daubechies	6	6.0	1.0
Daubechies	8	6.1	1.1
Daubechies	10	6.5	1.0
Daubechies	20	6.9	0.97
Daubechies	30	7.0	0.90
Vaidyanathan	24	7.0	1.1
Coiflet	6	6.0	1.05
Coiflet	12	6.0	1.04
Coiflet	18	6.1	1.05
Coiflet	30	6.6	1.01
Symmlet	8	6.0	1.10
Symmlet	16	6.0	1.05
Burt Adelson	6	5.9	1.01
Beylkin	18	5.9	1.02

Table 2: Results of the tests for converting the sentence *sa1.wav* of the directory */timit/test/dr1/mdab0* into *sa1.wav* of the directory */timit/test/dr1/mjsw0*. The tests report the mean (M) and standard deviations (S) of the rates given by the volunteers participating in the tests. A rate of 0 means the morphed speech is not similar at all to the corresponding original one, while a rate of 10 means they are indistinguishable.

REFERENCES

[1] C.H. Wu, C.C. Hsia, T.H. Liu, J.F. Wang. *Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis.,* IEEE Trans. on Audio, Speech and Language Processing, v.14, n.4, July 2006, p.1109-1116.

[2] A. Mouchtaris, J. Van der Spiegel, P. Mueller. *Non-parallel training for voice conversion by maximum likelihood constrained adaptation*. Proc. of the Int. Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings(ICASSP'04), v.1, 17-21, May 2004, p.I 1-4.

[3] G. Strang, T. Nguyen.. *Wavelets and filter banks*. Wellesley-Cambridge Academic Press, Wellesley, 1997.

[4] P. S. Addison. *The Illustrated wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medice and Finance*. Edinburg-UK: Institute of Physics Publishing, 2002.

[5] J. S. Walker. *A Primer on Wavelets and Their Scientific Applications*. Whashington-USA: Chapman and Hall/CRC, 1999.

[6] A. Jensen, A. Cour-Harbo. *Ripples in Mathematics: The Discrete Wavelet Transform*. New York-USA: Springer-Verlag, 2000.

[7] J.R. Williams, K. Amaratunga. *Introduction to Wavelets in Engineering*. Int. Journal for Numerical Methods in Engineering, 37(1994), 2365-2388.

[8] Steven W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing*. editora: Newnes, 2002.

[9] R.C. Guido, L.S. Vieira, S. Barbon, et al. *A Fractal and Wavelet Based Approach for Audio Coding*. Proc. of the Int. Symposium on Multimedia, 2006. Proceedings(ISM'06), december 2006, p.253-256.

[10] L. Arslan, D. Talkin. *Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum*. Proc. EUROSPEECH, p. 1347-1350, 1997.

[11] Y. Stylianou, O. Cappe, E. Moulines. *Statistical methods for voice quality transformation*. Proc. EUROSPEECH, p. 447-450, 1995.

[12] A. Kain, M.W. Macon. *Spectral voice conversionfor text-to-speech synthesis*. Proc. ICASSP-98, 1:285-288, 1998.

[13] C.Orphanidou, I.M. Moroz, S.J. Roberts. *Voice morphing using the generative topographic mapping*. Proc. of the International Conference on Computer, Communication and Control Technologies, 1:222-225, 2003.

[14] C.Orphanidou et al. *Multiscale voice morphing*. Pre-print submitted to Pattern Recognition Letters, Elseiver, on September 2006, to appear in 2007.

[15] S. Haykin. *Redes Neurais: princípios e prática*. Porto Alegre: Bookman, 2001.

[16] Linguistic Data Consortium. *TIMIT Data Base*. http://www.ldc.upenn.edu, january 2007.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Estudos teórico e prático sobre o emprego da transformada wavelet contínua na detecção passiva em águas rasas

José Roberto Motta da Silva[1], Mariane Rembold Petraglia[1] e Carlos Eduardo Parente Ribeiro[2]
[1]PEE/COPPE, [2]PENO/COPPE
Rio de Janeiro, Rio de Janeiro, 21.941-947, Brasil
motta@pads.ufrj.br, mariane@pads.ufrj.br, parente@peno.coppe.ufrj.br

RESUMO

Este trabalho investiga a aplicação da transformada wavelet contínua (CWT) na detecção passiva de uma fonte sonora no mar, na maior distância possível. Quanto mais rápido o alvo for detectado, maior tempo disporá o comando para tomar as ações necessárias. Este estudo é desenvolvido em duas partes: a primeira apresenta a teoria da transformada wavelet contínua, e a segunda apresenta os resultados obtidos com a CWT utilizando sinais simulados, comparando-os com os resultados obtidos com a transformada de Fourier de tempo curto (STFT).

INTRODUÇÃO

O equipamento para a detecção dos navios que se deslocam primordialmente submersos é denominado "sonar" (SONAR é o acrônimo em inglês de "sound navigation and ranging", que significa navegação e medição de distância por meio do som).

Os sonares pertencem a dois grandes grupos:

- os ativos, cuja característica principal é a emissão de um pulso sonoro, cujo eco de retorno detalhará não só as informações pertinentes dos alvos detectados como também das características ambientais existentes no local;
- os passivos, cuja característica principal é a de extrair as informações características dos alvos detectados e das condições ambientais a partir dos sinais captados nos sensores de recepção.

Diversas técnicas de análise espectral têm sido estudadas com a finalidade de serem aplicadas na detecção de alvos [1] por sistemas sonares passivos de submarinos à longa distância em águas rasas [10].

As seguintes técnicas foram analisadas em [7]: detecção quadrática, função de correlação, espectro de Welch, cepstrum, transformada de Fourier em tempo curto (STFT) e transformada wavelet continua (CWT).

Neste artigo será apresentada a teoria da CWT, bem como sua aplicação na detecção de alvos utilizando sinais de sonares passivos. Serão apresentados os resultados comparativos da detecção com a CWT e com a STFT, as técnicas que se destacaram das demais técnicas.

Os cenários básicos escolhidos para os testes assumem:

a) um navio na superfície apresentando uma assinatura acústica composta de sinais de faixa larga e ruído aleatório aditivo (ruído branco gaussiano). Utilizou-se também um sinal de faixa larga substituindo o ruído aleatório, com a finalidade de representar o ruído próprio do submarino;

b) um navio na superfície emitindo uma ou algumas freqüências (faixa estreita) na faixa ótima de propagação em águas rasas [4] com um ruído aleatório aditivo (ruído branco gaussiano);

c) sem navio na superfície, somente com ruído ambiente, representado por um ruído branco gaussiano adicionado a um ruído aleatório diferente do primeiro.

Os sinais resultantes simulam sinais gravados por um submarino com hidrofones omnidirecionais (antenas de um sistema sonar passivo), nos diferentes cenários.

Assume-se, então, uma situação inicial sem alvo (fonte) presente, impõe-se a presença de um sinal e testam-se com as duas técnicas de análise espectral.

A TRANSFORMADA WAVELET CONTÍNUA

O termo wavelet, proposto por Grossmann e Morlet, evoca uma onda do mar evanescente (de curta duração e cuja amplitude tende para zero), o seu valor médio sendo nulo [5]. A forma geral da convolução que caracteriza a transformada wavelet contínua (CWT) de uma série temporal x(t) é a seguinte:

$$C(a,b) = \int_{-\infty}^{\infty} x(t)|a|^{-1/2}\,\overline{\Psi}\left(\frac{t-b}{a}\right)dt \qquad (1)$$

onde $\overline{\Psi}(t)$ é o conjugado complexo de $\Psi(t)$, que é uma função denominada wavelet "mãe" ou "janela" da CWT.

A função janela, aplicada ao sinal x(t) na Eq. (1), levará em conta o tempo de atraso ou avanço (b) e, ainda, introduzirá um fator de escala no tempo (a), inversamente proporcional à freqüência.

A função que define a wavelet "mãe" de Morlet, doravante denominada apenas de Morlet, que é a mais indicada para detectar os sinais "monótonos" de uma assinatura acústica, é a seguinte [3], [5]:

$$\psi(t) = \pi^{-1/4}\left(e^{-jm_o t} - e^{-m_o^2/2}\right)e^{-t^2/2} \qquad (2)$$

onde $m_o = \pi\,(2/\ln 2)^{1/2}$ é a freqüência central, assegurando que a razão entre o primeiro e o segundo máximo da função $\Psi(t)$ seja aproximadamente igual a ½.

VANTAGENS DA TRANSFORMADA WAVELET CONTÍNUA

Enquanto que a STFT é uma conveniente generalização da Transformada de Fourier [8], ainda assim possui alguns problemas. Para apreciar este fato, considere dois sinais. O primeiro sinal x(t) é um sinal de alta freqüência, em que muitos ciclos são capturados pela janela empregada na STFT. Para o segundo, x(t) é de baixa freqüência, de tal forma que muito poucos ciclos estão dentro da janela. Assim, a precisão da estimativa da STFT é pobre para baixas freqüências, e cresce com o aumento da freqüência. Isto pode ser entendido interpretando esta transformada como um banco de filtros, em que todos os filtros têm faixas de passagem iguais ao invés da largura de faixa aumentar com a freqüência central [8].

Uma maneira conceitual de fazer filtros com faixas de passagem diferentes é substituir a janela w(t) por uma que seja função da freqüência e do tempo, de tal modo que a largura no domínio temporal da janela seja maior (i.e., a largura de faixa na freqüência seja mais estreita) com a diminuição da freqüência de análise. Desta maneira, a janela captura aproximadamente o mesmo número de cruzamento por zeros para qualquer sinal de entrada senoidal com freqüências arbitrárias. Além disso, quando a janela for larga no tempo, é desejável se usar tamanho de deslocamentos maiores ao mover a janela.

Isto significa que existe um compromisso entre a localização temporal e a freqüencial. Para tornar esta idéia mais precisa, a duração eficaz (valor médio quadrático) de um sinal foi introduzida por Denis Gábor, em 1946, em seu conhecido trabalho "Theory of communication" [2]. Assim, sejam D_t e D_f duas quantidades não negativas e definidas pelas seguintes equações:

$$D_t^2 = \frac{1}{E_w}\int_{-\infty}^{\infty} t^2 w^2(t)\,dt \quad D_f^2 = \frac{1}{2\pi E_w}\int_{-\infty}^{\infty} \Omega^2 \left|W(j\Omega)\right|^2 d\Omega \quad (3)$$

onde E_w é a energia da janela, isto é, $E_w = \int w^2(t)\,dt$, para w(t) real. Diz-se que D_t é a duração eficaz temporal e D_f é a duração eficaz freqüencial da janela. O produto $D_t D_f$ não pode ser pequeno. O resultado deve ser maior ou igual a 0,5, com igualdade acontecendo se e somente se $w(t) = Ae^{-\alpha t^2}$, $\alpha > 0$. Desta forma, a melhor janela é a gaussiana de duração infinita.

A Figura 1(a) representa a resolução tempo-freqüência para a STFT por grades uniformemente espaçadas e a Figura 1(b) representa a resolução tempo-freqüência para a CWT por grades não uniformemente espaçadas.

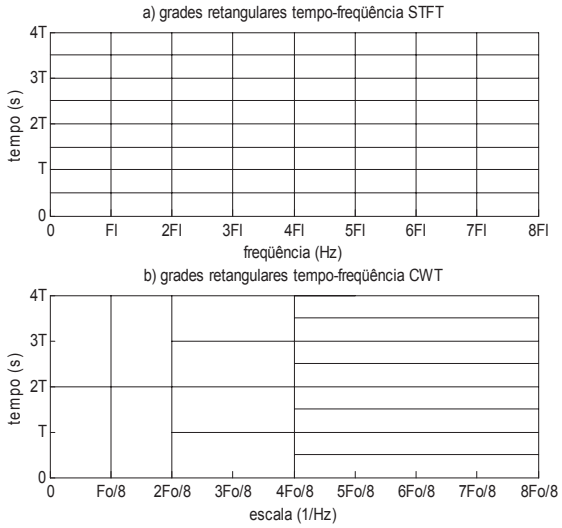


Figura 1: Grades retangulares tempo-freqüência da STFT e da CWT.

Isto significa que o espaçamento freqüencial é menor nas freqüências mais baixas, e o correspondente espaçamento temporal é maior. Para apreciar este fato, observa-se que o sinal x(t) é multiplicado por janelas de diferentes comprimentos, com faixas passantes não uniformes. Tais faixas não uniformes são bem apropriadas para processamentos de sinais sonoros, por causa da resolução decrescente do ouvido humano para as maiores freqüências.

GERAÇÃO DA FUNÇÃO-MÃE DE MORLET

Uma Morlet particular foi gerada a partir de um sinal senoidal complexo com 5 ciclos enquadrado numa janela de Gauss, com freqüência fundamental igual a 400 Hz e freqüência de amostragem igual a 4000 Hz, obtendo-se um sinal complexo que pode ser expandido, a partir da fórmula de Euler, numa parcela referente à componente real, representada por $\Psi_{cos}(t)$, e numa parcela referente à

componente imaginária, representada por $\Psi_{sen}(t)$. A Figura 2 mostra as componentes de $\Psi(t)$ geradas.

O sinal utilizado para ilustrar os resultados obtidos com os algoritmos apresentados nesta seção foi um ruído aleatório r(t) de 500 pontos; na metade desse ruído adicionou-se o sinal $\Psi\cos(t)$ de 400 Hz. A Figura 3 mostra o sinal resultante.

ANÁLISE DE SÉRIES TEMPORAIS PARA A DETECÇÃO

No estudo dos sonares passivos, destaca-se a análise no domínio da freqüência, como forma de se detectar os sinais desejados.

Dos seis cenários escolhidos para teste, três são descritos neste trabalho:

a) Cenário 1: composto de um sinal de cavitação gerado por um navio e sua propagação até a distância desejada. O sinal foi amostrado em 32.000 Hz, decimado [7] para 4.000 Hz e filtrado utilizando um filtro passa faixa entre 250 e 850 Hz.

b) Cenário 2: composto por um sinal de faixa estreita que se propaga por modos (uma das metodologias correntemente utilizadas para descrever a propagação do som no mar para águas rasas), surgindo no limiar de uma detecção e que é simulado pela wavelet "mãe" cosseno de Morlet, $\psi_{cos}(t)$. Como, quase sempre, em águas rasas, existe uma freqüência ótima de propagação [4], foi considerada a existência de uma freqüência discreta que se compõe com o ruído limitador. O sinal é composto então de $\psi\cos(t)$, na freqüência de 570 Hz, do ruído r(t). Obteve-se, assim, um sinal com um registro de 8.192 pontos, e procurou-se testar a eficácia da CWT e da STFT na detecção desse sinal.

c) Cenário 4: corresponde à ocorrência por tempo limitado de freqüências discretas, também por adição da $\psi\cos(t)$ de freqüências diferentes, entre 250 e 850 Hz a r(t), em momentos selecionados aleatoriamente. Procurou-se, assim, simular a possível ocorrência de freqüências ótimas de propagação em diversos instantes ao longo do registro.

Na entrada da unidade de processamento de sinais, considerada para a montagem dos testes de detecção, utilizou-se, primeiramente para o cenário 2, sinais com diferentes valores de relação sinal/ruído, dada em (dB) por $S/N = 10 \log(\sigma^2(A\ \Psi_{cos}(t) + r(t))/\ \sigma^2(r(t))$; em que A é um fator de escala utilizado para se conseguir as relações S/N na entrada iguais a 0, 2, 4, 6 e 8 dB, onde $\sigma^2(X) = E\left((X-\mu)^2\right)$, é a variância de X. Para o Cenário 4, foram preparados conjuntos de S/N na entrada iguais aos do Cenário 2; e foram adicionadas $\Psi\cos(t)$ com diferentes freqüências.

A relação S/N na entrada, escolhida para cada caso, é um conjunto de valores iguais a 0, 2, 4, 6 e 8 dB, para os Cenários 2 e 4.

Detecção Quadrática

A detecção quadrática é usada com freqüência na detecção passiva de sinais acústicos submarinos com características desconhecidas.

O resultado na saída do detector quadrático [1] pode ser representado para sinal ou ruído como $\frac{1}{T}\int_0^T x^2 dt$ o que, para um processo de média zero, corresponde à variância σ^2 do processo. Usa-se então, como resultado dessa detecção, a relação $10 \log(\sigma^2(\text{sinal})/\sigma^2(\text{ruído})$ que será a referência básica para a avaliação das demais técnicas de processamento de sinais.

Detecção com a CWT

Na detecção com a CWT busca-se a detecção localizada de ocorrências do sinal ao logo do registro de dados.

A Figura 4 mostra os resultados da CWT para dois Cenários 1 e 2; observa-se a grande quantidade de picos na transformada no primeiro caso. Esse resultado é típico para a transformada de um processo de faixa larga como é o do Cenário 1; avaliou-se a possibilidade de usar como resultado para cada freqüência, a média das amostras com maiores valores, a média de todas as amostras ou ainda, o valor máximo correspondente. A média de todas as amostras tende a diminuir a S/N do resultado, porque incorpora correlações da $\Psi_{cos}(t)$ com seções de ruído apenas. Neste trabalho, optou-se por utilizar a média dos vinte maiores valores.

Detecção com a Transformada de Fourier de tempo curto (STFT)

A detecção pela STFT pode revelar a presença de sinais de forma localizada, à semelhança da CWT. É calculado o espectro [9] para os 8.192 pontos do sinal e do ruído.

RESULTADOS PARA DETECÇÃO À LONGA DISTÂNCIA

No Cenário 2, para os níveis de sinal inserido, a única detecção possível é com a CWT, como se pode ver na Figura 4(b); uma ocorrência isolada de um sinal de freqüência discreta não é facilmente percebida pelas outras técnicas, que se utilizam da média das amostras de todo o segmento de 8.192 pontos. Vê-se na Figura 5(a) que o sinal é absolutamente igual ao ruído a menos da ocorrência próxima ao ponto 1.000 onde foi inserida a componente de freqüência discreta. O espectro de $\psi_{cos(}t) + r(t)$ na Figura 5(b) mal consegue discernir alguma alteração nas proximidades da freqüência do sinal inserido. A conclusão é a de que a ocorrência se dissipa nas médias calculadas pelo espectro.

No Cenário 4, observa-se na Figura 6(a) o sinal $\psi_{cos}(t) + r(t)$ no domínio do tempo mostrando apenas 4 das 8 inserções. A ocorrência de oito sinais de freqüências diferentes já permite um desempenho um pouco melhor na STFT, apesar do espectro na Figura 6(b) mal sobrepujar o ruído.

As Figuras 7 e 8 mostram os ganhos da CWT e da STFT, em relação à detecção quadrática para diferentes valores de S/N. Pode-se observar que o desempenho da CWT é superior ao da STFT para todas as relações sinal/ruído, exceto no Cenário 1 devido à estrutura de freqüências existente em todo o registro, dificultando a detecção localizada feita por essas duas técnicas.

CONCLUSÃO E PERSPECTIVAS

Para se alcançar o propósito da pesquisa em pauta, de uma conclusão sobre as técnicas de processamento de sinais mais apropriadas para a detecção de fontes sonoras por sonares passivos, seria conveniente contar com um

banco sintético de sinais acústicos com ruído controlado, composto de uma grande quantidade de dados dos tipos em estudo, de preferência idêntico ao utilizado em [6]; em seguida, após uma avaliação metódica e profunda com os dados simulados, aprofundar a pesquisa utilizando as técnicas de processamento de sinais com um banco de sinais acústicos reais, de alta confiabilidade, obtidos em experimentos controlados. Mesmo com a falta destes dados, com os resultados apresentados no presente trabalho, ficou constatada pela experimentação com alguns sinais simulados, que a utilização da transformada wavelet contínua com wavelet "mãe" de Morlet mostrou-se apropriada para a detecção antecipada de sinais sonoros de faixa estreita ou no limiar de detecção.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] .A. B. Baggeroer, "Sonar signal processing", Applications of digital signal processing. Englewood Cliffs, NJ, USA: Prentice- Hall, 1978. Cap. 6.

[2] D. Gabor, "Theory of communications", J. Inst. Elec. Eng. (London), v. 93, pp. 429 - 457, November 1946.

[3] A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelet of constant shape". SIAM Journal on Mathematical analysis, vol. 15, pp. 723 – 736, July 1987.

[4] F. B. Jensen and W. A. Kuperman, "Optimum frequency of propagation in shallow water environments", JASA, vol. 73, pp. 813 – 819, March 1983.

[5] P. C. Liu, "A fifteen minutes introduction of wavelet transform and applications", in *Proceedings of ASCE 1999 International Water Resources Engineering Conference*, (Washington, USA), AES, August 1999.

[6] R. Raghaven, "Wavelets in signal detection and identification comparative signal processing technology evaluation". Final Report, 23 set 1991 – 22 ago 1994, Lockheed Missiles and Space Co. Palo Alto, CA, dez 1994.

[7] J. R. M. da Silva, "Detecção e localização acústica em águas rasas usando transformada wavelet contínua e outras técnicas espectrais", tese de doutorado, Universidade Federal do Rio de Janeiro, Programa de Engenharia Elétrica – COPPE, Março 2006.

[8] P. P. Vaidyanathan, *Multirate systems and filter banks.* Englewood Cliffs, NJ, USA: Prentice- Hall, 1993.

[9] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms", IEEE transactions on Audio and Electroacoustics, v. AU-15, pp. 70 –73, June 1967.

[10] W. Soares-Filho, J. R. M. da Silva e S. N. Wolf, "Measurements and modeling of multitone, low-frequency acoustic data from Arraial do Cabo Range" Final Report, Naval Research Lab., Washington, DC, October 1993.

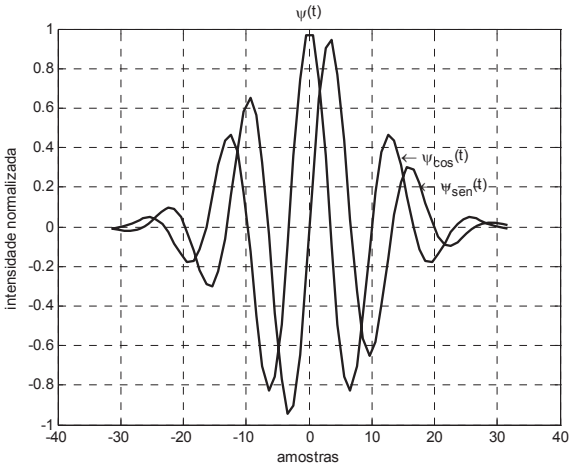


Figura 2: Gráfico das componentes da função-mãe de Morlet.

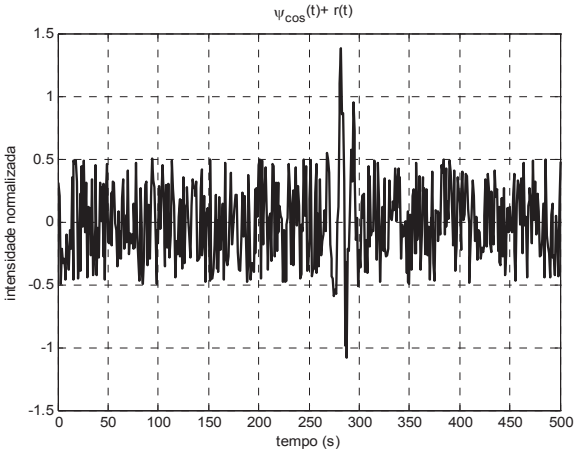


Figura 3: Sinal Ψ_{cos} (t) + r (t) usado no teste dos algoritmos.

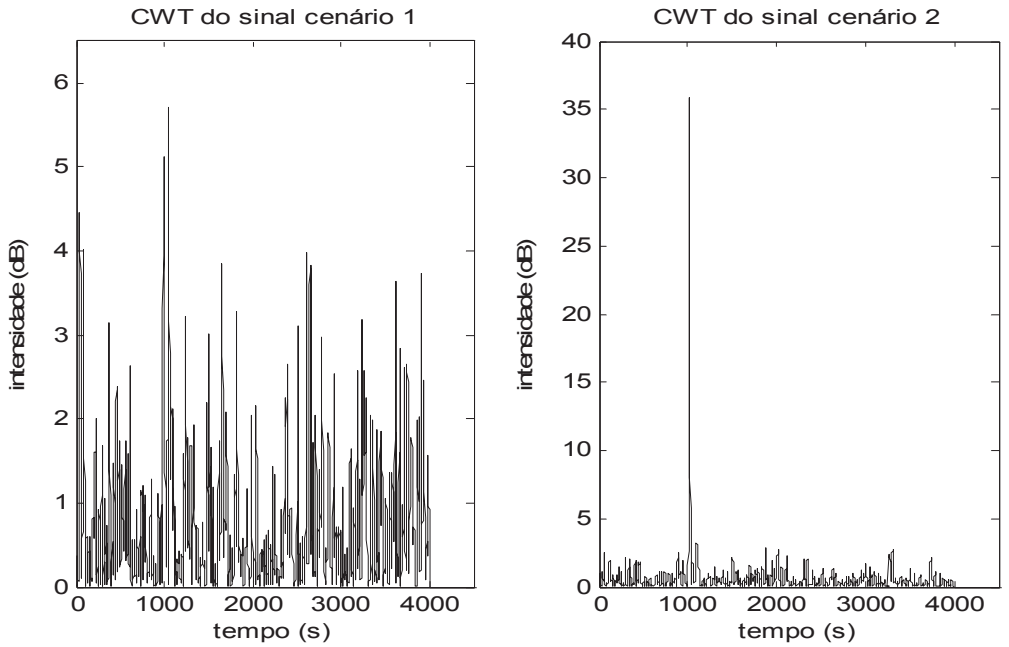


Figura 4: Resultados da detecção pela CWT para os Cenários 1 e 2.

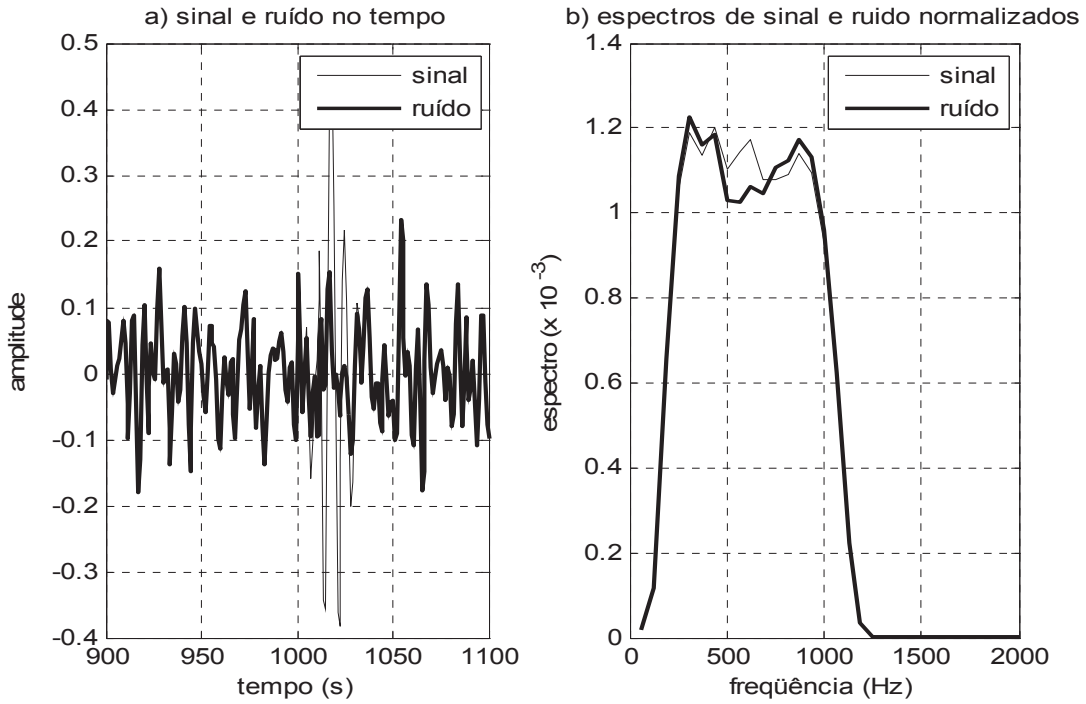


Figura 5: a) Representação de $\psi_{cos}(t)$ + r(t) e r(t) no tempo e b) espectros de $\psi_{cos}(t)$ + r(t) e r(t) para o Cenário 2.

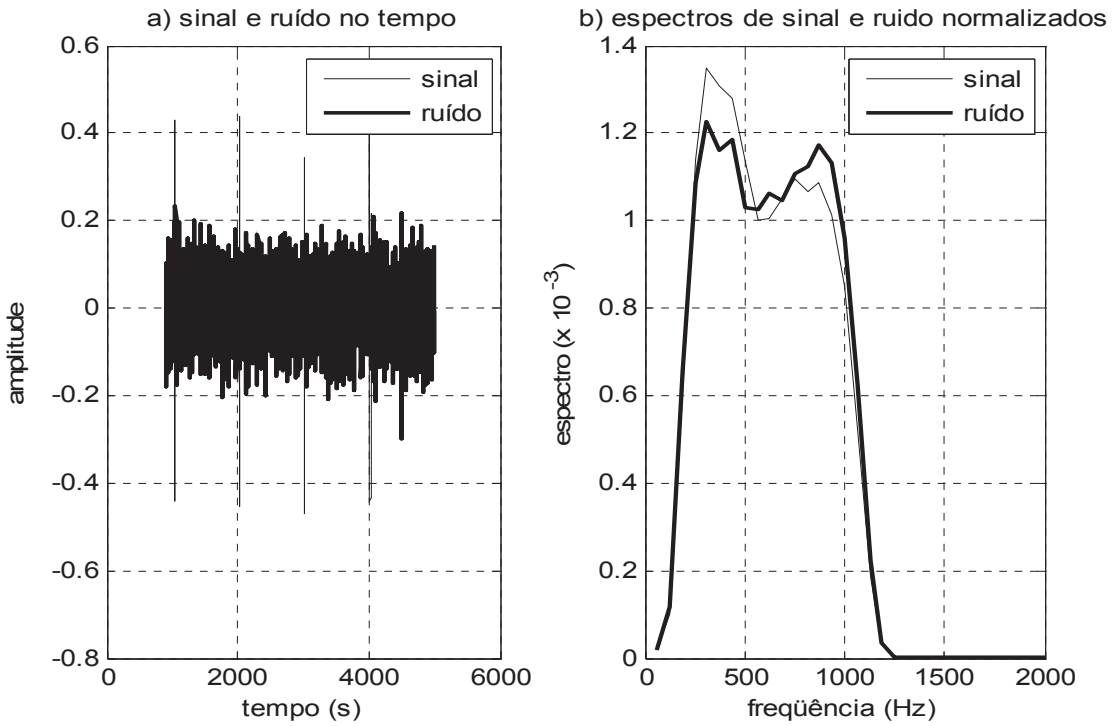


Figura 6: a) Representação de $\psi_{cos}(t)$ + r(t) e r(t) no tempo e b) espectros de $\psi_{cos}(t)$ + r(t) e r(t) para o Cenário 4.

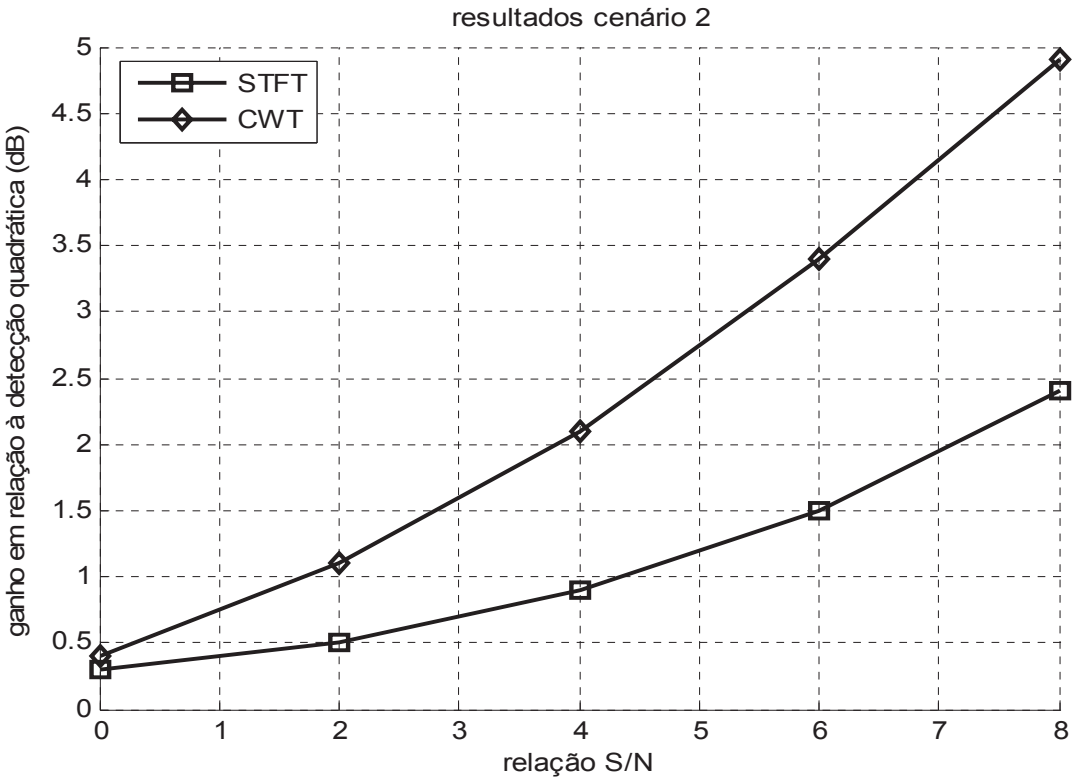


Figura 7: Resultados para o Cenário 2 – $\psi_{cos}(t)$ + r(t).

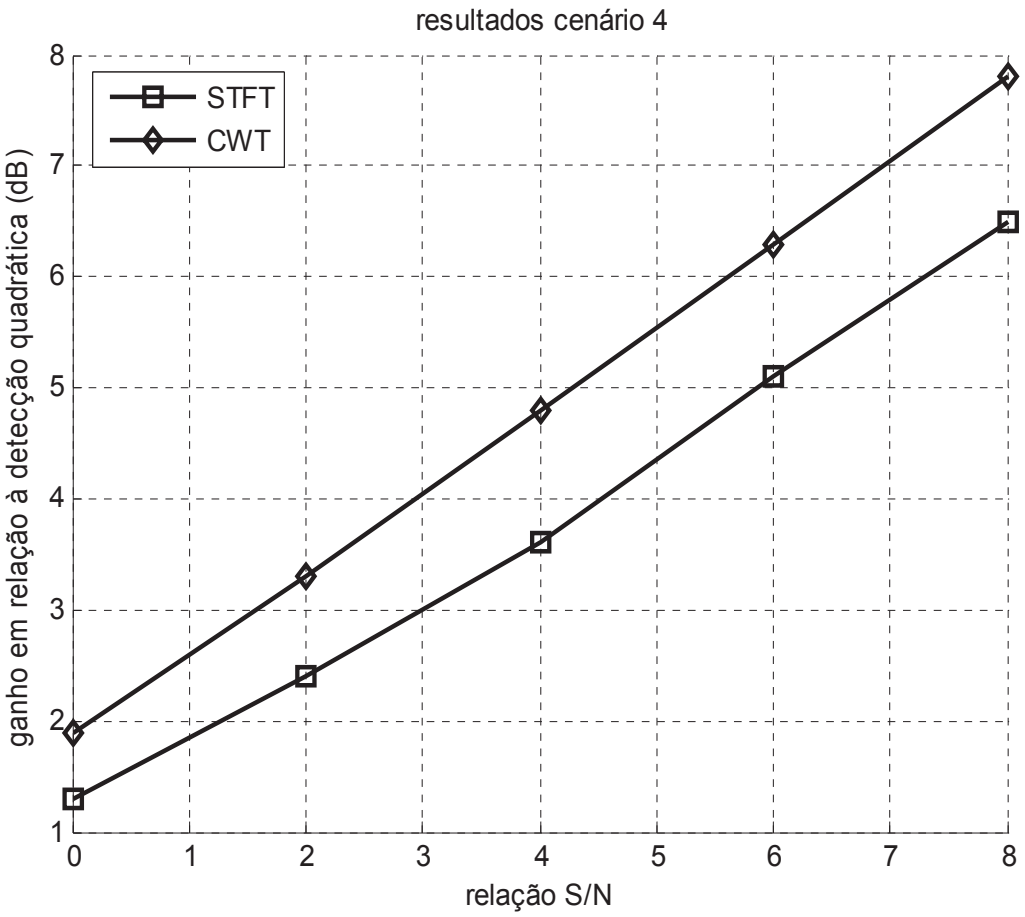


Figura 8: Resultados para o Cenário 4 – $\psi_{cos}(t) + r(t)$.

Sessão 2

Análise, síntese, percepção auditiva e computação musical

(Analysis, synthesis, auditory perception and computer music)

Sociedade de Engenhariade Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Transcrição automática de sinais de áudio monofônico baseada em quadro do Artigo

Tiago Fernandes Tavares[1], Jayme Garcia Arnal Barbedo[1,2] e Amauri Lopes[1]
[1]Departamento de Comunicações – FEEC – UNICAMP
Caixa postal 6101 – CEP 13.083-970 – Campinas – SP
[2]Harvard School of Applied Sciences – Harvard University
ZIP 02138 – Cambridge – MA – USA
ttavares@fee.unicamp.br, {jgab, amauri}@decom.fee.unicamp.br

RESUMO

Este artigo apresenta um novo algoritmo para transcrição automática de sinais de áudio, o qual utiliza quadros de tamanho variável de forma a otimizar o compromisso entre a resolução temporal do processo de transcrição e a menor freqüência identificável pelo algoritmo. O processo de determinação do tamanho do quadro se apóia em características específicas do método Yin, embora possa ser modificado para operar em conjunto com outros métodos de extração da freqüência fundamental. Os resultados obtidos demonstram que a técnica pode ser aplicada com sucesso na transcrição de sinais monofônicos. A extensão do método a sinais polifônicos será abordada em estudos futuros.

0 INTRODUÇÃO

Processos de transcrição automática de música têm por objetivo a construção de alguma forma de notação musical a partir de um sinal sonoro digitalizado empregando o processamento de sinais por computador. Quando lidamos com áudio monofônico, a determinação da nota musical tocada num determinado intervalo de tempo envolve a determinação da freqüência fundamental da onda sonora nesse intervalo [1,2]. Por esse motivo, o processo de transcrição automática de áudio normalmente começa com a divisão do sinal analisado em quadros de tamanho conhecido, seguida pelo cálculo da freqüência fundamental em cada um deles. Sabendo os tempos de início e término de cada freqüência fundamental, pode-se determinar durações de notas e, assim, obter uma notação musical simplificada (estilo MIDI) correspondente à execução analisada [1,2].

Os algoritmos mais comuns para detecção de freqüência fundamental precisam, para seu funcionamento, que o quadro tenha pelo menos duas vezes a duração do maior período analisado [3,4,5]. O aumento na duração do quadro, porém, significa que um trecho mais longo de áudio será

considerado uniforme quanto à freqüência fundamental, fazendo com que o limite inferior de duração de notas detectáveis seja também modificado. Desta forma, a divisão do sinal de áudio em quadros implica em um compromisso entre a menor freqüência e a menor duração de uma nota detectável. Neste artigo, é proposto um algoritmo para variar o tamanho do quadro buscando otimizar o compromisso citado.

Foi estudado o comportamento de um método de detecção de freqüência fundamental já existente quando aplicado sobre um quadro de duração menor que o dobro do período fundamental do sinal analisado. Conhecendo esse comportamento, pode-se determinar quando o quadro de análise tem duração muito curta para que o método de detecção de freqüência fundamental atue corretamente sobre o trecho avaliado. Ao detectar a ocorrência desse comportamento, o algoritmo decide que a duração do quadro deve ser aumentada. Em contrapartida, sempre que o período fundamental detectado for menor que a metade da duração do quadro, este pode ser reduzido para que se ajuste à onda analisada.

1 DETECÇÃO DA FREQÜÊNCIA FUNDAMENTAL

O primeiro passo para a construção de um transcritor automático de áudio monofônico é a escolha de um método de detecção de freqüência fundamental. Existem vários métodos, entre eles o *Harmonic Product Spectrum* (HPS)[4], a detecção baseada em autocorrelação[1,3,4], análise de coeficientes cepstrais[3] e o método Yin[3,5], que, por apresentar o melhor desempenho[5], foi escolhido como base para para este trabalho.

O método Yin, cujo nome deriva de *Yin-Yang*, determina a freqüência fundamental a partir da função diferença, definida pela Equação 1.

$$d_t[n]=\sum_{k=1}^{N}(x[k]-x[k+n])^2 \qquad (1)$$

A função diferença mede o quanto um sinal é diferente dele mesmo deslocado no tempo. Sinais periódicos, quando deslocados de tempo igual ao seu período fundamental, se assemelham muito ao sinal original, fazendo com que a função diferença apresente mínimos locais em pontos correspondentes a períodos fundamentais da onda analisada.

Como a função pode apresentar mais de um mínimo local, sendo apenas o primeiro deles o desejado – excluindo-se o mínimo referente ao deslocamento zero – o método Yin propõe que seja calculada a *função diferença absoluta acumulada normalizada (FDAAN)*, definida pela Equação 2.

$$\begin{gathered} d'_t[n]=1 \; para \; n=0; \\ d'_t[n]=\frac{d_t[n]}{\frac{1}{n}\sum_{j=1}^{n} d_t[j]} \; para \; n\neq 0 \end{gathered} \qquad (2)$$

A primeira expressão da Equação 2 elimina o primeiro vale da função diferença. Para encontrar o mínimo, então, procura-se o primeiro valor de *n* para o qual a *FDAAN* é um mínimo local menor que um certo limite absoluto (definido experimentalmente). É possível, porém, que nenhum ponto satisfaça as condições de parada. Nesse caso, a busca retorna o índice do menor valor da *FDAAN*.

Em qualquer circunstância, o ponto que foi determinado como mínimo, junto a seu antecessor e seu sucessor, são interpolados por uma parábola. Adota-se como mínimo da função o valor mínimo dessa parábola, com o objetivo de aumentar a resolução da detecção de freqüência.

2 A DURAÇÃO DO QUADRO DE ANÁLISE

A transcrição baseada no método Yin baseia-se no fato de que, para pequenos trechos, um sinal de áudio monofônico gerado por um instrumento harmônico é periódico. Ao se dividir o sinal em quadros com duração suficientemente pequena, pode-se extrair a freqüência fundamental em cada um deles. Esse procedimento permite que seja inferido o início e o final de cada nota musical executada, permitindo que seja obtida a notação musical correspondente.

Para determinar a duração do quadro, alguns aspectos devem ser levados em consideração. Quando é utilizado o método Yin, a menor freqüência fundamental detectável é aquela correspondente a um período igual à metade da duração do quadro, de forma que um quadro muito pequeno levará à perda de notas graves. Por outro lado, o aumento na duração do quadro faz com que um trecho maior de áudio seja analisado a cada iteração, reduzindo a resolução temporal do processo de transcrição e aumentando o esforço computacional demandado pelo algoritmo de detecção de freqüência fundamental.

Visando otimizar o compromisso entre a resolução no tempo e o alcance em freqüência, foi desenvolvido um algoritmo capaz de variar a duração do quadro de acordo com a demanda de cada trecho do sinal analisado. Esse algoritmo se apóia em características específicas do método Yin, como será visto adiante, embora possa ser redesenhado para outros métodos de extração de freqüência fundamental.

A redução da duração do quadro pode acontecer sempre que o período fundamental detectado tiver duração menor que a metade da duração do quadro, mantendo a condição de funcionamento do algoritmo. O aumento da duração do quadro depende de fatores que devem ser analisados mais cautelosamente. Quando é aplicado o método Yin, a busca por um mínimo local na *FDAAN* pode não encontrar valor abaixo do limite estabelecido, o que significa que a onda analisada não apresenta grande semelhança consigo mesma quando deslocada no tempo, caso em que é retornado o índice do mínimo global da função. Essa inexistência de semelhança pode significar que a duração do quadro não é suficiente para conter dois períodos fundamentais da onda analisada. A partir dessa característica, pode-se determinar que quando o método Yin retorna um mínimo global acima do limiar adotado o quadro deve ter sua duração aumentada.

A quantidade de aumento ou diminuição da duração quadro é, em princípio, arbitrária, embora quantidades muito grandes possam levar o algoritmo a oscilações na duração do quadro e quantidades muito pequenas possam fazer a variação aplicada tornar-se cada vez mais imperceptível. Neste trabalho, a duração é multiplicada ou dividida por dois, de forma a variar o limite inferior de freqüência detectável em uma oitava sempre que isso se fizer necessário.

O processo de transcrição deve ser iniciado com a escolha de uma duração inicial para o quadro. Uma vez que essa duração convergirá para um valor considerado ideal pelo algoritmo de variação para cada nota musical detectada, essa

duração pode ser escolhida livremente dentro de limites razoáveis, da mesma ordem de grandeza dos períodos fundamentais esperados. Para cada quadro analisado, é extraída a freqüência fundamental e registrada sua duração. Se a aplicação do método Yin retornar o mínimo global da FDAAN, determina-se que o tamanho do quadro deve aumentar. O resultado obtido para a freqüência fundamental do quadro é descartado e a análise é realizada novamente, desta vez com um quadro com a nova duração. Se o valor retornado foi um mínimo local, o quadro é deslocado adiante num passo igual à metade de sua duração e então, se foi encontrado um período fundamental menor que um quarto da duração do quadro, este é reduzido à metade. Para cada análise, registra-se a duração do quadro utilizado e a freqüência fundamental obtida, obtendo-se uma seqüência que permite determinar a freqüência fundamental do sinal sonoro a cada instante.

3 PÓS-PROCESSAMENTO

Para o agrupamento correto dos resultados do processo de detecção de freqüência fundamental e a eliminação de eventuais erros, é necessária a aplicação de um algoritmo de pós-processamento.

O algoritmo de pós-processamento proposto inicia-se com a conversão das freqüências fundamentais reconhecidas para o valor MIDI da nota musical correspondente através da Equação 3.

$$N = round\left(12\log_2\left(\frac{f_o}{440}\right)\right) + 69 \qquad (3)$$

A operação *round* corresponde ao arredondamento do valor calculado para o inteiro mais próximo.

Pode-se então agrupar a seqüência de notas de forma que elementos subseqüentes apresentando a mesma nota são substituídos por um elemento relacionado à mesma nota e cuja duração é a soma das durações dos elementos originais. Essa substituição permite mais simplicidade nas análises posteriores.

Para a eliminar eventuais erros do processo transcritivo, é preciso conhecer a origem dos erros mais comuns.

O erro mais comum decorre do ataque, estado transitório presente no começo da execução de uma nota. Nesse período, a onda musical está em processo de estabelecimento no corpo do instrumento, levando o processo de detecção de freqüências fundamentais a apresentar falhas. A partir do conhecimento sobre o tempo de ataque relacionado ao instrumento, nota e técnica utilizada, podemos determinar a duração mínima de notas musicais detectáveis. Esse tempo varia entre dois instrumentos diferentes e pode ser modificado, embora o uso de um tempo fixo de 40ms tenha apresentado bom desempenho para todos os instrumentos testados. Diante desse aspecto, classifica-se como ruído qualquer freqüência fundamental com duração inferior a 40ms.

Embora erros quase sempre sejam decorrentes do ataque, eles podem ocorrer também durante a execução da nota. Quando uma nota de duração menor que o limite inferior estabelecido está entre duas outras notas iguais, sua altura é modificada para a altura das notas adjacentes. Caso contrário, a nota errônea é classificada como um ataque e é substituída por silêncio de mesma duração. Após esse processo, a seqüência obtida é novamente agrupada. São excluídas da nova seqüência obtida as notas restantes cuja duração é insuficiente para apresentar uma sensação de tom [6]. Neste trabalho, a duração adotada foi de 50ms.

4 RESULTADOS

Na etapa de testes, o algoritmo de transcrição com variação do tamanho do quadro foi implementado em MATLAB, adicionando-se limites superior e inferior para o tamanho do quadro em 2048 e 256 amostras, respectivamente. Seu desempenho foi comparado ao de outros transcritores baseados no método Yin, mesmo algoritmo de pós-processamento e quadros de duração fixa em 256, 512, 1024 e 2048 amostras. Todos os sinais analisados nessa etapa foram amostrados a 44,1kHz, e estão disponíveis na base de dados de arquivos de som da universidade de Iowa [7]. Essa base de dados contém notas, tocadas por diversos instrumentos, em seqüência crescente de semitons e com tempo de silêncio significativo entre cada uma delas. Essa característica a torna ideal para uma primeira verificação de erros devidos à atuação sobre diferentes tons e timbres.

Os resultados, expressos na Tabela 1, foram classificados em notas corretas (cuja nota estimada corresponde à nota executada em cada intervalo de tempo), perdidas (notas que foram perdidas por conseqüência de uma duração inadequada do quadro), erradas (notas que não foram detectadas por ou que foram detectadas de forma errada), e, em adição a essa estatística, foram adicionadas as notas fantasmas, aquelas que o algoritmo transcritor detecta embora não existam realmente. A freqüência de notas fantasmas (F.N.F.) foi computada através divisão do número de notas fantasmas pelo número de notas corretas obtidas.

Considerando o índice de notas corretamente detectadas, os resultados mostram que o transcritor baseado em quadros de duração variável tem desempenho semelhante ao desempenho do transcritor com quadros de duração fixa em 2048 amostras. Outro aspecto a ser levado em consideração é que, com a variação do tamanho da quadro, pequenos erros durante o processo de transcrição de notas mais agudas se limitaram a intervalos de tempo proporcionais ao período fundamental da nota avaliada, podendo, por conseqüência, ser facilmente removidos. Ao se fixar o quadro em um tamanho suficientemente grande para abranger toda a gama de freqüências da música – 2048 amostras – pequenas flutuações na execução de notas agudas tornam-se mais significativas, resultando no aparecimento de mais notas fantasmas.

Tabela 1: Resultados da transcrição aplicada à base de dados da Universidade de Iowa

Quadro	**Var.**	**256**	**512**	**1024**	**2048**
Notas avaliadas	1163	1163	1163	1163	1163
% correta	88,91	43,59	69,91	81,94	89,25
% perdida	0,60	49,44	22,18	8,68	0,60
% errada	10,49	6,96	7,91	9,37	10,15
F.N.F.	8,03	7,69	10,70	12,49	13,97

Por tratar-se de uma base de dados extensa, o tempo de execução também foi registrado, em minutos, de forma a

verificar o ganho real de eficiência. Os algoritmos rodaram em um computador Athlon XP 2200+ com 236Mb de memória RAM. Os tempos de execução, em minutos, estão mostrados na Tabela 2. Verifica-se que o transcritor com quadros de duração variável apresentou eficiência substancialmente maior que o transcritor com quadros de tamanho fixo em 2048 amostras.

Tabela 2: Tempo de execução durante o teste

Quadro	**Var.**	**256**	**512**	**1024**	**2048**
Tempo (min)	78,94	43,70	49,83	69,88	107,35

Na segunda etapa de testes, algumas melodias foram submetidas aos transcritores de quadro variável e de quadro fixo em 2048 amostras. As melodias foram compostas de trechos de dez a vinte segundos de músicas conhecidas (versão do Bolero de Ravel, executada em um trompete, Asa Branca e Greensleeves, executadas em uma flauta doce, e Brasileirinho, executado em um violão), em gravações com considerável ruído de captação. Todas os sinais foram amostrados a 44,1kHz. As notas foram classificadas em corretas (cuja duração e nota, pela Equação 3, foram estimadas corretamente), perdidas (que não foram encontradas pelo sistema) e com a duração errada (que a altura foi estimada corretamente, mas não a altura). Foram também contabilizadas as notas fantasmas. Os resultados estão expressos na Tabela 3.

Tabela 3: Desempenho na transcrição de melodias gravadas

Quadro	**Variável**	**2048**
Total de notas avaliadas	88	88
Total de notas corretas	82	85
Duração errada	5	2
Notas perdidas	1	1
Notas fantasmas	1	12

Como pode ser verificado, o desempenho de um transcritor baseado no algoritmo proposto é semelhante ao de um transcritor com duração de quadro fixa em 2048 amostras, apresentando, porém, tempo de execução 30% menor.

5 CONCLUSÃO

Este artigo abordou o problema da transcrição automática em áudio monofônico. Foi apresentado um algoritmo para variação da duração do quadro, com o objetivo de otimizar o compromisso entre o limite inferior de freqüência detectável e a resolução temporal do processo.

Os resultados alcançados mostram que a variação da duração do quadro pode levar a ganhos significativos para o algoritmo de transcrição, não só no desempenho geral como no esforço computacional demandado.

Futuramente, pretende-se estudar a aplicabilidade da variação da duração dos quadros em algoritmos mais complexos, como transcritores para áudio polifônico.

6 REFERÊNCIAS

[1] N. Trevillato, J. G. A. Barbedo, A. Lopes, "Transcrição automática de sinais de áudio", anais do X Simpósio Brasileiro de Computação Musical, pp. 291-294, 2005.

[2] L. P. Clarisse, *et all*, "An auditory model based transcriber for singing sequences" , Proc of 3rd International Conference on Music Information Retrieval, ISMIR,'02, 2002.

[3] D. Gerhard, "Pitch extraction and fundamental frequency: history and current techniques" Technical report TR-CS 2003-06, University of Regina, 2003.

[4] F. Hamidi-Toosi, J. Laska "Pitch Detection for the Next Millenium and Beyond", www.ews.uiuc.edu/~laska/ece320/files/Report.pdf, 2004.

[5] A. Cheveigné, H. Kawahara, "YIN, a fundamental frequency estimator for speech and music" J. Acoust. Soc. Am. Vol. 111 No. 4, pp. 1917-1930, 2002.

[6] H. F. Olson, "Music, Physics and Engineering" ed. Dover, 1967.

[7] MIS – Musical Instrument Samples, http://theremin.music.uiowa.edu/MIS.html .

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 5º Congresso da AES Brasil
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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 101165-2520, 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.

Modelo Empírico da Sonoridade da Clarineta: Análise e Síntese Sonora

Luís Carlos de Oliveira[1], Fábio Furlanete[2], Ricardo Goldemberg[3], Jônatas Manzolli[3]
[1]FEEC e NICS (UNICAMP)
Campinas, SP, CEP: 13083-970, Brasil
[2]IA (UEL e UNICAMP)
Londrina, PR, CEP 86051-990, Brasil
[3]IA e NICS(UNICAMP)

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

ABSTRACT

In this work we developed an instrument for representation of the synthesis of a clarinet. It is the final stage of a series of experiments where we aimed to reproduce the sound of the instrument with PureData and in real time. The empirical data was obtained from the sound of a mechanical blowing chamber that operates independently of the action of the musician. We utilized the "Fractional Factorial Design" for analysis of the data and implemented the results in a model based on additive synthesis.

INTRODUÇÃO

Muitos trabalhos dedicados à síntese sonora de instrumentos musicais como, por exemplo, BORIN et al. [1], KEEFE [2] e SMITH [36], abordam o assunto através da modelagem física. Apesar dos avanços tecnológicos na direção de hardwares mais eficientes, o esforço computacional destes modelos ainda é muito alto. Este fato estimulou a condução deste trabalho numa direção oposta, isto é, obter a síntese de tais instrumentos a partir de modelo totalmente empírico.

Desde 2003, a natureza das sonoridades de instrumentos de sopro, principalmente da clarineta, é o foco desta pesquisa. Foi utilizado um mecanismo de simulação experimental que tem comportamento físico análogo ao envolvido na execução de instrumentos de sopro. Durante a pesquisa, percebeu-se que existem muitos fatores que são determinantes na construção e no entendimento de um modelo que possa elucidar todas as relações e variáveis intrínsecas ao processo de geração sonora dos instrumentos musicais.

Frente a esta complexidade, optou-se por utilizar um método de pesquisa denominado Projeto Fatorial de Experimentos, que possibilitou comparar e verificar a importância relativa dos diversos fatores envolvidos na produção sonora do instrumento analisado. Todavia, ficou claro que o controle de todas as variáveis de um sistema experimental é um problema complexo, o que estimulou a ampliação do escopo da pesquisa no sentido de utilizar, também, simulação computacional para depois, com a mesma, possibilitar a interpretação do modelo experimental. Neste momento, pareceu relevante iniciar um estudo da síntese sonora que mais se aproxima do modelo experimental utilizado.

Dos resultados anteriores determinou-se um modelo linear, de fácil manipulação algébrica, que possibilitou implementar o modelo mais simples de síntese musical, que é a síntese aditiva. Em conjunto com o emprego do

ambiente de linguagem de programação gráfica *Pure Data*, pode-se ouvir em tempo real os resultados obtidos.

A trajetória deste trabalho inicia-se com a descrição do método experimental utilizado e apresentam os modelos obtidos para as notas estudadas.

Uma descrição visual das subrotinas em PD é apresentada paralelamente aos resultados experimentais.

Uma rápida descrição dos resultados obtidos e das novas possibilidades musicais atingidas estão incluídas no final do texto.

APARATO E PROCEDIMENTO EXPERIMENTAIS

Optou-se em não utilizar um músico clarinetista na obtenção dos dados experimentais, pois sua experiência profissional poderia interferir na coleta dos dados. Utilizou-se então, um aparato experimental, já descrito em OLIVEIRA et al. [4, 5 e 6], cujos elementos principais estão apresentados nas Figuras 1 e 2.

Na Figura 1 mostramos o compressor, a caixa acrílica, a clarineta, manômetros e tubulações de conexão. Na Figura 2 está indicado o detalhe do mecanismo de contato com a palheta ("mordedura").

O procedimento adotado, também indicado naqueles trabalhos, pode ser resumido da seguinte forma: As notas desejadas são obtidas através de vedação dos furos do instrumento com pequenas borrachas. Inicialmente, a "mordedura" é posicionada de modo a bloquear a saída do ar transmitido pelo compressor. Lentamente a "mordedura" é flexibilizada possibilitando a saída do ar através da palheta. Quando a nota emitida está estabilizada (o ataque é eliminado) procede-se à medida da freqüência e amplitude durante alguns segundos (mais ou menos trinta a quarenta).

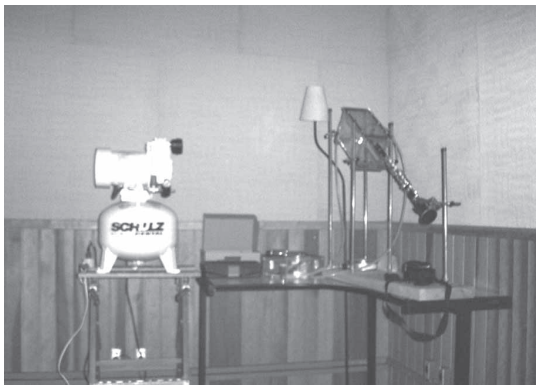


Figura 1: Visão Geral do Aparato Experimental

A unidade de captação de dados consiste em um microfone colocado próximo à clarineta e conectado a um computador onde está instalado o software SOUND FORGE 4.5. O microfone é unidirecional dinâmico (cardióide) e tem resposta de freqüência de 50 a 15000 Hz. A taxa de amostragem utilizada foi de 44 kHz e a configuração utilizada foi de 16 bits.

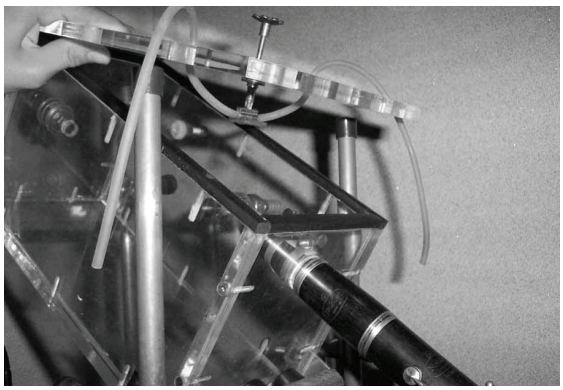


Figura 2: Detalhe do Contato com a Palheta ("mordedura")

Projeto Fatorial de Experimentos

Será designado ***fator*** como sinônimo para *variável independente* e ***resposta*** para *dependente*. Neste projeto os fatores são analisados fazendo-os variar entre dois níveis (–1 e +1). Experimentos no ponto central (0) foram repetidos três vezes para determinação do erro experimental. Grosseiramente, um Projeto Fatorial é uma expansão por Taylor, no nosso caso linear, a partir do ponto central nos diversos fatores. Isto significa que os termos quadráticos e de ordem superiores da série de Taylor foram eliminados.

Os fatores deste Projeto Fatorial foram: 1)Volume vazio do tanque pulmão (x_1), 2)Dureza da palheta (x_2), 3)Posição de contato na palheta (x_3), 4)Abertura da boquilha (x_4), 5)Área de contato com a palheta (x_5) e 6)Quantidade de material absorvente sonoro (x_6).

Como respostas, (Y_f e Y_a), foram analisadas as freqüências e amplitudes da fundamental e respectivos componentes espectrais de três notas da clarineta: (**E_2**, **C_3** e **G_3**) correspondendo a nota da região de baixo *chalumeau* (**E_2**,) e do alto *chalumeau* (**C_3** e **G_3**).

A elaboração de um Projeto Fatorial de Experimentos está minuciosamente detalhada em BOX [4]. Aqui é mencionado apenas que foi utilizado um Projeto Fatorial Fracionado com Resolução **III**. Ser fracionado significa que foi realizado parte (1/8) de um Projeto Fatorial Completo que, neste caso, para 6 fatores variando em dois níveis seriam 2^6=64 experimentos. Então, realizamos (1/8)*64=2^{6-3}=8 experimentos. A resolução indica que ao ser realizado parte do projeto completo nossa "perda" de informação está em fundir o efeito de um fator principal (**I**) com interações entre dois fatores (**II**), **I+II=III**.

Na tabela 1 estão descritos os níveis dos fatores utilizados neste experimento. A boquilha **A** tem a menor abertura enquanto a **C**, a maior.

Neste projeto em particular, o padrão de fusão (*confounding patterns*) foi **4=12**, **5=13** e **6=23**. Isto significa que o efeito do fator **4** foi confundido com a interação **12** (entre os fatores **1** e **2**), o efeito do fator **5** foi confundido com a interação **13** (entre os fatores **1** e **3**), etc.

Na região grave da clarineta, também denominada **chalumeau**, escolhemos a nota com a clarineta totalmente fechada, correspondendo à nota **E_2**. Esta nota refere-se à clarineta, portanto para o piano corresponde à nota **D_2**.

Resumindo, foram realizados, então, apenas 11 experimentos para verificar a influência de seis fatores, sendo três deles para a determinação do erro experimental. Finalmente, os experimentos foram sorteados para determinar a ordem de execução.

A título de exemplo, é apresentado o projeto para a nota **E_2**. Para as demais notas os procedimentos foram análogos. A Tabela 2 resume o que foi mencionado anteriormente. Ela indica que o primeiro ensaio realizado foi o ensaio nº 5. Sendo seguido pelo ensaio nº 7, nº 6 e assim sucessivamente. No ensaio nº 5 o fator **1** apresenta nível –1, e consultando a Tabela 1 sabe-se que o volume vazio do tanque pulmão corresponde a 60% do volume total. Para o fator **2** o nível também é –1 e a Tabela 1 indica que a palheta utilizada foi de nº 2. Analogamente, a posição da mordedura na palheta é a mais externa; a boquilha utilizada foi a C; a área de contato da mordedura com a palheta foi de 0,3 cm^2 e não foi utilizada estopa para absorção sonora.

FATORES	-1	0	+1
1)Volume Vazio do Tanque Pulmão(%)	60	65	70
2)Dureza da Palheta (Nº)	2	2,5	3
3)Posição da Mordedura na Palheta	Interna	Centro	Externa
4)Boquilha	A	B	C
5)Área de Contato com palheta (m^2)	$3x10^{-5}$	$1x10^{-4}$	$1,4x10^{-4}$
6)Quantidade de Estopa (kg)	0	0,015	0,030

Tabela 1: Níveis dos Fatores Utilizados nos Experimentos

ENSAIO Nº	FATORES						ORDEM
	1	2	3	4	5	6	
1	-1	-1	-1	+1	+1	+1	7
2	+1	-1	-1	-1	-1	+1	8
3	-1	+1	-1	-1	+1	-1	11
4	+1	+1	-1	+1	-1	-1	4
5	-1	-1	+1	+1	-1	-1	1
6	+1	-1	+1	-1	+1	-1	3
7	-1	+1	+1	-1	-1	+1	2
8	+1	+1	+1	+1	+1	+1	10
9	0	0	0	0	0	0	5
10	0	0	0	0	0	0	9
11	0	0	0	0	0	0	6

Tabela 2: Projeto Fatorial de Experimentos Fracionado com Resolução **III** para a região *chalumeau.*

Na análise espectral feita através do SOUND FORGE 4.5 foram testados vários "janelamentos" e o de Blackmann-Harris mostrou ser o mais apropriado. Em todas as análises os dados foram normalizados em –3,0 dB.

Tanto para a freqüência (Hz), Y_f, como para a amplitude (dB), Y_a, o modelo linear obtido, truncando-se os termos quadráticos e de ordem superiores da série de Taylor, é da forma:

$$Y_i = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6$$
$$i = f, a \qquad (1)$$

As tabelas 3 e 4 mostram os coeficientes (a_i) obtidos para a nota **E_2**, respectivamente, para a freqüência e amplitude. Analogamente, as tabelas 5 e 6 são os coeficientes para a nota **C_3** e as tabelas 7 e 8 para a nota **G_3**.

Y_f	a_o	a_1	a_2	a_3	a_4	a_5	a_6
E_2	146±2	-0,2	-1,0	-1,0	0,2	0,2	1,0
2	295±4	2,4	0,4	-1,9	-0,9	-1,6	0,4
3	440±4	-1,5	1,5	2,2	0,8	-1,5	0,0
4	588±4	0,9	1,6	0,4	-1,6	-1,9	0,4
5	735±2	-1,6	0,4	0,9	0,4	-1,6	0,4
6	885±4	0,0	0,5	1,5	0,8	-2,8	-0,8
7	1031±5	-0,2	2,5	0,5	-0,2	-3,8	-1,0
8	1178±5	-1,5	2,8	2,2	0,0	-4,0	-1,2
9	1325±4	-1,0	2,5	1,8	0,2	-3,0	-1,0
10	1473±6	-0,6	2,9	3,4	0,9	-4,2	-2,2
11	1620±5	-1,2	2,2	1,6	0,9	-3,6	-0,4
12	1770±7	-0,4	2,4	3,2	1,2	-5,2	-0,9

Tabela 3: Coeficientes do modelo empírico da freqüência da nota **E_2** (clarineta) e freqüência de seus componentes espectrais

Y_a	$-a_o$	a_1	a_2	a_3	a_4	a_5	a_6
E_2	21±1	-1,6	1,4	0,2	-0,9	-3,2	-0,2
2	59±1	0,4	-0,2	1,6	-2,6	-6,4	-0,4
3	23±1	-0,3	1,0	-0,8	-0,3	-2,0	-0,3
4	47±3	0,5	1,8	-0,3	-0,3	-3,2	-1,0
5	32±2	0,8	-1,0	-1,5	-1,6	2,8	1,0
6	34±1	0,6	1,9	-1,9	-1,2	-0,4	0,9
7	35±2	0,6	-0,9	-0,2	-1,2	0,6	0,2
8	28±1	-1,6	0,6	-0,4	1,6	-1,4	-0,2
9	42±7	1,0	-1,8	-0,3	-2,0	1,0	1,8
10	36±1	1,5	-1,0	-1,8	0,3	2,5	0,5
11	41±5	2,5	-2,2	-0,2	-0,5	-3,0	1,2
12	36±1	2,2	0,6	-2,2	0,4	4,2	0,6

Tabela 4: Coeficientes do modelo empírico da amplitude da nota **E_2** (clarineta) e amplitude de seus componentes espectrais

Y_f	a_o	a_1	a_2	a_3	a_4	a_5	a_6
C_3	231±4	-1,0	1,0	0,0	-1,2	-2,2	2,2
2	466±6	-3,4	-1,2	3,4	-1,2	1,4	1,2
3	699±6	-2,2	0,0	1,0	-1,2	-2,2	2,0
4	936±8	-3,2	2,2	4,2	-3,2	-3,2	4,2
5	1168±7	-2,8	0,5	3,8	-0,5	-1,8	1,5
6	1401±8	-2,2	-1,0	4,2	0,0	-3,2	0,0
7	1637±10	-3,4	-1,2	5,6	-2,2	-4,4	4,4
8	1870±11	-4,6	0,2	4,4	-1,2	-5,4	3,4
9	2104±11	-4,2	0,0	4,5	-2,2	-4,2	6,5
10	2337±14	-7,6	-3,4	8,9	-1,2	-4,4	4,4
11	2573±10	-4,9	0,6	4,9	-1,6	-5,9	1,6
12	2808±15	-7,8	-2,2	8,8	-2,2	-4,2	5,2

Tabela 5: Coeficientes do modelo empírico da freqüência da nota **C_3** (clarineta) e freqüência de seus componentes espectrais

Y_a	$-a_o$	a_1	a_2	a_3	a_4	a_5	a_6
C_3	30±3	1,7	-0,2	3,7	-2,2	-1,9	-0,7
2	60±2	0,8	-2,3	2,8	-2,3	0,3	1,3
3	28±3	1,8	0,5	2,3	-2,3	-1,0	-0,8
4	55±4	-2,2	-1,0	-1,8	-1,5	0,3	1,5
5	22±2	-0,3	-0,3	-0,3	0,3	1,3	0,8
6	41±4	2,7	4,9	-5,4	-1,9	5,4	2,7
7	32±5	-0,5	0,8	-1,3	-1,3	-2,3	-1,0
8	39±1	4,9	-1,2	-1,2	-1,2	-3,2	-1,2
9	38±4	-0,2	-3,9	-4,2	-4,2	-2,7	0,7
10	38±8	-1,0	-2,3	2,3	4,3	2,3	1,5
11	37±7	-0,3	0,8	0,0	0,0	0,0	-0,5
12	42±5	1,5	-0,8	4,3	4,3	1,8	-3,5

Tabela 6: Coeficientes do modelo empírico da amplitude da nota **C_3** (clarineta) e amplitude de seus componentes espectrais

Y_f	a_o	a_1	a_2	a_3	a_4	a_5	a_6
G_3	1053±*16*	-0,5	-4,2	6,8	0,5	4,5	-6,8
2	2120±*13*	-1,0	6,5	2,8	-2,8	-6,5	8,5
3	3175±*15*	0,2	0,2	7,8	-0,2	-7,8	7,2
4	4243±*21*	-0,2	0,0	7,2	0,0	-15	7,5
5	5305±*25*	-2,2	-2,5	8,8	2,5	-16	14
6	6362±*31*	-1,0	-1,0	18	8,5	-18	12
7	7431±*35*	-8,8	5,8	14	1,5	-21	16
8	8488±*36*	-0,2	3,6	18	3,9	-26	15
9	9523±*112*	-5,9	1,4	24	6,4	-31	21
10	10613±*45*	-4,8	-4,8	25	-2,8	-25	20
11	11678±*54*	-7,0	4,2	26	7,0	-31	26
12	12713±*102*	1,8	5,5	39	9,2	-32	17

Tabela 7: Coeficientes do modelo empírico da freqüência da nota **G_3** (clarineta) e freqüência de seus componentes espectrais

Y_a	$-a_o$	a_1	a_2	a_3	a_4	a_5	a_6
G_3	22±2	0,4	1,4	0,2	0,2	-0,2	0,9
2	22±1	-0,8	-1,0	1,6	-2,2	-1,0	-1,2
3	27±6	-2,9	-0,2	-0,8	0,4	0,6	-0,6
4	39±12	-2,8	1,5	-0,3	-1,5	0,0	1,8
5	47±10	-1,5	-1,5	-1,5	2,2	-1,0	-3,5
6	58±6	-1,6	0,6	-1,9	-1,4	-7,4	0,4
7	53±6	0,2	2,9	-0,2	-2,4	-0,2	-3,4
8	60±6	-2,2	5,9	-4,6	3,4	4,4	-1,2
9	71±4	0,0	0,0	-2,0	-3,8	-2,2	-1,8
10	66±6	-1,6	1,2	0,6	-4,4	-0,4	2,4
11	66±4	0,0	-1,5	-0,2	-5,2	3,5	-0,5
12	78±7	-2,2	-0,2	-3,6	-5,4	4,6	3,2

Tabela 8: Coeficientes do modelo empírico da amplitude da nota **G_3** (clarineta) e amplitude de seus componentes espectrais

SÍNTESE SONORA USANDO O PD

O *Pure-Data*, abreviadamente PD, é um ambiente de programação gráfica em tempo real adequado para aplicação em áudio, vídeo e processamento gráfico. É um ramo da família de linguagem de programação Max (Max/FTS, ISPW Max, Max/MSP, jMax, etc.) originalmente desenvolvido por Miller Puckette e outros pesquisadores do IRCAM (*Institut de Recherche et Coordination Acoustique/Musique*, Paris) . O PD possui um conjunto de *patchs* disponíveis para áudio, tal como o oscilador aqui empregado.

Para a síntese sonora foi empregado o método de síntese aditiva usando o modelo empírico obtido pela Equação 1. A Figura 3 mostra o *patch* em PD responsável pela aplicação da Equação 1 para o cálculo da freqüência e amplitude de cada componente espectral. Por exemplo, para a nota **E_2**, os valores dos parâmetros a_i para a freqüência e para a amplitude estão indicados nas Tabelas 3 e 4, respectivamente. Analogamente, os parâmetros da freqüência e amplitude para a nota **C_3** estão nas Tabelas 5 e 6 e para a nota **G_3** nas Tabelas 6 e 7.

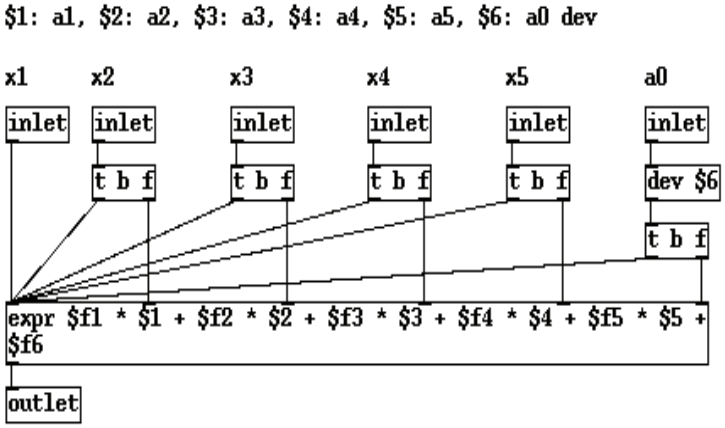


Figura 3: "Patch" correspondente ao cálculo da freqüência e amplitude de cada componente espectral (parameter.pd)

O desvio (erro experimental) para cada freqüência foi incorporado no modelo. Assim, quando o usuário deseja tocar a nota **E_2**, um conjunto de parâmetros (os a_i da freqüência e amplitude) é acionado. A freqüência da fundamental assim como a freqüência dos componentes espectrais será um valor no interior do desvio de cada freqüência central. Por exemplo, o terceiro componente espectral da nota **E_2** terá como freqüência um valor aleatório entre (440 ± 4)Hz. Este valor é calculado no *patch* dev.pd, mostrado na Figura 4.

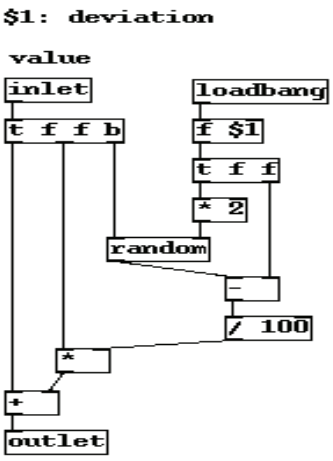


Figura 4: “Patch” para o cálculo da freqüência no interior de cada desvio (dev.pd).

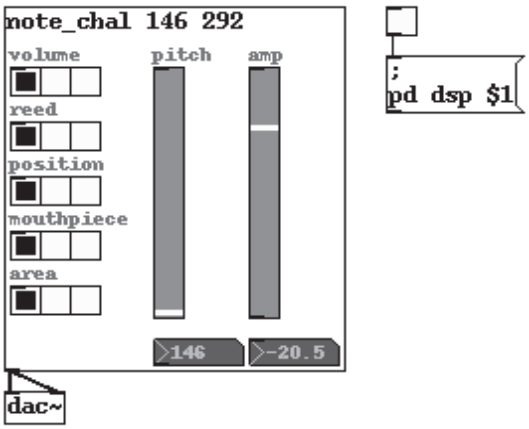


Figura 5: “Patch” de controle do instrumento (junior.pd)

O “instrumento” musical, denominado “júnior” está configurado no *patch* junior.pd, indicado na Figura 5 para a nota $\mathbf{E_2}$ (146 Hz) da região *chalumeau*. Nele pode-se alterar os valores dos fatores conforme indicados na tabela 1 e ouvir a sonoridade resultante.

Como os parâmetros obtidos são válidos apenas para a nota $\mathbf{E_2}$ (146 Hz) da região *chalumeau*, supomos que eles são constantes para esta região e extrapolamos uma oitava acima apenas para ilustração de um novo instrumental para possível aplicação musical. Hipótese similar foi aplicada às demais notas.

Uma vez determinado o valor da freqüência e amplitude de cada componente espectral, estes valores são conduzidos ao oscilador de cada componente espectral para posterior soma no ambiente PD. A Figura 6 ilustra este resultado com o *patch* partial.pd.

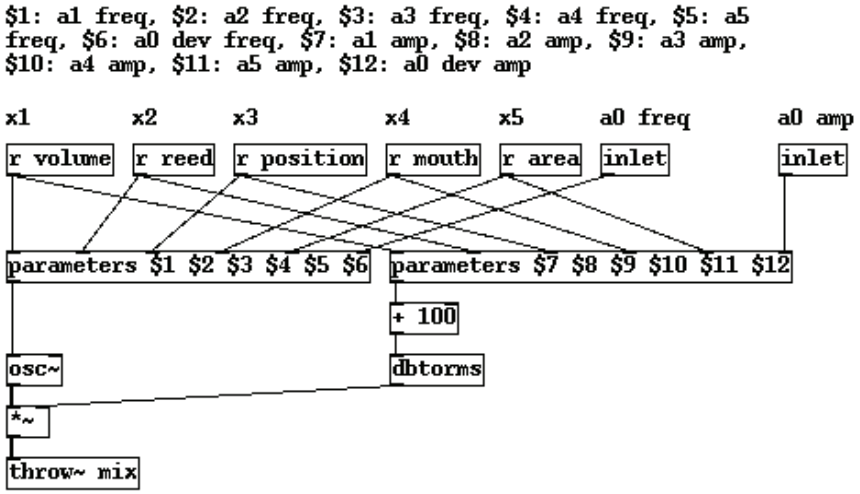


Figura 6: “Patch” para o cálculo da freqüência e amplitude de cada componente espectral para a entrada do oscilador (partial.pd).

As relações entre as freqüências de cada componente espectral pela fundamental, e suas respectivas amplitudes, da nota $\mathbf{E_2}$ (146 Hz), foram extrapoladas para uma oitava acima pelo mesmo motivo descrito acima. Este resultado pode ser visualizado na Figura 7 correspondendo no PD ao *patch* note_chal.pd. Os coeficientes indicados são para a região *chalumeau*. Os controles de ação estão mostrados na figura 8.

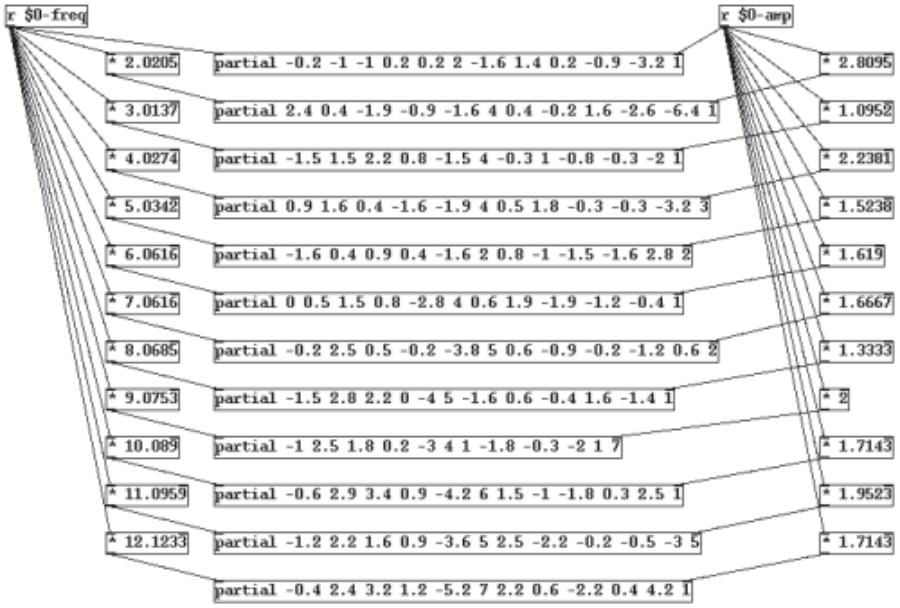


Figura 7: “Patch” para o cálculo da resultante da síntese aditiva na região *chalumeau* (note_chal.pd).

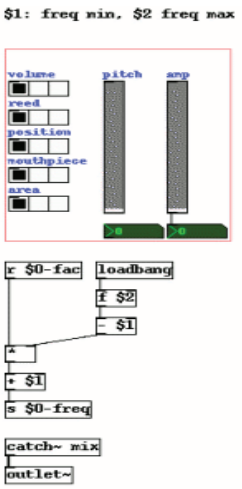


Figura 8: Controles para a ação do “patch” note_chal.pd.

CONTINUIDADE E CONCLUSÕES

Este trabalho fecha o ciclo da série de resultados anteriores OLIVEIRA et al. [4, 5 e 6]. Pode-se afirmar também que este conjunto inicial de experimentos teve caráter exploratório. No entanto, ele revelou a possibilidade de levar a cabo o procedimento experimental para a construção de um “instrumento” musical sintético.

Naturalmente tem-se que eliminar algumas dificuldades encontradas durante o percurso. O aparato experimental deve sofrer algumas modificações para facilitar sua operação. Por exemplo, o modo de variação do volume interno, o contato com a palheta, etc.

Apesar disso pode-se destacar a validade do modelo empírico na síntese sonora e a potencialidade do ambiente PD para síntese musical em tempo real. A partir dos modelos obtidos, constroem-se os arquivos *.wav correspondente a cada um dos ensaios. A comparação com os dados experimentais pode ser feita a partir da percepção auditiva, fazendo tocar os arquivos obtidos, ou fazendo uma “medida” da diferença entre estes dois sinais usando, por exemplo, a métrica do ℓ^2. Os resultados auditivos mostraram-se bastante razoáveis. Faz-se necessário realizar uma “medida” da diferença dos sinais.

A seqüência deste trabalho deve incluir um refinamento do modelo empírico, destacando os fatores que têm maior influência nos resultados. Conseqüentemente, o modelo empírico será mais simples e o esforço computacional será menor. Deve-se ampliar os resultados para outras notas das regiões consideradas e possivelmente obter modelos de segunda ordem para futuras comparações qualitativas.

Os resultados aqui obtidos, ainda que de caráter exploratório, abrem uma nova fronteira para a composição musical contemporânea. Existe a possibilidade de criar novas sonoridades baseadas em instrumentos musicais reais variando-se apenas alguns parâmetros e fatores do modelo.

Referências

[1] BORIN, G., De POLI, G., SARTI, A. (1992). Algorithms and Structures for Synthesis Using Physical Models. Computer Music Journal. 16(4): 30-42.

[2] KEEFE, D.H. (1992). Physical Modeling of Wind Instruments. Computer Music Journal. 16(4): 57-73.

[3] SMITH, J.O. (1996). Physical Modeling Synthesis Update. Computer Music Journal. 20(2): 44-56.

[4] OLIVEIRA, L.C, GOLDEMBERG, R., MANZOLLI, J. (2005a). Estudo Experimental da Sonoridade Chalumeau da Clarineta através de Projeto Fatorial (I), Anais da IX Convenção Nacional da AES, SP.

[5] OLIVEIRA, L.C, GOLDEMBERG, R., MANZOLLI, J. (2005b). Estudo Experimental da Sonoridade Chalumeau da Clarineta através de Projeto Fatorial (II), Anais do XV Congresso da ANPPOM, RJ

[6] OLIVEIRA, L.C, GOLDEMBERG, R., MANZOLLI, J. (2006). Estudo Experimental da Sonoridade Chalumeau da Clarineta através de Projeto Fatorial (I), Anais da IX Convenção Nacional da AES, SP.

[7] BOX, G.E.P.; HUNTER, W.G.; HUNTER, J.S. (1978). Statistics for Experimenters – An Introduction to Design, Data Analysis and Model Building. John Wiley & Sons, NY.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, São Paulo, SP

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

Avaliação do Bounded-Q Fast Filter Bank na Transcrição Automática de Música

Filipe C. da C. B. Diniz,[1] Luiz W. P. Biscainho[1] e Sergio L. Netto[1]

[1] Universidade Federal do Rio de Janeiro, COPPE, Laboratório de Processamento de Sinais
Caixa Postal 68504, Rio de Janeiro, RJ, CEP 21941-972, Brasil

filiped@lps.ufrj.br, wagner@lps.ufrj.br, sergioln@lps.ufrj.br

RESUMO

Este artigo apresenta um estudo do desemepenho da ferramenta de análise "*Bounded-Q Fast Filter Bank*"como base para um sistema de transcrição automática de música (TAM). Esse banco de filtros possui alta seletividade, espaçamento não-uniforme dos filtros e complexidade computacional reduzida, o que o torna atrativo para uso em TAM. São utilizados sinais musicais com características escolhidas para avaliar a resposta do sistema completo a problemas específicos relacionados com essa aplicação.

0 INTRODUÇÃO

A transcrição de música consiste em escrevê-la em uma partitura ou similar a partir de sua execução. Tradicionalmente, é realizada por uma pessoa com elevado nível de treinamento musical, apenas com o uso de sua audição.

Por questão de comodidade e praticidade, hoje a trancrição musical tende a ser realizada de forma automatizada, possivelmente por meio de algum *software*, a partir de uma gravação. A este processo, dá-se o nome de transcrição automática de música (TAM) [8].

Em termos gerais, a TAM pode ser dividida em 4 etapas:

1. Segmentação do sinal musical no domínio do tempo.
2. Identificação das notas presentes nos segmentos.
3. Reconhecimento do timbre das notas musicais identificadas.
4. Avaliação da coerência dos resultados obtidos.

A base para esse sistema é a análise espectral dos sinais de áudio. Uma das formas de a realizar é utilizar um banco de filtros [12] que separe o sinal em faixas estreitas na freqüência. As saídas dos filtros descrevem ao longo do tempo a distribuição espectral do sinal sob análise.

O presente trabalho afere a capacidade de análise do *bounded-Q fast filter bank*(BQFFB), recentemente proposto [3, 4] num contexto de TAM. O artigo está assim organizado: na Seção 1, revisam-se soluções para

análise espectral por bancos de filtros; o BQFFB é rapidamente descrito na Seção 2; na Seção 3 são descritos, realizados e avaliados alguns experimentos com o BQFFB aplicado a sinais musicais de características específicas; por fim, as conclusões são resumidas na Seção 4.

1 ANÁLISE ESPECTRAL POR BANCOS DE FILTROS

Um tipo de banco de filtros muito difundido é a *sliding Fast Fourier Transform* (sFFT), cuja implementação requer apenas uma multiplicação complexa por amostra de saída [5]. A sFFT corresponde a uma FFT [2] aplicada sobre janelas deslocadas de amostra em amostra sobre o sinal de entrada. Vista sob outra perspectiva, ela equivale a um banco de filtros cujas saídas, a cada nova amostra de entrada, são as raias da FFT calculada sobre o bloco corrente. Esse esquema resulta numa descrição dinâmica do espectro do sinal. Sua desvantagem está na baixa seletividade dos filtros associados (cerca de 13 dB em relação aos canais adjacentes), obtidos a partir de filtros-base de ordem muito reduzida [10].

Para atacar esse problema, uma nova ferramenta de análise foi proposta em [10]. Nesta nova técnica, denominada de *fast filter bank* (FFB), combina-se a estrutura modular da FFT com filtros-base de ordem maior, gerando filtros longos, porém com muitos coeficientes nulos (interpolados), como na técnica de *frequency-response masking* (FRM) [9] para projeto de filtros digitais não-recursivos. Com isso, obtém-se um banco de filtros de complexidade reduzida, mas alta seletividade (cerca de 56 dB entre canais adjacentes), que permite maior distinção entre as várias componentes do sinal musical, e conseqüente melhor identificação das notas. O espaçamento freqüencial dos filtros do FFB é constante, o que o torna ineficiente em TAM: na escala de temperamento igual, predominante na música ocidental [13], o espaçamento freqüencial entre as fundamentais das notas musicais aumenta geometricamente com a altura. Assim, se o espaçamento entre canais do FFB for dimensionado para baixas freqüências, será superdimensionado para altas freqüências, gerando uma quantidade desnecessária de dados para análise. Em contrapartida, se o espaçamento entre canais do FFB for escolhido de acordo com os requisitos de altas freqüências, acarretará resolução insuficiente em baixas freqüências.

No domínio da transformada, foi proposto em Brown [1] um sucedâneo da FFT com espaçamento geométrico entre as componentes espectrais de análise, a chamada *constant-Q transform* (CQT). Contudo, esta compartilha a baixa seletividade da FFT. Tal dificuldade foi superada por Graziosi et al. [6] [11] no chamado CQFFB, que combina o espaçamento geométrico da CQT com a alta selectividade do FFB. Embora implementando um banco de filtros altamente seletivos espaçados geometricamente, a CQFFB perde a característica de baixa complexidade computacional do FFB.

Uma outra transformada propõe um espaçamento quase-geométrico entre as componentes espectrais: analisam-se as oitavas (distribuidas de forma geométrica) subdivididas por filtros com espaçamento uniforme. Esse esquema, chamado *bounded-Q transform* (BQT) [7], também herda a seletividade reduzida da FFT, na qual se baseia. Associada ao FFB, a BQT foi a base do *bounded-Q fast filter bank* (BQFFB), que é objeto deste trabalho.

2 BOUNDED-Q FAST FILTER BANK

A versão de alta seletividade da BQT é denominada de *bounded-Q fast filter bank* (BQFFB), e foi apresentada por Diniz et al. [3] [4].

A implementação do BQFFB consiste em submeter o sinal a ser analisado, inicialmente, a uma CQFFB de 10 canais com espaçamento geométrico de 2, correspondendo às 10 oitavas percebidas pelo sistema auditivo humano, entre 20 Hz e 20 kHz. Essa separação inicial possui alta seletividade e, por apresentar poucos canais, moderado custo computacional. Cada oitava é decimada para que seu espectro ocupe toda a faixa de freqüências; nesta nova escala, é submetida a um FFB de 32 canais, resultando numa separação linear de no mínimo um quarto de tom entre os sub-canais. Tal resolução é suficiente para distinguir freqüências afastadas de meio tom, que é, convencionalmente, o menor afastamento entre notas da escala musical de igual temperamento. Esse procedimento pode ser visualizado na Figura 1.

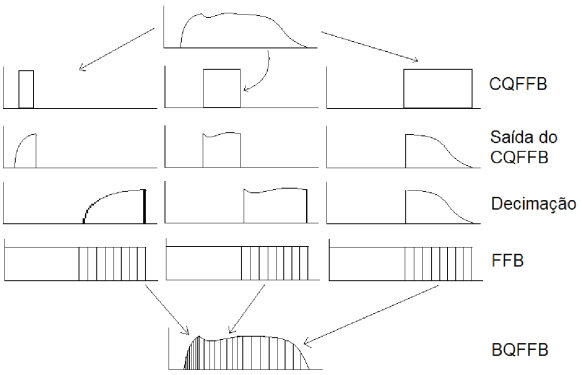


Figura 1: Procedimento para implementação de uma BQFFB a partir de um CQFFB seguido de um FFB.

Dessa maneira, tem-se uma ferramenta que une a propriedade de alta seletividade do FFB com a de espaçamento quase-geométrico da BQT. Deve-se notar, contudo, que para sua implementação, o BQFFB associa um CQFFB (que realiza a separação inter-oitavas) com um FFB (que realiza a separação intra-oitavas). Isso é análogo a implementar a BQT associando uma CQT seguido de uma FFT.

O resultado final é uma ferramenta de análise espectral que consegue conjugar espaçamento geométrico, alta seletividade e complexidade moderada. Uma tabela

comparativa com as propriedades de cada ferramenta de análise espectral e uma análise de complexidade podem ser encontradas em [3].

3 EXPERIMENTOS

Nesta seção, são realizados testes para investigar a capacidade de análise do BQFFB para sinais musicais reais de características variadas. Foram utilizados os seguintes sinais de teste:

- Exemplo 1 - Sinal com freqüências baixas: trecho de Os Elefantes, de O Carnaval dos Animais, de Saint-Saëns.
- Exemplo 2 - Sinal com freqüências altas: trecho do 3º Movimento do Concerto para Flautim RV 443, de Vivaldi.
- Exemplo 3 - Sinal com passagem rápida: trecho de O Vôo do Besouro, de Rimski-Korsakoff, em arranjo para piano de Rachmaninoff.
- Exemplo 4 - Sinal com passagem lenta: trecho do 3º Movimento da Marcha Fúnebre, 3º Movimento da Sonata nº 2 para piano de Chopin.
- Exemplo 5 - Sinal contendo glissando: trecho da Rhapsody in Blue, de Gershwin.

3.1 Visualização dos resultados

Os resultados dos experimentos descritos nesta seção precisam ser adequadamente apresentados, para permitirem avaliar o desempenho do sistema. Uma métrica largamente usada na literatura associada é a contagem das notas presentes na partitura do sinal original que foram corretamente detectadas no sinal analisado. Esse será o critério aplicado na Seção 3.7 a todos os experimentos, de forma comparativa.

Para facilitar a identificação das notas, foi feita uma normalização no sinal de saída dos filtros dentro de cada oitava do BQFFB, dividindo-se as amplitudes das saídas destes filtros pelo valor máximo de amplitude na oitava correspondente. Isso faz com que se perca a proporcionalidade entre as oitavas, o que, contudo, não altera a capacidade de detecção de notas do sistema como um todo.

Em seguida, para cada oitava, recorre-se à representação gráfica de uma matriz em que cada linha representa a saída de um filtro e cada coluna representa um instante de tempo. As amplitudes são codificadas em tons de cinza: o branco corresponde a amplitude zero e o preto, a amplitude máxima. Essa matriz, conhecida como sonograma, fornece uma representação tempo-freqüência de fácil visualização.

Contudo, esse tipo de figura ainda é deficiente para representar os transitórios dos filtros. Para contornar esse problema, lança-se mão de um gráfico tridimensional conhecido como *waterfall*. Ele exibe uma representação tempo-freqüência menos compacta, favorecendo a análise do regime transitório.

Para facilitar a visualização dos resultados obtidos com o BQFFB, foi feita uma análise preliminar para determinar em que oitavas o espectro do sinal musical estava concentrado. Tal análise levou em conta que instrumentos que estavam presentes na execução de cada trecho. E a seguir, também por economia de espaço, são representadas apenas as oitavas com informação significativa. Na ordenação das 10 oitavas analisadas, a décima contém as freqüências mais altas.

3.2 Exemplo 1 - Freqüências baixas

Este experimento tem como objetivo estudar o desempenho do BQFFB em baixas freqüências. O sinal usado neste Exemplo 1 (ver lista acima) é uma peça onde se pode ouvir uma linha melódica executada por um contrabaixo, instrumento capaz de gerar notas que vão, tipicamente, do E1 (43,65 Hz) até o B3 (246,94 Hz). A partitura do trecho da música aqui utilizado é visualizada na Figura 2(a).

As oitavas 2, 3 e 4 do sonograma deste sinal são mostradas na Figura 2(b), enquanto que o comportamento transitório da oitava 3 é representado na Figura 2(c).

Pela Figura 2(b), pode-se verificar que a melodia executada pelo contrabaixo se concentra nas oitavas 2 e 3. Na oitava 4 há apenas parciais superiores das notas emitidas pelo instrumento. A análise do gráfico de *waterfall* da Figura 2(c) mostra que a detecção precisa do início e do fim das notas fica efetivamente comprometida: o diagrama apresenta um aspecto suavizado e pouco contraste temporal entre notas consecutivas.

Isso se dá porque os filtros das oitavas inferiores do BQFFB, cujas freqüências centrais são mais baixas, apresentam banda muito estreita e, por isso, ordem mais elevada. Como conseqüência, seus transitórios de saída se tornam muito espalhados no tempo, dificultando com isso a detecção de início (*onset*) ou de fim (*offset*) de cada nota a ser identificada. Visto sob outro prisma, o transitório, sendo concentrado no tempo, envolve informação de alta freqüência, que os filtros das oitavas baixas não conseguem detectar.

3.3 Exemplo 2 - Freqüências altas

Este exemplo verifica o desempenho do BQFFB na identificação de notas de altas freqüências. O sinal deste Exemplo 2 é uma peça onde se pode ouvir uma linha melódica executada por um flautim (também chamado de *piccolo*). Esse instrumento é capaz de gerar notas que vão, tipicamente, do C5 (523,25 Hz) até o D8 (4698,60 Hz). A partitura do trecho da música analisado nesse experimento é mostrada na Figura 3(a).

Para a peça deste Exemplo 2, montaram-se o sonograma das oitavas de 5 a 10, visto na Figura 3(b), e o gráfico *waterfall* da oitava 7, dado na Figura 3(c).

Obsevando a Figura 3(b), pode-se verificar que a melodia executada pelo flautim se concentra nas oitavas

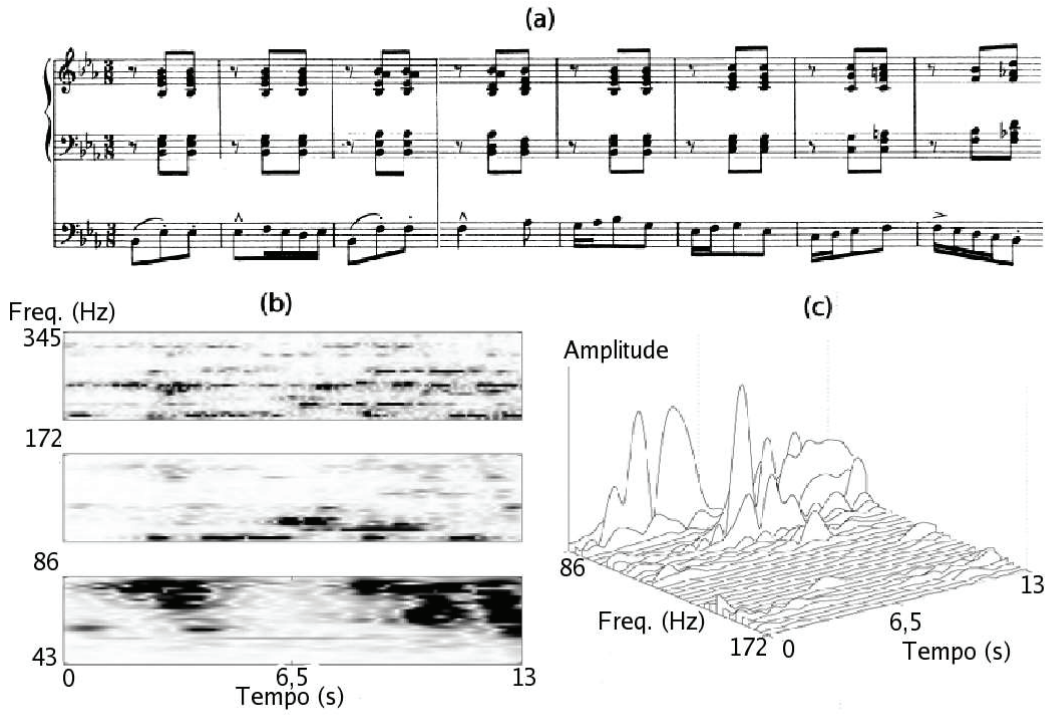


Figura 2: (a) Partitura do trecho da peça O Carnaval dos Animais usado no Exemplo 1. (b) Sonograma das oitavas 2, 3 e 4 no Exemplo 1. (c) Dinâmica temporal da oitava 3 no Exemplo 1.

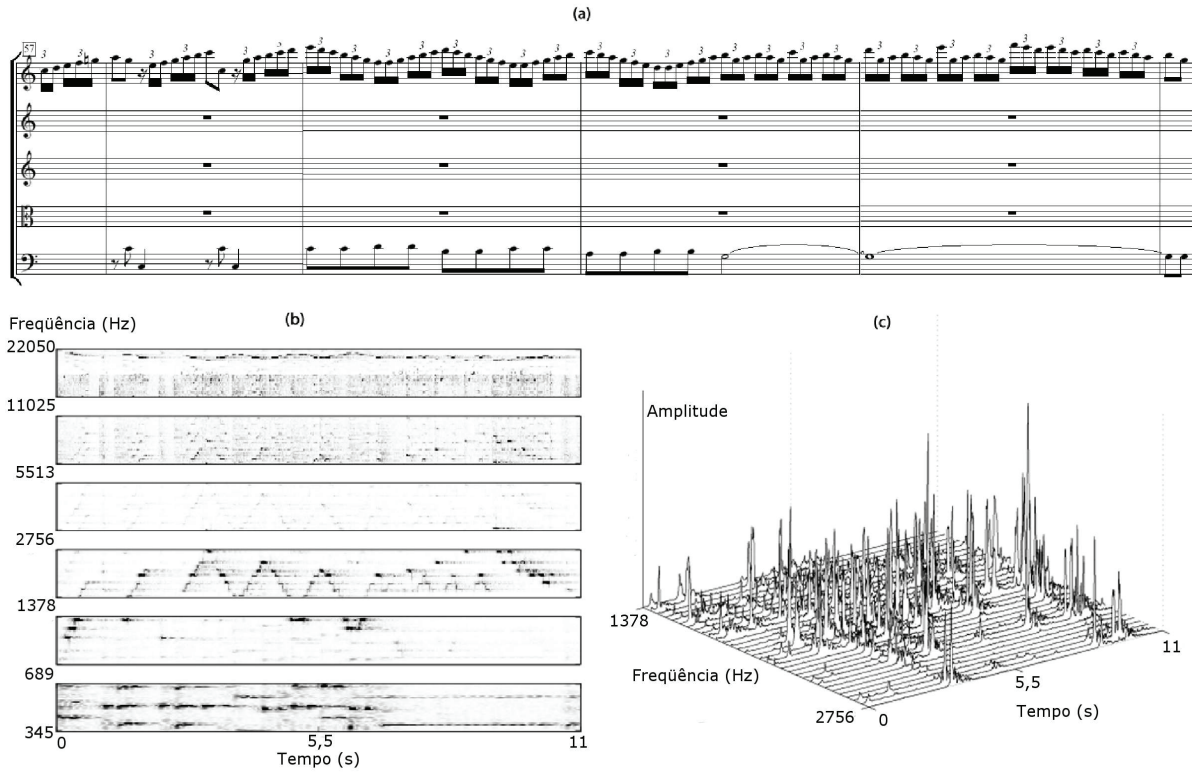


Figura 3: (a) Partitura do trecho do Concerto para Flautim usado no Exemplo 2. (b) Sonograma das oitavas de 5 a 10 no Exemplo 2. (c) Dinâmica temporal da oitava 7 no Exemplo 2.

6, 7 e 8, sendo fácil perceber-se o subir e descer da melodia. Nesse caso é possível, ainda, observar boa separação temporal das notas, se comparadas com os resultados do Exemplo 1, sendo possível precisar seus instantes de *onset* e de *offset*.

3.4 Exemplo 3 - Passagens rápidas

Este experimento analisa o desempenho do B*Q*FFB sobre um sinal com passagens rápidas, ou seja, com rápida mudança entre as notas, que soam durante um curto intervalo de tempo.

O sinal usado neste Exemplo 3 é uma peça que apresenta notas tocadas ao piano em grande velocidade pela mão direita (notas de freqüências mais altas), enquanto a mão esquerda executa o acompanhamento (em freqüências mais baixas). A partitura do trecho analisado nesse exemplo é exibida na Figura 4(a).

O sonograma das oitavas de 4 a 8 deste sinal é representado na Figura 4(b), enquanto que a representação dinâmica da oitava 7 é vista na Figura 4(c).

Por essas figuras, pode-se verificar que a melodia se concentra nas oitavas 6, 7 e 8, enquanto que o acompanhamento se concentra na oitava 4. Com base no que foi explicado nos dois exemplos anteriores, tais notas rápidas recaem no caso de sinais transitórios. Estando associadas, nesse caso, a freqüências relativamente altas, não apresentam problemas de detecção. De fato, o acompanhamento na faixa média ainda consegue ser bem descrito, pois ocorre em notas de maior valor, isto é, mais lentas.

Isso permite traçar limites para o bom funcionamento do sistema.

3.5 Exemplo 4 - Passagem lenta

Neste exemplo analisamos o desempenho do B*Q*FFB para um sinal com passagens lentas, ou seja, com mudança lenta entre notas, que soam durante períodos mais prolongados.

O sinal usado neste exemplo, cuja partitura é vista na Figura 5(a), é uma peça que apresenta notas na região grave do piano, executadas em andamento muito lento.

A Figura 5(b) representa o sonograma das oitavas de 2 a 7 do sinal usado neste exemplo, enquanto que a Figura 5(c) representa o desenvolvimento dinâmico da oitava 3 deste sinal. Por essas figuras, verifica-se que a melodia e seus harmônicos se espraiam pelas oitavasde 3 a 7, e que é possível fazer a separação temporal de todas as notas. Deve-se ressaltar que só foi possível distinguir notas de tão baixas freqüências porque elas foram tocadas de forma lenta.

3.6 Exemplo 5 - Trinado e *Glissando*

Este experimento tem como objetivo estudar o desempenho do B*Q*FFB sobre um sinal que apresente efeitos de trinado (repetição alternada e rápida de duas notas vizinhas) ou *glissando* (seqüência de notas executadas em rápida sucessão, de modo a não serem individualmente articuladas, mas sim tocadas em um movimento contínuo). As notas envolvidas nestes dois efeitos tendem a cair em filtros adjacentes e, por serem tocadas rapidamente, devem apresentar os mesmos problemas associados a sinais de passagem rápida.

O sinal usado neste exemplo é uma peça que apresenta ambos os recursos, como indicado pela partitura dada na Figura 6(a).

O sonograma das oitavas de 3 a 9 é visto na Figura 6(b), enquanto que as Figuras 6(c) e 6(d) representam as dinâmicas das oitavas 4 (com o trinado) e 7 (com o *glissando*), respectivamente.

Pela Figura 6(b), pode-se verificar que o *trinado* é executado na primeira metade do trecho na oitava 4, e o *glissando* na segunda metade do trecho, envolvendo as oitavas 4, 5 e 6. Ambos os recursos foram devidamente detectados. No gráfico *waterfall* da Figura 6(c), pode-se ver as notas bem definidas alternando-se no trinado. Já o *glissando* fica muito bem caracterizado na Figura 6(d), onde se pode acompanhar as notas deslizando ao longo dos diferentes canais.

3.7 Contagem de notas

Para avaliar a capacidade de análise do sistema, realizou-se a contagem das notas coincidentes entre as prescritas na partitura e as reconhecidas no sinal como executadas pelo instrumentista (obviamente, escolheram-se interpretações suficientemente acuradas). Nesse processo, são identificados algumas questões:

1. Quando considerar que uma nota foi efetivamente tocada? Isso pode ser feito aplicando-se um limiar à saída dos filtros, acima do qual decide-se pela presença da nota.

2. Como decidir se, em um instante que a saída do filtro ultrapassa o limiar estabelecido, trata-se de uma nota tocada ou de um parcial de uma outra nota? Analisando-se as outras parciais associadas.

3. Como mapear os 32 canais de cada oitava nos 12 semitons da escala temperada? Quantizando-os para as fundamentais mais próximas.

Como o presente trabalho é uma etapa inicial do estudo da transcrição automática de música, o processo de contagem de notas foi feito de forma manual, com base na leitura direta dos sonogramas e dos gráficos tipo *waterfall*. Dessa maneira, foi possível montar a Tabela 1, onde pode-se ver o percentual de acertos entre as notas reconhecidas como executadas e as descritas na partitura. Deve-se lembrar que as notas nesta etapa foram detectadas de forma subjetiva, levando-se em consideração os parciais e buscando apenas verificar a presença das notas contidas na partitura. O sistema

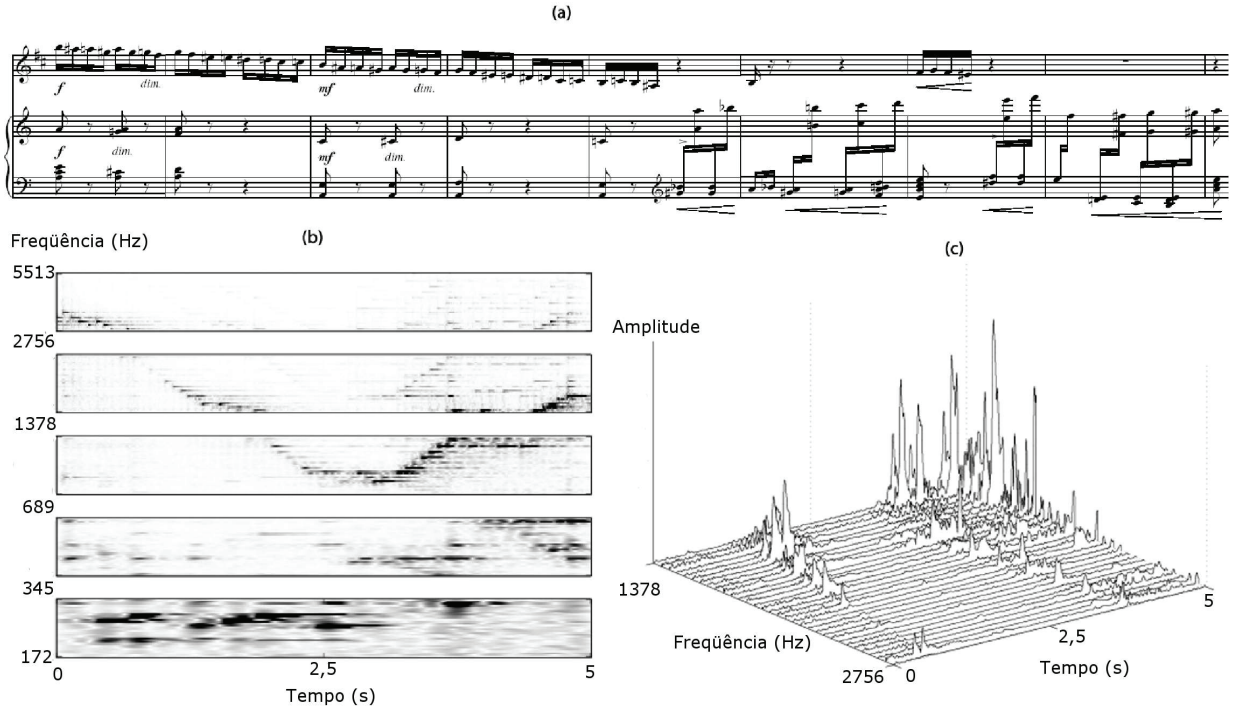


Figura 4: (a) Partitura do trecho de O Vôo do Besouro usado no Exemplo 3. (b) Sonograma das oitavas de 4 a 8 no Exemplo 3. (c) Dinâmica temporal da oitava 7 no Exemplo 3.

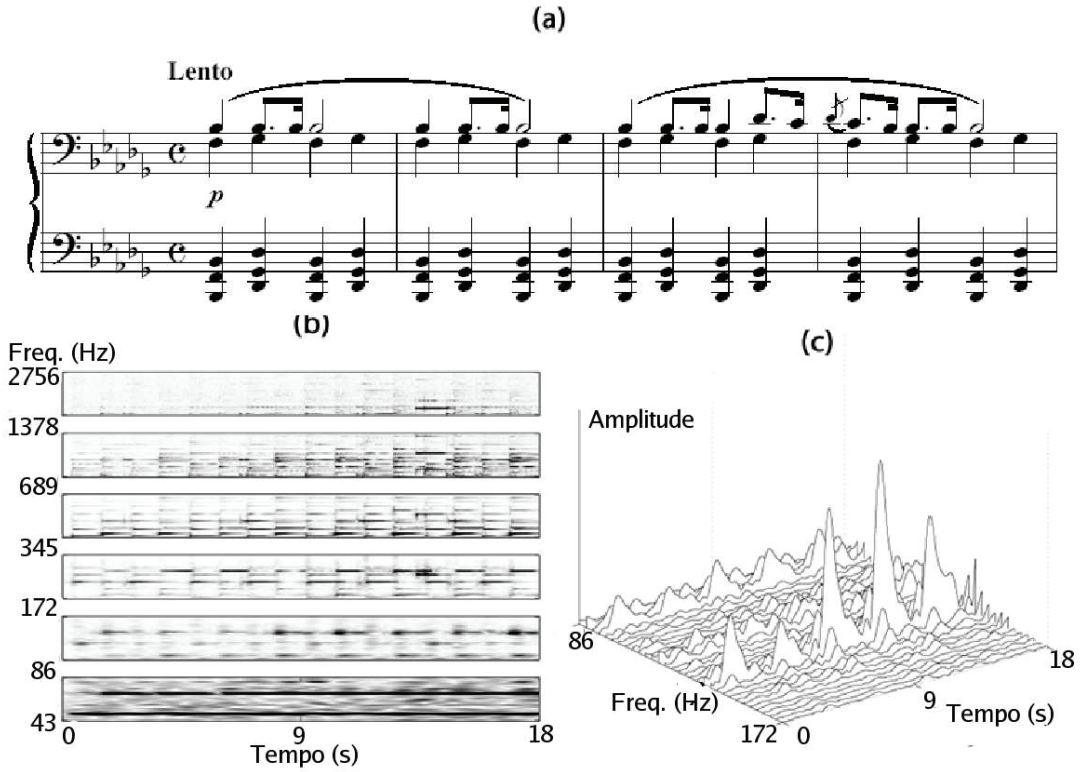


Figura 5: (a) Partitura de trecho da Marcha Fúnebre usado no Exemplo 4. (b) Sonograma das oitavas de 2 a 7 no Exemplo 4. (c) Dinâmica temporal da oitava 3 no Exemplo 4.

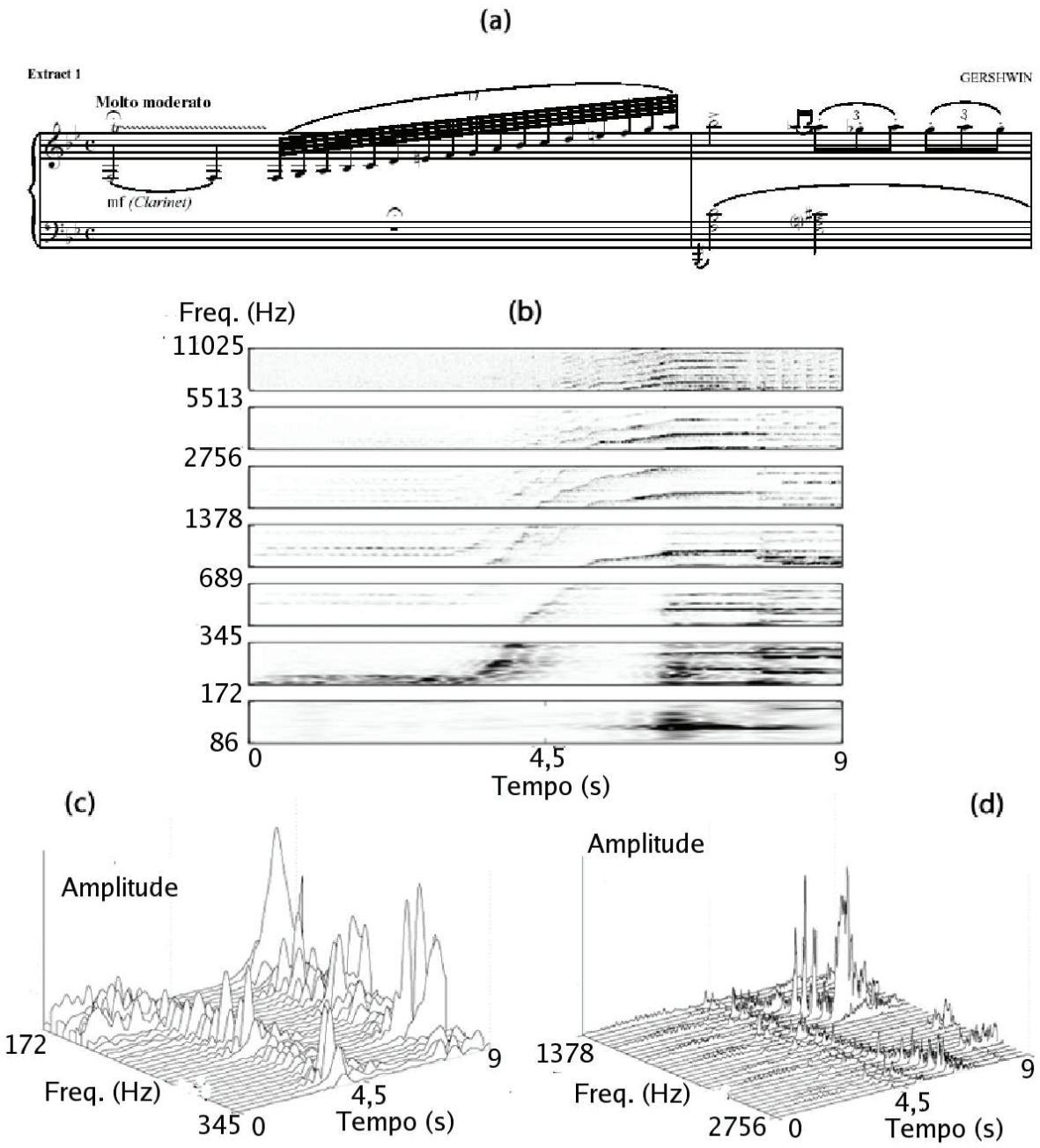


Figura 6: (a) Partitura de trecho de "Rhapsody in Blue" usado no Exemplo 5. (b) Sonograma das oitavas de 5 a 10 no Exemplo 5. (c) Dinâmica temporal da oitava 4 no trinado do Exemplo 5. (d) Dinâmica temporal da oitava 7 no *glissando* do Exemplo 5.

completo de TAM incluirá uma heurística sofisticada para esse fim.

Pode-se notar que o pior desempenho do sistema se deu nas freqüências mais baixas, onde ainda não há um bom compromisso entre as resoluções temporais e freqüenciais. Também como ja observado qualitativamente na experiência de altas freqüências, tal região do espectro apresentou o melhor desempenho. O desempenho médio na experiência de passagens lentas se deveu à presença de acordes ricos e respectivas parciais, enquanto que as passagens rápidas tiveram sua detecção prejudicada devido ao fato de a resolução temporal não ter sido suficiente (a acentuada reverberação presente na gravação certamente contribuiu para problema).

Já que o mapeamento entre os canais do banco de filtros e as freqüências das notas ainda está em um estágio inicial, foi admitida uma tolerância de 1 semitom e efetuada nova contagem. Percebe-se que, considerando tal tolerância, os indicadores de desempenho sobem, o que sugere que o mapeamento pode ser

Experiência	Percentual de acertos exatos	Percentual de acertos +/- 1 semitom
Freqüências baixas	36,67 %	73,34 %
Freqüências altas	62,90 %	80,64 %
Passagens rápidas	41,93 %	51,61 %
Passagens lentas	60 %	60 %
Trinado	100 %	100 %
Glissando	100%	100 %

Tabela 1: Percentuais de acerto de notas para as experiências realizadas.

melhorado reduzindo-se o número de canais por oitava. É importante lembrar que, como existem 32 canais mas apenas 12 semitons por oitava, há em média quase 3 canais por semitom. Formas mais elaboradas de mapeamento serão objeto das próximas etapas da pesquisa. Esse mapeamento, aliado a uma forma robusta de detecção de *onset* e *offset* das notas e a um sistema que permita discriminar automaticamente notas verdadeiras de parciais elevadas será o fundamento de um sistema completo para transcrição automática de música.

4 CONCLUSÃO

Neste artigo foi feita uma análise do desempenho do método de análise BQFFB como base de um sistema de transcrição musical automática (TMA). Ao sistema foram submetidos sinais musicais com características escolhidas para avaliar sua resposta a problemas específicos envolvidos em TMA. Foi possível verificar que o BQFFB apresenta um bom desempenho nas altas freqüências, permitindo detectar com precisão o início e o fim das notas (a detecção automática de *onset* e *offset* será implementada no sistema completo de TAM). Contudo, também foi visto que o desempenho em baixas freqüências fica prejudicado, em alguns casos específicos, dado que as estreitas bandas de passagem dos filtros com freqüências centrais baixas não são capazes de comportar as largas bandas dos transitórios. Efeitos como o trinado e o *glissando* foram bem analisados, com resultados bastante positivos. Este estudo ilustra o funcionamento de um sistema TMA como um todo e serve como base para a aperfeiçoamento do sistema de TMA usando a BQFFB.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] J. C. Brown, "Calculation of a constant Q spectral transform," *Journal Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[2] J. W. Cooley and J. W. Tukey, "An algorithm for the machine computation of complex Fourier series," *Mathematics of Computation*, vol. 19.

[3] F. C. C. B. Diniz, I. Kothe, S. L. Netto and L. W. P. Biscainho, "High-Selectivity Filter Banks for Spectral Analysis of Music Signals," *EURASIP Journal Advances on Signal Processing*, vol. 2007, pp. 1–12, paper ID: 94704, 2007.

[4] F. C. C. B. Diniz, I. Kothe, L. W. P. Biscainho, and S. L. Netto, "A bounded-Q fast filter bank for audio signal analysis," in *Proc. IEEE International Telecommunications Symposium*, pp. 971–975, Fortaleza, Brazil, Sept. 2006.

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

[6] D. B. Graziosi, C. N. dos Santos, S. L. Netto, and L. W. P. Biscainho, "A constant-Q spectral transformation with improved frequency response," in *Proc. IEEE International Symposium on Circuits and Systems*, vol. V, pp. 544-547, Vancouver, Canada, May 2004.

[7] K. L. Kashima and B. Mont-Reynaud, "The bounded-Q approach to time-varying spectral analysis," Stanford University, Dep. of Music, Tech. Rep. STAN-M-28, 1985.

[8] A. Klapuri and M. Davy (Editores), *Signal Processing Methods for Music Transcription,* Nova Iorque, EUA: Springer-Verlag, 2006.

[9] Y. C. Lim, "Frequency-response masking approach for the synthesis of sharp linear phase digital filters," *IEEE Trans. Circuits and Systems*, vol. 33, no. 4, pp. 357–364, 1986.

[10] Y. C. Lim and B. Farhang-Boroujeny, "Fast filter bank (FFB)," *IEEE Trans. Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 39, no. 5, pp. 316–318, 1992.

[11] C. N. dos Santos, S. L. Netto, L. W. P. Biscainho, and D. B. Graziosi, "A modified constant-Q transform for audio signals," in *Proc. IEEE International Conference on Audio, Speech, and Signal Processing*, vol. 2, pp. 469–472, Montreal, Canada, May 2004.

[12] P. P. Vaidyanathan, *Multirate Systems and Filter Banks* Upper Saddle River, USA: Prentice-Hall, 1992.

[13] D. William and E. Brown, *Theoretical Foundations of Music*, Belmont, USA: Wadsworth, 1978.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Evaluation of Threshold-Based Algorithms for Detection of Spectral Peaks in Audio

Leonardo de O. Nunes,[1] Paulo A. A. Esquef,[2] and Luiz W. P. Biscainho[1]

[1] LPS - PEE/COPPE, UFRJ
Caixa Postal 68504 - Rio de Janeiro, RJ, 21941-972, Brazil

[2] Instituto Nokia de Tecnologia
Av. Torquato Tapajós, 7200 - Manaus, AM, 69048-660, Brazil

lonnes@lps.ufrj.br, paulo@indt.org.br, wagner@lps.ufrj.br

ABSTRACT

Peak picking algorithms play an important role in audio analysis and synthesis methods based on sinusoidal modelling. Peak detection aims at selecting only those peaks corresponding to genuine resonant components present in the signal, while rejecting noise-induced ones. This paper investigates the performance of four threshold-based schemes for peak detection: two-pass split window, low-order autoregressive model, a nonlinear recursive filter and a stochastic spectrum estimator. The results of performance comparisons measured under a common metric are reported for a set of specific experimental setups.

0 INTRODUCTION

Sinusoidal modelling (SM) represents an audio signal as a sum of amplitude and frequency (phase) modulated sinusoids. SM finds use in a variety of audio-related applications, such as speech and audio synthesis, automatic transcription of music, and audio coding. First introduced for speech analysis in [1] and for audio signals in [2] SM has been expanded [3] and modified [4] to suit its different applications.

This paper considers the classic analysis method proposed in [1, 2]. A key-task to carry out a successful SM analysis is the so-called spectral peak picking. In this stage an algorithm has to decide whether a certain spectral peak corresponds to a genuine resonance in the input signal or is a spurious occurrence. The latter can happen due to either intrinsic characteristics of the spectral analysis or additive noise in the input signal.

A common peak picking approach considers an energy-based selection criterion in the frequency domain. For instance, spectral peaks whose energy exceeds a certain threshold can be segregated as genuine peak occurrences.

This paper evaluates a selection of methods that yield frequency dependent threshold curves, within a

peak detection system. A suitable test environment, composed by a test signal generator, a configurable peak detection system, and a performance meter, is created. The objective assessment on how well the investigated methods work can be obtained through performance measurements taken over the peak detection system.

After this introduction, the paper briefly outlines the sinusoidal analysis system studied. Then, it describes the adopted peak picking strategy and its possible processing configurations. In the sequel, the paper explains the structure and setup of the experimental system devised to assess the performance of the tested methods. Finally, the measured performance indicators are presented and discussed.

1 SINUSOIDAL MODELLING

Sinusoidal modelling describes an audio signal $x(t)$ as a sum of L sinusoids, possibly modulated in amplitude- and (phase- or) frequency

$$x(t) = \sum_{l=1}^{L} A_l(t) \sin \Psi_l(t), \tag{1}$$

$$\Psi_l(t) = \Psi_l(0) + \int_0^t \omega_l(u) du. \tag{2}$$

The continuous nature of the amplitude $A_l(t)$ and angular frequency $\omega_l(t)$ leads to a computationally intractable problem. In order to simplify the analysis, Eq. (1) is commonly replaced by a discrete model

$$x[n] = \sum_{l=1}^{L} A_l[n] \sin \Psi_l[n], \tag{3}$$

which can be further considered short-time stationary in amplitude and frequency. That is, for a given partial l, and assuming that $A_l[n]$ and $\Psi_l[n]$ possess much narrower bandwidth than that of the signal under analysis, the approximations $A_l[n] \approx A_l$ and $\Psi_l[n] \approx \Omega_l n + \Psi_l[0]$, where A_l and Ω_l are constant values, hold true during a time interval of N samples.

Fig. 1 illustrates the typical processing stages involved in the analysis part of a sinusoidal modelling system [5].

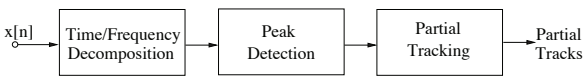


Figure 1: The three essential processing stages of a sinusoidal analysis system.

The 'time/frequency decomposition' stage performs a discrete-time Short Time Fourier Transform (STFT) [3] on the audio signal, i.e.,

$$X[k,m] = \mathrm{STFT}(x[n,m]) \tag{4}$$

$$= \frac{1}{N} \sum_{n=0}^{N-1} w[n] x[n+mH] e^{-jk\frac{2\pi}{N}n}, \tag{5}$$

where $w[n]$ is a window function of length N, e.g., the Hamming window, k is the frequency bin index, m denotes the analysis frame index, and H represents the hop size of the analysis frame along time. The output of the first stage is the magnitude spectrum of $x[n,m]$, which is defined by $S[k,m] = |X[k,m]|$.

The 'peak detection' stage receives $S[k,m]$ as input and selects only those peaks that correspond to stationary sinusoidal components of the signal present in frame m. The amplitude and frequency of the detected peaks can be finely estimated by a number of dedicated methods [2, 6]. Finally, a 'partial tracking' procedure is responsible for matching peaks across consecutive frames, in order to form the so-called partial tracks. These tracks contain information regarding the life-cycle of the sinusoidal components present in the signal under analysis.

Since the computational complexity of the 'partial tracking' algorithm increases with the number of selected peaks, the 'peak detection' stage should output only the most prominent spectral peaks. Another reason for a more judicious selection of spectral peaks is to avoid the formation of too short-living partial tracks due to noise-induced peaks. This way, a more precise model for the sinusoidal components of the signal can be obtained. Next section describes the peak picking strategy adopted in this work.

2 PEAK PICKING STRATEGY

Peak picking in audio spectrum can be carried out by several means, such as via threshold-based techniques [1, 2], analysis-by-synthesis schemes [4], and model-based analysis [7].

The present research focuses on threshold-based peak picking approaches. For that purpose, a general peak picking strategy is employed. It can be divided into two main parts: a spectral pre-processing stage and a peak selection algorithm, as depicted in Fig. 2. The latter will be fixed for all test cases investigated hereafter. On the contrary, the pre-processing stage is the variable element of the peak detection system: the system performance will be assessed w.r.t four different methods for spectral pre-processing. All these methods, as well as the peak selection criterion, are described in the subsequent sections.

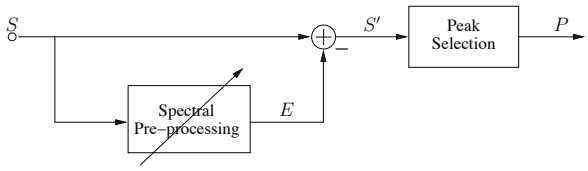


Figure 2: Block diagram illustrating the strategy adopted in the peak detection system.

2.1 Spectral Pre-Processing

2.1.1 Motivation

In threshold-based schemes, the usual selection criterion consists of choosing those spectral peaks that ex-

ceed a pre-defined energy level. While simple, such criterion is ineffective when dealing with general audio signals, which typically exhibit a spectral pattern whose energy decreases with frequency [8]. As a result, genuine low-energy spectral peaks in the high-frequency range may be discarded as spurious occurrences, as seen in Fig. 3.

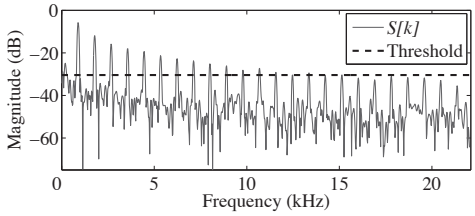


Figure 3: Inadequacy of applying a constant threshold to an observed spectrum of a noisy harmonic signal.

The pre-processing stage attempts to overcome the aforementioned difficulty compensating the observed spectral tilt. In a perhaps oversimplified manner, the spectral tilt can be compensated by an estimate of the colouring profile associated with the additive background noise that corrupts the signal of interest. Alternatively, an equally plain solution resorts to an estimate of the spectral envelope.

The ensuing sections describe four methods that have been designed to provide estimates for both the spectral envelope and the noise spectral profile. Either way, the desired estimates are referred to as $E[k]$.

2.2 Spectral Tilt Estimators

2.2.1 a) Two-Pass Split Window

The TPSW filtering has been originally proposed for noise spectrum estimation in sonar systems [9]. The procedure consists of three stages. In the first stage the sampled magnitude spectrum $S[k]$ of a given input signal is filtered through a split window defined as [9]

$$h^{\mathrm{sw}}[n] = \begin{cases} 0, & |n| < M^{\mathrm{sw}} \\ 1, & M^{\mathrm{sw}} \leq |n| < N^{\mathrm{sw}}, \end{cases} \tag{6}$$

with $0 \leq M^{\mathrm{sw}} < N^{\mathrm{sw}}$, being N^{sw} and M^{sw} positive integers that control, respectively, the length of the split window and the gap size.

The output of the first stage $\tilde{S}[k]$ is then modified according to the following criterion

$$\hat{S}[k] = \begin{cases} S[k], & \text{if } S[k] \leq \alpha\tilde{S}[k] \\ \tilde{S}[k], & \text{if } S[k] > \alpha\tilde{S}[k], \end{cases} \tag{7}$$

where $\alpha \geq 1$ is a parameter related to the ability to reject peaks in the observed spectrum.

In final stage, the modified spectrum $\hat{S}[k]$, which is supposed to be free of prominent peaks, is filtered through a conventional moving average filter, with the same length as that of the split window. The output of this third stage corresponds to the desired estimate, $E^{\mathrm{tpsw}}[k]$. Note that the FIR filters used in first and third stages are normalized for unity DC gain.

In the TPSW procedure, $S[k]$ is considered within the range between from 0 to π. In order to avoid boundary effects during the filtering, it is extended by about 20% of its original size. For that, part of the spectrum at both extremities of $S[k]$ is mirrored. Moreover, filtering delays are compensated in order to guarantee the synchronism between input and output. This is accomplished by taking only the central part of the correlation from the filtering results. In the end, the extensions appended to $S[k]$ are discarded in order to restore its original size.

When it comes to the choice of the split-window parameters, the smoothness of the estimate increases with N^{sw}. The value of M^{sw} should be set appropriately chosen as to yield the window gap as large as the bandwidth of a typical peak present in $S[k]$. The value of α is related to the peak rejection capability. It should be chosen small enough to guarantee that the component $\alpha\tilde{S}[k]$ is below the average magnitude of the peaks and large enough to place the component $\alpha\tilde{S}[k]$ above the noise floor. Typically, choosing $2 \leq \alpha \leq 8$ provides satisfactory results.

2.2.2 b) Low-Order AR Estimation

Autoregressive (AR) models are widely used in audio signal processing, such as vocal tract estimation in speech processing [10, 11].

Here, the idea is to take the spectral *envelope* estimate of a given audio signal for the desired spectral tilt curve. Thus, the procedure consists of fitting a low-order AR model to the time-domain signal $x[n]$ associated with $S[k]$. The desired envelope $E^{\mathrm{ar}}[k]$ is the magnitude spectrum of the estimated AR model.

In mathematical terms, a relaxed assumption is that $s[k]$ is governed by the following AR model

$$x[n] = \sum_{u=1}^{q} a[u]x[n-u] + r[n], \tag{8}$$

where q is the (insufficient) model order, $a[u]$ are the model coefficients, and $r[n]$ is the modeling error.

The model parameters can be estimated via any standard AR estimator, such as the Yule-Walker and Burg methods. Once the model $A(z) = \left[1 - \sum_{u=1}^{q} a_u z^{-u}\right]^{-1}$ is available, the desired spectral envelope is obtained by $E^{\mathrm{ar}}[k] = |A(e^{j\omega_k})|$, where $\omega_k = \frac{2\pi k}{N}$, with N being the length of DFT buffer used in the spectral analysis.

2.2.3 c) Nonlinear Recursive Smoothing Filter

In [12] a nonlinear recursive smoothing filter (NRSF) is proposed for estimating the spectral profile of coloured noise in audio spectra. The filter, devised under the assumption that the power spectrum density of the noise component would vary slowly over frequency, works by limiting in modulus the first derivative (or slew rate) of the spectrum samples, with respect

to frequency. The nonlinear recursive filter that implements the solution is given by

$$E^{\text{nrsf}}[k] = E^{\text{nrsf}}[k-1]\beta^{\text{sign}\left(S[k]-E^{\text{nrsf}}[k-1]\right)}, \qquad (9)$$

where $S[k]$ is the observed magnitude spectrum of the signal under test; $E^{\text{nrsf}}[k]$ is the desired spectrum estimate; $\text{sign}(x) = 1$ if $x \geq 0$ and $\text{sign}(x) = -1$ if $x < 0$; and β is a constant slightly larger than unity.

The parameter β can be expressed as $\beta = 1 + \lambda$. In theory, λ should be chosen as to exceed the maximum slew rate associated with the power density spectrum of the noise component. In practice, as can be seen from Eq. (9), the value of λ controls the forgetting factor of the filter. Thus, the larger λ, the higher the variance of $E^{\text{nrsf}}[k]$ becomes. Choosing $\lambda = 0.05$ provides a sufficiently smooth $E^{\text{nrsf}}[k]$ [12].

When dealing with recursive schemes, filter initialization should be considered carefully. An improper initialization can bias the initial samples of $E^{\text{nrsf}}[k]$ and thus degrade the overall performance of the algorithm. One possible solution is to extend $S[k]$ at the boundaries, as described earlier in Section 2.1.2.a, and initialize the recursion with $E^{\text{nrsf}}[k-1] = 0$. Of course, the spectrum extension should be long enough for the influence of a wrong initialization be mitigated. Afterwards, the estimated values of $E^{\text{nrsf}}[k]$ outside the original range of $S[k]$ are discarded. Alternatively, $E^{\text{nrsf}}[k-1]$ can be set as the median value among the first C samples of $S[k]$. In both cases, the original frequency range of $S[k]$ is from 0 to π.

2.2.4 d) Stochastic Spectrum Estimation

The stochastic spectrum estimator (SSE) is another non-linear stochastic spectrum estimator that has been introduced by [13].

Assuming $S[k]$ as defined before, the SSE method consists of the four steps described below:

1. Pass $S[k]$ through a three-tap moving average filter, in order to obtain a $S^1[k]$ possibly free of null magnitude samples;
2. Compute $R[k] = \frac{1}{S^1[k]}$;
3. Obtain $R^1[k]$ as a smoothed out version of $R[k]$, by computing a cyclic convolution between $R[k]$ and an N^{sse}-tap moving average FIR filter;
4. Compute the desired estimate as $E^{\text{sse}}[k] = \frac{1}{R^1[k]}$.

As with the TPSW method, the smoothness of $E^{\text{sse}}[k]$ increases with the value of N^{sse}. Moreover, all FIR filters used in the procedure are normalized to force the DC gain equal to 0 dB. Differently from the TPSW filtering, in the SSE scheme $S[k]$ should be considered within the whole range between 0 and 2π, in order to make effective the use of the cyclic convolution. It should be noticed, however, that one could employ $S[k]$ within the range between 0 and π, provided that this $S[k]$ is sufficiently extended at both extremities via the mirroring scheme mentioned in Section 2.1.2a.

2.2.5 Performance Comparison

In order to illustrate the qualitative performance of the previously described methods, they are tried over a test-signal, under the processing setups detailed below.

- Input signal: 2048 samples of a noisy harmonic signal with fundamental frequency equal to 1 kHz, sampled at 44.1 kHz, and windowed by a 2048-sample Hann window;
- S[k]: magnitude of a 2048-point DFT applied to the input signal;
- TPSW filtering: $N^{\text{sw}} = 51$, $M^{\text{sw}} = 8$, and $\alpha = 4$;
- AR method: $q = 10$;
- SSE method: $N^{\text{sse}} = 101$;
- NLRF method: $\beta = 1.01$.

The attained results are summarized in Fig. 4. It can be observed that all methods succeed in catching the overall shape of the spectrum. It is also worth noticing that, with exception of $E^{\text{ar}}[k]$, the remaining spectral envelopes follow closely the local average of the noise floor. The suspended $E^{\text{ar}}[k]$ is nothing to worry about, since those estimates will not be used directly as variable magnitude thresholds for the reference spectrum. Rather, they will serve to compensate the spectral tilt of the reference spectrum, before proceeding to the peak detection stage.

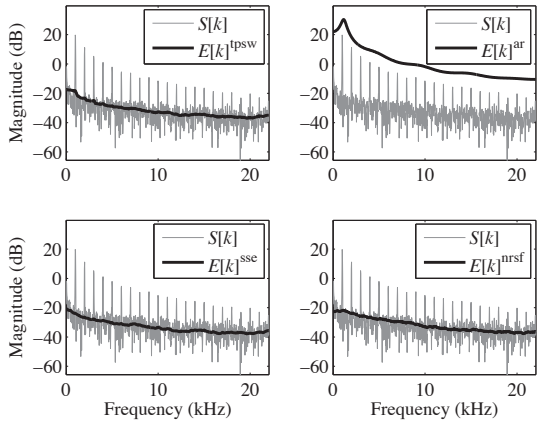


Figure 4: Qualitative performance comparison among the methods under evaluation.

2.3 Peak Selection Criterion

As depicted in Fig. 2, peak selection is performed over $S'[k]$, which is a tilt-compensated version of $S[k]$. Note that the compensation is carried out in the logarithmic scale. Fig. 5 compares an original $S[k]$ and its tilt-compensated version $S'[k]$, in an example where $E^{\text{sse}}[k]$ was used. Now, it is clear that a constant magnitude threshold can be employed to discriminate genuine from spurious spectral peaks.

As regards the adopted peak selection criterion, assume first the set $k \in \{2, 3, \ldots, (N/2-1)\}$ of $S'[k]$ bin indices. For all elements of k, collect in a sub-set $\mathcal{P}_m$ the indices k_{peak} that satisfy simultaneously the following conditions:

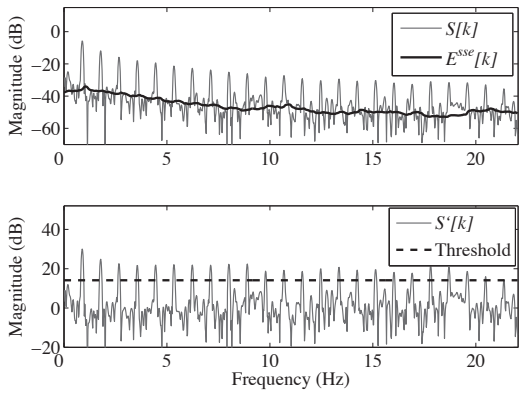


Figure 5: Comparison between the original (top) and tilt-compensated spectra (bottom).

1. $S'[k] > S'[k-1]$;
2. $S'[k] > S'[k+1]$;
3. $S'[k] > d\mu$;

In condition 3, d is an empirically chosen multiplier and μ is an estimate of the standard deviation of the noise observed in $S'[k]$. The selected indices k_{peak} contain the desired bin indices associated with the detected peaks in $S'[k]$.

The first two conditions above define which samples of $S'[k]$ could be qualified as peaks, regardless of being genuine or spurious. The third condition sets the value of the discrimination threshold, which sets the minimum energy a peak should possess to be qualified as a genuine peak.

The value of μ can be obtained by any estimator robust enough to provide reliable estimates for the standard deviation of data series, despite the presence of outliers. It was found experimentally that, when dealing with spectra which are densely populated by genuine peaks (outliers here), the median operator tends to over-estimate the standard deviation of the noise. Fortunately, with exception of the AR-based technique, the very spectrum pre-processing methods presented in Section 2.1.2 are competitive alternatives to aid the estimation of the noise standard deviation.

Among the available options for computing μ, the SSE method was found to be the least affected by the presence of genuine peaks. This is justified not only by the SSE's own formulation, but also experimentally. Thus, the adopted solution was to estimate μ as

$$\mu = \text{mean}(\bar{E}^{sse}[k]), \tag{10}$$

where $\bar{E}^{sse}[k]$ is the curve output by the SSE method to any tilt-compensated spectrum $S'[k]$.

As for the value of d, one assumes that the pre-processing was successful in 'whitening' the noise component, which can be considered Gaussian. Moreover, μ is believed to be a reliable estimate for the standard deviation of the noise, as observed in the frequency domain. In such case, setting $2 \leq d \leq 5$, assures a confidence interval greater than 95% that the spurious peaks will fall below the adopted threshold.

In reality, for a given signal-to-noise ratio (SNR), the larger the number of genuine peaks present in the signal, the less they tend to stand out from the noise floor, due to energy sharing among peaks. This favors the occurrence of detection errors and requires a more careful selection of d. On the contrary, placing the selection threshold is easier when dealing with a few genuine spectral peaks, even for low SNRs.

The aforementioned condition motivates the following strategy to set the value of d.

1. Calibrate d in order to assure a satisfactory detection performance considering a scenario with SNR as low as 10 dB and a spectrum densely populated with genuine peaks;
2. Attribute the previously found value of d to $d_{\min}$;
3. Compute $\rho = 1 \sqcap \left(\frac{\max(S'[k])-\mu}{10}\right)^{0.5}$, where $a \sqcap b$ stands for 'maximum between a and b';
4. Make $d = \rho d_{\min}$.

Although adequate for cases with many spectral peaks, the value of $d_{\min}$ tends to be too low for $S'[k]$ bearing few peaks of interest. As a consequence, the occurrence of false detections is favored.

In the computation ρ, the quantity $(\max(S'[k]) - \mu)$ can interpreted as the available room in magnitude between the spectral maxima and the average level of the noise floor. Thus, if this room is larger than 10 dB, the multiplier $\rho > 1$ contributes to raise the threshold by about half the magnitude room excess. Otherwise, $\rho = 0$ and $d = d_{\min}$. The former situation is likely to happen when there are few genuine peaks in $S'[k]$, avoiding the occurrence of false alarms.

3 SIMULATION SETUP

This section describes the experimental setup used to evaluate how well the peak detection system performs. As depicted in Fig. 6, the test setup consists of a signal generator, whose output $x[n]$ is fed to a STFT analyser, which in its turn provides the magnitude spectrum $S[k]$ to the peak detection system (see Fig. 2). Moreover, a performance meter assesses the peak detection results in quantitative terms, by counting how many of the selected peaks are in fact correctly detected, according to a reference indicator.

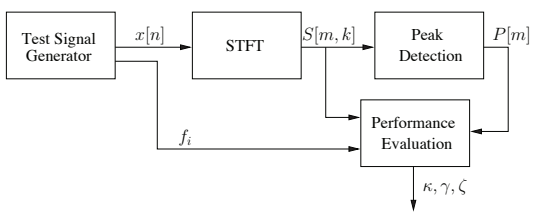


Figure 6: Setup devised to assess the performance of the peak detection system.

3.1 3.1 Test Signal Generation

The test signal, sampled at 44.1 kHz, is composed of two components: one deterministic and another stochastic. The former consists of a sum of harmonic signals with fundamental frequencies $f_{0,i}$. For each $f_{0,i}$, which is chosen randomly between 200 Hz and 1 kHz, with uniform probability, harmonics are generated up to the Nyquist frequency. Moreover, the amplitude of the harmonics can be either constant or decaying over frequency.

As for the zero-mean additive stochastic component, it can be either white or pink noise. In mathematical terms the test signal is defined as

$$x[n] = \sum_{i=0}^{I} \sum_{j=1}^{J_i} a_{ij} \cos\left(\frac{2\pi j f_{0,i} n}{F_s}\right) + \sigma\eta[n]. \qquad (11)$$

The double summation in Eq. (11) represents the deterministic part of the signal, where F_s is the sampling frequency, I is the number of fundamental frequencies in the mixture, and $J_i = \left\lfloor \frac{F_s}{2f_{0,i}} \right\rfloor$, with $\lfloor . \rfloor$ denoting 'the greatest integer less than or equal to', defines the number of harmonics associated with a given $f_{0,i}$. The amplitude of the harmonics a_{ij} is set either to unity or decreases with frequency according to $a_{ij} = \frac{1}{jf_{0,i}}$.

The second part of Eq. (11) represents the noise component, being $\eta[n]$ with $0 \leq n \leq (N-1)$, one realization of a stochastic process having constant or $1/f$ power spectrum density, and σ a variable that controls the noise power in order to force a desired SNR.

3.2 Peak Detection Setup

The computation of $S[k,m]$ and $S'[k,m]$ is identical to the procedures and parameters described as in Section 2.1.3. Those processing parameters were empirically tuned across the different methods as to render fair the comparisons among their performance.

As regards the peak selection algorithm, the main parameter to be set is the multiplier $d_{\min}$ of the noise standard deviation. It was found that $d_{\min} = 2$ is a suitable choice. The length of the second SSE filter was set to $N^{\text{sse}} = 150$.

3.3 Performance Evaluation

Peak detection performance is assessed basically by means of counting the number of correctly detected peaks and that of false alarms. Here, the main issue to consider is the observed *presence* of a given peak, regardless of whether it was detected with precisely estimated magnitude and frequency. In any case, it is necessary to define the conditions upon which a peak can be considered correctly detected.

On the measurement side, the peak observation domain is $S[k,m]$. Thus, the frequency of any observed peak can only lie in one of the available frequency bins, i.e., kF_s/N, for $0 \leq k < N/2$. On the reference side, the frequencies of the peaks in the test signal can be set in a frequency grid as fine as desirable. In order to ensure a meaningful peak detection performance assessment, the reference domain should be made compatible to the measurement domain. The following sections examine the issue in more detail.

3.3.1 3.3.1 Reference Domain Alignment

First, let the set Φ contain all the frequencies ϕ_i associated with the deterministic part of the input signal. The primary goal is to foresee which bins in the observation domain would be more strongly activated by those frequencies ϕ_i. Thus, each element ϕ_i is quantized to the nearest frequency bin, i.e., $\bar{\phi}_i = k_i F_s/N$, for $k_i = \text{round}(\phi_i N/F_s)$, where N is the length of the STFT analysis buffer and F_s the sampling frequency.

Gathering all non-repeated occurrences of k_i in a set $\mathcal{K}$, one can define a preliminary reference vector, whose elements are defined as

$$r[k] = \begin{cases} 1, & \text{if } k \in \mathcal{K} \\ 0, & \text{otherwise} \end{cases}, \text{ for } 0 \leq k < \frac{N}{2}. \qquad (12)$$

Note, however, that $r[k] = 1$ may not necessarily coincide with observed peaks of any kind in $S[k,m]$. This is because the position of an observed local maximum in $S[k,m]$, due to a given ϕ_i, is influenced not only by the corrupting noise, but also by the presence of the other ϕ_i.

The key-point here is to build a reference mask that corresponds to peak occurrences truly observable in the measurement domain. It is then evident that frequency quantization alone is an insufficient criterion to align the reference and measurement domains.

A more adequate domain alignment can be created by means of an auxiliary binary vector, the elements of which are defined as

$$c[k] = \begin{cases} 1, & \text{if } k \in O_m \\ 0, & \text{otherwise} \end{cases}, \text{ for } 0 \leq k < \frac{N}{2}, \qquad (13)$$

where O_m is a set with cardinality $|O_m|$ containing all bin indices associated with *observed* peaks in $S[k,m]$, either genuine or spurious. Then, the aligned reference vector is obtained by

$$r_a[k] = r_0[k] \oplus r_{-1}[k] \oplus r_{+1}[k], \qquad (14)$$

with $r_0[k] = r[k] \wedge c[k]$, $r_{-1}[k] = r[k-1] \wedge c[k]$, and $r_{+1}[k] = r[k+1] \wedge c[k]$, for $1 \leq k < (N/2 - 1)$, where the symbols $\wedge$ and $\oplus$ stand for the boolean operations 'AND' and 'XOR', respectively.

The parts of $r_a[k]$ can be interpreted as follows:

- $r_0[k]$ nullifies $r[k]$ if $c[k] = 0$, i.e., when active bins in the reference are not observed peaks;
- $r_{-1}[k]$ moves an active bin in $r[k]$ to the adjacent one on the left side if $c[k] = 0$ but $c[k-1] = 1$;
- $r_{+1}[k]$ moves an active bin in $r[k]$ to the adjacent one on the right side if $c[k] = 0$ but $c[k+1] = 1$.

Table 1: Performance indicators obtained from **Test 1**.

	SNR (dB)														
	0	5	10	15	20	0	5	10	15	20	0	5	10	15	20
TPSW	85.6	98.9	99.3	99.5	99.6	17.5	4.6	0.5	0.3	0.2	0.68	0.94	0.99	0.99	0.99
AR	88.9	99.7	99.8	99.8	99.8	22.5	13.4	1.4	0.2	0.1	0.66	0.86	0.98	1.00	1.00
NRSF	83.3	97.1	99.2	99.3	99.2	15.0	1.7	0.7	0.4	0.29	0.68	0.95	0.99	0.99	0.99
SSE	85.6	97.9	99.2	99.3	99.3	18.1	2.5	0.7	0.5	0.3	0.68	0.95	0.99	0.99	0.99
No method	38.3	63.0	84.9	93.6	98.5	6.3	2.3	0.9	0.4	0.2	0.32	0.61	0.84	0.93	0.98
Metric	γ					ζ					κ				

Table 2: Performance indicators obtained from **Test 2**.

	number of $f_{0,i}$ (I)														
	3	4	5	6	7	3	4	5	6	7	3	4	5	6	7
TPSW	97.2	91.6	85.7	78.0	71.0	3.0	3.5	3.6	3.8	4.0	0.94	0.88	0.82	0.74	0.67
AR	98.3	92.5	86.4	78.4	71.4	4.5	4.3	3.9	3.7	3.5	0.94	0.88	0.83	0.75	0.68
NRSF	91.4	84.6	79.4	72.5	66.4	2.1	3.6	4.5	5.3	5.8	0.89	0.81	0.75	0.67	0.61
SSE	94.3	88.5	82.9	75.5	69.0	2.0	2.9	3.3	3.8	4.1	0.92	0.86	0.80	0.72	0.65
No method	48.4	38.2	33.1	25.6	24.0	2.2	2.8	3.3	3.9	4.5	0.46	0.35	0.30	0.25	0.21
Metric	γ					ζ					κ				

In practice, the indices k for which $r_a[k] = 1$ indicate the reference spectral locations against which to confront those of the detected peaks in $S[k,m]$. Once $r_a[k]$ is obtained, the percentage of correctly detected peaks can be computed as

$$\gamma = \frac{G}{Q}, \tag{15}$$

where $Q = \sum_{k=0}^{\frac{N}{2}} r_a[k]$ is the count of all peaks in the reference vector and $G = \sum_{k=0}^{\frac{N}{2}} g[k]$ is the count of all correctly detected peaks, with

$$g[k] = \begin{cases} r_a[k], & \text{if } k \in \mathcal{P}_m \\ 0, & \text{otherwise} \end{cases}, \tag{16}$$

where $\mathcal{P}_m$ is a set of cardinality $|\mathcal{P}_m|$ containing the indices of all detected peaks from $S'[k,m]$.

The percentage of false alarms can be computed as

$$\zeta = \frac{|\mathcal{P}_m| - G}{|O_m| - Q}, \tag{17}$$

where $|\mathcal{P}_m| - G$ is the count of all incorrectly detected peaks and $|O_m| - Q$ is the count of all observable peaks in $S'[k,m]$ that should have remained undetected.

In addition, one can define the meter

$$\kappa = \gamma - \zeta, \tag{18}$$

which aggregates the two previous measures into a single performance indicator. Ideally, $-1 \leq \kappa \leq 1$ and a perfect detection is achieved when $\kappa = 1$. A value of κ close to 1 indicates the occurrence of more correctly detected than false alarms. On the contrary, a value of κ close to -1 indicates a larger count of false alarms in comparison to that of correctly detected peaks.

4 RESULTS

All four methods discussed previously were tested with the same test signals, in order to investigate their behaviour under identical conditions. Moreover, the results when no spectral tilt compensation is applied are also calculated and displayed as 'no method'. The simulation setup and its processing parameters are the same as those described in the previous section.

Each test signal was designed to accommodate 10 analysis frames. For each test, 500 realizations were generated. The reported results correspond then to the average performance indicators measured for each frame and for each signal. For convenience the values of γ and ζ are displayed in %.

Test 1 aims at assessing peak detection performance under different SNRs. The chosen test signal consists of a single $f_{0,i}$ $(I = 1)$, with $a_{ij} = \frac{1}{jf_{0,i}}$, immersed in pink noise. The SNR is set from 0 to 20 dB in steps of 5 dB. The attained results are seen in Table 1.

The values of γ show that most of the genuine peaks are correctly classified even in low SNR conditions. As expected, the lower the SNR, the higher the values of ζ, showing a tendency to misclassify spurious peaks. All methods yielded similar values of κ, evidencing that all perform equally well, under the tested conditions. By contrast, the systematic lower values of κ when spectral tilt compensation is not applied demonstrate the effectiveness of its use.

Test 2 measures peak detection performance w.r.t. the number of harmonic signals present in the mixture. Setup: test signals containing $3 \leq I \leq 7$ fundamental frequencies, with $a_{ij} = \frac{1}{jf_{0,i}}$, immersed in pink noise, and SNR $= 10\ dB$. Table 2 summarizes the results.

It can be observed that γ tends to decrease with the increase of I. As for ζ, with exception of the AR method, it tends to increase with I. Among the tested methods, the NRSF method achieves the lowest performance whereas the AR scores are the highest, revealing some robustness to the increase of I.

Test 3 strains the methods by replacing pink with white Gaussian noise. This time, $I = 3$, $a_{ij} = \frac{1}{jf_{0,i}}$, and SNR = 10 dB. The results are organized in Table 3. As one could anticipate, the performance of all methods decreases. The AR method showed a high γ value, since $E^{\mathrm{ar}}[k]$ tends to decay with frequency, favoring the detection of more genuine peaks immersed in noise. A side-effect is an also high grade for ζ that reduces the value of κ. Overall, according to κ, the TPSW method achieves the highest performance.

Table 3: Performance indicators obtained from **Test 3**.

	Method				
	TPSW	AR	NRSF	SSE	No method
γ	64.2	76.9	44.1	41.6	38.9
ζ	11.0	31.5	2.1	2.1	0.2
κ	0.53	0.46	0.42	0.39	0.39

5 CONCLUSION

This work investigated the effect spectral tilt compensation on the performance of a threshold-based peak detection system. For this, four spectral tilt estimators were examined. Moreover, an experimental environment, composed of a test signal generator, the configurable peak detection system, and a performance evaluation block, was designed.

The test results clearly favors the use of spectral tilt compensation within a peak detection system. However, none of the methods stand out as contributing to an overall superior peak detection performance. In part, the achieved well-balanced performance among the methods can be attributed to the sensible choice of the processing parameters, as well as to the proposed heuristics used to set the threshold level.

6 ACKNOWLEDGEMENTS

The authors would like to thank *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) and *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro* (FAPERJ) for funding this work.

REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on Sinusoidal Representation," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[2] J. O. Smith III and X. Serra, "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation," in *Proc. Int. Computer Music Conference*, Champaign-Urbana, 1987.

[3] X. Serra, "Musical Sound Modeling with Sinusoids plus Noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. De Poli, Eds. Swets & Zeitlinger, 1997.

[4] E. B. George and M. J. T. Smith, "Analysis-by-Synthesis/Overlap-Add Sinusoidal Modeling Applied to the Analysis and Synthesis of Musical Tones," *J. Audio Eng. Soc.*, vol. 40, no. 6, pp. 497–516, June 1992.

[5] X. Serra and J. O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[6] F. Keiler and S. Marchand, "Survey of extraction of sinusoids in stationary sounds," in *Proc. 5th Conf. Digital Audio Effects*, Hamburg, Germany, 2002, pp. 51–58.

[7] M. H. Hayes, *Statistical Signal Processing and Modeling*, chapter 6, John Wiley & Sons, Inc., 1996.

[8] J. M. Grey and J. W. Gordon, "Perceptual effects of the spectral modifications of musical timbres," *J. Acoust. Soc. Am.*, vol. 61, pp. 1270–1277, 1978.

[9] W. A. Struzinski and E. D. Lowe, "A performance comparison of four noise background normalization schemes proposed for signal detection systems," *J. Acoust. Soc. Am.*, vol. 76, no. 6, pp. 1738–1742, Dec. 1984.

[10] R. W. Schafer and L. R. Rabiner, "Digital representations of speech signals," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 662–677, Apr. 1975.

[11] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[12] M. D. Macleod, "Nonlinear recursive smoothing filter and their ues for noise floor esimation," *IEEE Eletronics Letters*, vol. 28, no. 21, pp. 1952–1953, Oct. 1992.

[13] N. Laurenti, G. De Poli, and D. Montagner, "A nonlinear method for stochastic spectrum estimation in the modeling of musical sounds," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 531–541, Feb. 2007.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, São Paulo, SP

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

Avaliação de uma Representação Compacta por "Matching Pursuit" em um Esquema de Análise/Síntese de Áudio

Michel P. Tcheou,[1] Lisandro Lovisolo,[2] Luiz W. P. Biscainho[3] e Eduardo A. B. da Silva[3]

[1] UFRJ, LPS - PEE/COPPE, CP 68504, Rio de Janeiro, RJ, 21941-972, Brasil

[2] UERJ, DETEL/FEN, R. São Francisco Xavier 524, sala 5036A, Rio de Janeiro, RJ, 20559-900, Brasil

[3] UFRJ, LPS - DEL/Poli e PEE/COPPE, CP 68504, Rio de Janeiro, RJ, 21941-972, Brasil

pompeu@lps.ufrj.br, lisandro@uerj.br, wagner@lps.ufrj.br, eduardo@lps.ufrj.br

RESUMO

A decomposição de sinais por "Matching Pursuit" (MP) vem sendo aplicada em áudio. Na literatura, mostrou-se que um dicionário de janelas moduladas nas freqüências da escala musical cromática possibilita boa resolução para análise tempo-freqüência. Este trabalho implementa um esquema de análise/síntese de áudio utilizando MP cromático. Seu desempenho para diversas famílias de sinais é avaliado por uma medida objetiva descrita pelo ITU e por testes subjetivos informais, em função do número de coeficientes da representação. É sugerido, ainda, um critério para parada da decomposição.

0 INTRODUÇÃO

Recentemente, o estudo de técnicas capazes de obter representações compactas de sinais tem-se intensificado [1]. Essas técnicas têm obtido sucesso no processamento de sinais de áudio para fins de transcrição musical [2, 3] e codificação [4, 5]. A codificação de áudio, que é o enfoque deste trabalho, envolve diversas etapas tais como análise, mascaramento no tempo e na freqüência, quantização, codificação por entropia, quantização inversa e síntese. Para que o codificador apresente bom desempenho é fundamental utilizar um método de análise capaz de aproximar o sinal com o menor número possível de coeficientes.

Normalmente, essa aproximação é realizada através decomposições que aproximam os sinais que compõem um dado espaço, usando uma combinação linear de formas de onda pré-definidas (ou átomos) provenientes de um dicionário dito redundante por conter mais elementos que os necessários para gerar o espaço. As-

sim, pode-se escolher um subconjunto dos átomos de um dicionário redundante e associar-lhes diferentes pesos a fim de representar um dado sinal. Esse paradigma de representação de sinais é também conhecido por decomposição atômica. Existem diversos métodos que realizam decomposições atômicas; dentre eles, há o popular *Matching Pursuit* [6].

Uma representação compacta pressupõe alto grau de similaridade entre os elementos do dicionário e os fenômenos presentes no sinal. O uso de dicionários redundantes permite alcançar maior nível de compacidade na representação que o uso de bases ortonormais reais, tais como as *wavelets* discretas, visto que um sinal pode apresentar estruturas complexas que não são bem representadas utilizando-se somente uma classe de formas de onda. Por sua vez, o dicionário redundante pode ser formado por uma união de bases ortogonais, por transformadas redundantes (*wavelet packet* e *frames* de *Gabor*) ou por formas de ondas parametrizadas [4, 7, 8, 9].

A Figura 1 ilustra o esquema de codificação de sinais utilizando decomposições atômicas. O *codificador* analisa o sinal de forma a encontrar uma boa representação para este, com base em um dicionário D. Nesse dicionário são modelados os fenômenos existentes nos sinais que se deseja aproximar. Em seguida, os coeficientes e os índices dos átomos da representação do sinal são transmitidos/armazenados. O *decodificador*, com base no mesmo dicionário, sintetiza/reconstrói o sinal a partir dos átomos e pesos provindos do codificador.

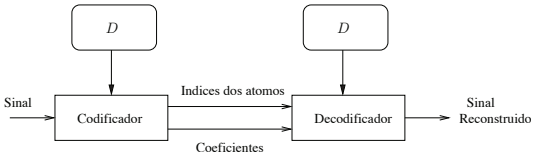


Figura 1: Codificação de sinais usando dicionários redundantes / representações atômicas

Se soubermos de que forma os sinais de uma classe são produzidos, podemos utilizar representações atômicas para tentar modelar o sistema físico que gera o sinal. Se conseguirmos um bom modelo e o codificarmos eficientemente, poderemos obter altas taxas de compressão. Esse é o contexto deste trabalho. Nele discutimos a utilização de representações atômicas obtidas usando o algoritmo conhecido por *Matching Pursuit* [6, 10] para sinais de áudio. Abordagens similares a esta já foram propostas na literatura [2, 3, 4, 5]. No presente trabalho, avaliam-se potencialidades e deficiências de tais métodos, e implementa-se um esquema simples de análise/síntese de áudio utilizando o MP cromático. Seu desempenho é avaliado para diversas famílias de sinais por uma medida objetiva descrita pelo ITU, em função do número de coeficientes da representação. É proposto, ainda, um critério para parada da decomposição.

A próxima seção apresenta aspectos relevantes de representações atômicas e como elas podem fornecer representações compactas e coerentes – relacionadas ao processo físico gerador dos sinais observados. A seção seguinte explica o dicionário aqui empregado, composto por senóides geradas a partir das freqüências da escala cromática. A seguir, detalha-se o algoritmo de decomposição empregado, que é uma variante do *Matching Pursuit*. Após isso, são apresentados resultados referentes ao desempenho do método para diversas famílias de sinais, que são também avaliadas usando o PEAQ [11]. Por fim, as conclusões são apresentadas.

1 REPRESENTAÇÕES ATÔMICAS

Numa representação atômica, um sinal $\mathbf{x}$ é descrito como uma combinação linear de poucas formas de onda selecionadas a partir de um conjunto pré-definido, o dicionário $\mathcal{D}$. Os sinais $\mathbf{g}_\gamma$ que compõem $\mathcal{D}$ são as formas de onda que podem ser utilizadas na decomposição, chamadas de átomos ou elementos. Algoritmos que obtêm representações atômicas de sinais escolhem um subconjunto de M elementos $\mathbf{g}_{\gamma(i)}$ do dicionário $\mathcal{D}$ que aproximam um sinal $\mathbf{x}$ na forma

$$\mathbf{x} \approx \hat{\mathbf{x}} = \sum_{i=1}^{M} \alpha_i \mathbf{g}_{\gamma(i)}, \quad \mathbf{g}_{\gamma(i)} \in \mathcal{D}. \tag{1}$$

Cada átomo $\mathbf{g}_{\gamma(i)}$ (indexado por $\gamma(i)$) utilizado na aproximação de $\mathbf{x}$ pode ser interpretado como uma característica intrínseca do sinal $\mathbf{x}$ de "peso" α_i. A representação mais compacta de $\mathbf{x}$ é aquela que utiliza o menor número de átomos [4] com distorção nula. Na prática, não se procura, necessariamente, o menor número de átomos, pois um pequeno conjunto de átomos que aproxime $\mathbf{x}$ com uma distorção aceitável pode ser suficiente [7, 8]. Assim, a compacidade de uma representação está relacionada ao número de termos M, e à medida que M cresce menos compacta é a representação.

Para que combinações lineares de M átomos de um dicionário $\mathcal{D}$ sejam capazes de representar qualquer sinal $\mathbf{x} \in \mathbb{X}$ com distorção $d(\mathbf{x}, M, \mathcal{D})$ arbitrariamente pequena, $\mathcal{D}$ tem que ser completo em $\mathbb{X}$ [10]. Assim, deverá existir pelo menos uma combinação linear de elementos de $\mathcal{D}$ que produza $\hat{\mathbf{x}} = \mathbf{x}$, $\forall\ \mathbf{x} \in \mathbb{X}$, isto é, $\mathcal{D}$ deve gerar o espaço $\mathbb{X}$. Dicionários são ditos sobre-completos ou redundantes quando possuem mais elementos que os necessários para tal fim. A redundância permite que diferentes combinações lineares de elementos de $\mathcal{D}$ expressem um mesmo $\mathbf{x}$.

Quando lidamos com processos físicos, o sinal observado é uma mistura de componentes $\mathbf{p}_i$, que representam fenômenos físicos, dada por

$$\mathbf{x} = \sum_i \beta_i \mathbf{p}_i + \mathbf{n}, \tag{2}$$

onde $\mathbf{n}$ é o ruído inerente à observação. Quanto mais parecidos forem os átomos $\mathbf{g}_{\gamma(i)}$ e seus pesos α_i utilizados na representação de um sinal $\mathbf{x}$ com as componentes $\mathbf{p}_i$ e seus pesos β_i presentes em $\mathbf{x}$, melhor será a

representação obtida para modelagem do sinal e reconhecimento de padrões, e a dizemos coerente.

Assim, vemos que para se poder obter uma representação compacta ou coerente, um dos aspectos-chave reside no dicionário, isto é, na definição de seus átomos.

2 DICIONÁRIO DE ESCALA MUSICAL CROMÁTICA

Os elementos de um dicionário redundante podem ser gerados escalonando-se, transladando-se e modulando-se uma função janela real $h(t)$. No caso de um dicionário de átomos complexos, cada átomo $\mathbf{g}_\gamma$ é definido como [6]

$$g_\gamma(n) = K_\gamma h\left(\frac{n-u}{s}\right) e^{i\xi n}, \tag{3}$$

onde K_γ é tal que $\mathbf{g}_\gamma$ possui norma unitária. O índice γ é composto pelo escalonamento s, pelo deslocamento no tempo u e pela freqüência de modulação ξ, ou seja, $\gamma = (s, u, \xi)$, onde $\gamma \in \Gamma = R^+ \times R^2$ e Γ é conjunto de todos os parâmetros do dicionário.

No trabalho original de Mallat e Zhang [6], propõe-se um dicionário formado por átomos gerados a partir de uma janela gaussiana, que possuem a melhor localização tempo-freqüência. Entretanto, esses átomos possuem baixo grau de similaridade com relação aos fenômenos presentes em sinais de áudio, comprometendo, assim, a compacidade na representação de sinais de áudio, fator fundamental para o bom desempenho de um codificador. Em [2], verifica-se que o uso de um dicionário composto por janelas de Hanning moduladas nas freqüências da escala musical cromática possibilita boa resolução para análise tempo-freqüência. Além disso, seus elementos são mais parecidos com os padrões de sinais de áudio. Dessa maneira, opta-se por utilizá-lo com o propósito de obter representações mais compactas.

O dicionário de escala musical cromática é obtido, de acordo com a Eq. (3), a partir de uma janela de Hanning $h(t)$ definida por

$$h(t) = 1 + \cos(2\pi t). \tag{4}$$

Os parâmetros de γ, s (escala) e u (de deslocamento no tempo), são discretizados da seguinte maneira (ver Eq. (3))

$$s = 2^j, \quad j \in \{j_0, \ldots, \log_2(N)\}, \tag{5}$$

$$u = p\left(\frac{s}{2}\right), \quad p \in \left\{0, \ldots, \frac{2N}{s} - 1\right\}, \tag{6}$$

onde N é comprimento do sinal. É importante destacar que evita-se que parte de um átomo fique "fora" do suporte do sinal (no caso, para $u = 0$) realizando-se a extensão periódica do sinal a ser decomposto. A freqüência, por sua vez, é discretizada da seguinte forma

$$\xi = \xi_0 2^{\frac{k}{12}}, \quad k \in \{0, \ldots, k_{\max}\}, \tag{7}$$

onde k é o índice da freqüência correspondente a um semitom da escala musical cromática entre a freqüência mínima ξ_0 e a freqüência máxima associada a $k_{\max}$. Com o intuito de representar sinais provenientes de instrumentos de afinação imperfeita, adicionam-se freqüências laterais em torno de cada semitom, resultando em uma escala mais fina

$$\xi \in \left\{\xi_0 2^{\frac{k-\delta}{12}}, \xi_0 2^{\frac{k}{12}}, \xi_0 2^{\frac{k+\delta}{12}}\right\}, \quad k \in \{0, \ldots, k_{\max}\}, \tag{8}$$

O parâmetro $\delta \in]0, 1/3]$ ajusta a razão custo-benefício entre sensibilidade e seletividade. Utilizou-se $\delta = 1/3$ para manter constante a razão geométrica entre as freqüências.

Observe que não é possível analisar de maneira eficaz componentes de freqüências que possuem menos de 1 ciclo dentro de uma escala s. Portanto para impedir a inserção de um número maior que o necessário de elementos no dicionário, seleciona-se, para uma dada escala, ξ_0 de modo que as freqüências apresentem mais que 1 ciclo.

Uma vez definidos os átomos que iremos empregar, passemos à descrição do algoritmo empregado para decompor os sinais.

3 ALGORITMO DE DECOMPOSIÇÃO

O método de decomposição proposto, inspirado em [12] e [13], utiliza o algoritmo de "Matching Pursuit" (MP) com base no dicionário de escala musical cromática. O MP realiza aproximações sucessivas de um sinal $\mathbf{x}$ buscando em um dicionário redundante a melhor aproximação, a cada iteração.

O sinal de áudio é dividido em quadros $\mathbf{x}_k$ (k é o índice do quadro) de tamanho N, ponderados por uma janela suavizadora, em saltos de H amostras. Aqui, usou-se a janela de Hanning com $H = N/2$, garantindo a reconstrução exata do sinal com a sobreposição dos quadros. Antes de se janelar o sinal, é necessário inserir $H/2$ amostras antes de seu início e após o seu fim. A decomposição é realizada em cada um dos quadros separadamente.

Dado que o sinal é real e passa-banda, os elementos do dicionário são sinais analíticos (ver Eq. (3)), e como supomos que a janela compreenderá pelo menos 1 ciclo (como determinado na seção anterior) então o resíduo inicial é ajustado como

$$r^0_{\mathbf{x}_k}(n) = h_N(n)\, x_k(n) + j\, h_N(n)\, \mathcal{H}\{x_k(n)\}, \tag{9}$$

onde $h_N(n)$ corresponde à janela de Hanning de comprimento N e $\mathcal{H}\{\cdot\}$ representa a transformada de Hilbert. É importante ressaltar que a convergência na aproximação de um sinal analítico garante a convergência da aproximação do sinal real correspondente [13]. A cada iteração da decomposição, o MP busca o átomo $\mathbf{g}_{\gamma(i)}$ de maior correlação com o resíduo da iteração anterior, $\mathbf{r}^{i-1}_{\mathbf{x}_k}$. O átomo escolhido é subtraído do resíduo, gerando um novo resíduo

$$\mathbf{r}^i_{\mathbf{x}_k} = \mathbf{r}^{i-1}_{\mathbf{x}_k} - \alpha_i\, \mathbf{g}_{\gamma(i)}, \text{ onde } \alpha_i = \left\langle \mathbf{r}^{i-1}_{\mathbf{x}_k}, \mathbf{g}_{\gamma(i)} \right\rangle. \tag{10}$$

Este processo se repete até que o energia do resíduo seja suficientemente pequena ou que algum outro critério de parada seja atendido. Ao final, obtém-se a seguinte aproximação linear de $\mathbf{x}_k$:

$$\hat{\mathbf{x}}_k = \Re e \left\{ \sum_{i=1}^{M} \alpha_i \, \mathbf{g}_{\gamma(i)} \right\}, \tag{11}$$

onde $\mathbf{g}_{\gamma(i)}$ é o átomo complexo do dicionário associado ao índice de parâmetros γ_i, α_i corresponde ao coeficiente complexo e M é o número de termos na representação. A seqüência de pares $\{\alpha_i, \gamma(i)\}$, $i = 1, \ldots, M$, forma o que denomina-se de Livro de Estruturas [6].

3.1 CRITÉRIO DE PARADA

Deseja-se que a decomposição seja interrompida quando a aproximação do sinal for suficientemente boa, significando que os átomos escolhidos para a aproximação são coerentes, ou seja, altamente correlacionados com as características intrínsecas do sinal. Normalmente, utiliza-se a norma do resíduo ou um número máximo de iterações como critério de parada [6, 4, 2]. A norma do resíduo é uma medida do erro de aproximação, contudo com base apenas nela não é possível averiguar se o resíduo ainda é correlacionado significativamente com algum átomo do dicionário [9]. Para isso, propõe-se o uso da taxa de aproximação [6, 14], definida como

$$\lambda(i) = \frac{\left| \left\langle \mathbf{r}_{\mathbf{x}}^{i-1}, \mathbf{g}_{\gamma(i)} \right\rangle \right|}{\|\mathbf{r}_{\mathbf{x}}^{i-1}\|}, \tag{12}$$

que mede o quanto do resíduo $\mathbf{r}_{\mathbf{x}}^{i-1}$ é aproximado na iteração i. A taxa de aproximação possui um comportamento decrescente exponencial, como é demonstrado em [6]. A derivada da envoltória dessa exponencial diminui à medida que o número de iterações aumenta até que $\lambda(i)$ convirja para um valor médio estacionário. Quando a taxa de aproximação atinge esse valor, poder-se dizer que não há mais elementos do dicionário que sejam coerentes com o resíduo, portanto a decomposição pode ser interrompida. Esse critério apresenta um bom desempenho quando aplicado a sinais elétricos [7, 8, 9].

4 RESULTADOS

Nos experimentos, realizam-se a análise/síntese, utilizando o MP cromático, de notas de diferentes famílias de instrumentos: clarinete (A3), piano (A3), tímpano (E3) e violão (A3). Os sinais possuem 1 segundo de duração e são amostrados a uma taxa de 44,1 kHz, logo têm 44100 amostras cada. O sinal é dividido em quadros de $N = 1024$ amostras, o que corresponde a 0,023 segundo, com sobreposição de $H = 512$ amostras. Dessa forma, têm-se um total de 88 quadros. O dicionário foi gerado para valores do parâmetro de escala $s \in \{4, 8, \ldots, 1024\}$, resultando em 95420 elementos. Definiu-se a escala de freqüências cromáticas variando de 27,5 Hz (A0) a 15,6 kHz (B9). Esse limite superior corresponde à freqüência da escala cromática mais próxima da média geométrica dos extremos da última oitava representável com a taxa de amostragem de 44,1 kHz.

Primeiramente, avaliou-se o esquema de análise/síntese proposto através do PEAQ (*Perceptual Evaluation of Audio Quality)*) [11]. O PEAQ é um método de medição objetiva da qualidade de sinais de audio descrito pelo ITU (*International Telecommunications Union*), originalmente destinado à avaliação de codecs com perdas. Ele utiliza diversas medidas de base psicoacústica que, combinadas, resultam em uma medida de qualidade relativa entre duas instâncias de um sinal – um sinal de referência (original) e um sinal de teste (reconstruído) [11]. Essa medida de qualidade é classificada nas seguintes faixas:

- **-4 a -3**: muito perturbador;
- **-3 a -2**: perturbador;
- **-2 a -1**: pouco perturbador;
- **-1 a 0**: não perturbador.

As Figuras de 2 a 5 apresentam resultados da avaliação do desempenho do MP cromático através do PEAQ. Todas apresentam o mesmo padrão, segundo o qual o avaliador, naturalmente, premia as representações mais completas. O comportamento anômalo para poucos coeficientes se explica pelo fato de o PEAQ se destinar à avaliação de áudio de alta qualidade: sinais muito prejudicados geram notas pouco confiáveis.

Observa-se para o clarinete (Fig. 2), o número mínimo de coeficientes para que o sinal reconstruído fosse avaliado como não perturbador (PEAQ entre -1 e 0) foi 14178; no caso do piano (Fig. 3), foram necessários 9776 coeficientes; no do tímpano (Fig. 4), 6903; e no do violão (Fig. 5), 6818. Em ordem decrescente de compactação, teríamos violão, tímpano, piano e clarinete.

Em seguida, fizeram-se testes informais de audição a fim de verificar o número mínimo de coeficientes considerados necessários para representar fielmente as notas de cada instrumento. Usou-se como referência o percentual da norma inicial de cada quadro que determinava o resíduo de parada. Os resultados são apresentados na Tabela 1, juntamente com o percentual de energia representado. O número máximo de iterações do algoritmo de decomposição foi fixado em 400 por quadro.

Os testes subjetivos atribuíram o maior nível de compactação ao violão, seguido do piano, do tímpano e do clarinete. Observa-se que, em todos os casos, as quantidades de coeficientes encontradas foram menores que as sugeridas pelo PEAQ, o que numa primeira análise poderia indicar pequeno grau de exigência dos ouvintes. Mas além disso, somente o

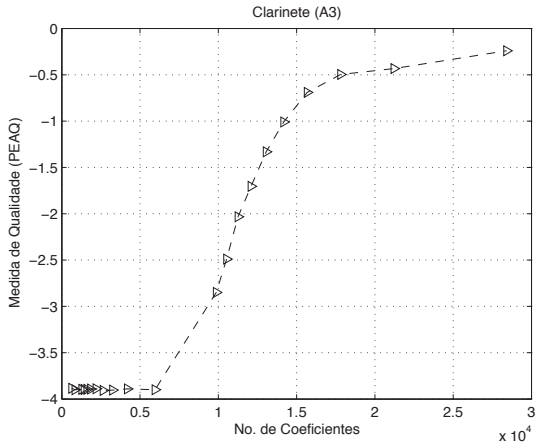


Figura 2: Avaliação de Qualidade através do PEAQ para diferentes números de coeficientes na representação do sinal de clarinete (A3).

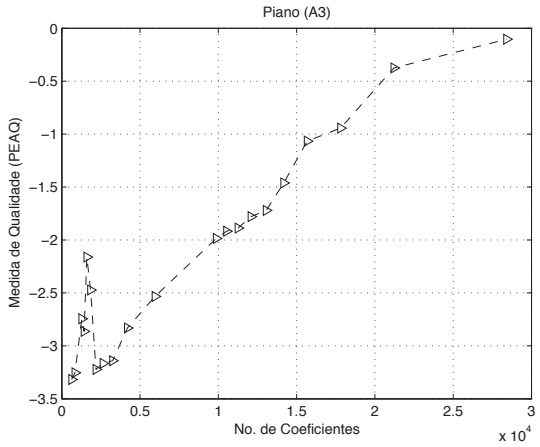


Figura 3: Avaliação de Qualidade através do PEAQ para diferentes números de coeficientes na representação do sinal de piano (A3).

Tabela 1: Resultados de testes informais de audição.

	Clarinete (A3)	Piano (A3)	Tímpano (E3)	Violão (A3)
Percentual da norma inicial como critério de parada (%)	1	2	1	4
Número mínimo de coeficientes	≈ 10000	≈ 3000	≈ 4500	≈ 800
Percentual de energia representado (%)	99,90	99,83	99,89	99,84

clarinete e o violão foram classificados na mesma ordem, quanto à compactação. A discrepância entre os resultados dos testes subjetivos e objetivos pode ser causada em parte pelo fato de o PEAQ ter sido treinado com base em codificadores de fundamentação psicoacústica. A adaptação dessa medida objetiva de qualidade para aplicação a sinais que sofreram outras formas de modificação é um tópico a ser investigado.

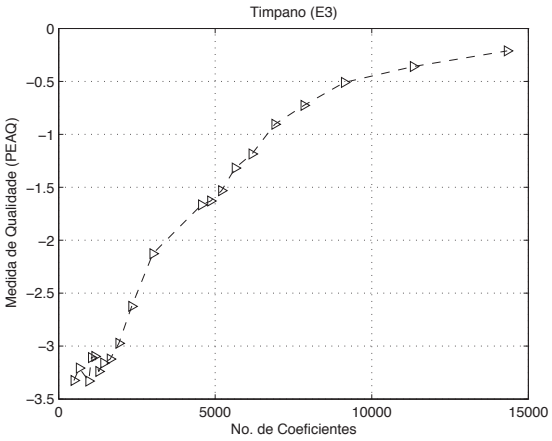


Figura 4: Avaliação de Qualidade através do PEAQ para diferentes números de coeficientes na representação do sinal de timpano (E3).

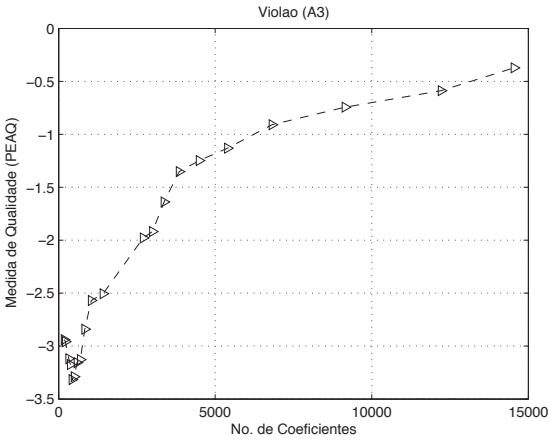


Figura 5: Avaliação de Qualidade através do PEAQ para diferentes números de coeficientes na representação do sinal de violão (A3).

As Figuras de 6 a 9 mostram as formas de onda originais das notas dos 4 instrumentos e o número de coeficientes usados, quadro a quadro, para reconstruir cada um deles no contexto da Tabela 1.

Nota-se que os trechos de baixa energia no início da Figura 6 (clarinete) e no final da Figura 9 (violão) precisaram de uma grande quantidade de coeficientes para serem representados. Isso sugere que nessas porções de sinal já resta pouca correlação entre o resíduo e os átomos do dicionário.

A Figura 10 ilustra o comportamento da norma do resíduo em função das iterações para dois quadros do sinal de clarinete: 4 e 50. Verifica-se o mesmo comportamento exponencial. Contudo, uma derivada reduzida da envoltória exponencial da norma indica que os átomos escolhidos do dicionário aproximam cada vez menos o resíduo. Isso explica que sejam utilizados mais átomos para o quadro 4, de baixa energia, que para o quadro

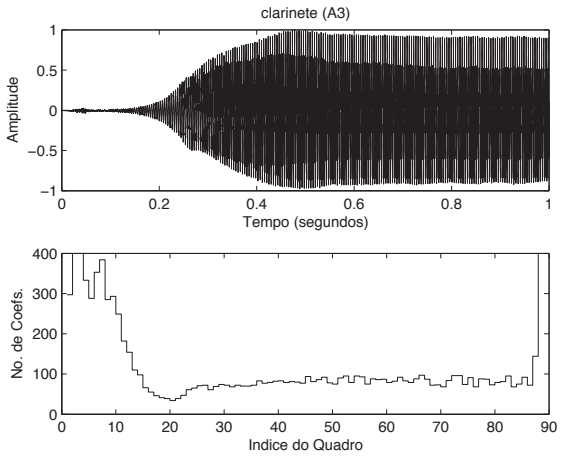


Figura 6: Sinal do clarinete (A3) e número de coeficientes na representação dos quadros.

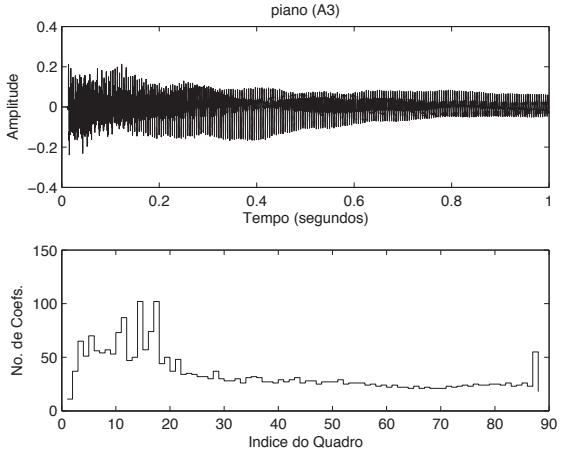


Figura 7: Sinal do piano (A3) e número de coeficientes na representação dos quadros.

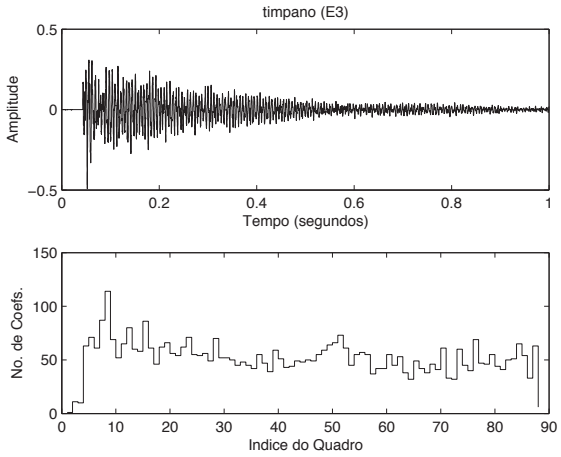


Figura 8: Sinal do tímpano (E3) e número de coeficientes na representação dos quadros.

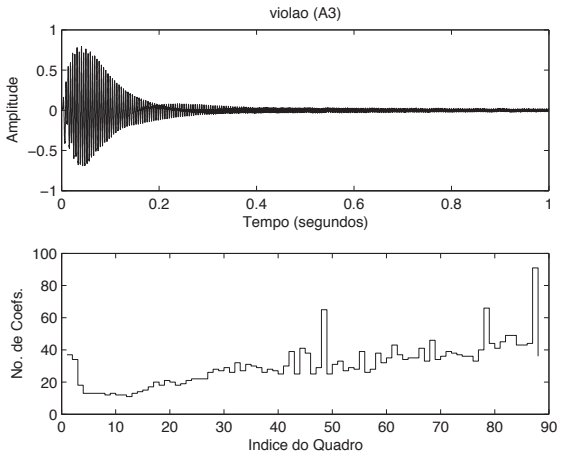


Figura 9: Sinal do violão (A3) e número de coeficientes na representação dos quadros.

50, de alta energia. Portanto, adotar um percentual da energia do resíduo inicial como energia do resíduo que decide a parada da decomposição não parece ser adequado.

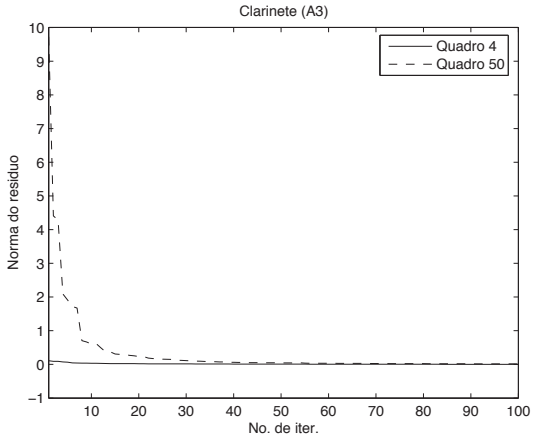


Figura 10: Comportamento da norma do resíduo nos quadros 4 e 50 do sinal de clarinete (A3)

.

A Figura 11 ilustra a taxa de aproximação (Eq. (12)) em função do número de iterações do MP. Observa-se que esta figura de mérito consegue traduzir a capacidade do dicionário de aproximar o resíduo a cada iteração. Essa capacidade diminui exponencialmente da mesma forma que a norma do resíduo. Quando a sua taxa de variação média se torna pequena é porque os elementos do dicionário passaram a ter baixa correlação com o resíduo, e, portanto, não é mais preciso prosseguir com a decomposição. Isso indica que pode ser interessante utilizar a taxa de aproximação como critério de parada para o MP cromático. Esse critério apresentou um bom desempenho quando aplicado a sinais elétricos [7, 8, 9], e é um tópico correntemente sob investigação.

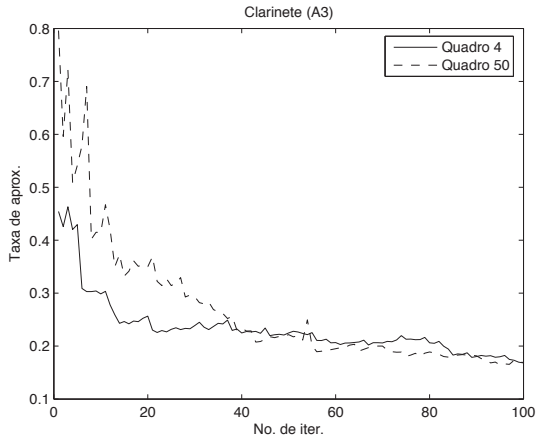


Figura 11: Taxa de aproximação dos quadros 4 e 50 do sinal de clarinete (A3)

5 CONCLUSÃO

Neste artigo, avaliaram-se potencialidades e deficiências do método de *Matching Pursuit* quando aplicado a áudio. Implementou-se um esquema de análise/síntese de sinais de áudio utilizando MP cromático. Seu desempenho para diversas famílias de sinais foi avaliado pelo *Perceptual Evaluation of Audio Quality* (PEAQ) em função do número de coeficientes empregados na representação e por testes informais de audição, com alguma discrepância, sugerindo a necessidade de se modificar o PEAQ para ampliar seu alcance. Por fim, verificou-se que a taxa de aproximação possui um bom potencial como critério de parada da decomposição.

Na continuação imediata deste trabalho, deverão ser realizados testes sistemáticos sobre um conjunto mais extenso de sinais, contemplando outros naipes e instrumentos, e com inclusão controlada de ruído. À medida que outros desenvolvimentos forem incorporados ao método apresentado, seu desempenho deve ser comparado ao de outros da literatura.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] J. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. on Information Theory*, vol. 50, no. 10, pp. 2231–2242, Out. 2004.

[2] O. Derrien, "Multi-scale frame-based analysis of audio signals for musical transcription using a dictionary of chromatic waveforms," in *IEEE ICASSP'06 - Int. Conf. on Acoustics, Speech, and Signal Processing*, Toulouse, França, Maio 2006, vol. 5, pp. 57–60.

[3] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Trans. on Signal Processing*, vol. 51, no. 1, pp. 101–111, Jan. 2003.

[4] M.M. Goodwin and M. Vetterli, "Matching pursuit and atomic signal model based on recursive filter banks," *IEEE Trans. on Signal Processing*, vol. 47, no. 7, pp. 1890–1902, Jul. 1999.

[5] L. Daudet, "Sparse and structured decompositions of signals with molecular matching pursuit," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1890–1902, Set. 2006.

[6] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dez. 1993.

[7] L. Lovisolo, E.A.B. da Silva, M.A.M. Rodrigues, and P.S.R. Diniz, "Efficient coherent representations of power systems signals using damped sinusoids," *IEEE Trans. on Signal Processing*, vol. 53, no. 10, pp. 3831–3846, Out. 2005.

[8] M.P. Tcheou, L. Lovisolo, E.A.B. da Silva, M.A.M. Rodrigues, and P.S.R. Diniz, "Optimum rate-distortion dictionary selection for compression of atomic decompositions of electric disturbance signals," *IEEE Signal Processing Letters*, vol. 14, no. 2, pp. 81–84, Fev. 2007.

[9] L. Lovisolo, M.P. Tcheou, E.A.B. da Silva, M.A.M. Rodrigues, and P.S.R. Diniz, "Modeling of electric disturbance signals using damped sinusoids via atomic decompositions and its applications," *EURASIP Journal on Advances in Signal Processing,*, vol. 2007, pp. Article ID 29507, 15 pp., 2007, doi:10.1155/2007/29507.

[10] S. Mallat, *A wavelet tour of signal processing*, Academic Press, 1998.

[11] P. Kabal, "An examination and interpretation of ITU-R BS.1387: perceptual evaluation of audio quality," Tech. Rep., TSP Lab Tech. Report, Dept. Elec. and Comp. Eng., McGill Univ., vers. 2, Montreal, Canadá, Dez. 2003.

[12] T.V. Verma and T.H. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *IEEE ICASSP'99 - Int. Conf. on Acoustics, Speech, and Signal Processing*, Phoenix, EUA, Mar. 1999, vol. 2, pp. 981–984.

[13] M.G. Christensen and S. van de Par, "Efficient parametric coding of transients," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1340–1351, Jul. 2006.

[14] G. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency approximations with matching pursuits," Tech. Rep., TR1994-657 Tech. Report, Computer Science Dept., New York Univ., Nova Iorque, EUA, Mar. 1994.

Sessão 3

Áudio espacial, sonorização multicanal e acústica de ambientes

(Spatial audio, multichannel sound and environmental acoustics)

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Modelagem Eficiente de HRTFs por Parametrização Espacial

Arnaldo S. Gunzi,[1] Fábio P. Freeland,[1] Luiz W. P. Biscainho[2] e Paulo S. R. Diniz[2]

[1] PEE/COPPE–UFRJ

[2] PEE/COPPE & DEL/Poli–UFRJ
Caixa Postal 68504, Rio de Janeiro, RJ, 21941-972, Brasil

asgunzi, freeland, wagner, diniz@lps.ufrj.br

RESUMO

A geração binaural de som tridimensional usando HRTFs (*Head-Related Transfer Functions*) requer que essas funções sejam medidas para algumas posições e interpoladas nas demais. Recentemente, generalizou-se para modelar HRTFs uma estrutura de atraso digital continuamente variável. O presente artigo reformula a obtenção dos parâmetros da estrutura generalizada, reduzindo memória e tempo de convergência gastos na otimização. Avalia-se o desempenho do método proposto contra o do precedente, e em função da ordem da estrutura e do tamanho das sub-regiões modeladas.

0 INTRODUÇÃO

O áudio 3D encontra aplicação nos contextos em que se queira aumentar a sensação de realidade do usuário, como em jogos eletrônicos, simuladores e aparelhos de *home-theater*.

A geração de áudio 3D envolve a projeção de um sinal monofônico numa determinada posição do espaço, ali localizando virtualmente uma fonte sonora. A isso se chama panoramização.

Dentre os diversos métodos de panoramização existentes, há os que produzem a ilusão tridimensional através de HRTFs (*Head-Related Transfer Functions*) ou respectivas HRIRs (*Head-Related Impulse Responses*). As HRTFs na freqüência e as HRIRs no tempo descrevem as alterações sofridas pelo som desde a posição da fonte sonora até cada orelha do ouvinte, em ambiente anecóico. Um sinal monofônico por elas processado parecerá provir de uma fonte virtual localizada naquela posição. Tipicamente, um conjunto de HRIRs é obtido em câmara anecóica: emitem-se sinais de teste por uma caixa acústica em diversas posições sobre uma casca esférica; e grava-se o sinal que atinge os microfones implantados em cada orelha de um manequim no centro da esfera. Um conjunto de medidas amplamente referenciado na literatura decorre do trabalho de Gardner e Martin [1]. Deve-se notar que as HRTFs são in-

dividuais de cada ouvinte, dependendo da conformação de seu tronco, sua cabeça e sua pina. O uso de medidas genéricas é bastante aproximativo.

Por não ser possível fazer as medições para todas as posições angulares, é necessário interpolar as posições que não estão no conjunto de medidas para se preservar a continuidade do som para fontes em movimento. A forma mais simples de se realizar essa interpolação é ponderar as HRTFs medidas nas posições mais próximas ao ponto em que se deseja posicionar a fonte virtual [2, 3]. Outras alternativas favorecem o caso de múltiplas fontes sonoras [4, 5].

Em [6], foi proposto um método de parametrização direta das HRIRs pelas variáveis angulares de posição. Esse método, doravante denominado estrutura Farrow, generaliza para duas variáveis a estrutura polinomial de atraso digital continuamente variável (CVDD) desenvolvida por Farrow [7]. Com aquele método, a modelagem conjunta das HRTFs de toda a esfera pode elevar excessivamente o custo computacional da otimização dos coeficientes.

A principal finalidade do presente trabalho é reformular a otimização da estrutura Farrow generalizada em busca de algoritmos eficientes que reduzam seus requisitos de memória e sua complexidade computacional. A Seção 1 revisa a generalização da estrutura Farrow. Na Seção 2, é proposta a formulação alternativa do problema de otimização da estrutura. Na Seção 3, avalia-se sob diversas configurações o desempenho da interpolação realizada com a estrutura Farrow resultate da otimização reformulada. Finalmente, as conclusões são apresentadas na Seção 4.

1 APROXIMAÇÃO DE HRTFS PELA ESTRUTURA FARROW

1.1 Estrutura Farrow Original

A estrutura de atraso digital continuamente variável (CVDD) [7] é capaz de simular um atraso fracionário τ (i.e., não múltiplo inteiro do período de amostragem T_s) em função de um único parâmetro $-\frac{1}{2} < \alpha \leq \frac{1}{2}$. A resposta na freqüência correspondente a esse sistema é

$$G(w, \alpha) = e^{-jw\tau}, \tag{1}$$

onde

$$\tau = T_s(m + \alpha), \quad m \geq 1, \text{ inteiro.} \tag{2}$$

A resposta na freqüência de um filtro com resposta ao impulso de comprimento finito (FIR) de ordem N pode ser escrita na forma

$$F(w) = \sum_{n=0}^{N} C_n e^{-jnwT_s}, \tag{3}$$

onde C_n é o n-ésimo peso do filtro. Fazendo cada coeficiente C_n um polinômio de ordem P na variável α, tem-se

$$C_n(\alpha) = \sum_{p=0}^{P} C_n^p \alpha^p. \tag{4}$$

Das Equações (3) e (4) é possível reescrever a resposta na freqüência do filtro como

$$F(w, \alpha) = \sum_{p=0}^{P} \alpha^p \sum_{n=0}^{N} C_n^p(\alpha) e^{-jnwT_s}. \tag{5}$$

Fazendo-se $\Omega = \omega T_s$, os coeficientes $C_n^p(\alpha)$ podem ser obtidos minimizando-se a função de custo

$$J = \int_{\Omega=-\pi}^{\pi} \int_{\alpha=-\frac{1}{2}}^{\frac{1}{2}} \left| F\left(\frac{\Omega}{T}, \alpha\right) - G\left(\frac{\Omega}{T}, \alpha\right) \right|^2. \tag{6}$$

A Figura 1 mostra a estrutura resultante para o CVDD.

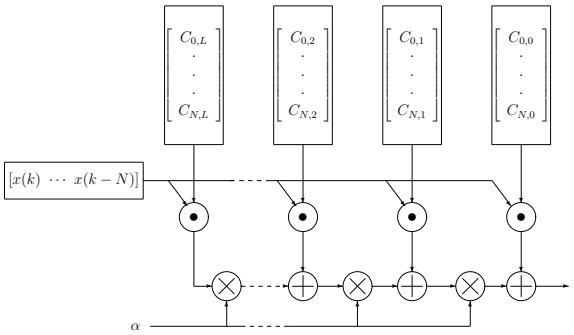


Figura 1: Diagrama em blocos da Estrutura Farrow Original.

1.2 Estrutura Farrow Generalizada

A generalização proposta em [6] para a estrutura Farrow consiste na aproximação de cada HRTF por uma função de transferência dependente de duas das variáveis esféricas de posição, θ (azimute) e ϕ (elevação). A terceira variável, relativa ao raio r, é considerada fixa, da mesma forma que foram medidas as HRIRs [1] utilizadas nesse trabalho.

É importante notar que, para se obter os coeficientes da estrutura, devem-se utilizar versões de fase mínima das HRIRs, sendo os excessos de fase tratados em separado [6].

Para se obter a nova formulação da estrutura Farrow, considera-se, inicialmente, que uma HRTF de ordem N é definida por

$$G(e^{j\Omega}, \theta, \phi) = \sum_{n=0}^{N} h_n(\theta, \phi) e^{-j\Omega n}, \tag{7}$$

onde $h_n(\theta, \phi)$ é o n-ésimo peso da HRIR da posição (θ, ϕ).

A estrutura Farrow generalizada aproxima cada peso $h_n(\theta, \phi)$ por um polinômio de ordem L em θ e ordem M em ϕ, como em

$$\hat{h}_n(\theta, \phi) \approx \sum_{l=0}^{L} \sum_{m=0}^{M} \theta^l \phi^m c_n^{l,m}, \tag{8}$$

onde $c_n^{l,m}$ são os coeficientes a otimizar. A resposta na freqüência associada é dada por

$$F(e^{j\Omega},\theta,\phi) = \sum_{n=0}^{N}\sum_{l=0}^{L}\sum_{m=0}^{M} \theta^l \phi^m c_n^{l,m} e^{-j\Omega n}. \quad (9)$$

A Figura 2 mostra um diagrama em blocos da estrutura resultante apenas para o canal esquerdo. A saída para o canal direito pode ser obtida simplesmente aplicando $-\theta$ aos resultados dos produtos internos (veja a Figura 1), aproveitando a simetria com relação a θ.

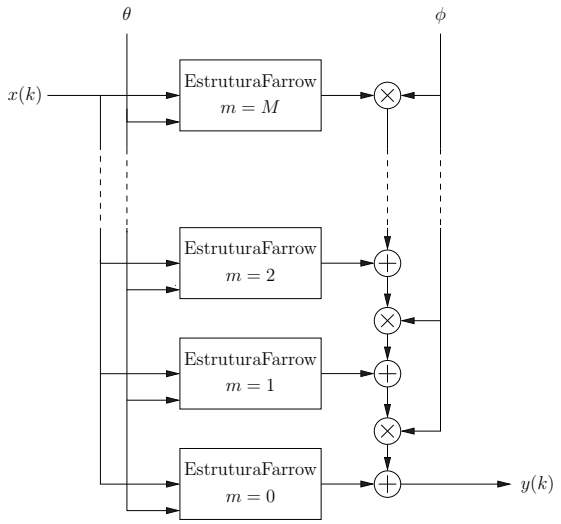


Figura 2: Diagrama em blocos da Estrutura Farrow Generalizada para 2 variáveis espaciais.

Para se obter os coeficientes $c_n^{l,m}$, pode-se minimizar o erro quadrático da aproximação, dado pela função de custo

$$J = \sum_{\Omega}\sum_{\theta}\sum_{\phi} \left| F(e^{j\Omega},\theta,\phi) - G(e^{j\Omega},\theta,\phi) \right|^2, \quad (10)$$

sendo $F(e^{j\Omega},\theta,\phi)$ e $G(e^{j\Omega},\theta,\phi)$ dadas pelas Equações (7) e (9), respectivamente.

1.2.1 Otimização Iterativa no Domínio da Freqüência

O método originalmente utilizado [6] para obtenção dos coeficientes $c_n^{l,m}$ realiza a otimização iterativa por um método *quasi*-Newton.

Considerando que haja HRIRs de ordem N medidas para A posições (θ,ϕ) e que se adotem ordens L e M para cada polinômio (ver Equação (8)), a otimização dos $B = (N+1)(L+1)(M+1)$ coeficientes envolve o cálculo de um vetor-gradiente com B elementos e uma matriz hessiana com B^2 elementos. Para o conjunto de HRIRs [1] utilizado neste trabalho, $A = 726$ e $N = 127$. Se $L = M = 6$, por exemplo, o problema consiste em calcular conjuntamente 6252 coeficientes; a hessiana tem cerca de 40×10^6 elementos e um cálculo de J envolve mais que 600×10^6 de multiplicações. Isso dá uma idéia dos custos computacional e de memória extremamente elevados inerentes a essa formulação, tornando-a muito suscetível a problemas numéricos, além de resultar em tempos de processamento exagerados.

Nessas condições, não foi possível utilizar estruturas de ordens maiores que $L = 6$ e $M = 6$; e ainda assim, foi preciso modelar a esfera em sub-regiões convenientemente ligadas. Isso levou à busca de alternativas, descritas na próxima seção.

2 REFORMULAÇÃO DA OTIMIZAÇÃO

É possível reformular a função a otimizar, transformando o problema original de ordem elevada em diversos problemas de ordem menor.

Forçando que a ordem das HRIRs estimadas seja igual à ordem N das HRIRs medidas, pode-se reescrever a Equação (10) combinada com as Equações (7) e (1) na forma

$$J = \sum_{\Omega}\sum_{\theta}\sum_{\phi} \Big| \sum_{n} \Big(\sum_{l}\sum_{m} \theta^l \phi^m c_n^{l,m} - \\ - h_n(\theta,\phi) \Big) e^{-j\Omega n} \Big|^2. \quad (11)$$

Aplicando-se a definição da DFT e o Teorema de Parseval $\sum_n |x(n)|^2 = \frac{1}{N+1}\sum_{\Omega}|X(\Omega)|^2$, a Equação (11) torna-se

$$J = (N+1)\sum_{n}\sum_{\theta}\sum_{\phi} \Big(\sum_{l}\sum_{m} \theta^l \phi^m c_n^{l,m} - \\ - h_n(\theta,\phi) \Big)^2. \quad (12)$$

Assim, para cada n, tem-se uma parcela quadrática da função J, conforme

$$J = (N+1)(J_0 + J_1 + ... + J_N), \quad (13)$$

onde

$$J_n = \sum_{\theta}\sum_{\phi} \Big(\sum_{l}\sum_{m} \theta^l \phi^m c_n^{l,m} - h_n(\theta,\phi) \Big)^2. \quad (14)$$

A minimização de cada parcela quadrática leva à minimização da função custo global.

Cada um dos $(N+1)$ sub-problemas envolve, agora, o cálculo de $B' = (L+1)(M+1)$ coeficientes. Para $A = 726$, $N = 127$ e $L = M = 6$, por exemplo, tem-se 49 coeficientes a otimizar e 2041 elementos na hessiana. Dessa forma, os requisitos de memória são bastante reduzidos e espera-se que a redução de ordem melhore bastante o condicionamento numérico do problema.

2.1 Otimização Direta no Domínio do Tempo

A redução da complexidade do problema inicial da Equação 10 ao ser reescrito nos sub-problemas de

minimização da Equação (14) já permite que se busque sua solução direta por inversão de matriz.

Fixando-se um dado n, cada elemento do gradiente da função de custo da Equação (14) em relação aos coeficientes é dado por

$$\frac{\partial J_n}{\partial c_n^{l_0,m_0}} = 2\sum_l \sum_m \Big(\sum_\theta \sum_\phi \theta^{l_0}\phi^{m_0}\theta^l\phi^m c_n^{l,m} \Big) - \\ -2\sum_\theta \sum_\phi \theta^{l_0}\phi^{m_0} h_n(\theta,\phi). \quad (15)$$

Igualando-se o gradiente a zero, tem-se

$$\sum_l \sum_m \Big(\sum_\theta \sum_\phi \theta^{l_0}\phi^{m_0}\theta^l\phi^m c_n^{l,m} \Big) = \\ = \sum_\theta \sum_\phi \theta^{l_0}\phi^{m_0} h_n(\theta,\phi). \quad (16)$$

Fazendo-se $0 \leq l_0 \leq L$ e $0 \leq m_0 \leq M$, chega-se a $(L+1)(M+1)$ equações em $(L+1)(M+1)$ incógnitas $c_n^{l,m}$. Na forma matricial, pode-se reescrever o problema como $\mathbf{Xc} = \mathbf{p}$, onde os elementos da matriz $\mathbf{X}$ são

$$x_{i,j} = \sum_\theta \sum_\phi \theta^{v(i)}\phi^{\mu(i)}\theta^{v(j)}\phi^{\mu(j)}, \quad (17)$$

e os dos vetores $\mathbf{c}$ e $\mathbf{p}$ são, respectivamente,

$$c_{i,n} = c_n^{v(i),\mu(i)}, \quad (18)$$

$$p_{i,n} = \sum_\theta \sum_\phi \theta^{v(i)}\phi^{\mu(i)} h_n(\theta,\phi), \quad (19)$$

com $0 \leq i,j \leq (L+1)(M+1)$ e tendo-se definido $v(i,j)$ como o quociente e $\mu(i,j)$ como o resto da divisão inteira $\frac{i,j}{L+1}$.

Dessa forma, para encontrar os coeficientes resta resolver

$$\mathbf{c}_n = \mathbf{X}^{-1}\mathbf{p}_n. \quad (20)$$

A matriz $\mathbf{X}$ de dimensão $(L+1)(M+1)$ é simétrica, e sua inversa pode ser obtida por métodos robustos adequados.

A otimização de todos os coeficientes $c_n^{l,m}$ da estrutura requer, então, que se resolva o problema de baixa ordem acima descrito $N+1$ vezes, variando-se $0 \leq n \leq N$. Mas a matriz $\mathbf{X}$, peça-chave da solução, traz em si outro fator importante para a simplicidade dessa reformulação: ela independe de n e, portanto, só precisa ser invertida uma vez; as soluções para cada n são obtidas trocando-se apenas o vetor $\mathbf{p}$.

O procedimento aqui apresentado permitiu a obtenção dos coeficientes da estrutura Farrow generalizada para a esfera toda, e para ordens maiores do que $L = 6$ e $M = 6$.

3 AVALIAÇÃO DE DESEMPENHO

Nesta seção serão confrontados os resultados das estruturas

- de ordens $L = 5$ e $M = 5$ cujos coeficientes foram otimizados na freqüência ou no tempo para uma fração da esfera;
- de ordens $L = 5$ e $M = 5$ cujos coeficientes foram otimizados no tempo para a esfera toda ou uma fração da esfera;
- de ordens $L = 5$ e $M = 5$ ou $L = 8$ e $M = 8$ cujos coeficientes foram otimizados no tempo para uma fração da esfera.

A otimização na freqüência é a da Subseção 1.2.1, e a otimização no tempo é a da Subseção 2.1. A fração da esfera é definida pelas faixas $-45 \leq \theta \leq 45$ graus e $-40 \leq \phi \leq 40$ graus.

Para ilustrar a questão da complexidade computacional, a Tabela 1 compara os tempos consumidos pela otimização medidos em cada caso, em relação ao de ordem $L = 5$, $M = 5$ com otimização no tempo na esfera toda. A redução do tempo de convergência da otimização de mesma ordem para uma região limitada da esfera foi aproximadamente proporcional à redução do número de HRTFs consideradas. A partir deste caso, o aumento da ordem resultou num aumento comparável à razão das ordens quadráticas. Por fim, a otimização na freqüência foi ordens de grandeza mais demorada.

Tabela 1: Tempo de processamento da otimização.

Método Otimização	Área de Otimização	Ordem Polinomial	Tempo Processamento
Freqüência	Região	$L = 5, M = 5$	$>> T$
Tempo	Região	$L = 5, M = 5$	$\approx T/2$
Tempo	Esfera	$L = 5, M = 5$	T
Tempo	Região	$L = 8, M = 8$	$\approx 2T$

3.1 Comparação entre as respostas na freqüência das soluções

Foram comparados os resultados obtidos nos quatro casos delineados acima. Como referência, utilizou-se o conjunto de funções medidas de [1]. Vale observar que a estrutura Farrow generalizada recalcula as funções nas posições com HRTFs medidas.

Definiram-se duas trajetórias que estivessem incluídas na região da esfera delimitada por $-45 \leq \theta \leq 45$ graus, $-40 \leq \phi \leq 40$ graus. Foram calculadas as HRTFs aproximadas com passos de 1 grau no azimute e na elevação ao longo dessas trajetórias, para todos os casos. No caso das funções medidas, efetuou-se sua interpolação bilinear [3] como referência.

As Figuras 3 e 4 mostram as respostas na freqüência para as diversas soluções ao longo dos caminhos (θ, ϕ)

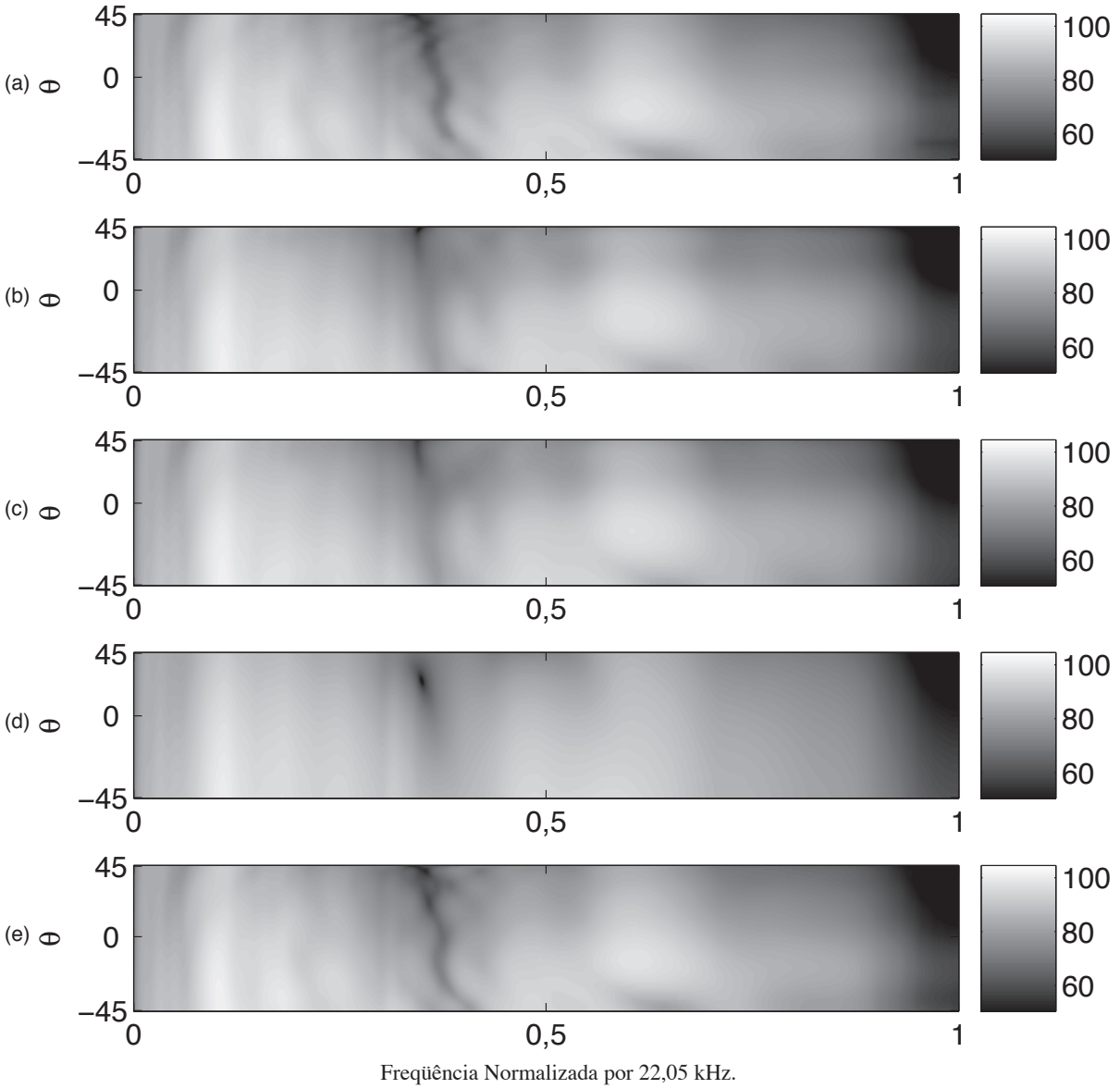


Figura 3: Mapas para o caminho horizontal de (-45,0) a (45,0) graus. (a) HRTFs originais, interpolação bilinear. Estrutura Farrow: (b) L=5, M=5, solução no tempo numa fração da esfera; (c) L=5, M=5, solução na freqüência numa fração da esfera; (d) L=5, M=5, solução no tempo na esfera toda; (e) L=8, M=8, solução no tempo numa fração da esfera.

horizontal de $(-45, 0)$ a $(45, 0)$ graus e diagonal de $(-40, -40)$ a $(40, 40)$ graus, respectivamente. Optou-se por trajetórias que pudessem ser parametrizadas por uma só variável angular, permitindo a visualização da amplitude das HRTFs em dB (níveis de cinza) em todo o espectro nas diversas posições do percurso num mesmo gráfico.

Primeiramente, pode-se destacar que a continuidade do comportamento das HRTFs é preservada em qualquer solução. Em seguida, a comparação entre os gráficos (b) e (c) confirma a equivalência das otimizações na freqüência e no tempo, esta apresentando a vantagem da menor complexidade de cálculo.

O confronto entre os gráficos (b) e (d) aponta para o fato de que, para uma mesma ordem, comparada à referência (a), a interpolação na esfera toda tende a ser mais suavizada que quando realizada numa região limitada. Nenhuma surpresa aqui, uma vez que a descrição global da esfera exige mais do mesmo número de coeficientes polinomiais. Assim, é preferível realizar a otimização por regiões, apenas com o cuidado adicional de preservar a continuidade nas transições entre regiões adjacentes [6], por exemplo impondo restrições às fronteiras.

Por último, examinando-se os gráficos (b) e (e), fica claro que com o aumento da ordem a estrutura Farrow torna-se capaz de delinear muito melhor os acidentes (em particular, os vales) das HRTFs em (a), em que pe-

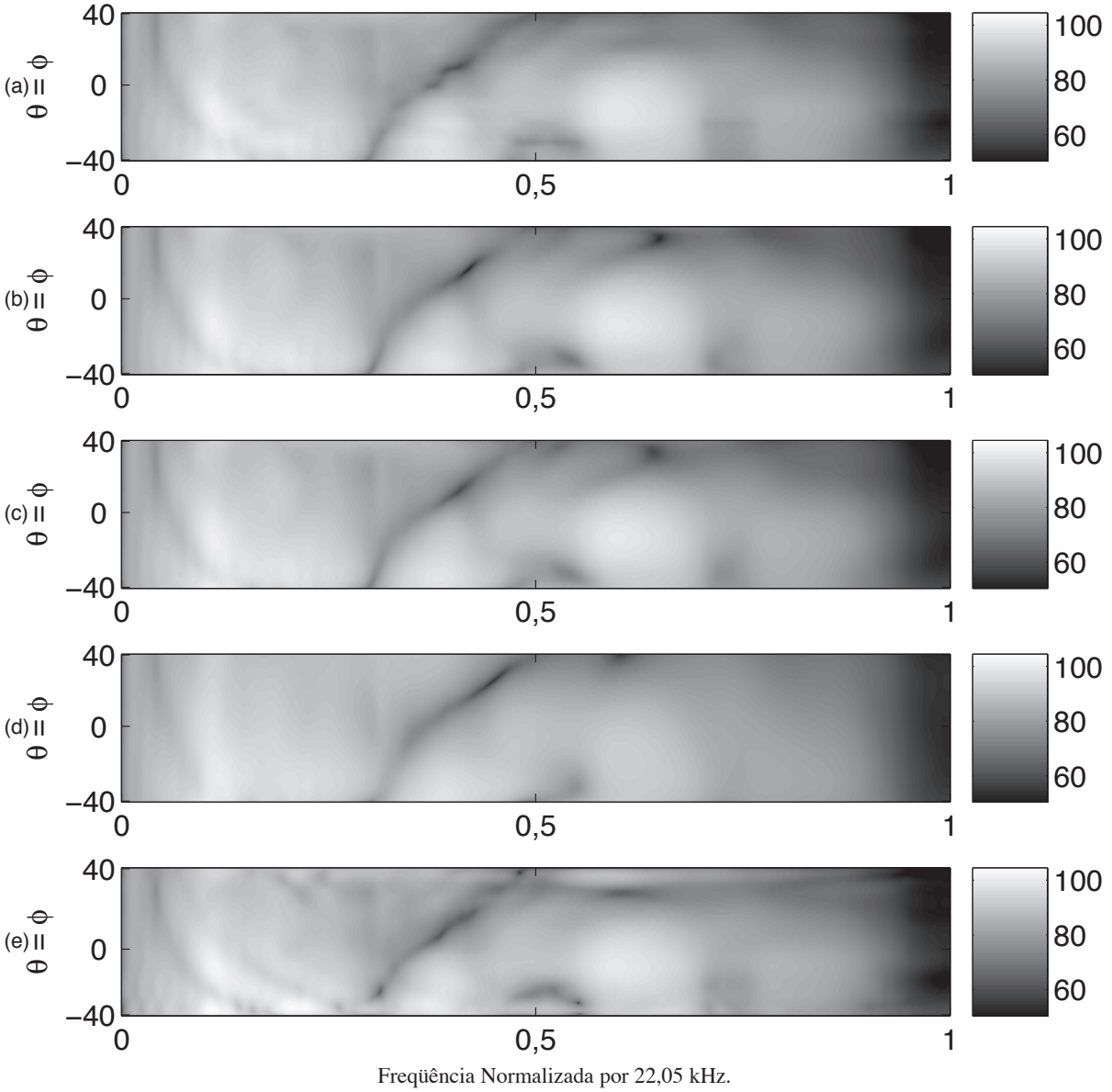


Figura 4: Mapas para o caminho diagonal de (-40,-40) a (40,40) graus. (a) HRTFs originais, interpolação bilinear. Estrutura Farrow: (b) L=5, M=5, solução no tempo numa fração da esfera; (c) L=5, M=5, solução na freqüência numa fração da esfera; (d) L=5, M=5, solução no tempo na esfera toda; (e) L=8, M=8, solução no tempo numa fração da esfera.

sem algumas diferenças de normalização. Nesse caso, contudo, é preciso ter em mente que o aumento da ordem eleva quadraticamente a complexidade do cálculo, podendo torná-lo sensível a erros numéricos a ponto de inutilizar a solução.

A fim de dar uma visão mais detalhada da aproximação realizada sobre as HRTFs, a Figura 5 mostra a resposta na freqüência das diversas soluções obtidas para as seguintes coordenadas do espaço: $(0, 0)$, $(0, -40)$ e $(45, 0)$ graus.

Confirmando as observações anteriores, a solução por região com ordens $L = 8$ e $M = 8$ foi a que melhor acompanhou as HRTFs originais; as soluções no tempo e na freqüência foram muito próximas; e a solução global foi a que mais se afastou do esperado.

4 CONCLUSÕES

A estrutura Farrow generalizada [6] aproxima HRIRs ao longo da esfera de referência com pesos polinomiais em função das variáveis angulares de posição. Nesse trabalho modificou-se o método de otimização desses polinômios, reformulando-o no tempo pelo Teorema de Parseval. Dessa forma, tornou-se possível a aproximação conjunta das HRIRs ao redor de toda a esfera de referência, bem como a utilização de ordens polinomiais mais elevadas que no trabalho original.

Simulações mostraram que as duas formulações são equivalentes, e que um pequeno aumento na ordem

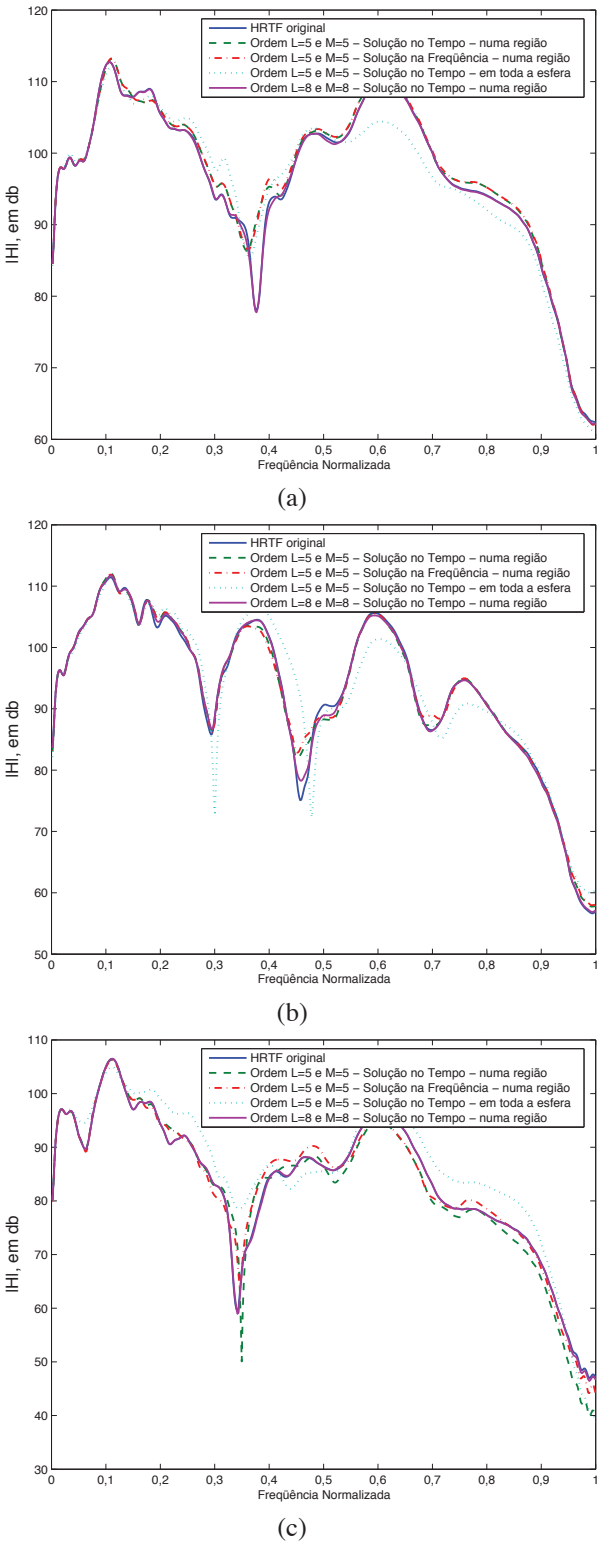


Figura 5: Comparação entre as respostas na freqüência para: (a) azimute 0 e elevação 0; (b) azimute 0 e elevação -40 graus; (c) azimute 45 graus e elevação 0.

pode melhorar bastante a aproximação; contudo, ainda é preferível realizar a aproximação em sub-regiões da esfera. Naturalmente, é necessária a realização de testes subjetivos sistemáticos do desempenho do sistema proposto, para fontes virtuais em movimento.

A continuação do trabalho inclui buscar elevar a ordem polinomial pelo uso de métodos mais robustos de cálculo; elaborar técnicas de baixa complexidade para garantir a transição suave entre regiões modeladas separadamente; e buscar novas formulações para a função de custo.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] B. Gardner and K. Martin, "HRTF measurement of a KEMAR dummy-head microphone," Tech. Rep. 280, MIT Media Lab Perceptual Computing, May 1994.

[2] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *Journal of the Audio Engineering Society*, vol. 47, no. 9, pp. 675–705, September 1999.

[3] F. P. Freeland, L. W. P. Biscainho, and P. S. R. Diniz, "Interpositional transfer function for 3d-sound generation," *Journal of the Audio Engineering Society*, vol. 52, no. 9, pp. 915–930, September 2004.

[4] J. Chen, B. D. Van Veen, and K. E. Hecox, "A spatial feature extraction and regularization model for virtual auditory display," in *Proc. of the ICASSP'93 - International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, USA, April 1993, IEEE, pp. 129–132.

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

[6] F. P. Freeland, L. W. P. Biscainho, and P. S. R. Diniz, "HRTF interpolation through direct angular parameterization," in *Proceedings of the ISCAS'07 - International Symposium on Circuits and Systems*, New Orleans, USA, May 2007, IEEE, To appear.

[7] C. W. Farrow, "A continuously variable digital delay element," in *Proc. of the ISCAS'88 - International Symposium on Circuits and Systems*, Helsinki, Finland, 1988, IEEE, pp. 2641–2643.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Desenvolvimento experimental de amplificadores multicanal para auralização

Renan Santos Vital[1] e Regis Rossi A. Faria[1]
[1]LSI – Laboratório de Sistemas Integráveis, Escola Politécnica da USP
Av. Prof. Luciano Gualberto, 158 tv.3, São Paulo, SP, 05508-900, Brasil
{renvit, regis}@lsi.usp.br

RESUMO
Neste artigo reportamos a experiência de desenvolvimento de um amplificador de oito canais, desenhado para a auralização imersiva de cenas sonoras 3D. Atualmente há equipamentos com seis canais e um incipiente mercado com oito canais, para os sistemas de decodificação envolvente 5.1 e 7.1, mas existem poucos amplificadores para uso profissional com número de canais superior a quatro em um único chassi. Isto mostra uma deficiência em técnicas e modelos de desenvolvimento de amplificadores multicanais para auralização multi-propósito. Apresentamos um panorama atual do mercado de amplificadores multicanais e resultados prospectivos.

0 INTRODUÇÃO

Este trabalho foi proposto para o projeto AUDIENCE [1][2] a partir de uma demanda por um amplificador multicanal com oito canais mais saída para *subwoofer* em um único chassi, para acionar uma matriz de alto-falantes e auralização de campos sonoros 3D.

O projeto tem objetivos que incluem desde a implementação de reprodutores para os tradicionais formatos e configurações *surround* (como 5.1 e 7.1 [5]) até o desenvolvimento e disponibilização de formatos mais sofisticados para a geração e reprodução de áudio 3D, ainda restritos às aplicações de alta-fidelidade e aos ambientes de pesquisa.

Para estes formatos não somente a reverberância e o envelopamento envolvente são importantes, mas também a localização dos objetos sonoros distribuídos numa cena tridimensional (percepção da diretividade do som) bem como a síntese do ambiente acústico. Nesta categoria incluímos o Ambisonics [3] e o *Wave-Field Synthesis* (WFS) [4], que são técnicas de produção de campo sonoro 3D mais precisas e realistas que o 5.1 e que exigem grande número de canais para uma projeção de alta fidelidade, utilizando para isso uma densa matriz de alto-falantes que precisa ser excitada por sistemas multiamplificadores.

No projeto AUDIENCE necessita-se de configurações de caixas acústicas e equipamentos amplificadores que diferem das existentes atualmente no mercado. Em sua maioria são comercializados sistemas de 2 canais ou são vendidos em forma de *home-theaters*, em pacotes combinados de amplificador + cabeamento + caixas, com limitações para se explorar novos conceitos de áudio espacial.

Torna-se inviável a utilização de por exemplo uma matriz de 256 alto-falantes, como é possível implementar com WFS, utilizando-se amplificadores de dois ou quatro canais cada, considerando-se os modelos profissionais existentes hoje, e que possuem uma boa qualidade de

reprodução sonora. Isso revela a necessidade de utilizarmos amplificadores que controlem uma quantidade maior de canais utilizando menos espaço.

1 ESTADO DA ARTE EM AMPLIFICAÇÃO MULTICANAL

O maior mercado de amplificadores multicanal atual é aquele explorado pelos desenvolvedores de *home-theaters* e receptores de áudio e vídeo que utilizam codificação multicanal proprietária, como os sistemas Dolby™[1] [7] e DTS™[2] [9] para som envolvente (*surround*) 5.1 ou 7.1 [13]. Tais equipamentos tipicamente possuem incorporados conjuntos amplificadores de potência com saídas diretas para 6 ou 8 caixas de som.

O mercado profissional de sistemas multicanais atende basicamente a (1) sonorização de cinemas, que busca soluções multiamplificadores compativeis com 6 ou 8 canais fontes [8], e (2) sonorização de grandes ambientes (*shows*, casas de teatro, etc.). Freqüentemente a solução mais comum é a utilização escalável de muitos módulos amplificadores estéreo multiplicando os sinais fontes, o que requer grande espaço físico e potência elétrica compatível. Por isso, esforços estão sendo empregados pela comunidade acadêmica e pelo setor produtivo afim de encontrar soluções eficientes para a construção de sistemas amplificadores multicanais.

Os fabricantes de equipamentos de áudio estão atualmente construindo aparelhos com 4, 6 ou 8 canais. Estes aparelhos estão sendo lançados com a promessa de atender às exigências de um mercado atento a características de qualidade e fidelidade, seja na linha profissional ou na linha doméstica. Estes aparelhos recebem um cuidado maior na sua confecção, o que naturalmente resulta em maior qualidade, mas este aspecto esbarra em limitações físicas e apelo do consumidor por uma alta relação qualidade/preço.

Os amplificadores devem ter características otimizadas, como: distorção harmônica total (THD) mínima, para uma saída de som com suas componentes harmônicas em relação a harmônica fundamental tendo um desvio mínimo; alta relação sinal ruído (SNR); ampla e plana resposta em freqüência, contemplando principalmente as freqüências audíveis; fator de amortecimento (*damping factor*) alto para baixas freqüencias e mínima distorção (mudança na forma de onda na saída, comparada com o sinal de entrada).

1.1 Estatísticas de mercado

De acordo com uma pesquisa da Unicamp de 2002 [21], o setor de bens eletrônicos de consumo, que contempla equipamentos de áudio e vídeo como aparelhos de som e televisores, representa 7,6% do mercado mundial de eletrônicos, que significa um mercado de US$ 107,5 bilhões. Segundo a mesma pesquisa, no Brasil, o mercado de produtos eletrônicos em geral, é que seja de US$ 25,4 bilhões, com os eletrônicos de consumo sendo 10,6% desse mercado, ou US$ 2,7 bilhões. Vemos na pesquisa industrial do IBGE [22] em 2004 que o conjunto de produtos que contempla microfones e seus suportes, alto-falante, amplificador de áudio e fones de ouvido representam 12.435.124 de peças fabricadas, com valor de produção de R$284.743.000,63, e venda de 11.652.413 unidades, com valor de venda de R$261.392.000,35. Isto representa 9,7% das vendas no cruzamento de informações com a pesquisa da Unicamp para o mercado de eletrônicos de consumo no Brasil. Os dados fornecem a perspectiva de um setor que têm um grande mercado nacional e mundial, e existem grandes empresas atuando num cenário de grande competitividade, principalmente no desenvolvimento de novas tecnologias e novos produtos.

1.2 Amplificadores e soluções existentes

Algumas empresas estão trabalhando para prover amplificadores multicanais com um alto nível de fidelidade acústica [6],[11],[23],[24],[25],[26],[27]. Estes aparelhos são no momento comercializados para o mercado doméstico em sua maioria, com soluções de até oito canais e para o mercado profissional com soluções de até seis canais, como por exemplo os equipamentos da empresa Powersoft™[3] [27], que fabrica amplificadores profissionais de até seis canais.

Para *shows* em espaços abertos utilizam-se mais aparelhos com apenas dois canais, independente da quantidade de instrumentos, pois é requisitado uma grande quantidade de potência ao mesmo tempo que a fidelidade sonora deve ser mantida em alto patamar, e portanto os amplificadores utilizados são aqueles com fontes individuais (para evitar contaminação e ruído) e alta dissipação de potência, sem elevar o estresse de seus componentes.

Para eventos em ambientes fechados e de menor volume, como cinemas e teatros, utlizam-se aparelhos com até quatro canais, pois a quantidade de potência requerida é menor e a qualidade de reprodução não será afetada.

Entre os sistemas utilizando centenas de canais de saída encontram-se os baseados em WFS. A parte mais crítica de tais sistemas se dá na complexa gerência de distribuição, sincronismo e amplificação de centenas de canais, em especial na multiplicação de cabos, do consumo de potência, e comprometimento de espaço físico.

Para a distribuição de áudio multicanal há protocolos e interfaces adequadas indicadas para aplicações lidando com número de canais distintos. A figura 1 mostra uma hierarquia de protocolos existentes, e o número de canais suportado por cada um.

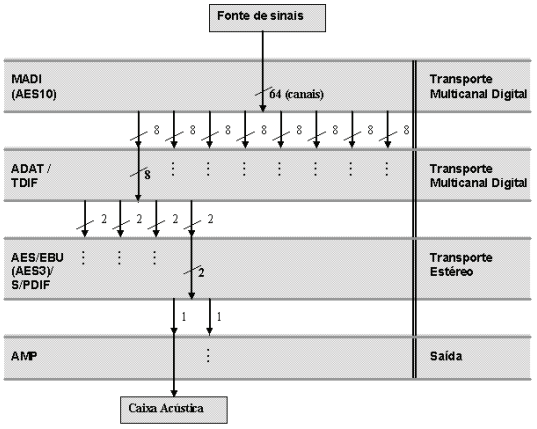


Figura 1 Hierarquia de Protocolos para Áudio Multicanal

[1] Dolby Laboratories Inc.

[2] DTS Inc.

[3] Powersoft Audio USA Inc.

No topo há soluções para distribuição de grande número de canais mas não soluções para amplificação final. A razão principal seria a impraticabilidade de construir dezenas de estágios amplificadores finais em chassis reduzidos. Nos níveis mais baixos de hierarquia, os equipamentos incorporam saídas analógicas de sinal, e muitos até saídas de potência. A grande maioria dos amplificadores de potência não possuem interfaces digitais, mas somente entradas e saídas analógicas de sinal, principalmente aqueles com melhores figuras de qualidade.

Encontramos ainda no topo protocolos capazes de transportar números elevados de canais. Para lidar com números maiores de canais, a AES normatizou o protocolo MADI (*Multichannel Audio Digital Interface*) capaz de transportar até 64 canais multiplexados em um único fluxo de bits [16]. Lançado em 1991 este protocolo é medianamente utilizado até o momento e não temos conhecimento de equipamentos que disponham de amplificação de saída para todos os canais.

Os protocolos mais populares para 2 canais são o S/PDIF [18] e o AES/EBU (AES3)[17][19]. São adotados amplamente em equipamentos de áudio e vídeo de consumo e profissionais, e utilizam uma via óptica ou coaxial para transmitir 2 canais não comprimidos[4].

Os equipamentos que lidam com 2 e mesmo alguns que lidam com 8 canais possuem saídas analógicas disponíveis. Freqüentemente vemos que, para construir sistemas de sonorização com alta densidade de canais de saída os desenvolvedores são forçados a usar vários protocolos, convertendo-se saídas de um para outro, num processo em cascata, até o acesso a cada canal individual.

Uma solução assim é utilizada por exemplo no sistema IOSONO™[5] [15], de sonorização espacial baseada em WFS, desenvolvido pelo Fraunhofer Institute, hoje em escala comercial. Nos enlaces mais altos são utilizados conversores MADI-AES ou MADI-ADAT, para distribuição digital em grupos de 8 ou grupos de 2 canais até matrizes especiais de 8 alto-falantes cada, onde os amplificadores finais estão embutidos.

Três soluções de saídas para alto-falantes são possíveis atualmente:

1. saída digital em 2 canais
2. saída analógica por canal, em nível de sinal de linha (0,3 ~ 4Vpp)
3. saída analógica por canal, amplificadas.

As duas primeiras soluções exigem caixas acústicas amplificadas, que recebem sinal digital (1) ou analógico (2), balanceado ou não. A terceira solução é para o uso de caixas acústicas passivas.

Entre os aparelhos com menos de quatro canais podemos citar o XM4080 da Yamaha™[6] [23], um amplificador para múltiplas utilizações de quatro canais e o ST 2.0 da Sankya™[7] [25], um aparelho de dois canais com THD menor que 0,03% e potência de saída a 4Ω de 250W por canal. A QSCaudio™[8] [11][12] produz amplificadores para eventos ao vivo de dois canais apenas, com opção para operação em ponte (*bridge*); para instalações permanentes existe aparelho amplificador para até oito canais, CX108V, com 100W por canal a 70V.

Entre os aparelhos com mais de quatro canais podemos citar os da Lab.gruppen™[9] [24], que fabrica amplificadores multicanais de alta potência como o C 48, de quatro canais com máxima potência de 4800W e SNR de 112dBA, e o C 20 de oito canais com potência de até 250 W por canal e SNR maior que 112 dBA, com THD a 1kHz e 1dB de no máximo 0,05% para os dois modelos.

Já a Rotel™[10] [26], que comercializa itens com 4, 5 ou 8 canais em um único chassi, com diferenças mínimas nas especificações na comparação com amplificador de um canal da mesma linha, como SNR que para 8 canais é de 116dB e para um canal é de 118dB e o THD máximo, que é o mesmo para os dois modelos, o que mostra a característica de refinamento sonoro atingida neste tipo de aparelho.

1.3 Limitações

Uma dificuldade na confecção de um amplificador multicanal é a alimentação. Existe o risco de contaminação de um canal por outro se ambos estiverem compartilhando uma mesma fonte. Idealmente o uso de uma fonte para cada canal é recomendado, caso que exigiria muito espaço para um aparelho amplificador multicanal. Então se faz necessário que a fonte compartilhada forneça energia de forma estável e contínua, situação que pode não ocorrer em aparelhos conectados a rede elétrica.

Neste quesito, fisicamente o aparelho ficará pesado se forem utilizados sistemas de alimentação independentes. Isto é o que acontece em sistemas profissionais de dois canais, tendo uma fonte de alimentação para cada um.

Para sistemas com mais canais, geralmente utiliza-se duas fontes para a alimentação do aparelho, pois o espaço utilizado para toróide+placas no chassi é muito grande.

1.4 Tendências do setor

Faria reporta em [10] a tendência à consolidação do multicanal em oposição aos sistemas estéreo somente. Isto aliado à clara demanda por qualidade sonora e crescente oferta de capacidade computacional aumentada pelas novas gerações de processadores e dispositivos DSP a custos acessíveis estabelecem nichos para a incubação de produtos inovadores voltado para a sonorização de cinemas, auditórios, casas de espetáculo, e em segundo lugar, produtos de consumo doméstico.

Funcionalidades mais sofisticadas como autocalibração dos modos multicanais às configurações reais de instalação dos alto-falantes são hoje *commodities* em equipamentos para som envolvente 5.1. A busca pelo conforto acústico, envolvimento e ambiência realista sugerem a adoção de técnicas multicanais orientadas à reconstrução de campos sonoros como a WFS.

1.5 Necessidades

Atualmente existem propostas de sistemas de sonorização sofisticados para produção de campos sonoros em larga escala, superando as limitações dos sistemas 5.1 (como os *sweet spot* críticos) e em verdade ampliando a

[4] ou, por meio de uma extensão da norma [20] tranportam até 6 canais comprimidos (amplamente utilizados nos receptores 5.1)
[5] IOSONO GmbH.
[6] Yamaha Co.
[7] Sankya do Brasil Eletrônica Indústria e Comércio Ltda.
[8] QSC Audio Products Inc.
[9] Lab.gruppen AB.
[10] Rotel Co Ltd.

qualidade da experiência espacial. Estes sistemas utilizam uma alta densidade de canais de saída, e necessitam de soluções eficazes para multi-amplificação.

A maioria dos receptores e decodificadores de A/V possuem conexões digitais de saída de áudio multicanal e somente saída analógica estéreo em nível de sinal de linha. É menos comum encontrar equipamentos com 6 ou mais saídas analógicas amplificadas, para ligação direto às caixas, e mais comum a utilização de módulos amplificadores externos, como em alguns *home-theaters*. A maioria dos amplificadores de potência utilizados em instalações profissionais são também estéreo, utilizados em pares para atender amplificação de múltiplos canais. As classes de amplificação mais comuns são A, B, AB e H, usadas na maior parte dos produtos comercializados [6],[11],[23],[24],[25],[26],[27].

A adoção de técnicas como o WFS para auralização de ambientes em progressiva substituição das atuais técnicas de *surround* para entretenimento como 5.1 aumentará a demanda por sistemas de amplificação com maior número de estágios de saída para canais individuais.

Precisa-se de equipamentos menores, com maior nível de integração e menor volume ocupado, capazes de amplificar 6, 8 ou mais canais num só gabinete. Isso facilitará o projeto e instalação de sistemas multiamplificados para uma diversidade de aplicações profissionais e domésticas. A adoção de amplificação em classe digital elimina a necessidade de fontes pesadas que usam transformadores toroidais e também a redução do tamanho dos dissipadores de calor, através do melhor gerenciamento térmico e menor consumo de potência.

Projetos orientados à integração de módulos são desejáveis para se permitir a construção de amplificadores com número de canais escalável. Isso requer uma arquitetura de sistema inerentemente escalável, por exemplo, utilizando-se barramentos expansíveis onde a adição de mais canais seja facilitada.

2 OBJETIVOS DO TRABALHO

Nos projetos desenvolvidos no LSI precisávamos adotar soluções de amplificação multicanal para 8 ou mais canais operando em sincronia que não estavam disponíveis no mercado. Os equipamentos existentes não ofereciam os tipos de conectores, número de saídas e funcionalidades necessárias para uma solução de amplificação multicanal profissional. Precisávamos por exemplo de conectores P10 (TRS) ou XLR, controle de volume individual por canal e saída para *subwoofer*. Embasados por esta necessidade iniciamos o projeto e a construção de um módulo protótipo de 8 canais atendendo os tais requisitos.

Primordialmente avaliamos o processo de construção de um amplificador multicanal e investigamos as dificuldades e facilidades neste tipo de desenvolvimento.

3 MÉTODOS

Metodologias para desenvolvimento de hardware são mais escassas e menos públicas como as existentes para desenvolvimento de software, sendo que na indústria o processo de concepção de hardware é feita com metodologia específica de cada empresa, não havendo um padrão mais utilizado para a montagem de equipamentos. É difícil até mesmo encontrar padrões dos órgãos de normas técnicas que contemplem uma metodologia geral para confecção de hardware. Elaboramos assim uma metodologia própria que resultou em uma qualidade de desenvolvimento de projeto mínima necessária para comparar com metodologias de software.

Os principais requisitos de desenvolvimento foram: (1) atender à demanda de sonorização existente em outros projetos do laboratório em tempo adequado, e (2) basear o projeto no uso e na integração de componentes "*commodities*" disponíveis. Para atender ambos os requisitos utilizamos módulos de amplificação de dois canais (30W/canal) disponibilizados pela empresa Sankya™, parceira do projeto AUDIENCE.

4 ETAPAS DE DESENVOLVIMENTO

O desenvolvimento do projeto consistiu das seguintes fases:

1. Análise de requisitos,
2. Análise de existência de módulos e componentes existentes integráveis,
3. Projeto do produto, dividido em etapas,
4. Aquisição de componentes,
5. Montagens e
6. Testes

O amplificador multicanal final foi construído em parceria com a Sankya™.

4.1 Complexidade / Descrição do Amplificador

O projeto do amplificador foi concebido dividindo-o em módulos, que somadas resultam no produto final. Tais módulos englobam o chassi, que envolve a carcaça do equipamento e define o leiaute do equipamento, a fonte de alimentação que envolve o toróide e a placa com o circuito, os conectores que envolve o conjunto de peças de entrada e saída de sinal, o cabeamento que envolve toda a ligação entre os conectores às placas e as placas amplificadoras, que envolve o *chip* amplificador, todos os componentes discretos utilizados na placa e a refrigeração. A descrição de cada etapa é feita a seguir.

Chassi - O chassi foi feito para atender os requisitos de ser pequeno e resistente, portanto é montável em rack, ocupando duas unidades (2U), o chassi acomoda a placa da fonte, o toróide, as placas amplificadoras, os dissipadores e os conectores de entrada e saída foram dispostos privilegiando o centro de gravidade no centro do chassi. A Sankya™ disponibilizou um chassi de amplificador estéreo para utilizarmos, e foram feitas readequações para acomodar número e tipos de conectores e botoeiras desejada, controle de intensidade dos oito canais na frente e conectores de entrada e saída para cada um dos canais, além do rearranjo dos porta fusíveis e cabo de alimentação, na parte traseira. A figura 2 mostra a frente e a traseira do equipamento projetadas.

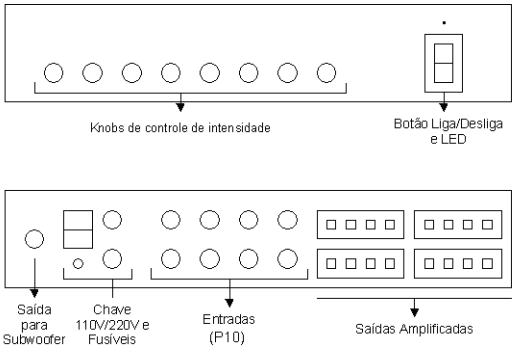


Figura 2 Frente (Acima) e Traseira (Abaixo) do amplificador projetado

Fonte - A fonte de alimentação constitui-se de um transformador toroidal que alimenta duas fontes independentes, que por sua vez alimentam quatro canais cada uma e têm os mesmos componentes para que a tensão de saída para a placa amplificadora seja semelhante. O usuário pode selecionar tensão de entrada de 110V ou 220V, de acordo com a rede elétrica local, e para cada seleção existe fusível de proteção. O transformador toroidal foi escolhido pois é melhor indicado para este tipo de função. Foram utilizados diodos retificadores configurados em ponte, para retificação completa e após foram colocados capacitores eletrolíticos de carga 4700μF para diminuir o *ripple* de saída. A Figura 3 mostra o layout da placa das pontes de retificação.

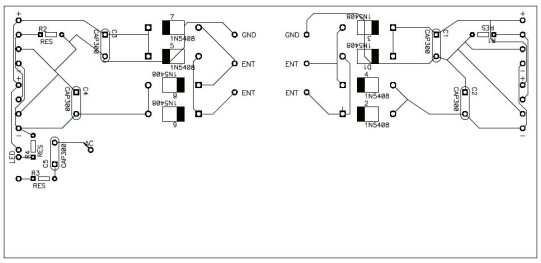


Figura 3 Esquemático da fonte de alimentação

Conectores - A conectorização foi desenhada para usar conectores profissionais, pois a aplicação pede um nível alto de fidelidade sonora, para evitar a degeneração das propriedades significativas do som na entrada e saída, como a relação sinal ruído. O tipo de conexão escolhida para a entrada foi P10 (TRS), para sistema não balanceado e mono; como conectores P10 são ainda sensíveis a mal-contatos, optou-se pela aquisição de conectores de qualidade, com duplo contato interno. Na etapa de interligação dos blocos, percebemos que a manipulação dos conectores de sinal de entrada estava influindo nas características elétricas do produto, revelando que qualidade do conector escolhido não condizia com a esperada.

Cabeamento - O cabeamento interno foi escolhido para que o sinal sonoro não tivesse a qualidade afetada no caminho percorrido, e impedir que eventuais perturbações externas, como campos eletromagnéticos indesejáveis, influenciassem na fidelidade. Condutores de tipos diferentes foram utilizados para o transporte de sinal de áudio em nível de linha e para sinal de áudio amplificado. Inicialmente previmos o uso de conexões blindadas para o transporte de áudio entre conectores de entrada, knobs potenciômetros e entrada dos módulos amplificadores, utilizando-se cabos de microfone para impedir que mesmo um pequeno ruído influísse diretamente no áudio.

Para sinal amplificado, da saída dos módulos amplificadores até os bornes de saída utilizamos condutores de cobre unifilares, com o sinal de maior potência. Verificamos posteriormente que o uso de cabos unifilares apresentavam vantagens frente à maior dificuldade de montagem, maior susceptibilidade a curtos, maior perda de qualidade e outras desvantagens do uso de condutores blindados, sendo estes, utilizados na montagem inicial, substituídos na montagem final.

Placas amplificadoras - Nas placas amplificadoras a admissão de entrada de sinal se dá em nível de linha. O uso de conectores soldados diretamente na placa se mostrou a melhor opção para o tipo de cabo utilizado, com um incremento mínimo de custo, e vantagens para a montagem, além de evitar chances de curtos-circuitos por soldá-los direto na placa amplificadora.

A amplificação adotada foi de classe AB, e o amplificador monolítico utilizado (*chip*) apresenta as seguintes características principais: potência máxima por canal de 30W, THD de 0,015% para potência de saída de 20W a 1kHz, máximo *Slew Rate* de 8V/μs a 20W, 8Ω e 70 kHz BW, *Power Supply Rejection Ratio* (PSRR) de 95dB, SNR de 100dB, proteção contra curtos, proteção térmica, tensão de alimentação de 16V a 60V e ganho em malha aberta de 90dB [14].

A placa amplificadora utiliza poucos componentes externos ao elemento amplificador, devido ao tipo de circuito integrado utilizado, que otimiza a quantidade de componentes no circuito.

O arrefecimento de calor é feito através do uso de dissipadores de alumínio que utiliza dissipação por ventilação natural, não forçada. As placas amplificadoras foram montadas juntas aos dissipadores respeitando uma configuração que os CI's tivessem garantidas a conectividade térmica e isolamento elétrico dos módulos amplificadores.

5 ANÁLISE DE CUSTOS

A figura 5 mostra um gráfico de composição de custos para a presente montagem.

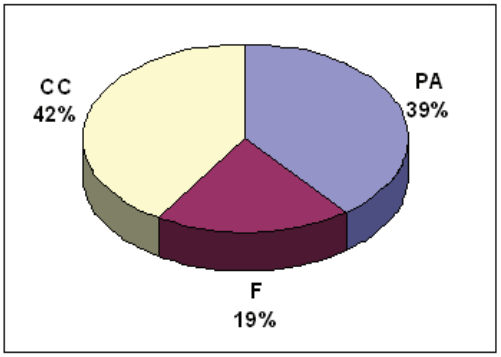


Figura 5 Composição de Custos do Amplificador Multicanal construído, com PA (Placa Amplificadora), F (Fonte) e CC (Conectores + Cabeamento)

A análise de custos revela a predominância de custos dos cabos e conectores de entrada e saída do sinal (CC), ligados a placa amplificadora, sobre o custo da placa em si. Em seguida, o elemento de maior participação nos custos globais é a placa amplificadora (PA), em especial o chip amplificador utilizado.

6 RESULTADOS

A figura 4 mostra o amplificador desenvolvido.

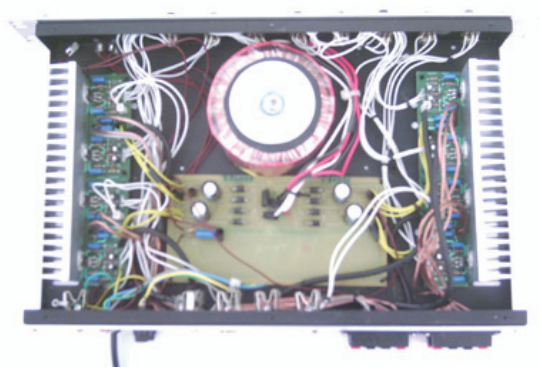


Figura 4 Amplificador Multicanal desenvolvido

Após a etapa de montagem foi elaborado um plano de testes em formato de *check-list*, dividido em 2 partes – testes antes e após a energização, para se determinar se todas as ligações estavam corretas e se todas as pendências haviam sido sanadas antes da segunda parte de testes, com o equipamento ligado.

6.1 Testes

A primeira parte dos testes foi mais longa que a segunda, pois exigiu uma análise detalhada de todo o aparelho. Os eventuais problemas encontrados foram solucionados rapidamente.

Na segunda parte, foram realizados testes de qualidade sonora através da audição de material polifônico, para atestar a separação de instrumentos, ausência de ruídos notáveis e artefatos de degradação. Sons musicais foram reproduzidos com diferentes níveis de amplificação até o máximo da excursão. Observamos em 25% dos canais a queima de *chips* sob volume alto, em condições de variação extrema e rápida da dinâmica. Para rastrear a fonte do problema foram refeitos os testes na placa amplificadora e na fonte de alimentação, principalmente nas placas com *chips* queimados, a procura de indícios da causa de problema.

Observamos nos canais queimados a correlação com apreciável distorção sonora presente nestes canais. A causa principal se deve a não conformidade de alguns circuitos integrados amplificadores à folha de especificações no que tange à capacidade de saída, provavelmente devido a lotes defeituosos em circulação no mercado.

A tabela 1 resume os itens considerados mais relevantes no *check-list* do plano de testes, correlacionados à montagem física dos módulos.

Tabela 1 Itens principais de testes

Elemento	Teste
Potenciômetros	
continuidade dos potenciômetros às placas de cada terminal	OK
cheque visual por curtos visíveis	OK
ligação de terra de cada potenciômetro ao chassi	OK
ligação do sinal de entrada a cada terminal de sinal do potenciômetro	OK
Terminais P10 (entradas)	
continuidade dos terminais de entrada aos potenciômetros	OK
conectividade dos conectores macho P10 nos jacks (conexão validade pelo som "click")	OK
cheque visual das partes não ligadas dos conectores	OK
cheque visual por curtos visíveis e qualidade da solda	OK
cheque com multímetro por curtos	OK
Terminais com borne (saídas)	
continuidade das saídas dos amplificadores aos bornes	OK
cheque visual por defeitos nos bornes	OK
cheque visual por curtos visíveis e qualidade da solda	OK
fixação no chassi	OK
Placas amplificadoras	
teste visual de caminho (comparando com esquemático das placas)	OK
teste de componentes soldados	OK
teste de fixação no módulo dissipadores	OK
cheque visual por pontos em risco de curto-circuito	OK
Fonte	
tensão de saída das fontes 1 e 2 está conforme esperado?	OK
polaridade e caminho de sinais	OK
continuidade até Vcc e GND das placas amplificadoras	OK
utilização 110V e 220V	OK
tensão de saída nos terminais do toróide	OK
fixação da placa	OK
inexistência de pontos de risco de curto-circuito(pedaços de fio livre,etc.)	OK
Montagem e acabamento	
Fios soltos	OK
disposição dos fios	OK
amarração (braçadeiras)	OK

6.2 Testes Futuros

Para uma próxima etapa, deverá ser usado um medidor acústico para auferir as características principais do equipamento, que englobam potência por canal, distorção harmônica total, resposta em freqüência, intermodulação e *slew rate*. Temos como patamares de referência os dados do *chip* escolhido, mostrado no item 4 (Etapas de Desenvolvimento), além dos dados da placa amplificadora fornecidos pela empresa ligada ao projeto, que tem THD de 0,01% e resposta em freqüência de 10 Hz a 30 kHz.

7 CONCLUSÕES

A confecção de amplificadores multicanal está em crescimento no mundo, pois há uma tendência dos fabricantes de equipamentos de áudio e vídeo em explorar o mercado consumidor doméstico, que desperta no usuário de som envolvente a vontade de se reproduzir uma cena da forma mais realista possível.

Projetamos e construímos um amplificador multipropósito de 8 canais mais uma saída de *subwoofer*, característica rara no mercado abastecido

predominantemente por equipamentos de no máximo 6 canais em um único chassi, que atendia aos requisitos de número de canais, e de conexões de entrada e saída para os experimentos de auralização do projeto AUDIENCE. Os conectores de alta qualidade utilizados, em conjunto com os cabos escolhidos e o cuidado no projeto constituem um detalhe importante no refinamento sonoro para um amplificador desse porte. Os testes de audição e de auralização realizados com Ambisonics em quatro e seis caixas apresentam resultados dentro do esperado, mesmo em fase inicial, atestando a qualidade sonora do equipamento.

O amplificador montado atende à finalidade proposta e propiciou um ganho de informação sobre o mercado e a atuação da indústria eletro-eletrônica, além de um melhor conhecimento do processo de produção, desde a fase de projeto até a etapa de testes, abrindo caminho para o amadurecimento e a proposição de metodologias para fabricação mais adequadas à produção de sistemas amplificadores multicanais multiuso, e que possam ser alinhadas a estratégias voltadas ao atendimento de necessidades de mercados incipientes aliadas ao uso de técnicas inovadoras (amplificadores digitais, arquiteturas escaláveis, menor dissipação térmica, menor volume e peso, opções de conectividade profissional, adequação de recursos e funcionalidades).

7.1 Agradecimentos

Os autores agradecem à empresa Sankya do Brasil Eletrônica Indústria e Comércio Ltda pela intensa colaboração e suporte no desenvolvimento do projeto.

8 REFERÊNCIAS BIBLIOGRÁFICAS

[1] Faria, R. R. A., AUDIENCE – *Audio Immersion Experience by Computer Emulation: Systems for Audio Immersion, Production and Reproduction of Spatial Audio*. Disponível em www.lsi.usp.br/interativos/neac/audience. Acesso em 25/03/2007.

[2] Faria, R. R. A et al. AUDIENCE – *Audio Immersion Experiences in the CAVERNA Digital*. Anais do 10º Simpósio Brasileiro de Computação Musical, Belo Horizonte, 2005. p.106-117.

[3] Ambisonic.Net, *where surround-sound comes to life*. Disponível em www.ambisonic.net/. Acesso em 30/03/2007.

[4] Corteel, E., Caulkins, T., *Sound Scene Creation and Manipulation using Wave Field Synthesis*. Disponível em recherche.ircam.fr/equipes/salles/WFS_WEBSITE/Documents/WFS_overview.pdf. Acesso em 21/03/2007.

[5] ITU-R. *Multichannel stereophonic sound system with and without accompanying picture*. – Rec. ITU-R BS.775-1. ITU, 1994. (rev. 2002)

[6] Crown Audio, Inc. *CTs Series Multi-channel*. White paper, 2005. Disponível em www.crownaudio.com/gen_htm/express/documents/136318.pdf

[7] Dolby Laboratories Inc. *5.1-Channel Music Production Guidelines*. White paper, issue 3, 2005. Disponível em www.dolby.com/assets/pdf/tech_library/4_Multichannel_Music_Mixing.pdf. Acesso em 25/03/2007.

[8] Dolby Laboratories Inc. *Digital Cinema: Observations and Considerations*, 1994. Disponível em www.dolby.com/assets/pdf/tech_library/1_DCinemaPhilosophy.pdf

[9] DTS Inc. Digital Entertainment. Disponível em www.dts.com. Acesso em 30/03/2007.

[10] Faria, R. R. A. *Auralização em ambientes audiovisuais imersivos*. Tese de Doutorado em Engenharia Eletrônica, Escola Politécnica da Universidade de São Paulo. São Paulo, 2005. 191 p.

[11] QSC Audio Products Inc. *The PowerLight 9.0PFC High-Power Innovations in Audio Amplifier Technology*, White paper, 1998. Disponível em www.qscaudio.com/support/library/papers/pl90wp.pdf. Acesso em 16/03/2007.

[12] Quilter, Patrick. *Amplifier Anatomy - Part 1 & 2*, Feb. 20, 1993. Disponível em www.qscaudio.com/support/library/papers/amptalk.pdf. Acesso em 25/03/2007.

[13] Toole, Floyd E. *Loudspeakers and Rooms for Multichannel Audio Reproduction*. 31 January, 2002. Disponível em www.harman.com/wp/pdf/Loudspeakers&RoomsPt1.pdf. Acesso em 27/03/2007.

[14] National Semiconductor Corp. *LM1875 20W Audio Power Amplifier*. Datasheet, July 2002. Disponível em www.national.com. Acesso em 28/03/2007.

[15] IOSONO GmbH The Art of Natural Sound. Disponível em www.iosono-sound.com. Acesso em 25/03/2007.

[16] AES Standard AES10-2003: *AES Recommended Practice for Digital Audio Engineering -- Serial Multichannel Audio Digital Interface (MADI) (Revision of AES10-1991)*. Disponível em www.aes.org/publications/standards/. Acesso em 25/03/2007.

[17] AES Standard AES3-2003. *AES standard for digital audio engineering - Serial transmission format for two-channel linearly represented digital audio data (Revision of AES3-1992, including subsequent amendments)*. Disponível em www.aes.org/publications/standards/. Acesso em 25/03/2007.

[18] International Electrotechnical Commission. *IEC 60958-3. Digital audio interface –Part 3:Consumer applications*. Disponível em webstore.iec.ch/webstore/webstore.nsf/artnum/036136.. Acesso em 29/03/2007.

[19] International Electrotechnical Commission. *IEC 60958-4. Digital audio interface –Part 4:Professional applications (TA4)*. Disponível em webstore.iec.ch/webstore/webstore.nsf/artnum/030575. Acesso em 29/03/2007.

[20] International Electrotechnical Commission. *IEC 61937-1. Digital audio - Interface for non-linear PCM encoded audio bitstreams applying IEC 60958 - Part 1: General*. Disponível em webstore.iec.ch/webstore/webstore.nsf/artnum/037433. Acesso em 29/03/2007.

[21] Sá, M. T. de V., *ESTUDO DA COMPETITIVIDADE DE CADEIAS INTEGRADAS NO BRASIL: impactos das zonas de livre comércio Cadeia: Bens Eletrônicos de Consumo Nota Técnica Final*, Campinas, Dezembro de 2002.

[22] Instituto Brasileiro de Geografia e Estatística - *IBGE Pesquisa Industrial – Produto 2004*. Disponível em www.ibge.gov.br/home/estatistica/economia/industria/pia/produtos/produto2004/. Acesso em 30/03/2007.

[23] Yamaha Co. Disponível em www.global.yamaha.com/index.html. Acesso em 30/03/2007.

[24] Lab.gruppen AB. Disponível em www.labgruppen.com/. Acesso em 30/03/2007.
[25] Sankya do Brasil Eletrônica Indústria e Comércio Ltda. Disponível em www.sankya.com.br/. Acesso em 30/03/2007.
[26] Rotel Co. Ltd. Disponível em www.europe.rotel.com/. Acesso em 30/03/2007.
[27] Powersoft Audio USA Inc. Disponível em www.powersoft-audio.com. Acesso em 30/03/2007.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Implementação do Método do Traçado de Raios e das Fontes Virtuais para a simulação acústica de salas

Márcio H. de Avelar Gomes[12], João G. Dedecca[1], Rhenan G. do Nascimento[1]
[1]Departamento de Arquitetura e Construção da Faculdade de Engenharia Civil, Arquitetura e Urbanismo, Universidade Estadual de Campinas
Campinas, S.P., Brasil
[2]Eco Acústica – Vibrações e Acústica Ltda.
Florianópolis, S.C., 880302-002, Brasil
marcio@ecoacustica.com.br

RESUMO
Este trabalho descreve dois métodos baseados em princípios geométricos para a simulação acústica de ambientes. O Método do Traçado de Raios e o Método das Fontes Virtuais para salas de geometria poliédrica arbitrária foram implementados na UNICAMP. Resultados de simulações realizadas com ambos os programas são comparados com aqueles obtidos em medições acústicas de duas salas. As vantagens e desvantagens de cada método são discutidas, tendo em vista o cálculo de parâmetros acústicos e a posterior auralização das salas. A implementação de um módulo para a auralização dos resultados também é descrita.

0 INTRODUÇÃO

O projeto AcMus, desenvolvido na Universidade de São Paulo sob a coordenação dos Professores Fernando Iazzetta (Escola de Comunicação e Artes) e Fábio Kon (Instituto de Matemática e Estatística) prevê a implementação de uma plataforma computacional que inclua ferramentas para medição e simulação de salas (veja [1]). O presente trabalho descreve o desenvolvimento de dois métodos de simulação, a serem incluídos na plataforma. Além dos métodos de simulação computacional, uma ferramenta para a auralização das salas simuladas também foi desenvolvida e é aqui brevemente descrita.

Ambos os métodos implementados são baseados em princípios geométricos para descrever a propagação sonora. Neste caso, fenômenos tais como interferência construtiva ou destrutiva de ondas não são levados em consideração. As hipóteses assumidas para a teoria acústica geométrica são razoáveis em situações onde o comprimento de onda é suficientemente menor do que as dimensões da sala, ou, sob outra ótica, onde a densidade modal é alta. Tal situação pode ser verificada através do cálculo da "freqüência de Schroeder", a partir de uma relação que envolve o volume da sala e o Tempo de Reverberação.

Os métodos implementados foram o Método do Traçado de Raios e o Método das Fontes Virtuais. Apesar de semelhanças na formulação de ambos, a Resposta Impulsiva resultante contém diferenças que podem ser significativas, especialmente quando se tem o objetivo de "auralizar" os resultados. Tais diferenças se referem à resolução temporal da resposta, como pode ser visto na apresentação dos resultados.

A "auralização" de salas tem se tornado uma ferramenta conveniente para vários objetivos, desde a simples verificação do resultado de um projeto acústico até a

elaboração de estudos psicoacústicos. O processamento de sinais para realizar uma auralização é baseado em operações de convolução. Diferentes métodos para realizar tal operação foram implementados e testados, com relação à qualidade do resultado e ao tempo de processamento.

Para verificar a implementação dos métodos de simulação computacional, resultados obtidos através dos programas desenvolvidos foram comparados com medições realizadas em várias salas (uma comparação apresentada aqui). Tais comparações foram feitas em termos de Respostas Impulsivas e parâmetros acústicos objetivos.

1 MÉTODOS DE SIMULAÇÃO ACÚSTICA

1.1 Método do Traçado de Raios

O Método do Traçado de Raios é o mais tradicional, entre aqueles baseados nos princípios geométricos de propagação sonora. Neste método, o percurso de raios acústicos emitidos pela fonte sonora é seguido, até que sua energia seja desprezível, após a perda durante a propagação e as reflexões nas superfícies internas à sala (atenuadas através do coeficiente de absorção da superfície).

Em cada trecho de propagação, verifica-se se o raio atinge uma esfera receptora. Quando isso acontece, a energia do raio e a duração do percurso são registrados para a posterior construção da Resposta Impulsiva.

Para gerar as direções iniciais dos raios a serem seguidos, vários métodos podem ser utilizados, aleatórios ou determinísticos. No programa implementado, utilizou-se um método determinístico baseado na subdivisão geodésica de um icosaedro regular (conforme descrito em [2] e [3]).

Em métodos geométricos de simulação, efeitos como difusão ou difração podem ser modelados de forma simplificada. No método implementado, as reflexões nas superfícies são modeladas como se ocorressem de forma simétrica (especular) ou através de um modelo simples para a reflexão difusa. Para isso, um "coeficiente de espalhamento" deve ser informado como dado de entrada. Quando um raio incide sobre uma superfície, um número aleatório entre 0 e 1 é gerado. Tal número é comparado com o coeficiente de espalhamento daquela superfície e, caso ele seja menor, a reflexão ocorre em um ângulo também aleatório. Caso contrário, a reflexão ocorre de forma simétrica (com relação à direção normal à superfície).

A partir do registro do tempo de chegada de cada raio à esfera receptora e de sua energia, determina-se a Resposta Impulsiva para aquela posição.

1.2 Método das Fontes Virtuais ("Image-Sources")

O Método das Fontes Virtuais é, a princípio, uma alternativa para determinar o caminho dos raios refletidos pelas superfícies. A idéia básica do método é encontrar a posição das "imagens" da fonte sonora, com relação às superfícies, em uma analogia com a ótica. Uma vez encontradas as posições das imagens, o problema se reduz a determinar a contribuição de cada fonte sonora (original e imagens) para a resposta acústica em um determinado ponto (ouvinte ou receptor).

Apesar das semelhanças entre os dois métodos, existem diferenças relevantes entre eles. Uma delas é relativa à resolução temporal da Resposta Impulsiva resultante. No Método das Fontes Virtuais, essa resolução pode ser tão alta quanto se queira. Se o usuário desejar obter como resultado um sinal (Resposta Impulsiva) com resolução de 0,02 ms, equivalente a uma freqüência de amostragem de 44100 Hz, basta incluir este valor como dado de entrada. Isto significa que o momento de chegada das reflexões pode ser previsto com maior precisão. O Método do Traçado de Raios possui uma limitação com relação a este aspecto e a resolução da resposta é limitada pelo diâmetro da esfera receptora. Uma esfera de 10 mm de diâmetro, por exemplo, resultaria em uma Resposta Impulsiva com resolução de 0,03 ms, aproximadamente.

Outro aspecto importante do Método das fontes Virtuais é que este gera um algoritmo cujo tempo de processamento cresce exponencialmente com a ordem de reflexão (número de vezes que cada raio é refletido). Este fato impõe limitações ao uso do método, que é restrito à simulação de salas pequenas e pouco reverberantes.

Uma das tarefas que consomem tempo é a verificação da "visibilidade" da fonte sonora ao ouvinte. Uma grande parte das fontes virtuais é invisível ao ouvinte, mas o teste de "visibilidade" deve ser aplicado a todas. Uma alternativa para minimizar este problema é proposta em [4] e consiste em pré-calcular as fontes sonoras "visíveis" ao ouvinte através do Método do Traçado de Raios. No projeto aqui descrito tal implementação híbrida dos dois métodos ainda não foi feita, mas está prevista.

2 AURALIZAÇÃO

Desde que modelos em escala reduzida começaram a ser usados no processo de projeto acústico de ambientes, desejou-se possibilitar a escuta do resultado previamente à execução das obras. Os avanços tecnológicos e dos métodos de simulação acústica das últimas décadas viabilizaram a realização deste desejo.

É possível atualmente criar amostras sonoras de como a música ou a fala será ouvida em uma sala ainda inexistente, com alto grau de realismo. A técnica que torna isso possível é chamada de "auralização", um nome alusivo à contrapartida ótica da visualização realista de um ambiente ainda na fase do projeto.

Uma definição bastante citada para a técnica é proposta em [5] e pode ser traduzida como: "Auralização é o processo de tornar audível, através de modelagem física ou matemática, o campo sonoro de uma fonte no espaço, de forma a simular a experiência de audição binaural em uma certa posição do espaço modelado."

Em termos matemáticos o processo pode ser representado através da integral de Duhamel, que consiste na operação de convolução da Resposta Impulsiva entre a fonte sonora e o receptor com um sinal de áudio anecóico. Tal Resposta Impulsiva, neste caso, deve ser binaural, isto é, possuir dois canais de áudio que correspondem ao sinal acústico que deve atingir o ouvido esquerdo e direito.

A Resposta Impulsiva binaural pode ser medida através de uma cabeça artificial ou real, com microfones posicionados na entrada dos canais auditivos ou na posição do tímpano. Quando se trata de simulação acústica, esta resposta deve ser obtida de outra forma. Nesse caso, cada componente da Resposta Impulsiva simulada deve passar

por uma operação de convolução com uma Resposta Impulsiva binaural medida em câmara anecóica. Toma-se, para isso, uma resposta medida em uma direção correspondente à chegada deste componente ao receptor. Portanto, usa-se um banco de dados de Respostas Impulsivas relativas à cabeça, com resolução espacial determinada pela malha de medição. Tais bancos de dados são chamados de bancos de HRTF (de "Head Related Transfer Function", correspondente às respostas impulsivas no domínio da freqüência).

Um programa foi feito no âmbito deste projeto para determinar Respostas Impulsivas binaurais a partir daquelas simuladas e realizar a convolução com arquivos anecóicos de som. Para isso, utilizamos dois bancos de HRTF: um, cedido pelo "Institut für Technische Akustik" da Universidade de Aachen (RWTH-Aachen) e outro, para uso público, provido pelo "Media Lab" do "Massachussets Institute of Technology" [6].

Como a operação de convolução é básica no processo para a obtenção da Resposta Impulsiva binaural a partir de simulação acústica, quatro diferentes métodos foram implementados e testados (para detalhes, veja [8]). Destes, concluiu-se que um método no domínio do tempo pode ser vantajoso em casos onde a Resposta Impulsiva tem poucos componentes não nulos. Tal método usa uma propriedade da operação de convolução para acelerar o processo:

$$\delta(t-t_0)*s(t)=s(t-t_0), \qquad (1)$$

onde δ(t) é a função Delta de Dirac.

Além deste, outro método eficiente é baseado no algoritmo chamado de FFTW [7]. Os outros métodos testados foram: 1) implementação direta da operação de convolução e 2) implementação da operação de convolução através de algoritmo tradicional de Transformada Discreta de Fourier.

3 RESULTADOS

Para verificar a implementação dos métodos de simulação descritos, resultados de simulação de várias salas foram comparados com medições (feitas de acordo com a norma ISO 3382). Uma dessas salas foi o auditório da Faculdade de Engenharia Civil, Arquitetura e Urbanismo da UNICAMP, cujas principais dimensões são representadas na Figura 1.

Detalhes da construção são omitidos na figura. A área inclinada do piso é construída, na realidade, em patamares que abrigam fileiras de cadeiras levemente estofadas. Uma das paredes laterais e a parede ao fundo são construídas com blocos de concreto e são pintadas. A outra tem aproximadamente metade de sua área construída da mesma forma e outra fechada por janelas, que eventualmente são cobertas por uma cortina plástica. A parede frontal também tem os blocos de concreto pintados aparentes e possui uma lousa instalada. O piso é revestido com "paviflex" e o teto é de concreto pintado. O volume total do auditório é de aproximadamente 690 m3.

A Tabela I apresenta os coeficientes de absorção atribuídos a cada superfície. Esses coeficientes são equivalentes a esta característica para a banda de freqüência de 1000 Hz (1/1 oitava).

Tabela I Coeficientes de absorção usados na simulação computacional.

Área	*Material*	*Coeficiente de absorção*
Parede lateral direita, parede ao fundo	Bloco de concreto pintado	0,07
Parede frontal	Bloco de concreto pintado+lousa	0,07
Parede lateral esquerda	Bloco de concreto pintado+cortina plástica	0,3
Área entra palco e assentos	Paviflex	0,05
Palco	Paviflex+móveis	0,2
Público	Paviflex+assentos levemente estofados	0,5
Teto	Concreto pintado	0,05

A Resposta Impulsiva simulada através do Método do Traçado de Raios é apresentada na Figura 2, comparada com aquela medida para a mesma posição. Observa-se que o aspecto de ambas é similar. Note que a resolução da resposta simulada é limitada pelo diâmetro da esfera receptora, estabelecido em 0,1 m (correspondente a um intervalo de tempo de aproximadamente 0,3 ms).

Um dos parâmetros para o programa de Traçado de Raios é o número de raios a serem seguidos. No caso analisado este número foi de 100.000 raios.

A Figura 3 mostra um detalhe do mesmo tipo de comparação, agora para o resultado obtido com o Método das Fontes Virtuais.

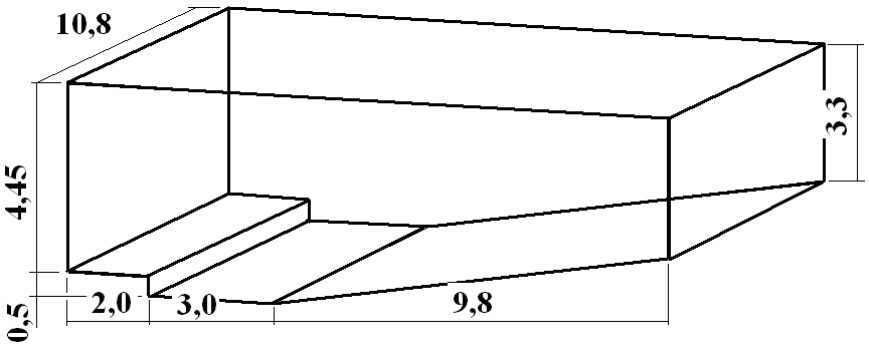


Figura 1 Representação do auditório utilizado na comparação. Dimensões em metros.

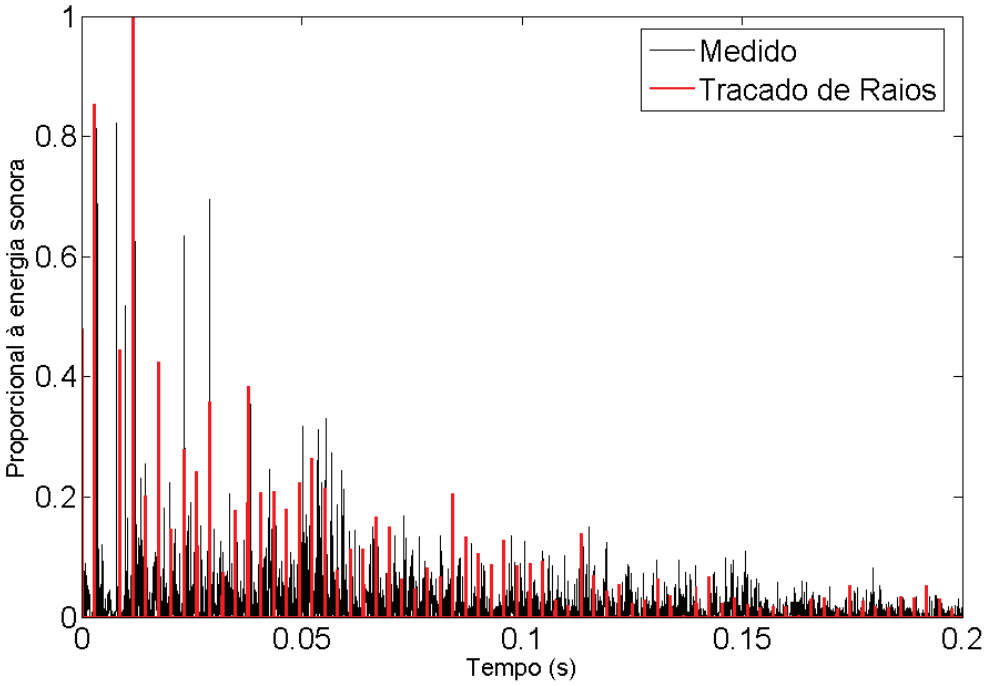


Figura 2 Respostas Impulsivas medida e simulada através do Método do Traçado de Raios.

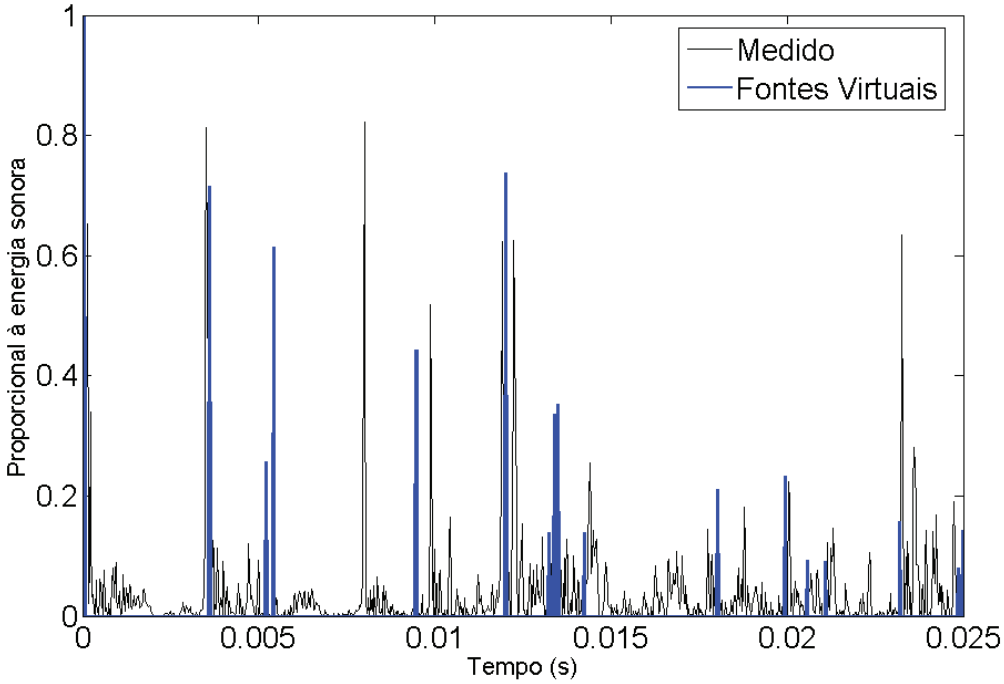


Figura 3 Detalhe do trecho inicial das Respostas Impulsivas medida e simulada através do Método das Fontes Virtuais.

Observa-se, na Figura 3, a semelhança entre as primeiras reflexões presentes na medição e na simulação realizada através do Método das Fontes Virtuais, onde a ordem máxima de reflexão considerada foi 10.

Tal ordem de reflexão, no entanto, não é suficiente para garantir que a Resposta Impulsiva seja determinada de forma correta. De acordo com previsões baseadas em [8], a ordem de reflexão a ser considerada neste caso deve ser maior do que 60. Isso, no entanto, seria inviável, devido ao alto tempo computacional necessário para o processamento através deste método. A Figura 4 mostra uma comparação entre o decaimento sonoro obtido a partir da Resposta Impulsiva simulada através do Método das Fontes Virtuais e aqueles previstos para diferentes ordens de reflexão.

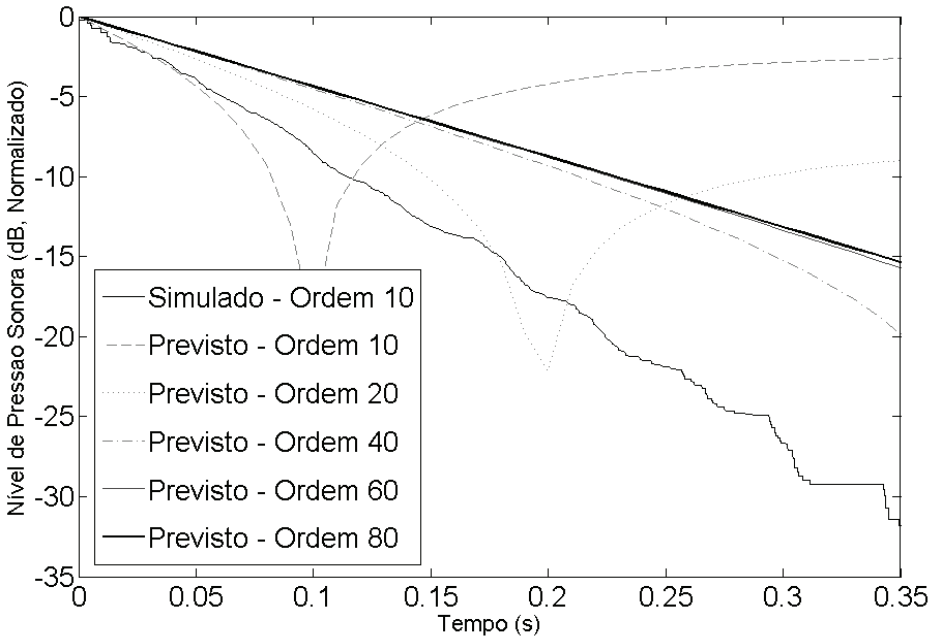


Figura 4 Curva de decaimento simulada através do Método das Fontes Virtuais, com ordem de reflexão igual a 10, comparada àquelas previstas para diferentes ordens de reflexão.

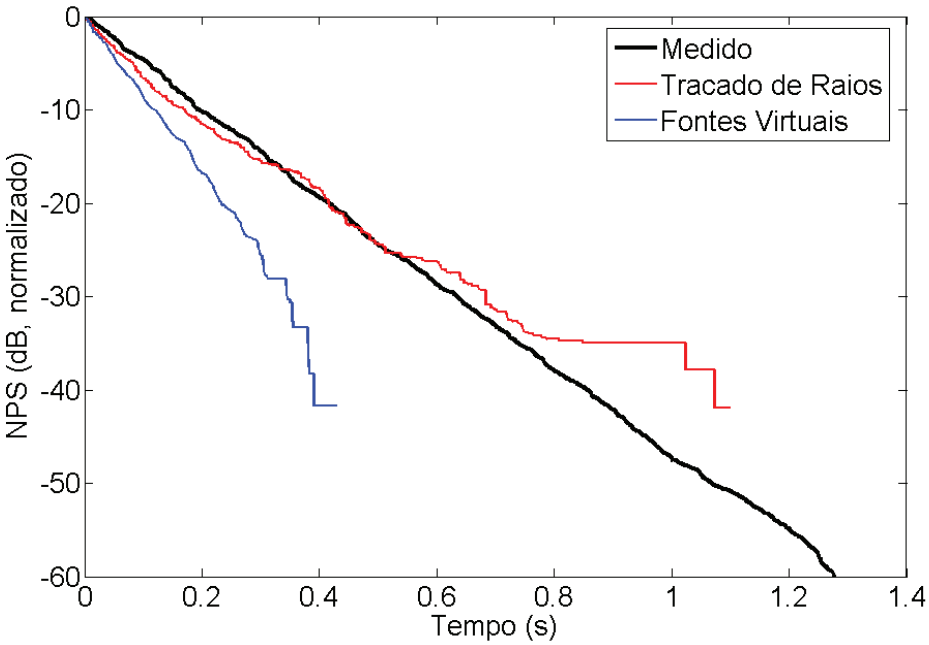


Figura 5 Curvas de decaimento simuladas e medida.

A partir das Respostas Impulsivas pode-se determinar curvas de decaimento e parâmetros relativos ao Tempo de Reverberação da sala. As curvas de decaimento obtidas a partir das simulações e da medição são apresentadas na Figura 5. Os parâmetros obtidos das simulações e das medições são apresentados na Tabela II.

O parâmetro "Early decay Time" refere-se ao tempo de reverberação determinado a partir dos trecho entre 0 e -10 dB, da curva de decaimento. Da mesma forma, T_{10} e T_{20} referem-se ao tempo de reverberação determinado entre -5 e -15 dB e -5 e -25 dB, respectivamente.

Tabela II Resultados relativos à reverberação, obtidos a partir da Resposta Impulsiva medida e simuladas para o Auditório da FEC (1000 Hz, banda de 1/1 oitava).

	Medido	***Traçado de Raios***	***Fontes Virtuais***
EDT (s)	1,33	1,05	0,69
T_{10} (s)	1,36	1,31	0,79
T_{20} (s)	1,55	1,77	0,74

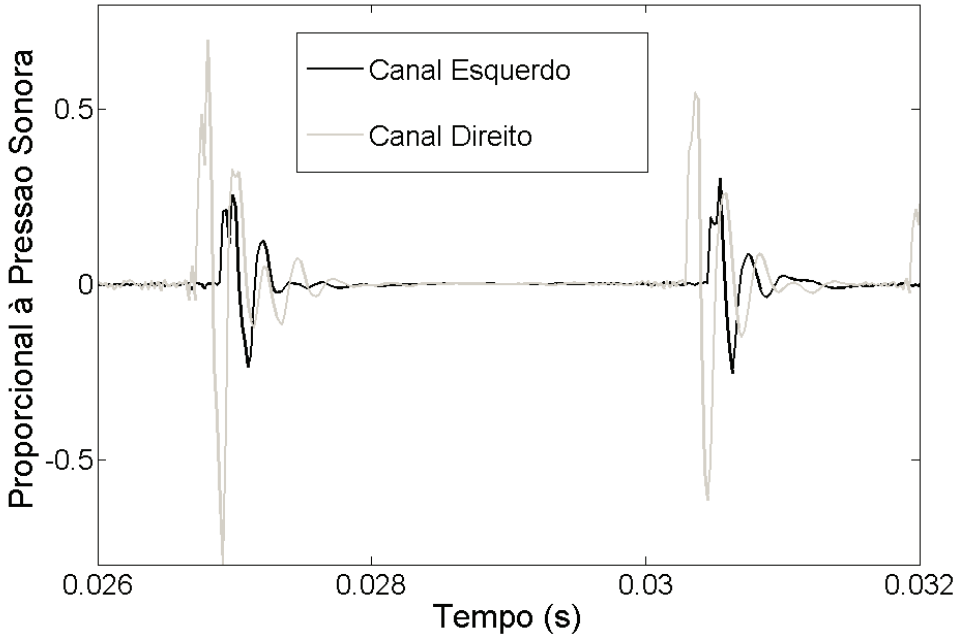


Figura 6 Detalhe da Resposta Impulsiva binaural (simulada).

Para ilustrar o processo de auralização, a Figura 6 mostra o som direto e a primeira reflexão contidos em uma Resposta Impulsiva simulada, após a convolução com o filtro de HRTF. Neste caso, a fonte sonora se encontra na região à direita e à frente do ouvinte. Os sons chegam primeiro ao ouvido direito e depois ao ouvido esquerdo, como é de se esperar.

4 CONCLUSÕES

A Tabela II indica que o Método do Traçado de Raios proporciona resultados semelhantes aos medidos, mas com uma discrepância maior para EDT. Apesar do programa ter a capacidade de considerar efeitos de reflexão difusa, tal recurso não foi utilizado nesta simulação. Novas simulações serão realizadas para avaliar se os resultados para este parâmetro se aproximam mais do valor medido.

Os resultados obtidos através do Método das Fontes Virtuais ainda não são satisfatórios. Verificações feitas com decaimentos esperados, determinados teoricamente, indicam que este foi corretamente implementado e que o desvio percebido é conseqüência da baixa ordem de reflexão utilizada nas simulações (veja [8]). Este fato limita a aplicação do método em salas com muitas superfícies e com Tempo de Reverberação elevado.

Percebeu-se, portanto, que o Método do Traçado de Raios é capaz de prever o comportamento geral da Resposta Impulsiva. Há uma limitação, no entanto, baseada na resolução máxima alcançada, relacionada ao diâmetro do receptor, dada por:

$$\Delta t = d/c, \qquad (2)$$

onde d é o diâmetro do receptor e c é a velocidade do som no meio.

Por outro lado, a Resposta Impulsiva obtida através do Método das Fontes Virtuais, apesar da limitação imposta pela impossibilidade de considerar ordens de reflexão mais altas, mostra o potencial do método em prever de forma mais precisa a distribuição das primeiras reflexões.

5 COMENTÁRIOS FINAIS

Pequenas alterações e ajustes têm sido feitos nos programas descritos. Uma forma híbrida, conforme mencionado, será implementada em breve, para viabilizar o uso do Método das Fontes Virtuais em um universo maior de problemas. O programa que utiliza o Método do Traçado de Raios vem sendo aprimorado para considerar a diretividade das fontes sonoras.

Testes subjetivos deverão ser feitos para verificar se a diferença relativa à resolução das Respostas Impulsivas simuladas através dos diferentes métodos é perceptível à audição humana.

Os programas descritos serão em breve disponibilizados na página eletrônica do AcMus e no endereço www.e-science.unicamp.br/ppa , com seus respectivos tutoriais.

6 AGRADECIMENTOS

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado de São Paulo pelo financiamento do projeto, aos coordenadores do Projeto AcMus, à Profa. Stelamaris Rolla Bertoli por acolher e dar apoio ao projeto, e ao Prof. Michael Vorlaender e Tobias Lentz, do "Institut für Technische Akustik", da Universidade de Aachen, por ceder o banco de dados de HRTF para a realização de parte deste projeto.

7 REFERÊNCIAS

[1] Figueiredo, F., Iazzetta. F., Parâmetros Acústicos em Salas de Música: análise de resultados e novas interpretações. Anais do 4.o Congresso de Engenharia de Áudio. São Paulo: Audio Engineering Society - Seção Brasil, v. 1, p. 66-71, 2006.

[2] Lewers, T., A combined beam tracing and radiant exchange computer model of room acoustics. Applied Acoustics, v.38, p.161-173, 1993.
[3] Tenenbaum, R. A., Camilo, T.S., Torres,J.C.B., Gerges, S.N.Y., Hybrid method for numerical simulation of room acoustics with auralization: Part 1 - Theoretical and numerical aspects. Journal Of The Brasilian Society Of Mechanical Sciences And Engineering, Rio de Janeiro, 2006.
[4] Vorlaender, M., Simulation of the Transient and Steady-state Sound Propagation in Rooms Using a New Combined Ray-tracing/Image-source Algorithm. Journal of the Acoustical Society of America, No.86(1), p.172-178, July 1989.
[5] Kleiner, M., Dalenbäck, B.-I., Svensson, P., Auralization – An Overview, Jounal of the Audio Engineering Society 41, p.861-875, 1993.
[6] Gardner, W., Martin, K., página eletrônica http://sound.media.mit.edu/KEMAR.html, acessada em janeiro de 2007.
[7] Frigo, M., Johnson, S.G., The Design and Implementation of FFTW3. In Proceedings of IEEE, vol.93, no.2, p.216-231, 2005.
[8] Gomes, M.H.A., Bertoli, S.R., Dedecca, J.G., Implementação de Métodos para a Simulação Acústica e Auralização de Salas, Acústica e Vibrações – Revista da Sociedade Brasileira de Acústica, a ser publicado em maio de 2007.

Sessão 4

Tratamento, restauração e avaliação de qualidade de áudio

(Audio treatment, restauration and quality assessment)

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Plugin VST para Remoção de Ruído de Fundo

Alan Freihof Tygel[1] e Luiz Wagner Pereira Biscainho[1]

[1] LPS – PEE / COPPE & DEL / POLI, UFRJ
Caixa Postal 68504, Rio de Janeiro, RJ, 21941–972, Brasil

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

RESUMO

Este trabalho implementa na forma de *plugin* VST (*Virtual Studio Technology*) três técnicas para redução de ruído de fundo em sinais de áudio: subtração espectral, filtragem de Wiener e *wavelet shrinkage*. O *plugin* tem quatro modos de operação: o primeiro extrai informações do ruído e os outros três implementam os métodos de tratamento. O funcionamento se dá em tempo real, e o usuário pode controlar dois parâmetros: o tamanho da janela de análise e um parâmetro α para ajuste fino da quantidade de redução de ruído. Além disso, foram implementados um pré- e um pós-equalizador para complementar o processo de restauração.

0 INTRODUÇÃO

Gravações de áudio estão freqüentemente sujeitas a ruídos indesejados. Seja por defeitos nos transdutores de entrada (microfones), na mídia (vinil, fita, CD etc.), nos amplificadores ou nos transdutores de saída (caixas acústicas), as chances de se ouvir uma música com imperfeições que não estavam presentes no momento da geração do sinal é bem alta.

Na era do áudio digital, a representação em bits reduziu bastante esses problemas. Surgiu, então, o novo desafio de digitalizar gravações da era analógica reduzindo os defeitos inerentes aos discos analógicos e fitas magnéticas. Os defeitos mais comuns em gravações analógicas podem ser agrupados em três tipos: ruído impulsivo, pulsos longos e ruído de fundo.

O ruído de fundo, também chamado de *hiss*, tem características que tornam seu tratamento mais complicado do que o dos outros dois. A primeira delas é que ele está presente todo o tempo, de modo que é necessário interferir em todo o sinal, o que torna mais alta a probabilidade de se provocar degradações inexistentes anteriormente. Além disso, geralmente possui banda larga, impossibilitando, assim, o uso de um filtro linear fixo.

Este trabalho tem como objetivo implementar um sistema para tratamento de ruído de fundo na forma de *plugin* VST[1], contemplando três métodos de redução de ruído da literatura [1] e equalização auxiliar. A meta

[1] Virtual Studio Technology. *VST is a trademark of Steinberg Media Technologies GmbH.*

final é construir uma plataforma que possibilite a adição posterior de novos métodos de tratamento, incluindo os que operam no domínio do tempo. Portanto, aqui o foco não está nos métodos e no seu desempenho, mas sim na implementação.

No contexto deste trabalho:

- ***Plugin*** = *Software* cujo funcionamento depende de um *software* hospedeiro. Ele recebe as amostras de áudio, processa-as e as devolve ao *host*.
- ***Host*** = Programa de edição e gravação de áudio, normalmente multipista, que permite associar um *plugin* a uma determinada pista, fazendo com que as amostras do sinal contidas nela passem pelo *plugin* antes de irem para a saída. O *host* pode ser compatível com o protocolo VST.

O uso de *plugins* é hoje muito difundido no mercado de áudio digital. Eles são vantajosos para os desenvolvedores, por permitirem que estes não mais precisem se preocupar com aspectos de entrada e saída, que normalmente são específicos de cada sistema; e para os usuários, que podem se especializar no uso de uma plataforma de edição (*host*) e escolher apenas os *plugins* que lhes forem úteis.

Existem algumas alternativas de protocolos de *plugins* disponíveis, além do VST: AU (Audio Unit), RTAS (Real Time Audio Suite), LADSPA (Linux Audio Developer's Simple Plugin API) e DirectX. Considerando o fato de se tratar de um protocolo aberto, aceito pela maior parte dos *hosts*, e multiplataforma, o VST é a melhor escolha entre os protocolos citados.

Dentre os *hosts* mais difundidos, somente um deles, o mais caro e mais usado em estúdios profissionais, não é compatível com o VST. Entretanto, existem plugins *wrappers* que contornam este problema, fazendo uma espécie de interface entre protocolos.

O artigo está assim organizado: a Seção 1 descreve os blocos que constituem o sistema, dele fornecendo uma visão geral; na Seção 2, são apresentados os métodos de restauração implementados, e na Seção 3 os equalizadores; após a discussão da implementação do *plugin* via protocolo VST na Seção 4, faz-se uma avaliação dos resultados na Seção 5; por último, a Seção 6 lista possíveis continuações do trabalho.

1 VISÃO GERAL DO SISTEMA

A Figura 1 mostra um diagrama em blocos do sistema. Ele pode ser dividido em três partes principais: um pré-equalizador, a aplicação da técnica de tratamento e um pós-equalizador. Além disso, há um bloco auxiliar capaz de calcular o espectro de um sinal e gravá-lo em um arquivo. Caso haja um trecho do sinal contendo apenas ruído, a informação de seu espectro poderá ser usada na restauração.

Além dos ganhos nas diversas bandas dos equalizadores, o usuário deverá fornecer mais quatro parâmetros:

- **Modo de operação**

  Subtração espectral, filtragem de wiener, *wavelet shrinkage* ou análise de ruído.

- **Tamanho da janela**

  O tamanho da janela de análise utilizada nos métodos de restauração e na análise do ruído.

- **Intensidade de remoção do ruído (α)**

  Um parâmetro de ajuste, que regula a intensidade da atuação do método de tratamento de ruído de fundo ao sinal. Se $\alpha = 0$, nenhum tratamento é realizado. Aumentando-se esse valor, o método retira cada vez mais ruído; a partir de certo ponto, o próprio sinal começa a ser degradado.

- **Tipo de ruído a ser utilizado**

  O usuário escolhe se quer utilizar o ruído analisado ou supor ruído branco, nos métodos espectrais.

Desta forma, espera-se conseguir uma boa usabilidade do *plugin*, fornecendo ao usuário diversos parâmetros e mostrando os resultados sem latência perceptível. Uma vez que o processamento é realizado à medida que o programa *host* reproduz o sinal musical, pode-se caracterizar, com alguma licença de terminologia, o funcionamento do sistema de restauração como sendo "em tempo real".

2 MÉTODOS DE TRATAMENTO DE RUÍDO DE FUNDO

Os métodos de tratamento de ruído de fundo aqui utilizados baseiam-se naqueles abordados em [1]: dois são baseados no espectro [2] e um, na *discrete wavelet transform* (DWT) [3].

2.1 Métodos baseados no espectro

Escrevendo-se o sinal contaminado como

$$y[n] = x[n] + r[n], \tag{1}$$

onde $x[n]$ é o sinal original e $r[n]$ é o ruído aditivo, podemos admitir que $x[n]$ e $r[n]$ são amostras dos processos aleatórios $X[n]$ e $R[n]$, não-correlacionados e de média zero. Assim, o espectro de densidade potência do processo corrompido é dado por:

$$S_{YY}(\Omega) = S_{XX}(\Omega) + S_{RR}(\Omega), \tag{2}$$

onde $S_{XX}(\Omega)$ é o espectro de densidade potência do processo original e $S_{RR}(\Omega)$ é o espectro de densidade potência do processo de ruído.

A estimativa de $S_{YY}(\Omega)$ pode ser obtida através da análise do próprio sinal. A estimativa de $S_{RR}(\Omega)$ pode ser obtida ou através da análise de um trecho contendo apenas ruído (tipicamente no começo ou no final da gravação) ou então considerando-se que o ruído é branco e possui espectro constante.

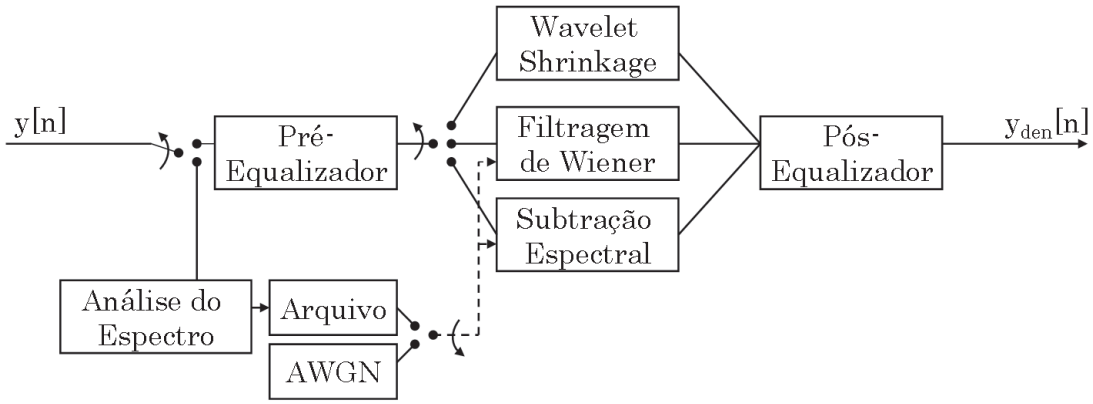


Figura 1: Diagrama em blocos simplificado do sistema de redução de ruído de fundo, onde $y[n]$ é o sinal corrompido e $y_{\mathrm{den}}[n]$ é o sinal tratado. O diagrama não inclui a entrada de parâmetros externos aos blocos.

2.1.1 Considerações práticas

A estimativa prática da densidade espectral de potência $S_{YY}(\Omega)$ de um processo aleatório $Y[n]$ pode ser feita facilmente, se valer a hipótese de que durante um intervalo de W amostras o processo $Y(\Omega)$ é estacionário no sentido amplo e ergódico. Separa-se um trecho de W amostras do sinal $y[n]$ multiplicando-o por uma janela suavizadora $w_H[n]$ (de Hamming, neste trabalho) com esse comprimento; e aplica-se a DFT (*discrete Fourier transform*) ao resultado, mais uma eventual normalização:

$$Y[k] = \mathrm{DFT}\left(\frac{y[n]w_{\mathrm{H}}[n]}{M}\right). \tag{3}$$

Nesse caso, M é o tamanho da DFT, ajustado para a primeira potência de 2 igual ou superior a W para permitir o uso de um algoritmo rápido (FFT - *fast Fourier transform*). Uma versão amostrada de $S_{YY}(\Omega)$ pode, agora, ser aproximada pelo espectro de energia $|Y[k]|^2$ desse segmento de $y[n]$.

Caso o usuário opte por analisar o ruído, basta que selecione um trecho suficientemente longo contendo somente ruído. Por um processo similar ao descrito acima, o *plugin* calcula uma estimativa de $S_{RR}(\Omega)$ pela média dos valores de $|R[k]|^2$ calculados por 100 janelas.

Caso o usuário deseje considerar o ruído como branco e utilizar um valor fixo para estimar $S_{RR}(\Omega)$, faz-se na primeira janela uma estimativa da potência do sinal contaminado e adota-se um valor constante para $|R[k]|^2$ 20 dB abaixo daquele valor. Este valor é arbitrário e o usuário podera ajustá-lo através do parâmetro α. Caso o início do trecho contenha amostras nulas, não ocorrerá o tratamento. Este problema poderá ser solucionado com a utilização de métodos de estimação do espectro do ruído sem a necessidade de haver trecho contendo somente ruído [4].

As subseções 2.1.2 e 2.1.3 utilizam as variáveis aqui definidas. O processamento é realizado em blocos com sobreposição de 50%.

2.1.2 Subtração espectral

Da equação (2), podemos imediatamente escrever:

$$S_{XX}(\Omega) = S_{YY}(\Omega) - S_{RR}(\Omega). \tag{4}$$

Esta é a base do algoritmo, que foi implementado da seguinte maneira:

$$|Y_{\mathrm{den}}[k]| = \sqrt{|Y[k]|^2 - \min\left(|Y[k]|^2, \alpha|R[k]|^2\right)}, \tag{5}$$

$$\angle Y_{\mathrm{den}}[k] = \angle Y[k], \tag{6}$$

sendo $Y_{\mathrm{den}}[k]$ a DFT do sinal restaurado. O fator $\alpha \geq 0$ permite que o usuário regule a redução de ruído empiricamente. No caso em que $\alpha|R[k]| > |Y[k]|$, o ganho é forçado a zero. E o sinal restaurado é obtido pela DFT inversa:

$$y_{\mathrm{den}}[n] = \mathrm{IDFT}(Y_{\mathrm{den}}[k]). \tag{7}$$

Deve-se notar que este método não estima a fase do sinal restaurado, uma vez atua somente sobre o módulo do sinal. Reutiliza-se, por isso, a própria fase do sinal degradado. Este é um ponto fraco da subtração espectral.

2.1.3 Filtragem de Wiener

O filtro de Wiener é muito utilizado em aplicações de eliminação de ruído. Para o problema descrito no início da Seção 2.1, ele é o filtro que minimiza o erro quadrático médio da estimativa de $x[n]$ a partir de $y[n]$, e tem função de transferência

$$H[k] = \frac{S_{XX}(\Omega)}{S_{YY}(\Omega)} = \frac{S_{YY}(\Omega) - S_{RR}(\Omega)}{S_{YY}(\Omega)}. \tag{8}$$

Implementou-se o filtro de Wiener na forma abaixo:

$$Y_{\mathrm{den}}[k] = Y[k]\left[1 - \min\left(1, \frac{\alpha|R[k]|^2}{|Y[k]|^2}\right)\right], \tag{9}$$

sendo $Y_{\text{den}}[k]$ a DFT do sinal restaurado. O fator $\alpha \geq 0$ permite que o usuário regule a redução de ruído empiricamente. No caso em que $\alpha|R[k]| > |Y[k]|$, o ganho é forçado a zero. E o sinal restaurado é obtido pela DFT inversa:

$$y_{\text{den}}[n] = \text{IDFT}(Y_{\text{den}}[k]). \tag{10}$$

Novamente, o sinal restaurado reaproveita a fase do sinal degradado.

2.2 *Wavelet shrinkage*

A *Discrete Wavelet Transform* (DWT) [3] representa um sinal contínuo no tempo pela ponderação de versões escaladas e transladadas de uma função denominada *wavelet*-mãe, geralmente de suporte finito. Os coeficientes de ponderação que compõem a transformada são discretos, pois os fatores de escala e as parcelas de deslocamento também o são. Quanto mais a função *wavelet*-mãe se parecer com o sinal, mais compactamente ele será representado pela DWT, isto é, menos coeficientes com valores significativos serão necessários para aproximá-lo. Sob algumas hipóteses de aproximação, os coeficientes da DWT podem ser calculados diretamente a partir da versão amostrada do sinal. E finalmente, a transformação assim definida pode ser convenientemente implementada em blocos, convertendo um vetor com N amostras do sinal $y[n]$ num outro vetor com N amostras de sua versão transformada $D[k]$. Assim, com certa liberdade de notação, pode-se representar o resultado da DWT do segmento correspondente do sinal $y(t)$ como:

$$D[k] = \text{DWT}(y[n]). \tag{11}$$

Remetendo mais uma vez ao problema descrito no início da Seção 2.1: o método *wavelet shrinkage* se baseia no fato de que se uma *wavelet*-mãe representa bem o sinal $x[n]$ e mal o ruído $r[n]$ (o que é particularmente verdade se este é branco e descorrelacionado do sinal), os coeficientes elevados da DWT correspondente devem estar associados ao sinal, e os resíduos de pequena amplitude devem estar associados ao ruído. Propõe, então, a aplicação de um limiar δ sobre os coeficientes da DWT, tal que:

$$\begin{aligned} \text{se } D[k] &\geq \lambda, D'[k] = D[k] \\ \text{se } D[k] &< \lambda, D'[k] = g(D[k]) \end{aligned}$$

onde $D'[k]$ são os coeficientes após a *wavelet shrinkage*, e $g(\cdot)$ é uma função atenuadora tal que $|g(D[k])| < |D[k]|$. Se $g(D[k]) = 0$, esse procedimento é chamado de *hard-thresholding*.

Neste trabalho, implementou-se *hard-thresholding*, sendo o limiar calculado segundo

$$\lambda = \sigma\sqrt{2 \ln W}, \tag{12}$$

sendo W o tamanho de cada bloco do sinal e σ o desvio-padrão do ruído. O processamento é feito em blocos com sobreposição de 50%.

Analogamente ao que se fez para os métodos espectrais, neste *plugin*, λ é pré-multiplicado por $\alpha \geq 0$ para permitir que o usuário ajuste empiricamente o limiar. Naturalmente, após a obtenção dos coeficientes $D'[k]$, obtém-se o sinal tratado $y_{\text{den}}[n]$ pela DWT inversa:

$$y_{\text{den}}[n] = \text{IDWT}(D'[k]). \tag{13}$$

3 PRÉ- E PÓS-EQUALIZADORES

O sinal de áudio degradado freqüentemente possui distorções em seu espectro, em relação ao som efetivamente emitido por ocasião da gravação, principalmente se digitalizado a partir de uma mídia analógica. Assim, um equalizador sempre deve fazer parte de um sistema de restauração. Além de atuar no equilíbrio das freqüencias originais, os equalizadores também podem ajudar na restauração propriamente dita.

Em muitos casos de gravações degradadas, a informação espectral do áudio de interesse está restrita a uma faixa de freqüências, enquanto o ruído de fundo se espalha por todo o espectro. O pré-equalizador pode ser utilizado para retirar do sinal faixas de freqüência que contenham somente ruído. Isto pode permitir a redução da intensidade da intervenção automáica no sinal.

Os métodos de tratamento descritos acima podem gerar efeitos não desejados no sinal. No caso da subtração espectral e do filtro de Wiener, que atuam diretamente no domínio da freqüência, pode ocorrer o chamado ruído musical. Este tipo de ruído é causado por pequenos picos isolados deixados no espectro como resíduo do tratamento; ele é especialmente perturbador por conter componentes tonais. O pós-equalizador pode amenizar ou até mesmo eliminar este problema.

Os dois métodos baseados na freqüência já realizam uma DFT, como mostrado acima. Assim, se implementados nesse domínio, os equalizadores adicionam pouquíssima complexidade computacional ao sistema, apenas pela aplicação de ganhos nas bandas da DFT. Foi a solução adotada.

Para cada equalizador, são dados ao usuário 10 controles de ganho, cada um correspondendo a uma oitava do espectro. As faixas de atuação de cada ganho são obtidas dividindo-se a freqüência máxima de 22,05 kHz (metade da freqüência de amostragem do sinal de entrada) sucessivamente por dois. Os ganhos fornecidos pelo usuário são, então, atribuídos aos canais da FFT contidos em cada faixa. Para evitar mudanças abruptas, é feita uma suavização da curva pela simples média dos 3 valores adjacentes:

$$g'[k] = \frac{g[k-1] + g[k] + g[k+1]}{3} \tag{14}$$

onde $g[k]$ é o ganho fornecido pelo usuário e $g'[k]$ é o ganho que será de fato atribuído ao k-ésimo canal da DFT. A Figura 2 exemplifica o esquema acima para uma DFT de tamanho 2048. Vale ressaltar que os ganhos são fornecidos em dB, numa faixa de -12dB a

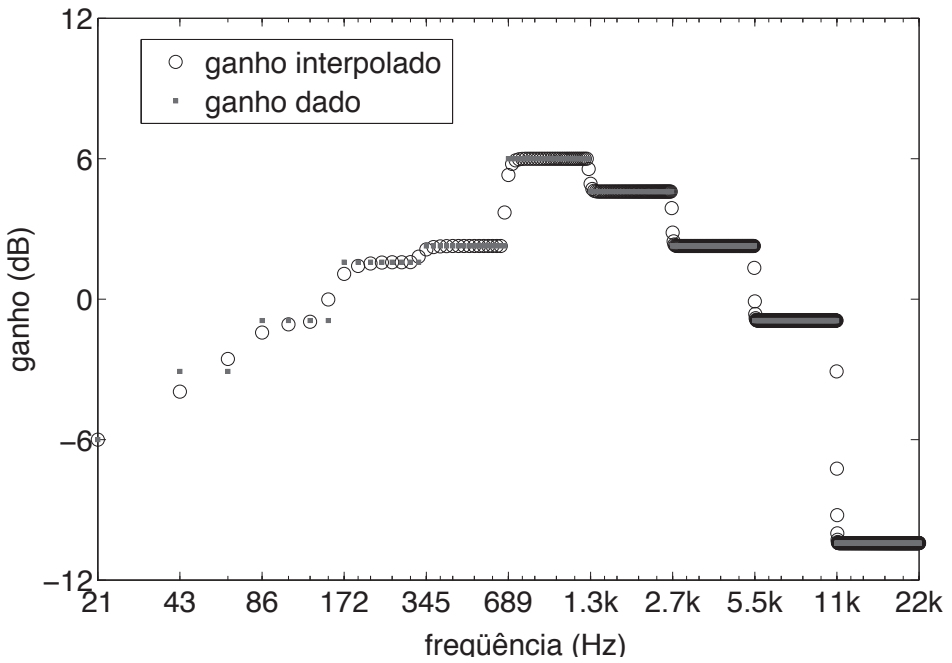


Figura 2: Ganhos do equalizador - exemplo.

+12dB, o que resulta num ganho ou atenuação na amplitude de até 4 vezes.

4 IMPLEMENTAÇÃO

Além do fato de o VST ser um dos padrões do mercado, outro ponto contruibuiu para a decisão da sua escolha para implementação deste trabalho. O *Software Development Kit* (SDK), código necessário para se desenvolver o *plugin*, pode ser obtido e usado livremente [5], ao contrário do que ocorre com outros padrões. Além disso, um *plugin* que utilize o SDK pode ser distríbuido ou comercializado sem restrições, apesar de o VST ser uma marca registrada por empresa privada.

Existe, porém, um pequeno problema em relação à licença do SDK. Seu código não pode redistribuído, mesmo que inalterado. Como todo *plugin* VST contém o código do SDK, a distribuição desses *plugins* como *open-source* (de código aberto) fica dificultada. Isso pode se constituir num entrave aos novos programadores.

Um comentário adicional: a comunidade de desenvolvedores é bem ativa, e os próprios mantenedores do padrão participam dela. O forum de discussão associado constitui uma grande base de dados onde dúvidas podem ser sanadas [6].

A seguir, é feita um descrição dos tópicos de implementação de maior interesse, neste trabalho.

4.1 Função *processReplacing*

Esta é a principal função de um *plugin* VST. O *host* a chama repetidas vezes durante a reprodução do sinal sonoro, fornecendo como argumentos um ponteiro para as amostras da entrada, outro para as de saída e o número de amostras N_s que podem ser acessadas.

O número N_s de amostrar que chega a cada chamada da função não é fixo. Ele pode variar, dependendo do *host* ou até mesmo dentro de um mesmo *host*. Por outro lado, todo o processamento precisa ser feito em janelas cujo tamanho W é escolhido pelo usuário. Assim, a menos que $N_s = kW$, k inteiro, devemos armazenar amostras entre as chamadas da função *processReplacing*.

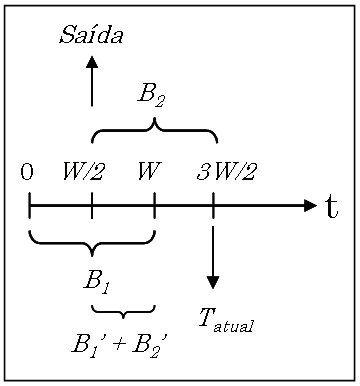


Figura 3: B_1, B_2: *Buffers*; Efeito da sobreposição de janelas. B_1': Segunda metade do *Buffer* 1 processado; B_2': Primeira metade do *Buffer* 2 processado;

Outro ponto a ser considerado é a sobreposição de 50% entre as janelas de análise. As amostras de uma janela só podem enviadas à saída depois que a janela seguinte tiver sido processada, para que a soma das parcelas sobrepostas seja efetuada. Isto resulta num atraso

intrínseco ao sistema de W amostras, como mostrado na Figura 3.

Para lidar com essa situação, foram implementados dois *buffers*. Enquanto o início do primeiro é preenchido com as amostras que chegam, o segundo recebe as mesmas amostras, mas começa a ser preenchido pela posição central. Isto garante que sempre teremos as mesmas amostras na primeira metade de um *buffer* e na segunda metade do outro. Como o janelamento foi feito utilizando a janela de *Hamming*, garante-se que sem processamento a soma dos *buffers* sobrepostos reconstrói perfeitamente o sinal. A seguir, a descrição do algoritmo correspondente:

```
n1 = 0;
n2 = W/2;

for (i = 0; i<Ns; i++){
  buffer1[n1] = entrada[i];
  buffer2[n2] = entrada[i];

  if (n1 == W){
    processe(buffer1);
    for(j = 0; j<W/2; j++)
      saida[j] = buffer1processado[j]
      + buffer2processado[j+W/2];

    n1 = 0;
  }

  if (n2 == W){
    processe(buffer2);
    for(j = 0; j<W/2; j++)
      saida[j] = buffer2processado[j]
      + buffer1processado[j+W/2];

    n2 = 0;
  }

  n1++;
  n2++;
}
```

4.2 Interface Gráfica

A interface gráfica do sistema, mostrada na Figura 4, foi desenhada utilizando a biblioteca VSTGUI [7]. Esta é escrita em C++ e foi especialmente desenvolvida para uso em *plugins* VST, possuindo classes para a criação de controles como *sliders*, *knobs* ou *switches*.

Neste *plugin*, foram utilizados *sliders* verticais para os equalizadores, para o parâmetro α e para tamanho da janela de análise. Para a escolha do modo de funcionamento e do tipo de ruído a ser utilizado, foram usados *switches* horizontais.

4.3 Transformada rápida de Fourier

A equalização e os métodos de subtração espectral e filtragem de Wiener são realizados no domínio da

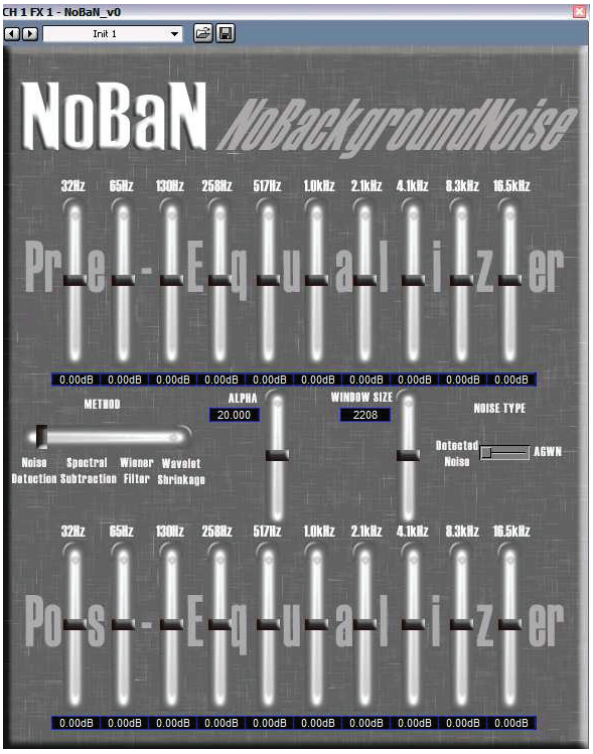


Figura 4: Interface gráfica do *plugin*

freqüência. Para tal, utilizou-se a biblioteca FFTW [8], que realiza o algoritmo da FFT (*fast Fourier transform*, versão rápida da DFT) de maneira otimizada para cada processador.

O tamanho da janela de análise é definido pelo usuário, no entanto a DFT é realizada sempre com o tamanho da menor potência de 2 acima de W, para diminuir sua complexidade. Vale ressaltar que a DFT implementada pela biblioteca não realiza nenhuma normalização; por isso suas amostras de saída foram divididas pelo tamanho da transformada.

4.4 *Discrete wavelet transform*

O código para implementação da DWT foi retirado do livro *Numerical Receipes In C* [9]. Foi utilizada a função-mãe *Daubechies 44*. Assim como no caso da FFT, o tamanho do vetor de entrada deve ser uma potência de 2 para que seja usado um algoritmo mais eficiente.

Um próximo ajuste no sistema seria deixar outras funções-mãe disponíveis para que o usuário possa escolher a mais adequada ao sinal.

5 RESULTADOS

O *plugin* aqui descrito foi utilizado para realizar alguns testes de redução de ruído sobre sinais reais. Obteve-se operação em tempo real para os três métodos, e o resultado dos ajustes dos parâmetros pôde ser ouvido *online*. Os equalizadores também se mostraram de grande valia, já que o simples resgate do

equilíbrio espectral de um sinal degradado já proporciona uma sensação auditiva mais agradável. Além disso, houve casos em que a forte incidência de ruído nas extremidades do espectro pôde ser amenizada diretamente com o pré-equalizador, sem afetar sensivelmente o sinal.

De acordo com testes subjetivos informais, os resultados do processamento pelos diversos métodos confirmaram as expectativas. A subtração espectral e o filtro de Wiener mostraram resultados próximos. Aumentando-se α, para sinais com SNRs mais baixas, percebe-se primeiro a ocorrência de ruído musical, devido ao resíduo de ruído, que forma pequenos picos no espectro. Após um certo valor de α, o ruído musical é eliminado, ao custo de se impor distorção ao sinal, notadamente um abafamento, já que se eliminam componentes de alta freqüencia, que normalmente possuem menor energia. Fica claro que a eliminação do ruído musical é uma meta a atingir, permitindo trabalhar com menores valores de α.

Apenas como ilustração, as Figuras 5, 6 e 7 mostram um exemplo simples de processamento por um método espectral. A Figura 5 mostra um trecho de uma gravação de clarinete. A este sinal foi adicionado ruído branco gaussiano com média zero de modo a simular uma razão sinal-ruído de 20dB, gerando o sinal da Figura 6. O sistema analisou o ruído e efetuou o tratamento utilizando a subtração espectral com uma janela de 1024 amostras e um fator α = 4. O resultado pode ser observado na Figura 7.

O método *wavelet shrinkage* mostrou-se pouco eficaz em termos da rejeição de ruído, numa primeira seqüência de testes. Contudo, dois fatores a tornam uma solução bastante promissora:

- É possível buscar uma família de *wavelets* a empregar na DWT que seja mais adequada ao tipo de sinais sob processamento.
- Por atuar eliminando parcelas da aproximação, o tipo de efeito causado sobre o sinal pode soar mais natural, ou menos incômodo, do que os gerados pelos outros métodos. Isso irá depender da família de *wavelets* utilizada.

O *plugin* está disponível em versão alfa, juntamente com instruções de uso e instalação, em http://www.lps.ufrj.br/∼alan.

6 TRABALHOS FUTUROS

Dispondo-se do sistema-base para os métodos de restauração, o próximo passo é aprimorar a remoção de ruído propriamente dita, o que passa por duas fases.

Na primeira, é possível tornar mais flexíveis e robustas as técnicas já implementadas. Pode-se, por exemplo: permitir a escolha do tamanho da DFT, de outras janelas que não a de Hamming, do fator de superposição entre blocos, de outras funções atenuadoras na "wavelet shrinkage" etc.; atacar especificamente

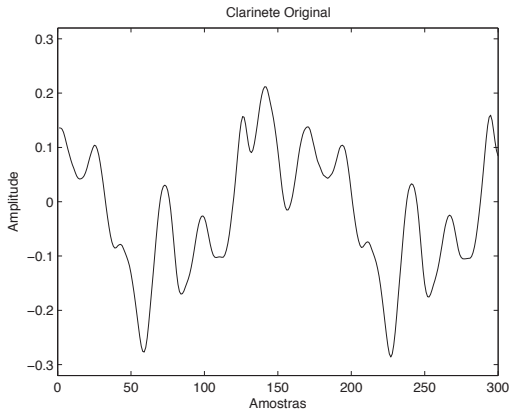


Figura 5: Trecho de uma gravação de clarinete.

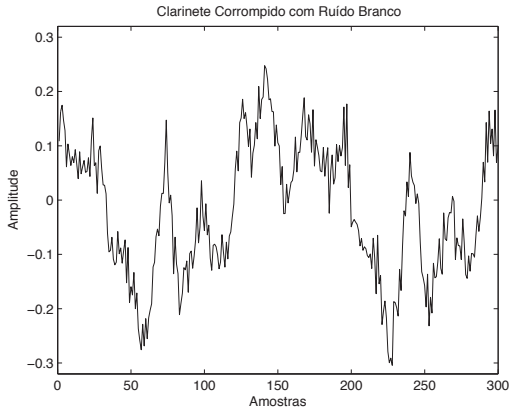


Figura 6: Clarinete corrompido com ruído branco (AWGN).

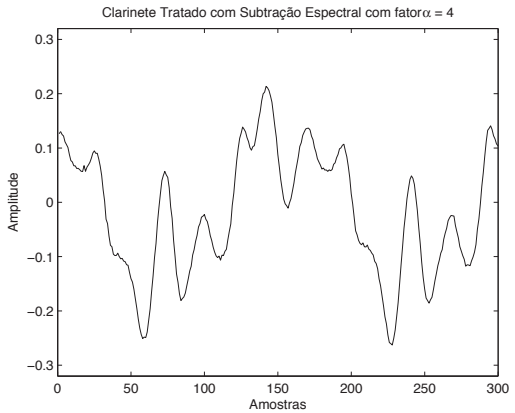


Figura 7: Clarinete tratado utilizando a subtração espectral, com fator α = 4.

o ruído musical, característico dos métodos espectrais de redução de ruído, com a aplicação de uma média ou mediana entre *frames* contíguos do sinal [2], ou com o método de Ephraim-Malah [10].

Na segunda, serão acrescentados métodos alternativos de redução de ruído. Em [4], por exemplo,

descreve-se um algoritmo para detecção da envoltória espectral do ruído que pode ser combinado aos métodos espectrais para evitar que se precise dispor de um trecho contendo somente ruído.

Outro campo bastante promissor parece ser o da análise senoidal: nesse caso, a filosofia é retirar o sinal original, e não o ruído, do sinal corrompido [11]. Além disso, há a possibilidade de se usar métodos no domínio do tempo, como os baseados na filtragem de Kalman [12].

Em todas estas propostas, é preciso não perder de vista o compromisso entre remoção de ruído e degradação do sinal original.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Biscainho, L. W. P., *Restauração Digital de Sinais de Áudio Provenientes de Gravações Musicais Degradadas,* Tese de Doutorado, PEE/COPPE, UFRJ, 2000, Rio de Janeiro, Brasil. Disponível em `http://www.pee.ufrj.br/teses/?Resumo=2000127001`.

[2] Godsill, S. J. and Rayner, P. J. W., *Digital Audio Restoration, A Statistical Model Based Approach*, Springer, 1998.

[3] Burrus, C. S., Gopinath, R. A., Guo, H. *Introduction to Wavelets and Wavelet Transforms*, Prentice Hall, 1998.

[4] Laurenti, M., De Poli, G., Montagner, D., *A Nonlinear Method for Stochastic Spectrum Estimation in the Modeling of Musical Sounds*, IEEE Transactions on Audio, Speech and Language Processing 15(2), 531–541 (2007).

[5] Steinberg Media Technologies (VST): http://www.steinberg.net/324_1.html

[6] Repositório da lista de discussão sobre VST: http://lists.steinberg.net:8100/

[7] VSTGUI: http://sourceforge.net/projects/vstgui/

[8] Frigo, M. and Johnson, S. G., *The Design and Implementation of FFTW3*, Proceedings of the IEEE 93 (2), 216–231 (2005). Invited paper, Special Issue on Program Generation, Optimization, and Platform Adaptation.

[9] Press, W. H., Teukolsky, S. A., Vetterling, W. T. e Flannery, B. P., *Numerical Recipes in C, The Art of Scientific Computing*, 2nd. ed., Cambridge, 1988.

[10] Cappé, O., *Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor.* IEEE Trans. Speech and Audio Processing, 2(2), 345–349, 1994.

[11] Master, A., *Physical Modeling and Sinusoidal Modeling for Noise and Artifact Elimination.* Music 421 / EE 367B CCRMA Technical Report, Stanford University 2001, Stanford, USA.

[12] Bari, A., Canazza, S., De Poli, G., Mian, G. A., *Some key points on restoration of audio documents by the Extended Kalman Filter.* Proc. of Diderot Forum on Mathematics and Music, Vienna, pp. 37–47, 1999.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

O PEAQ Aplicado a Sinais de Áudio Contaminados por Ruído

Rafael A. de Jesus[1] e Luiz W. P. Biscainho[1]

[1] UFRJ, LPS – PEE/COPPE
Caixa Postal 68504, Rio de Janeiro, RJ, 21941-972, Brasil

rjesus@lps.ufrj.br, wagner@lps.ufrj.br

RESUMO

Este trabalho é motivado pela necessidade de avaliação de métodos de restauração de áudio através de medidas objetivas da qualidade dos resultados obtidos. O PEAQ é um avaliador objetivo de base psicoacústica definido pela ITU que mede a diferença pecebida entre um sinal de referência e sua versão submetida a um codec com perdas. Aqui investiga-se o uso do PEAQ na medida da qualidade de sinais contaminados por ruído não originário de codificação, pela comparação de seus resultados com os de testes subjetivos realizados em condições controladas.

0 INTRODUÇÃO

Métodos confiáveis de avaliação de qualidade de áudio são objeto de pesquisa há décadas. Numa primeira fase, foram motivados pela necessidade de se avaliar a fidelidade dos sistemas de armazenamento e reprodução de áudio, conforme iam sendo desenvolvidos. Na era digital, a necessidade de se avaliar a transparência dos métodos de compressão de áudio com perdas deu nova motivação à área. Contudo, o horizonte de aplicação dos métodos de avaliação de qualidade se estende a outras áreas, que vão da restauração de gravações musicais históricas à difusão de áudio pela Internet.

Na medida em que são baseados na avaliação direta do som por pessoas, os chamados métodos subjetivos de avaliação de qualidade de áudio produzem resultados mais acurados e coerentes com a qualidade percebida pelo ouvinte médio. Já foram desenvolvidas padronizações para esse tipo de avaliação [1], que entretanto recaem em duas dificuldades: reunir um grupo numeroso de indivíduos com audição seleta para realizar os testes e garantir certas condições repetíveis para os testes e o ambiente. Tais fatores tornam os métodos subjetivos custosos e demorados.

Esse problema motivou o desenvolvimento de métodos objetivos que realizassem a avaliação de qualidade dispensando a participação de pessoas, ou seja, através do processamento computacional dos sinais de áudio. Medidas diretas tradicionais como a razão sinal-

ruído (SNR) são incapazes de discernir ruído audível de ruído inaudível. Por isso, os métodos objetivos de avaliação de qualidade de áudio foram aprimorados pela inserção de critérios psicoacústicos: a meta é avaliar a qualidade percebida, emulando o julgamento humano. Objeto da literatura desde a década de 80 [2], esses métodos usualmente medem a diferença percebida entre o sinal a avaliar e um sinal de referência (em vez de sua diferença absoluta). Em 1998, a ITU (*International Telecomunication Union*) padronizou na recomendação BS.1387 [3] o avaliador PEAQ (*Perceptual Evaluation of Audio Quality*), combinando o trabalho de diversos pesquisadores na área.

O PEAQ se tornou referência na avaliação objetiva de sinais de áudio de alta qualidade. Assim como a maior parte dos avaliadores objetivos envolvidos no seu desenvolvimento [4], foi definido para avaliar a transparência de codificadores perceptivos de áudio, pela quantificação das distorções perceptíveis inseridas pelo processo de codificação. Sendo um sistema resultante de treinamento, o PEAQ garante elevada correlação com a avaliação subjetiva desse tipo de distorção. Mas por esse mesmo motivo, pode mostrar-se inadequado em outros contextos que requeiram a avaliação perceptiva da qualidade do som.

O objetivo deste trabalho é avaliar o desempenho do PEAQ na avaliação de sinais contaminados com ruído não gerado por codificação. Sua principal motivação é verificar a adequação do PEAQ para a avaliação de métodos de restauração desses sinais. O artigo está organizado da seguinte forma: na Seção 1, é apresentada a estrutura do PEAQ; na Seção 2, são descritos os testes realizados, cujos resultados são apresentados na Seção 3; a Seção 4 sumariza as conclusões.

1 A ESTRUTURA DO PEAQ

Como a maioria dos métodos objetivos de avaliação de qualidade de áudio, o PEAQ é um avaliador com referência, i.e., avalia um sinal de áudio através da sua comparação com um sinal de referência.

Conforme pode ser visto na Fig. 1, o PEAQ realiza primeiramente o mapeamento tempo-freqüência dos sinais; em seguida, aplica a saída do mapeamento ao modelo psicoacústico, que fará a análise perceptiva dos sinais, considerando diversos efeitos como o limiar absoluto de audibilidade e o mascaramento; no terceiro bloco, o resultado do modelo psicoacústico é processado de diversas formas, gerando as chamadas MOVs (*Model Output Variables*); esse conjunto de variáveis alimenta a rede neural (pré-treinada) que fornecerá a nota final. Deve-se ressaltar que nos três primeiros blocos o processamento do sinal de teste e do sinal de referência é feito de maneira independente.

A rede neural incluída no sistema é treinada para um extenso conjunto de arquivos degradados de áudio. À entrada da rede são apresentadas as MOVs calculadas para um dado arquivo e à sua saída, a nota média que a este foi atribuída em testes subjetivos formais. A

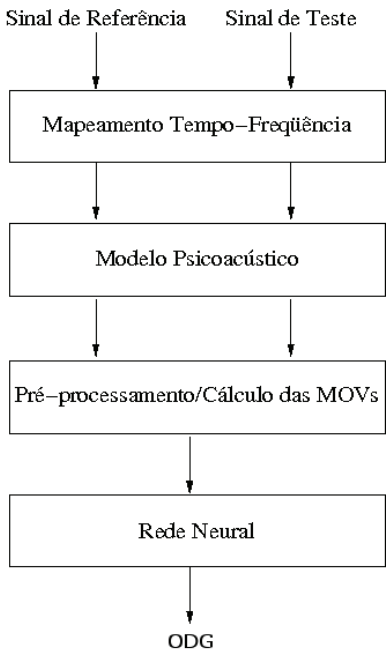


Figura 1: Diagrama simplificado dos blocos do PEAQ.

nota que a rede fornece segue a escala ODG (*Objective Difference Grade*) de diferença em relação a uma referência. A ODG corresponde à SDG (*Subjective Difference Grade*), que se refere a testes subjetivos e cuja graduação (com os respectivos significados) é mostrada na Tab. 1. Obviamente, o treinamento da rede foi realizado através das SDGs.

Tabela 1: Notas e respectivos significados da SDG/ODG.

Nota	Descrição da diferença
0	Imperceptível
-1	Perceptível, porém não incômoda
-2	Pouco incômoda
-3	Incômoda
-4	Bastante incômoda

O PEAQ possui dois modos de implementação, chamados básico e avançado. O modo básico é menos complexo, quatro vezes mais rápido que o avançado, e se destina a aplicações em tempo real; o modo avançado envolve processamento mais pesado, mas fornece resultados mais acurados, chegando a uma correlação de 95% com os resultados dos testes subjetivos [5].

O fato de o PEAQ ser um sistema fixo, especializado pelo seu treinamento em sinais degradados por ruído de codificação, faz supor que seu desempenho na avaliação de outros tipos de ruído possa não ser aceitável.

2 TESTES REALIZADOS

A fim de se avaliar o desempenho do PEAQ na análise da qualidade de sinais de áudio contaminados

com diversos tipos de ruído, serão comparados os resultados por ele fornecidos com os resultados de testes subjetivos realizados em condições controladas buscando seguir orientações da Recomendação BS.1116-1 do ITU-R [1]. Devido à dificuldade de cumprir todas as orientações expressas naquela recomendação, os testes não puderam segui-las estritamente, o que torna os resultados obtidos apenas indicativos.

Vinte ouvintes jovens sem histórico de problemas de audição foram submetidos a um teste *double-blind triple-stimulus with hidden reference* (duplamente cego com estímulo triplo e referência oculta). Primeiramente apresentaram-se aos ouvintes amostras dos ruídos utilizados para a contaminação dos sinais. No passo seguinte os ouvintes avaliaram as amostras, tendo sido primeiramente apresentados ao sinal original e em seguida a dois outros sinais (um deles, igual ao original). Os ouvintes foram informados de que um dos dois sinais era igual ao original e portanto deveria receber nota máxima.

2.1 ESCOLHA DOS ARQUIVOS

As amostras de áudio escolhidas para servir como sinais de referência para o teste deveriam atender às seguintes características:

- serem representantes de uma classe abrangente de sinais de áudio;
- não estarem originalmente contaminados com ruído.

De acordo com os critérios apresentados, selecionaram-se os 5 arquivos descritos a seguir, extraídos de gravações digitais de alta qualidade:

Voz masculina falada–VMF: Trecho de 15 segundos onde um crítico fala em língua inglesa sobre um pianista. A escolha desse arquivo é motivada pelo fato de a avaliação de fala tender a ser distinta da de música e pela presença de trechos de silêncio. Vale ressaltar que a utilização de um sinal em língua estrangeira objetivava que os ouvintes se concentrassem menos no que estava sendo dito e mais na qualidade do sinal.

Samba–SMB: Trecho de 15 segundos de música predominantemente percussiva. Esse arquivo foi escolhido por ser característico da música popular brasileira, e pelo fato de a percussão tender a se confundir com o ruído, e por se tratar de um *crescendo*, com o acréscimo gradual de instrumentos.

Sivuca–SVC: Trecho de 15 segundos de uma peça para acordeom. Esse arquivo, também representativo da música popular brasileira, foi escolhido por possuir trechos de silêncio e pelo timbre rico em harmônicos do instrumento.

Vôo–VOO: Trecho de 15 segundos de uma peça para piano. Esse arquivo foi escolhido por se tratar de música para piano, extremamente rápida e articulada.

Voz feminina cantada–VFC: Trecho de 10 segundos de canto *a capella* por uma intérprete de língua inglesa. A escolha desse arquivo se explica por conter voz numa forma musical.

O número e a escolha dos arquivos que seriam utilizados em cada um dos testes foi feita de maneira a manter um compromisso entre lidar com alguma diversidade de características e manter a duração dos testes dentro dos limites aceitáveis.

2.2 PROCESSAMENTO DOS ARQUIVOS

Definiram-se, então, 4 categorias de testes, a saber:

Testes com ruído de fundo: O ruído aleatório de fundo é uma forma de degradação comum a todos os sistemas de medida e gravação de sinais analógicos. Em sinais de áudio, é normalmente percebido pelo ouvinte como um *hiss* e é originário dos ruídos oriundos dos circuitos elétricos, de não-idealidades das mídias de armazenamento (com destaque para os característicos do armazenamento em fitas magnéticas) e de ruído ambiente da gravação. O efeito combinado dessas fontes normalmente é encarado como um único ruído, possuindo geralmente componentes significativas em todas as freqüências de áudio [6].

Para a contaminação dos sinais com ruído de fundo, gerou-se artificialmente por *software* ruido pseudo-aleatório branco gaussiano. Contaminaram-se aditivamente os sinais VMF, SMB e SVC com três níveis diferentes deste ruído, de forma a obter razões sinal-ruído de 40, 45 e 50 dB.

Testes com cliques: O termo 'clique' normalmente é utilizado fazendo referência a um tipo localizado de degradação, comum em diversas mídias de áudio. Numa classificação simplificada, consideram-se cliques os defeitos de duração curta (até unidades de ms) e amplitude moderada que ocorrem em instantes aleatórios. As causas dos cliques vão desde pequenas deposições ou ranhuras em discos até falhas na mídia digital. Podem soar isolados como pequenos estalidos ou, mais comumente, como um 'fritar' aparentemente contínuo. Num disco antigo de 78 rpm não é raro haver cerca de 2000 cliques por segundo de gravação [6].

Obtiveram-se os cliques que seriam adicionados aos sinais a partir de duas gravações reais de áudio contaminadas. Esse procedimento, em lugar do uso direto de sinais naturalmente contaminados, se deveu ao fato do PEAQ necessitar de um sinal de referência. Os dois sinais apresentavam uma contaminação por cliques de $0{,}6\%$ e 5% de suas

amostras. Cada um deles foi inserido com dois ganhos distintos, calibrados manualmente para produzir sensações equivalentes a dois níveis subjetivos de ruído: 'alto' e 'baixo'. Os arquivos SVC e VOO foram os escolhidos para a realização deste conjunto de quatro testes.

Testes com pulsos longos: Característicos de discos analógicos com arranhões profundos ou até mesmo partidos, os pulsos longos possuem sua forma de onda dependente das características eletro-mecânicas do sistema de reprodução, gerando, tipicamente, uma rápida descontinuidade de alta amplitude seguida de um transitório de decrescente de baixa freqüencia, que pode durar centenas de ms. A sensação auditiva desse tipo de ruído se aproxima de um estalo de alta intensidade.

Modelou-se um pulso longo característico conforme explicado em trabalho anterior de um dos autores [7]. Em seguida, ele foi aplicado aos sinais SVC e VOO em quatro versões. Como uma seqüência de pulsos com freqüência de aparecimento simulando um arranhão profundo num disco de vinil de 45rpm e amplitude constante, em dois níveis , 'baixo' ou 'alto'; e simulando o início ou o fim de um arranhão profundo num disco de vinil de 30rpm, ou seja, com sua amplitude diminuindo e aumentando, respectivamente.

Teste com ruídos múltiplos: Em grande parte dos registros de áudio contaminados com ruído, não se encontra apenas um único tipo de ruído, e sim uma combinação de diversos deles. Para avaliar o PEAQ também nessa situação real, contaminou-se o sinal VFC com os três tipos de ruídos utilizados nos testes anteriores com dois níveis subjetivamente escolhidos de contaminação, chamados 'alto' e 'baixo'. Em ambos os níveis a ponderação dos três tipos de ruído foi feita empiricamente visando a equilibrar sua percepção.

3 RESULTADOS

Para se obter uma visualização dos resultados que permitisse uma análise mais fácil, separou-se o total de 27 testes nas quatro categorias descritas na seção anterior. Os resultados de cada uma serão discutidos tanto dentro de cada categoria quanto entre categorias.

As Figs. 2, 3, 4 e 5 mostram, para os testes de cada uma das categorias, o diagrama *box-and-whisker*[1], a média da distribuição (representada com o símbolo '$\times$') e o resultado obtido através do PEAQ (representada com o símbolo '$\circ$'). No topo de cada uma das figuras se encontra indicado, quando necessário, a que arquivos se referem os testes representados.

[1] O diagrama box-and-whisker ilustra uma distribuição identificando os pontos onde se encontram o primeiro quartil, a mediana, o terceiro quartil e os pontos de máximo e mínimo estimados – que não podem ultrapassar 1,5 vezes o tamanho da caixa (distância entre o primeiro e o terceiro quartis). Os pontos além desses limites (*outliers*) são indicados com o símbolo '$+$'.

3.1 Testes com ruído de fundo

Na Fig. 2, que apresenta os resultados obtidos para os testes com ruído de fundo, é possível observar que:

- Para o arquivo VMF, o resultado absoluto do PEAQ foi muito distante do dos testes subjetivos. Dois fatores talvez possam explicar isso: a ocorrência de silêncios, que, apesar de tratados pelo PEAQ, podem perturbar seu desempenho – o que requer melhor investigação; o fato de se tratar de voz falada, que os ouvintes podem avaliar de forma menos exigente que sinais musicais, já que no contexto da comunicação verbal se dá mais importância à compreensão da mensagem que à fidelidade do sinal.

- Salvo num caso, o PEAQ tendeu a dar notas menores que os ouvintes, mas com alta correlação.

- O arquivo SMB tendeu a receber notas subjetivas melhores, na média, que os outros sinais. Por outro lado, a ocorrência de *outliers* foi exagerada, especialmente para SNR alta. Isso se explica pela dificuldade de distinção entre o sinal percussivo e o ruído aditivo. O PEAQ teve boa coincidência com as medidas subjetivas.

3.2 Testes com cliques

Na Fig. 3, que apresenta os resultados obtidos para os testes com cliques, é possível observar que:

- O desempenho do PEAQ para os dois arquivos testados foi bastante parecido, e razoavelmente satisfatório. Isso indica que a presença ou ausência de silêncio perde importância na avaliação da distorção gerada por cliques; o PEAQ parece realizar uma avaliação aceitável deste tipo de distorção.

- Salvo num caso, o PEAQ tendeu a dar notas maiores que os ouvintes.

- A acurácia do PEAQ em relação ao julgamento humano diminuiu com o aumento do nível de ruído.

- Não houve diferença no desempenho do PEAQ na avaliação dos arquivos com diferentes percentuais de contaminação ($0{,}6\%$ ou 5% das amostras). Isso indicou a esperada prevalência dos critérios psicoacústicos sobre a diferença das formas de onda.

3.3 Testes com pulsos longos

A Fig. 4 apresenta os resultados obtidos para os testes com pulsos longos. Ao contrário dos dois conjuntos de testes anteriores, os resultados obtidos não indicam um padrão regular evidente. O que se pode observar é:

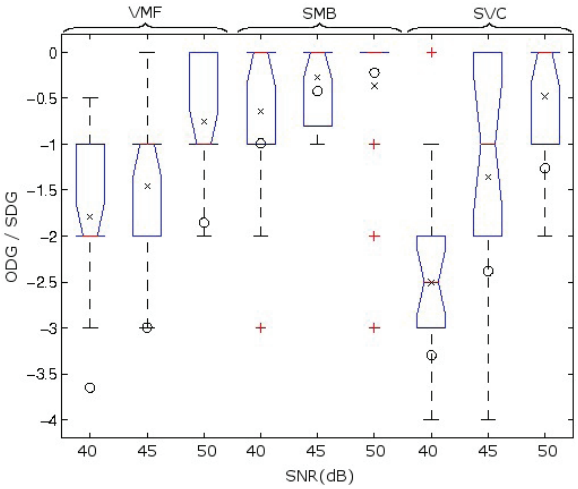


Figura 2: Resultados dos testes com ruido de fundo.

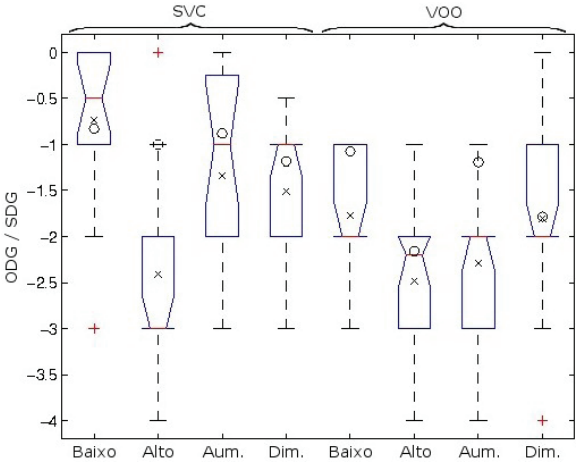


Figura 4: Resultados dos testes com pulsos longos.

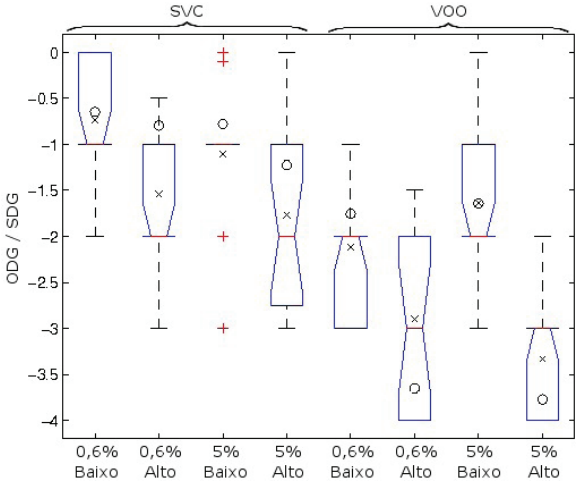


Figura 3: Resultados dos testes com cliques.

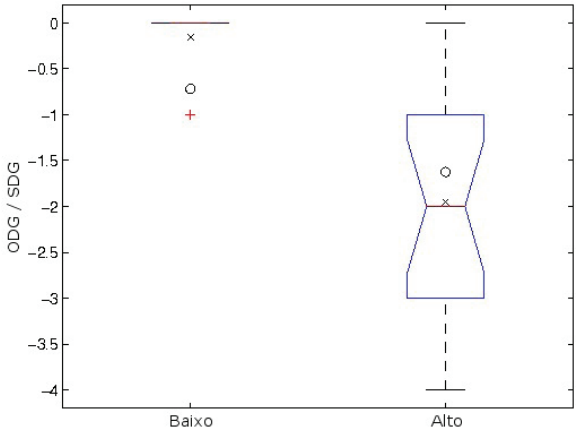


Figura 5: Resultados dos testes com ruídos múltiplos.

- Somente em dois dos oito testes o PEAQ teve um desempenho ruim. Como os dois são diferentes (num há pulsos de nível alto no sinal (SVC) e no outro, pulsos de nível crescente no sinal (VOO), não é possível traçar nenhuma causa possível para isso.

- Salvo num caso, o PEAQ tendeu a dar notas maiores que os ouvintes.

3.4 Testes com ruídos múltiplos

Conforme se pode observar nos resultados ilustrados na Fig. 5:

- Apesar de a diferença ter sido pequena, o resultado do PEAQ na avaliação de ruídos múltiplos foi mais distante dos testes subjetivos para nível baixo que para nível alto de ruído.

- Ocorreu uma extraordinária diferença entre os espalhamentos das avaliações subjetivas nos dois casos.

Nessas condições, e considerando que foi realizado para um único sinal, este teste não forneceu informações de interesse.

4 CONCLUSÕES

Como conclusão geral dos testes realizados, apesar de apresentar diversos resultados acurados, o PEAQ não pôde ser considerado um método confiável para julgar sinais contaminados por ruído, em geral. Sua tendência a se aproximar mais dos testes subjetivos para menores níveis de contaminação é coerente com o fato de ter sido definido para avaliar sinais de alta qualidade; isso também pode explicar o fato de nos casos mais fortemente contaminados o PEAQ atribuir notas menores que os avaliadores. Porém, no contexto deste trabalho ocorre freqüentemente a necessidade de se lidar com níveis elevados de ruído, e.g. quando se trata

da avaliação indireta de métodos para restauração de gravações históricas, que pode ser feita através de sinais contaminados artificialmente.

Apesar da distância absoluta entre o PEAQ e as notas subjetivas, o caso de ruído de fundo apresentou a melhor correlação dentre todos os testes; isso indica que, devidamente recalibrado, o PEAQ pode ser usado para avaliar sinais contaminados com *hiss*. O fato de o PEAQ realizar uma média entre as notas dos diversos *frames* do sinal pode torná-lo inadequado para avaliar coerentemente problemas de ocorrência rara, como pulsos longos; alguns fatores cognitivos associados à localização temporal devem ser levados em conta.

Naturalmente, os testes realizados poderiam ter considerado uma quantidade maior de sinais de referência. Nota-se, também, que as avaliações subjetivas tenderam a apresentar variância elevada, em geral; isso pode refletir a falta de 'expertise' por parte dos avaliadores, que ainda poderiam ter sido pré-treinados. Esses aspectos decorreram do não-cumprimento estrito das orientações da Recomendação BS.1116-1, que redundaria em dificuldades consideráveis numa primeira etapa do trabalho. Então, apesar de não rigorosos, os resultados deste trabalho fornecem informações úteis àqueles que desejam utilizar o PEAQ no contexto proposto: pode-se afirmar que o uso do PEAQ, tal como definido, com ruído em geral só deve ser feito como um indicativo de tendência, não como um resultado absoluto.

Possíveis trabalhos futuros são: validar os testes realizados neste trabalho com a sua replicação em condições idênticas para a avaliação de arquivos degradados por codificação com perdas; examinar a calibração do PEAQ para ruído de fundo; retreinar a rede neural para categorias específicas de ruído; e considerar a utilização de outras MOVs, apropriadas aos tipos de ruído a considerar.

5 AGRADECIMENTOS

Os autores gostariam de agradecer: aos voluntários que realizaram os testes subjetivos; aos revisores, que muito contribuíram para a forma final do texto; ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e à Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), que financiaram este trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] ITU-R Rec. BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," International Telecommunications Union, Geneva, Switzerland, Oct, 1997.

[2] M. Karjalainen, "A new auditory model for the evaluation of sound quality of audio systems," in *Proc. of ICASSP'85 – IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tampa, USA, Mar. 1985, vol. 10, pp. 608–611.

[3] ITU-R Rec. BS.1387, "Method for objective measurements of perceived audio quality," International Telecommunications Union, Geneva, Switzerland, 1998.

[4] 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 Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, Jan./Feb. 2000.

[5] P. Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality," Tech. Rep., McGill University, TSP Lab, Department of Electrical and Computer Engineering, May 2002.

[6] M. Kahrs and K. Brandenburg, Eds., *Applications of digital signal processing to audio and acoustics*, Kluwer, Norwell, MA, USA, 1998.

[7] P. A. A. Esquef, L. W. P. Biscainho, and V. Välimäki, "An efficient algorithm for the restoration of audio signals corrupted with low-frequency pulses," *Journal of Audio Engineering Society*, vol. 51, no. 6, pp. 502–517, June 2003.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Sistemas de Áudio no Controle Ativo de Ruído

Bruno S. Masiero[1] e Vítor H. Nascimento[1]

[1] Universidade de São Paulo, Escola Politécnica, Laboratório de Processamento de Sinais
Av. Prof. Luciano Gualberto, trav. 3, n° 158, São Paulo, SP, 05508-970, Brasil

bruno.masiero@poli.usp.br, vitor@lps.usp.br

RESUMO

A primeira idéia que vem em mente quando falamos de sistemas de áudio é a reprodução/amplificação de sons. A idéia de utilizar alto-falantes e microfones para reduzir os níveis sonoros de um ambiente parece então contraditória. Mas desde meados do século XX a técnica de Controle Ativo de Ruído vem sendo constantemente estudada, apesar de até o momento existirem poucas soluções comerciais. A aplicação mais conhecida desta técnica são os fones de ouvido que buscam reduzir o ruído ambiente escutado pelo usuário, enquanto reproduzindo o sinal sonoro desejado pelo usuário. Outra aplicação já bastante desenvolvida é o controle de ruído emitido por tubulações, como chaminés industriais ou dutos de ar-condicionado. Mas algumas aplicações, apesar da intensa pesquisa realizada, ainda não apresentam soluções comercialmente viáveis, como é o caso do cancelamento do ruído de transformadores de potência.

Este artigo é divido em duas partes. Na primeira parte, realiza-se uma revisão histórica e teórica do controle ativo de ruído, descrevendo as tendências mais recentes nesta área. Na segunda parte descreve-se os experimentos realizados para o controle de ruído ativo em transformadores de potência em campo.

Parte I

1 INTRODUÇÃO HISTÓRICA

O resumo histórico que segue foi baseado no tutorial publicado por Nelson e Elliott em 1993 [1]. Segundo eles, as idéias básicas do controle ativo de ruído já são antigas, sendo a primeira patente deste tipo atribuída a Paul Lueg em 1936 (citado em [1]). Nessa patente é descrito um sistema que capta o campo sonoro com um microfone (microfone de referência), manipula este sinal eletronicamente e injeta o sinal em um duto por meio de uma fonte secundária localizada a jusante do microfone de referência. Lueg previa a aplicação desse aparato para o controle de ruído em dutos, considerando a propagação de uma onda sonora plana. A superposição da onda originada pela fonte primária com a onda gerada pela fonte secundária resultaria em uma interferência destrutiva e, conseqüentemente, silêncio na região a jusante da fonte secundária.

Em 1953, Olson e May (citado em [1]) publicaram um artigo descrevendo um "absorvedor sonoro eletrônico", para ser usado na redução de ruído em aeronaves ou automóveis, tentando gerar uma "zona de silêncio" próxima à cabeça dos ocupantes. Seu sistema dispunha de um microfone próximo a um alto-falante, ambos no encosto de cabeça do assento, e um controlador realimentado tentava minimizar a pressão sonora no microfone, gerando então a "zona de silêncio".

Na mesma época, William Conover (citado em [1]) trabalhava em um sistema de ANC para transformadores de potência. O ruído radiado por esses equipamentos possui algumas características espectrais que tornam propícia a aplicação da técnica ANC no seu controle. Conover realizava o controle das harmônicas manualmente, alterando ganho e fase da freqüência de interesse. Esse sistema necessita de ajustes constantes para corrigir as variações do ruído ocorridas por causa de variações de tensão da rede ou das condições meteorológicas. Conover comentou a possibilidade do uso de um sistema automático de controle, que só foi introduzido em aplicações ANC em 1975 por Kido (citado em [1]). O processamento digital de sinais (PDS) permite a realização, com relativa facilidade, de filtros adaptativos para o ajuste do sinal de cancelamento. Filtros adaptativos ajustam automaticamente seus coeficientes com o objetivo de minimizar um sinal de erro e podem ser realizados de diversas formas [2]. A forma mais comum de filtro adaptativo usado em aplicações de ANC é o filtro de resposta ao impulso finita (FIR) com adaptação por mínimos quadrados (LMS).

2 ASPECTOS PRÁTICOS

Não importa como o sistema de controle é implementado, o princípio acústico em que qualquer sistema de controle ativo de ruído se baseia é o mesmo: o princípio físico da *superposição*. Wright publicou uma série de artigos descrevendo um sistema de controle ativo de ruído ambiental por meio do uso de uma parede de alto-falantes, que geram um campo secundário que tenta cancelar um ruído em campo aberto por meio da superposição de ondas [6].

O princípio da superposição é válido para qualquer sistema linear, e a propagação de uma onda acústica de amplitude não absurdamente elevada (abaixo de 140 dB) pode ser considerado um processo linear. Em ANC, a mais provável causa de não-linearidade está presente nos alto-falantes de bobina móvel usados como fonte secundária, quando, por exemplo, eles são forçados a reproduzir um sinal senoidal de baixa freqüência com alta amplitude. A freqüência de excitação não será ouvida por causa da interferência destrutiva, mas as harmônicas superiores geradas pela não-linearidade do alto-falante nestas condições podem se tornar um incômodo. Em geral, com o uso de alto-falantes adequadamente projetados, a não-linearidade presente será pequena e poderá ser ignorada sem infligir grandes perdas ao sistema.

O sistema de Olson e May propunha que o microfone de erro ficasse posicionado próximo à fonte secundária. Com isso, a amplitude do sinal necessário para o cancelamento da pressão no ponto do microfone seria pequena, e pontos afastados da fonte secundária não seriam afetados. Essa arquitetura produziria atenuação apenas em uma região próxima aos transdutores (a "zona de silêncio"), deixando o campo distante praticamente inalterado. Isso permite a atenuação de baixas freqüências ao redor da cabeça de um passageiro, mas não é viável para a atenuação de uma fonte de ruído de grande porte. Se for possível posicionar a fonte secundária (ou um arranjo de fonte secundárias) tal que ela reproduza tanto temporal como espacialmente o campo gerado pela fonte primária, é possível obter um controle "global" do campo sonoro, ou ao menos controlar uma maior região.

Agora, tentar produzir um controle global do campo sonoro usando apenas *um* sensor de erro distante das fonte primárias e secundárias poderia ter um resultado catastrófico. Como o sensor de erro estará distante da fonte secundária, esta terá de ser alimentada com um sinal de amplitude elevada para conseguir cancelar o campo nesse ponto, e apesar de ocorrer um cancelamento da pressão sonora no microfone, corre-se o risco de no restante do campo acústico ocorrer um acréscimo significativo do nível de pressão sonora. Por causa deste fato é importante realizar simulações do sistema acústico para procurar otimizar o posicionamento dos alto-falantes e microfones tentando garantir uma atenuação "global" nesta região.

3 APLICAÇÕES

Não é de se esperar que em um futuro próximo, um sistema ANC universal esteja disponível nas prateleiras dos mercados, para simplesmente ser posicionado ao lado de uma fonte de ruído e ao ser ligado, como que em um passe de mágica, cancelar todo ruído emitido por esta fonte. Para cada situação de ruído (característica espectral, tamanho da fonte, campo sonoro aberto ou reverberante, etc) um tipo adequado de sistema ANC irá fornecer a melhor atenuação. E apesar de o desempenho do sistema eventualmente até melhorar com o acréscimo do número de alto-falantes ou microfones ou com o uso de um sistema de controle mais potente, existirão sempre limitações físicas para o nível de atenuação que será obtido em cada caso. A mais importante destas limitações é o fato que o sistema ANC só apresenta atenuação considerável quando a distância entre as fontes primárias e as fontes secundárias for menor do que o comprimento de onda de interesse [1]. Por esta razão as aplicações onde os sistemas ANC apresentam bom desempenho se limitam a problemas onde o ruído em baixas freqüências é dominante. O fato interessante é que técnicas passivas de atenuação não são tão eficientes em baixas freqüências. Isso leva obrigatoriamente à idéia de utilizar as duas técnicas em conjunto

para realizar atenuação numa faixa espectral a mais ampla possível.

Após muitos anos apenas como experimento em laboratório, os fones de ouvido com controle ativo do ruído externo estão se tornando itens comercialmente viáveis. Os primeiros fones de ouvido ativos foram desenvolvidos para ocupações específicas, como para motoristas de ambulância e carros de bombeiros. Esses equipamentos recebiam um sinal de referência, correlacionado com o sinal que se deseja atenuar, visto que se conhece exatamente a excitação enviada à sirene, e com isso uma boa atenuação pode ser atingida nestas freqüências específicas. Esses fones reproduziam apenas o "anti-ruído", ou seja, eram apenas protetores auriculares. Para algumas outras ocupações, como piloto de helicóptero ou pilotos de carros de corrida, era necessário que o fone reproduzisse sinais de voz além do anti-ruído. Este equipamento é o antecessor dos fones de ouvido encontrados hoje em dia no mercado. Esses fones de ouvido são projetados para reduzir o ruído externo (sendo este ruído cíclico ou aleatório), reproduzindo com qualidade o sinal sonoro desejado pelo usuário (música, rádio, televisão, etc). Em geral esses fones de ouvido são fones fechados, ou seja, vedam a orelha do ouvinte do meio externo. O material plástico geralmente usado para a casca do fone de ouvido consegue atenuar ruídos de alta freqüência, e complementa a atuação do sistema ANC. O sistema ativo em fones de ouvido costuma apresentar atenuações de 10-15 dB na faixa de freqüência até 500 Hz [3].

O controle de ondas planas que se propagam em dutos também já é comercialmente viável. Nesses sistemas, uma referência do ruído é extraída por um microfone de referência próximo à fonte primária ou por tacômetros, no caso de um duto de ventilação forçada, e os alto-falantes de controle e sensores de erro são posicionados já próximos à saída do duto onde se deseja atenuação. Por causa da extensão dos dutos, muitas vezes a faixa de freqüências que o sistema ANC consegue atuar é relativamente pequena quando comparado com um fone de ouvido ativo.

Sistemas para o controle de ruído em ambientes fechados também têm sido constantemente estudados. Existem sistemas para atenuação de ruído em salas que chegam a produzir atenuação de até 15 dB numa região específica usando uma única fonte de controle [3]. O desenvolvimento destes sistemas está voltado principalmente para a redução do nível de ruído no interior de cabine de aviões e carros de passeio. Para estas aplicações, faz-se necessário o uso de soluções multicanais, que passam a requerer algoritmos de controle mais eficientes. Até o momento, apesar das inúmeras tentativas, ainda não há modelos comerciais de sistemas ANC que sejam "itens de série" para esses equipamentos. Soluções de laboratório já conseguiram reduzir o ruído do motor dentro de um automóvel em 10 a 15 dB [3]. Mas o maior desafio está em reduzir o ruído do pneu e do vento, para os quais não é possível extrair um sinal correlacionado para ser usado como referência.

Como já mencionado, os transformadores de potência apresentam característica espectral ideal para o uso de sistemas ANC. Mas, por causa de seu tamanho, até hoje não foi possível se verificar um sistema que apresente uma atenuação adequada em todas as direções ao redor de um transformador. Foram verificadas atenuações de cerca de 15 dB em um azimute de 30° em 120 Hz [3]. O problema é que é necessário posicionar os sensores de erros no campo distante do transformador para se obter uma atenuação considerável, mas muitas vezes o transformador está localizado próximo a muros ou paredes corta-fogo, que impede o posicionamento do sensor no campo distante e reduz o nível de atenuação obtido.

A função de transferência entre um auto-falante e um microfone é denominada "caminho acústico", e deve ser bem estimada para garantir o estabilidade em um sistema ANC. Sua estimação é geralmente feita usando-se um filtro adaptativo na configuração de estimador, excitado por um ruído conhecido e descorrelacionado com o sinal de controle (geralmente usa-se ruído branco para estimar o caminho secundário). Mas a estimativa do caminho secundário passa a ficar comprometida quando os microfones estão posicionados longe dos auto-falantes, gerando mais uma restrição ao posicionamento dos transdutores.

Parte II

1 PESQUISA DESENVOLVIDA

Este artigo foi motivado pelo projeto de pesquisa e desenvolvimento em que os autores se encontram engajados, para o desenvolvimento de uma solução de controle ativo de ruído para um transformador de potência em uma subestação abaixadora na cidade de São Paulo. O objetivo inicial era obter atenuação de cerca de 10 dB na direção de uma das faces do transformador, desenvolvendo para isto um sistema de controle, especificando o sistema de áudio e definindo o posicionamento ótimo dos alto-falantes e microfones.

Os testes do sistema ANC montado, primeiramente testado em um gramado (campo aberto) e posteriormente na subestação de interesse, serão descritos a seguir.

2 EQUIPAMENTO UTILIZADO

A precisão do sistema de controle é de suma importância no desempenho global do sistema ANC. O uso de um filtro digital adaptativo como elemento de controle do sistema ANC pode fornecer a precisão que os sistemas ANC requerem. Para o sistema de controle, comprou-se a placa EZ-ANC da empresa Causal Systems, uma placa de processamento digital de sinais desenvolvida para aplicações ANC. Esta placa possui seis saídas para fontes de controle e seis entradas, sendo

uma entrada para o sinal de referência e cinco entrada para sensores de erro. A placa EZ-ANC possui o algoritmo de controle direto LMS filtrado (FxLMS) programado, sendo que a estimação do caminho secundário pode ser feita com ruído branco ou com o próprio sinal de controle, o segundo um método proprietário desta empresa. O usuário pode alterar todos os parâmetros do controle adaptativo via comunicação serial com um micro-computador. Pode-se alterar, entre outras coisas, o número de coeficientes e taxa de convergência do filtro adaptativo, ganho dos sinais de entrada e a taxa de amostragem dos conversores A/D e D/A. A placa EZ-ANC também permite que o sinal de referência seja filtrado digitalmente por um filtro FIR fornecido pelo usuário.

Para o sistema de áudio, adquiriram-se seis caixas acústicas, seis microfones, três amplificadores de potência, uma mesa mixadora e cabos.

As caixas acústicas adquiridas são de fibra de vidro. Utilizou-se este material já que as caixas acústicas terão de ficar expostas ao tempo, e têm de ser à prova de água. Os alto-falantes tiveram uma camada de silicone aplicada a seus cones para elevar sua resistência à água. As caixas acústicas foram dimensionadas para apresentarem melhor desempenho na faixa média de freqüência, faixa onde a maior parte da energia do ruído do transformador se concentra. Para alimentar estes alto-falantes, foram comprados três amplificadores de potência Z300 da StudioR.

Foram adquiridos microfones de medição ECM8000 da Behringer, com característica omnidirecional e resposta em freqüência plana na faixa espectral de interesse. Para ficarem expostos ao tempo, estes microfones irão precisar de alguma forma de proteção, como uma caixa metálica, por exemplo. Para a alimentação "Phanton Power" e pré-amplificação dos sinais do microfone, adquiriu-se uma mesa mixadora CSM 16 R da Ciclotron.

Como a placa EZ-ANC funciona com um algoritmo de controle direto, é necessário fornecer à placa um sinal de referência com a freqüência exata que se deseja atenuar na planta. Para o controle de ruído em transformadores, o sinal de referência é normalmente extraído da própria rede elétrica. Por causa das características físicas dos transformadores, a freqüência de vibração é o dobro da freqüência de excitação. Portanto é necessário um circuito que retire uma amostra da freqüência de rede e forneça na sua saída um sinal com o dobro desta freqüência e com amplitude dentro da faixa dinâmica dos conversores A/D. Foi montado um circuito para realizar esta operação. Primeiro o sinal da rede passa por um transformador de tensão, que reduz o sinal de 110 V para 5 V, além de desacoplar o circuito da rede. O sinal que sai do transformador passa então por uma ponte de diodos, cuja saída é aproximadamente o módulo (valor absoluto) da tensão de entrada, o que efetivamente dobra a freqüência do sinal, além de inserir uma componente de corrente contínua e harmônicas superiores. O sinal que sai da ponte de diodos ainda passa por um divisor de tensão, para que sua amplitude fique dentro da faixa dinâmica do conversor A/D e por um filtro passa-bandas (com freqüência de corte inferior em 10 Hz e superior em 800 Hz), para eliminar a componente de corrente contínua e as harmônicas superiores inseridas pela retificação de onda completa, além de ruído de alta freqüência.

3 TESTES DE LABORATÓRIO

O funcionamento do sistema ANC montado foi testado em um gramado da Escola Politécnica. Os testes foram realizados usando-se um dos alto-falantes como fonte de ruído senoidal, dois alto-falantes como fonte de controle, posicionados a cerca de 30 cm de cada lado da fonte primária, três microfones como sensores de erro, posicionados a cerca de 2 m da fonte de ruído, e um microfone foi usado para monitorar o campo acústico. O posicionamento esquemático dos transdutores é mostrado na figura 1. Para a estimação do caminho secundário, utilizou-se o método proprietário da placa EZ-ANC, que usa o próprio sinal de controle para estimar do caminho acústico secundário, sendo cada caminho secundário estimado por um filtro FIR de dez coeficientes. Para o filtro adaptativo de controle usado foi um filtro FIR de quinze coeficientes, com taxa de adaptação suficientemente baixa. Como sinal de referência usou-se o próprio sinal de excitação.

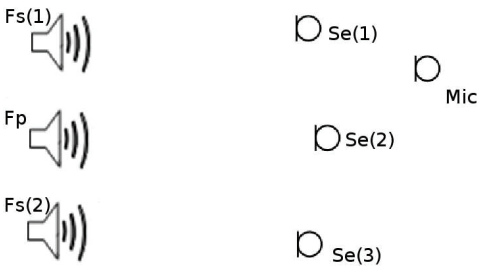


Figura 1: Posicionamento esquemático dos transdutores da fonte de ruído (Fp), das duas fontes de controle (Fs), dos três sensores de erro (Se) e do microfone de monitoramento (Mic).

Com o arranjo da figura 1, verificou-se atenuações na casa dos 10 dB para a freqüência de excitação. Foi possível observar a presença de harmônicas superiores[1] do sinal de excitação no sinal no microfone monitor, com considerável elevação em seus níveis quando o sistema ANC entra em funcionamento, como pode ser visto na tabela 1. Apesar da atenuação da freqüência de excitação, a sensação psicoacústica é mais desagradável com o ANC ligado, por causa da presença das harmônicas superiores no sinal audível.

[1] Em campo aberto é necessário trabalhar com níveis de excitação mais elevados que o nível do ruído de fundo, caso contrário o sinal

Tabela 1: Atenuação medida em um microfone monitor para controle ativo de ruído senoidal em campo aberto.

Freq. (Hz)	Atenuação (dB)
120	10.3
240	-2.6
360	-8.5

Vale ressaltar que em um dia de testes choveu e as caixas acústicas resistiram bem à água, um de seus requisitos de projeto.

4 TESTES NA SUBESTAÇÃO

Como última etapa do projeto, o sistema de controle de ruído projetado foi testado em campo, em uma das subestações da AES Eletropaulo, onde várias combinações de quantidade e posicionamento de alto-falantes e microfones foram testada (tanto os auto-falantes quando os microfones ficaram sempre posicionados a 1,5 m do solo), verificando-se a atenuação obtida tanto nos microfones de erro quanto em posições ao redor destes microfones. Para o sistema de controle foi usada a placa EZ-ANC, com a referência extraída da rede elétrica. Nos primeiros testes os alto-falantes usados estavam localizados em uma plataforma localizada diretamente à frente do transformador. Posteriormente eles foram todos colocados mais próximos ao transformador, diretamente sobre a base de concreto, próximo à base do tanque do transformador. A posição dos microfones variou bastante.

4.1 Uma Fonte de Controle e Dois Sensores de Erro

A disposição dos transdutores é mostrada esquematicamente na figura 2, assim como a posição relativa onde se mediu a pressão sonora. O auto-falante estavam cerca de 1 m a frente do transformador e os microfones estavam cerca de 6 m do transformador, distando cerca de 1,5 m entre si. Verifica-se, como esperado, que entre os dois microfones há uma região de atenuação. Foi surpreendente verificar que no lado externo dos microfones também ocorreu atenuação da pressão sonora. Já atrás dos microfones houve um acréscimo de pressão sonora, provavelmente efeito do muro que estava a cerca de 1 m deste ponto de medição. Os valores de atenuação medidos são mostrados na tabela 2.

4.2 Duas Fontes de Controle e Quatro Sensores de Erro

Aumentou-se a quantidade de transdutores, sendo agora dois auto-falantes e quatro microfones em uso. Os auto-falante estavam cerca de 1 m a frente do transformador, distando aproximadamente 1 m entre si, e

de excitação não será ouvido, o que acabou forçando o alto-falante a trabalhar numa faixa de resposta não-linear.

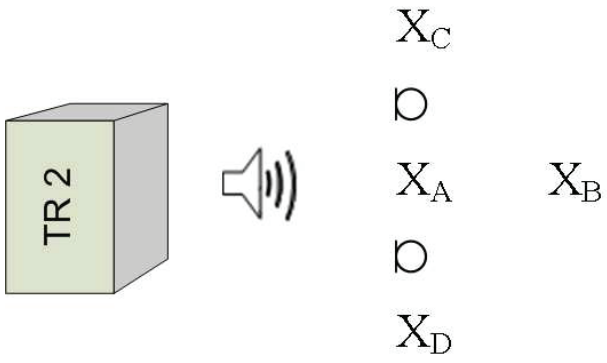


Figura 2: Posicionamento esquemático dos transdutores e dos pontos de medição (X) para uma fonte de controle e dois sensores de erro.

Tabela 2: Atenuação medida em cada um dos pontos considerados para um sistema ANC com uma fonte de controle e dois sensores de erro.

Ponto	Atenuação (dB)
X_A	10
X_B	-6
X_C	2
X_D	5

os microfones estavam cerca de 6 m do transformador, distando cerca de 1,5 m entre si. Esta disposição é mostrada esquematicamente na figura 3. Nesta configuração mediu-se apenas a atenuação próxima aos microfones de erro.

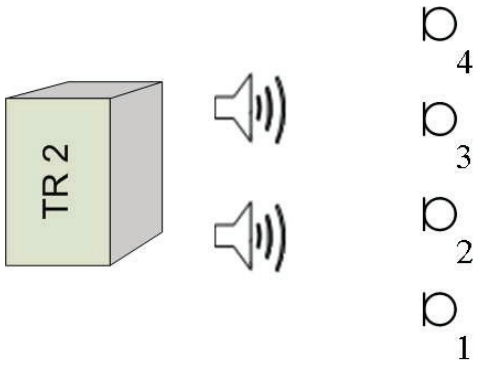


Figura 3: Posicionamento esquemático dos transdutores para duas fontes de controle e quatro sensores de erro.

O resultado não foi satisfatório, ou seja, as duas fontes de controle não estavam sendo capazes de atenuar a pressão sonora em todos os microfones de erro, mas apenas em dois deles, como mostrado na tabela 3, sendo que em um dos microfones verificou-se inclusive um

aumento do nível de ruído. O fato deste único microfone apresentar ganho de ruído, ao invés de atenuação, pode estar relacionado com o fato do nível de ruído do transformador medido neste microfone ser cerca de 20 dB menor do que o nível de ruído medido nos demais microfones. Com o sistema ANC ativo, o nível de ruído medido no microfone 1 ficou no mesmo patamar do nível de ruído medido nos microfones 2 e 4.

Tabela 3: Atenuação medida nos microfones de erro para um sistema ANC com duas fontes de controle e quatro sensores de erro.

Microfone	Atenuação (dB)
1	-12
2	10
3	0
4	8

4.3 Três Fontes de Controle e Quatro Sensores de Erro

Mantendo-se a posição dos microfones, aumentou-se novamente a quantidade de alto-falantes, sendo agora 3 fontes de controle em uso, todos distando cerca de 1 m do transformador e 1 m entre si. A configuração esquemática é mostrada na figura 4.

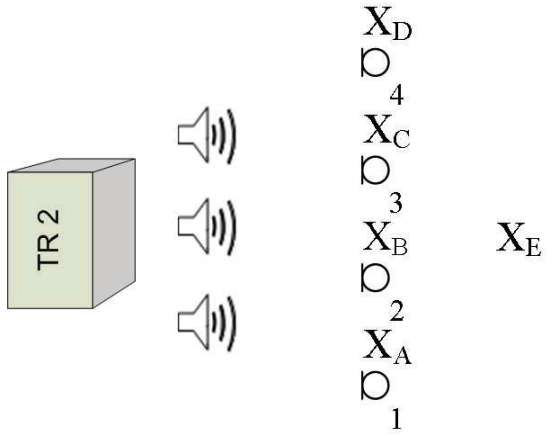


Figura 4: Posicionamento esquemático dos transdutores e dos pontos de medição (X) para três fontes de controle e quatro sensores de erro.

A atenuação nos microfones (exceto o microfone 2) ficou praticamente inalterada, apesar da adição de mais uma fonte secundária, como se verifica na tabela 4. Isto aconteceu porque o sistema de controle convergiu para coeficientes de valores muito pequenos para a nova fonte de controle, praticamente excluindo-a do sistema de controle de ruído. A provável razão para este fato é que os alto-falantes estavam posicionados muito próximos uns dos outros ou muito distante da fonte de ruído. Alterou-se então a posição dos alto-falantes, para verificar se esta suposição era verdadeira.

Tabela 4: Atenuação medida nos microfones de erro para um sistema ANC com três fontes de controle e quatro sensores de erro.

Microfone	Atenuação (dB)
1	-10
2	20
3	-3
4	10

Os alto-falantes, que até então estavam dispostos sobre uma plataforma de acesso ao painel foram reposicionados sobre base do transformador, com distância de aproximadamente 2,5 m entre eles. A posição dos microfones foi mantida inalterada. Verifica-se agora uma melhora considerável nos níveis de atenuação medidos (tabela 5) [2], evidenciando a influência do posicionamento dos transdutores no desempenho do sistema ANC, e justificando uma etapa de simulação para otimizar o posicionamento dos transdutores.

Tabela 5: Atenuação medida nos microfones de erro para um sistema ANC com as três fontes de controle reposicionadas e os quatro sensores de erro na mesma posição.

Microfone	Atenuação (dB)
1	2
2	11
3	15
4	20

Mediu-se também a pressão sonora em pontos entre os microfones, e foi possível verificar atenuação em todos os pontos entre os microfones, como mostrado na tabela 6, que mostra a atenuação obtida nos pontos de medição relativos à figura 4.

4.4 Quatro Fontes de Controle e Quatro Sensores de Erro

O último teste foi feito com um número igual de sensores de erro e fontes de controle. Os quatro alto-falantes foram posicionados sobre a base de concreto do transformador, distando cerca de 2,5 m entre si, enquanto que os microfones tiveram sua posição inalterada em relação ao último teste. Com um número igual de fontes de controle e sensores de erro, é esperado que todos os microfones tenham a pressão sonora ao seu redor anulada. O que foi verificado é que todos os

[2] O microfone 1, que apresenta pequena atenuação, está numa região onde o campo sonoro primário é bem menor que nos demais microfones, como já previamente comentado.

Tabela 6: Atenuação medida em cada um dos pontos considerados para um sistema ANC com três fontes de controle e quatro sensores de erro.

Ponto	Atenuação (dB)
X_A	8
X_B	5
X_C	5
X_D	5
X_E	-10

Tabela 7: Atenuação medida em cada um dos pontos considerados para um sistema ANC com quatro fontes de controle e quatro sensores de erro.

Ponto	Atenuação (dB)
X_A	-8
X_B	5
X_C	1
X_D	0
X_E	-5

microfones apresentaram atenuação, mas nenhum deles teve atenuação maior do que 20 dB. O microfone 1, que antes tinha uma atenuação de apenas 2 dB, nesta configuração apresenta atenuação de 10 dB.

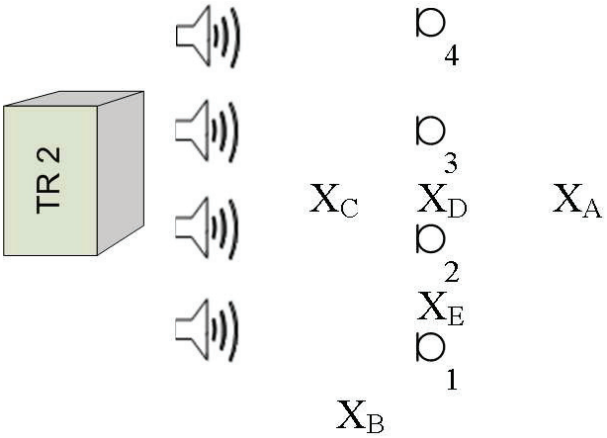


Figura 5: Posicionamento esquemático dos transdutores e dos pontos de medição (X) para quatro fontes de controle e quatro sensores de erro.

A pressão sonora foi medida em alguns pontos ao redor dos microfones, pontos estes esquematizados na figura 5. Os pontos X_D e X_E foram medidos 1m acima da linha horizontal dos microfones. A surpresa nestes testes foi verificar atenuação no ponto X_B, que já estava um pouco distante dos microfones. Os demais pontos apresentaram ou pouca atenuação (X_C e X_D) ou aumento da pressão sonora, como mostrado na tabela 7.

5 SIMULAÇÃO DO SISTEMA ACÚSTICO

O projeto da distribuição física dos transdutores é feito sempre buscando minimizar algum tipo de medida da perturbação que se deseja atenuar. Em geral, para problemas em campo aberto, o que se deseja minimizar é a potência sonora estimada em uma superfície de controle. Para isto, algoritmos de otimização devem realizar dois cálculos distintos: 1) calcular a máxima atenuação possível para um dado arranjo de transdutores e 2) achar o arranjo de transdutores ótimo. A primeira etapa é realizada baseada em simulações acústicas. A segunda etapa envolve algum tipo de busca coordenada, sendo algoritmos genéticos comumente usados para este fim.

Para determinar a máxima atenuação possível para um dado arranjo de transdutores, é necessário saber o valor da pressão gerada tanto pela fonte primária quanto pelas fontes secundárias em cada um dos sensores de erro. Neste projeto, o valor da pressão primária foi estimado baseado em medidas de vibração da parede do tanque do transformador. Como sons em baixa freqüência são geralmente irradiados de forma omnidirecional por alto-falantes, as fontes secundárias foram simuladas como fontes pontuais de irradiação esférica.

6 MODELAGEM DO CAMPO SONORO PRIMÁRIO

A maneira mais precisa de se determinar o campo sonoro primário produzido pelo transformador seria realizar uma medição em campo, o que é bastante difícil de se fazer no caso de um transformador de potência por causa de suas dimensões físicas e também por restrições de segurança.

Usry *et al.* propõem um método simplificado para a estimação do campo acústico produzido por transformadores de potência [5], baseado na equação de Helmholtz para fontes contínuas de som e nas equações de Pierce para a difração em quinas [4]. Este método considera cada elemento de área como se fosse um pistão em uma parede infinita e calcula a pressão gerada em um dado ponto pela onda direta vinda deste "pistão", a onda refletida no solo e as ondas difratadas nas quinas do transformador. Somando-se o efeito de cada elemento de área no ponto desejado, obtém-se a estimativa da pressão gerada pela fonte primária no ponto de observação.

O nível de vibração foi medido em cerca de cem pontos distribuídos pela superfície do transformador. Esta técnica se mostrou bastante atrativa no início por permitir que o nível de pressão fosse estimado em qualquer ponto ao redor do transformador, permitindo uma maior flexibilidade na escolha da posição dos sensores nas simulações (caso houvessem sido usados níveis medidos, as posições de sensores de erro e controle forne-

Tabela 8: Nível de Pressão Sonora medido no campo.

	1	2	3	4	5	6	7	8
D	61,3	66,5	63,5	71,5	63,0	60,0	49,2	55,8
C	63,5	64,0	59,6	66,6	62,7	70,0	66,7	54,0
B	64,2	65,1	63,0	65,8	62,8	67,5	62,2	54,1
A	60,8	65,1	56,1	64,3	66,1	68,7	63,0	62,5

Tabela 9: Nível de Pressão Sonora estimado pelo método de Usry.

	1	2	3	4	5	6	7	8
D	66,6	66,8	67,1	68,6	69,1	71,1	70,5	69,5
C	62,5	48,0	59,4	65,2	65,9	63,9	61,3	62,4
B	56,0	56,2	66,1	70,1	69,6	63,4	58,5	61,0
A	59,3	65,9	70,5	73,5	73,8	71,6	69,5	66,0

cidas aos algoritmos de otimização ficaria restrita aos pontos medidos).

Infelizmente os níveis de ruído estimados com este método não apresentaram boa correlação com os níveis medidos em campo. As tabelas 8 e 9 mostram respectivamente os níveis medidos e estimados em 120 Hz, a 6 m do transformador, entre o transformador e o prédio vizinho. As medições foram feitas em quatro cotas igualmente espaçadas na vertical e oito pontos igualmente espaçados na horizontal. A cota A é a mais baixa, a 1,5 m do chão e a cota D é a mais alta a 6 m do chão. O ponto 1 é o mais à esquerda e o ponto 8 o mais à direita de um observador voltado para o transformador. A tabela 9 mostra os níveis sonoros estimados pelo método de Usry para os mesmos pontos da tabela 8. Por meio de uma simples comparação ponto a ponto é possível constatar que as estimativas não são boas. Em especial, na coluna 8 da tabela 8, os valores medidos apresentam uma redução significativa se comparados com os demais pontos da tabela, mas nem mesmo uma tendência de redução pode ser verificada na coluna 8 da tabela 9.

7 CONCLUSÃO

Na primeira parte do artigo fez-se uma revisão histórica (baseada em [1]) e técnica do controle ativo de ruído, uma aplicação que utiliza equipamentos de sonorização para *reduzir* o nível sonoro de um ambiente, exatamente o contrário de sua aplicação usual de *amplificação* do nível sonoro em um ambiente. Mostrou-se que diversas aplicações para esta técnica vêm sendo desenvolvidas, sendo que algumas dessas soluções já estão disponíveis no mercado. Com o desenvolvimento das pesquisas nesta área, inovações nas técnicas de redução de ruído, principalmente nas indústrias aeronáutica e automobilística podem ser esperadas, elevando consideravelmente o conforto dos passageiros destes veículos.

Na segunda parte do artigo expôs-se os resultados de uma pesquisa sobre o controle ativo do ruído de um transformador de potência. O sistema ANC montado em uma subestação apresentou atenuação do campo sonoro próximo aos microfones de erro e em algumas regiões entre estes microfones. Faz-se agora necessário determinar o melhor posicionamento das fontes de controle e sensores de erro para que uma área qualquer, distante dos microfones de erro, possa ser atenuada, e não somente as regiões próximas aos microfones de erro.

A principal dificuldade encontrada neste projeto foi a caracterização do campo sonoro radiado pelo transformador. A falta de conhecimento sobre o campo gerado pelo transformador praticamente inviabiliza o uso de simulações para otimização do posicionamento dos transdutores.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] S. J. Elliott and P. A. Nelson. Active Noise Control. *IEEE Signal Processing Magazine*, pages 12–35, October 1993.

[2] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, 4rd edition, 2002.

[3] Sen M. Kuo and Dennis R. Morgan. Active Noise Control: A Tutorial Review. *Proceedings of the IEEE*, 87(6):943–973, June 1999.

[4] Allan D. Pierce. Diffraction of sound around corners and over wide barriers. *Journal of the Acoustic Society of America*, 55(5):941–955, May 1974.

[5] G. O. Usry, P. Saha, J. Hadden, and A. Pierce. Prediction of far-field sound radiation from transformers. *IEEE Transactions on Power Apparatus and Systems*, 99(1):358–364, January 1980.

[6] S. E. Wright and B. Vuksanovic. Active control of environmental noise. *Journal of Sound and Vibration*, 190(3):565–585, 1996.

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Fast Distortion vs. Level Measurements with smoothly Increasing Signal Amplitudes

Swen Müller e Nelson Melo do Espírito Santo
INMETRO – Instituto Nacional de Metrologia, Normalização e Qualidade Industrial
Duque de Caxias, RJ, Brasil
smuller@inmetro.gov.br, nmesanto@inmetro.gov.br

ABSTRACT

Distortion vs. level measurements are usually performed in discrete level steps with separate sine waves being fed sequentially to the device under test. In order to speed up this burdensome step-by-step analysis, a continuous level sweep gradually fading the amplitude up to the maximum level is more convenient. The distortion analysis can then be performed in contiguous or overlapping segments by means of the FFT. This paper discusses some of the merits and limitations of the gliding level method, which is applicable for both distortion and intermodulation measurements.

1. INTRODUCTION

Distortion-versus-level graphs provide very useful information about transducers, ADCs and DACs, amplifiers and electronic circuits in general. Using traditional methods, a measurement of these properties in the audio frequency range typically takes some 15 to 30 seconds on a state-of-the-art analyzer, depending on the level step granularity. The overall time obviously increases if the measurement is repeated at different frequencies.

A reduction of this time toll is desirable for all phases of product design, including the development phase (where the effects of circuit modifications need to be quickly verified), up to final quality control (where high throughput of tested devices is necessary).

In metrology, economizing time for this kind of test opens the possibility to repeat it at several frequencies or to simply reduce calibration costs.

The basic idea proposed in this work is trivial: In order to avoid the necessary settle time prior to each individual measurement, one single sine wave is sent to the DUT with a continuously increasing level, as if the fader position on a mixing desk were increased until the onset of clipping.

Of course this operation is not done manually, but by digitally pre-processing a software-generated sine wave before converting it to the analogue domain via a DAC. This means that the method is predestinated to fully-computerized analyzers using DDS (direct digital synthesis) in the generator section. It might even prove feasible for analog generators followed by a VCA (voltage controlled amplifier), though these devices often produce a good deal of distortion themselves. In contrast, modern Delta Sigma audio DACs nowadays compete with very low distortion analog generators, with THD values down to -115 dB at full scale level.

The remainder of this paper refers to computer-controlled analyzers with DDS and synchronously clocked DAC and ADC. The most essential tool of these analyzers is the FFT, normally working on records of 2^N data samples.

1.1. Windowing vs. Coherent sampling

The FFT, as is well known, delivers exact results only for signals that repeat continuously, with the FFT input data record length being equal to the period length. Otherwise, the calculated spectra are polluted with the famous spectral *leakage,* which originates from the discontinuity between left and right edge of the data record.

The leakage can be reduced with data windows, but at the cost of blurring spectral details, thus leading to loss of resolution. There are innumerous windows designed for the analysis of harmonic signals [1], trading main lobe width (the smaller, the better) with side lobe suppression (the higher, the better). The Albrecht 5-term window (±5 bins of main lobe width, 125 dB of side lobe suppression) [2] is a good candidate for harmonic analysis in the audio range and has been used in the examples of this work.

If a window is applied in FFT-analysis, all bins that reside in the main lobe must be summed up to obtain the relative energy of a specific component.

2. EXCITATION SIGNAL CHOICE

Before creating the excitation signal, suitable values for the desired FFT-order, signal frequency, level range and level granularity have to be chosen.

As holds for steady state measurements, a small FFT order allows for very fast testing, while a higher order improves spectral selectivity and decreases the relative height of the spectral bins containing uncorrelated noise.

For coherent sampling, the signal frequency is chosen from one of the values $k\ df$ defined in equation (1). A higher FFT order offers a finer grid of available frequencies.

For the level ramp used to shape the sine wave, different characteristics can be considered. The most practical and easiest to implement is the exponential progression, leading to a constant level increment per time $\partial L/\partial t$ in dB/s.

The total length of the excitation signal is a function of the level range L_{RANGE}, desired analysis level increment

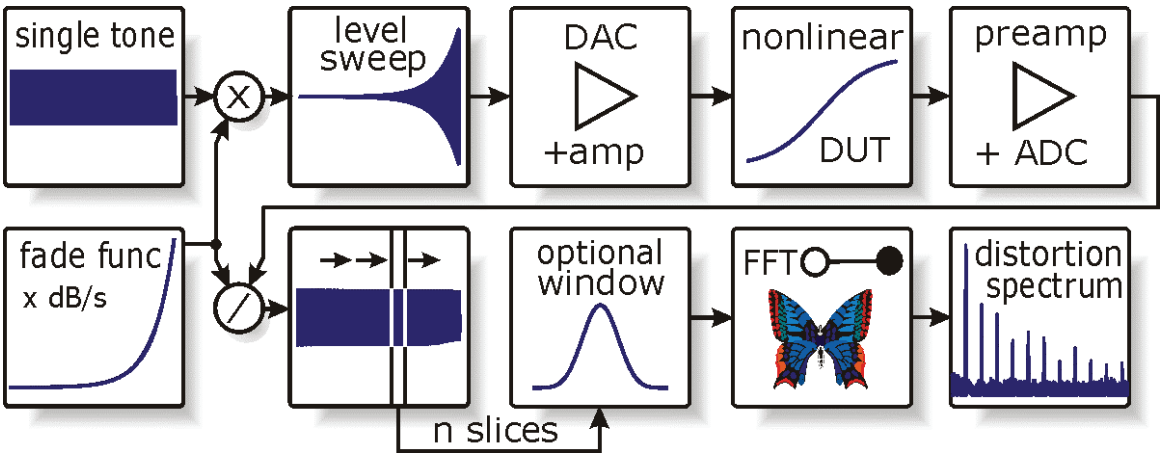


Fig. 1: Block diagram of distortion analysis with level sweep

In contrast, if generator and analyzer are synchronously clocked and the excitation signal is synthesized in the digital domain, *coherent sampling* may be applied to eliminate the need to window prior to the FFT. This means that the chosen excitation signal frequency has to match exactly one of the frequency bins of the FFT spectrum. It then is an integer multiple of the bin distance *df*, which depends on the sampling rate f_S and the FFT-order N:

$$df = \frac{f_S}{2^N} \qquad f_{SIG} = k \cdot df \qquad k = 1,2,3 \ldots 2^{N-1} \qquad (1)$$

This way, an integer number of signal periods will fall into the FFT interval, allowing error-free FFT analysis, with the energy of each component concentrated in just one bin.

Coherent sampling is very powerful and should be used whenever possible. It allows detecting the faintest tones even in close vicinity of very strong components. It also allows revealing tonal components whose level is well below the noise floor by either increasing the FFT order **N** or performing synchronous averages.

The method proposed here is based on coherent sampling, however, as we will see, windowing is still necessary under certain circumstances

ΔL, FFT-order N and overlap factor α (0: no overlap, 1: full overlap) between each FFT segment:

$$T_{SIG} = T_{FFT} \cdot \left(1 + \frac{L_{RANGE}}{\Delta L}(1-\alpha)\right) \text{ with } T_{FFT} = \frac{2^N}{f_S} \qquad (2)$$

The level range L_{RANGE} of the excitation signal must be somewhat larger than the intended analysis range. A dead time is required prior to the start of the analysis to eliminate artifacts originating from the sudden onset of the sine wave. A typical value for T_{DEAD} is the length T_{FFT} of an FFT interval. In this case, 1.5 times the level swept through in one FFT interval must be subtracted from the analysis start level:

$$L_{SIG_START} = L_{ANA_START} - 1.5 \cdot T_{FFT} \cdot \frac{\partial L}{\partial t} \qquad (3)$$

Likewise, 0.5 times the level swept through in one FFT interval must be added to the analysis end level to obtain the excitation signal's end level:

$$L_{SIG_END} = L_{ANA_END} + 0.5 \cdot T_{FFT} \cdot \frac{\partial L}{\partial t} \qquad \mathbf{(4)}$$

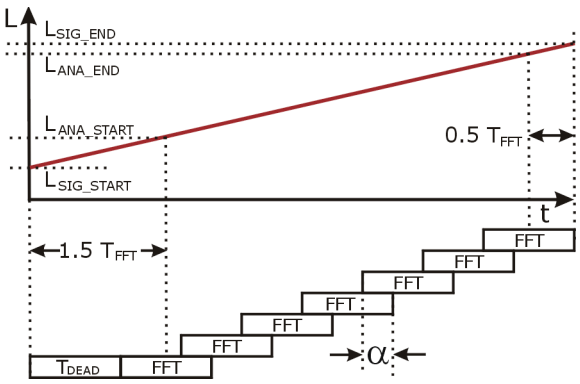


Fig. 2: Relationship between start and end levels of the excitation signal and the analysis.

If the DAC itself is the DUT, the signal end level L_{SIG_END} of course cannot rise above 0 dB FS (dB relative to fullscale). As can be deduced from Fig. 2, this means that the analysis can only be performed up to L_{ANA_END}, which lies $0.5 \cdot T_{FFT} \cdot \partial L / \partial t$ (half the level swept through in one FFT interval) below 0 dB FS. This means that at least the measurement for 0 dB FS must always be completed with the traditional method. If the overlap factor α is higher than 50%, some additional discrete-level measurements become necessary to complete the measurement..

2.1. Excitation signal synthesis

The sine can be generated with the help of the trigonometric support of the CPU floating-point unit, or more efficiently by extracting the imaginary part of a rotating complex pointer. Starting with *Re=1*, *Im=0*, the sample-to-sample angle increment of this rotating pointer is:

$$\partial\varphi = 2\pi \cdot \frac{f_{SIG}}{f_S} \tag{5}$$

The level ramp is applied by the simple process of multiplying each contiguous sample of the sine wave with a gain value *G(t)*, which is then multiplied with the gain increment factor G_{INC}. This factor is based on the sampling rate f_S and the desired fade speed *dL/dt* in dB/s:

$$G_{INC} = 10^{\frac{\partial L/\partial t}{20 \cdot f_S}} \tag{6}$$

The start gain *G(0)* depends on the excitation signal length T_{SIG} and the fade speed *dL/dt*:

$$G(0) = 10^{\frac{T_{SIG} \cdot \partial L/\partial t}{-20}} \tag{7}$$

The momentary level *L(t)* in dB FS is related to the elapsed time by the simple linear equation:

$$L(t) = 20\log\big(G(0)\big) + t \cdot \frac{\partial L}{\partial t} \tag{8}$$

Finally, the samples have to be requantized from the floating-point format (preferably 64 bit) to the integer word length of the DAC by multiplying them with 2^{Q-1}-1 (Q = converter's word length) and then properly applying 2 LSB pp TPDF (triangular probability density function) dither before truncating to the final integer format. This is an essential step to ensure that the excitation signal itself is free of harmonic distortion [3]. TPDF dither spanning amplitude values from –1..1 LSB can be generated comfortably by adding two (uniformly distributed) random numbers in the range ±0.5 LSB.

2.2. Analysis of the received level ramp

Obviously, the FFT cannot be applied directly to segments of the received signal ramp, as the level difference between left and right edge would lead to serious leakage. To avoid this problem, the received signal is first level-equalized by applying the inverse of the fade operation used to shape the excitation signal. In case of a perfectly linear DUT, this will re-establish the constant envelope of the original sine signal. In contrast, strong non-linearities can be detected visually as deviation from the ideal constant envelope, especially when it is viewed on a logarithmic amplitude scale with a small dB range (see Fig. 4). A certain amount of discontinuity between the edges of each individual FFT segment is always present since the amount of level reduction and the accompanying distortion are level-dependent.

The level-equalized DUT output signal is analyzed in narrow segments of 2^N samples. The segments are taken from left to right, with their start point incrementing according to the desired level resolution ΔL :

$$\Delta t = \frac{\Delta L}{\partial L/\partial t} \tag{9}$$

The slices may be adjacent or overlapping. When using windowed FFT analysis, at least 50% of overlap is recommended to avoid loosing details near the edges of each segment. In the case of stronger overlapping (corresponding to finer level steps), the final *distortion vs. level* curve will have a smoothed appearance because the same time signal features enter in various consecutive FFT segments.

Each segment is subjected to a separate FFT and the distortion components extracted and related to the fundamental. The results are graphed in a figure with the distortion values in % or dB on the y-coordinate versus the signal input level on the x-coordinate.

3. EXPERIMENTAL VERIFICATION

The viability of the gliding level approach has been tested under deliberately difficult conditions. A strongly non-linear resistor-diode network has been measured using a level range of 60 dB. The RD network (see Fig. 3) is composed of three legs: one with a bipolar suppressor diode (which is known to gradually start to conduct well below its threshold voltage of 30 Volt), and the other two with standard diodes to provide additional distortion above two different thresholds.

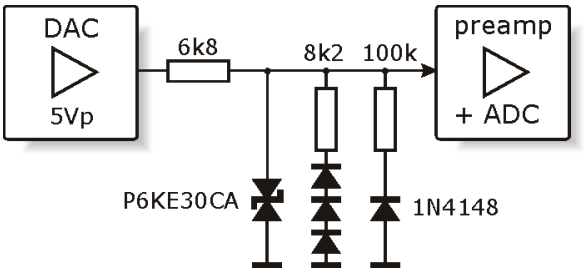


Fig. 3: Resistor-diode network used for distortion test

The THD is initially very low, but rises rapidly as the diodes begin to conduct, reaching almost 7% at the end of the level sweep (5V peak voltage). The resultant level loss can be clearly seen in Fig. 4, in which the amplitude of the level sweep after the fade compensation is displayed on a dB scale.

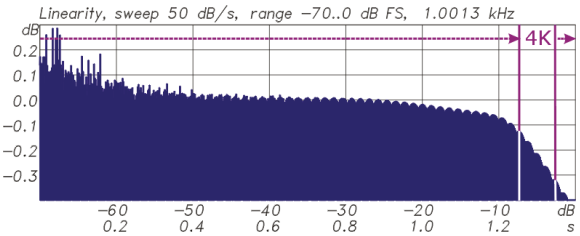


Fig. 4: Sweep at the output of the DUT after level- equalization, showing deviation from ideal

To further challenge the method, the level sweep has been made fairly fast. It spans a 70 dB level range (up to 5V peak) in only 1.4 seconds. The results for the level sweep were compared to traditional stepped level measurements made in 1 dB steps from -60 dB FS upwards.

For both methods, the FFT block size was 4K at 44.1 kHz sampling rate. According to equation (2), this leads to an overlap between neighboring segments of approximately 80%.

The biggest deviation from the traditional method can be expected in the last strip between –10 and 0 dB FS, where the amplitude of the fundamental decreases noticeably, while the amount of harmonic distortion increases dramatically. This entails level differences of fundamental and harmonics at the left and right edge of the analyzed segments.

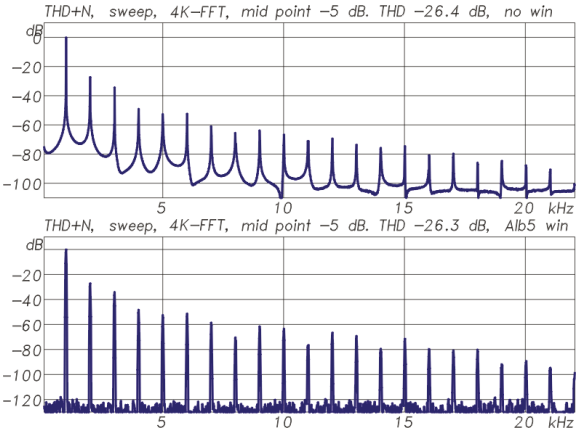


Fig. 5: 4K-FFT of level equalized sweep around –5 dB FS, without (upper plot) and with (lower plot) window.

The expected result is considerable leakage when applying the FFT without prior windowing. In fact, noteworthy leakage *is* present in the upper plot of Fig. 5, which represents the 4K segment around -5 dB input level marked in Fig. 4.

So windowing (result in lower plot of Fig. 5) seems to be mandatory for this stretch of the level sweep. Surprisingly, the THD value obtained from the non-windowed analysis differed by only 0.1 dB compared to the result with proper windowing. Nevertheless, some of the higher order harmonics revealed higher differences. If they shall be traced and displayed separately, windowing really is indispensable to estimate correct results.

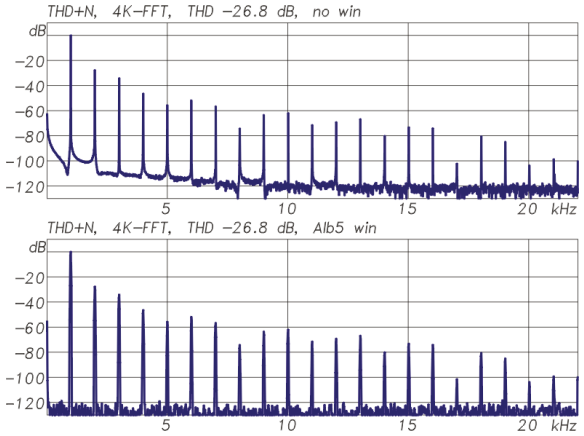


Fig. 6: 4K-FFT of steady -state signal with–5 dB FS level, without (upper plot) and with (lower plot) window.

We experienced another surprise when the THD measurement was carried out with a steady level of -5dB. Some unexpected leakage also showed up in this case! The reason for this is probably the slow warming of the resistors and diodes upon application of the signal. As proof, the leakage disappeared after a couple of seconds, indicating that thermal stability had been reached.

Despite of the presence of this *time-variance*, the THD value differed by only 0.5 dB between the traditional and new methods at -5dB input level. The difference is consistently less than 1 dB over the entire range of input levels. Fig. 7 shows the *THD vs. level* figure obtained with the sweep method, with and without windowing. Fig. 8 shows the results for traditional stepped level analysis.

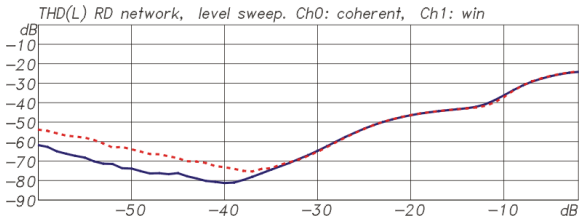


Fig. 7: THD vs. level measured with level sweep method. Solid: coherent analysis, dotted: windowed analysis.

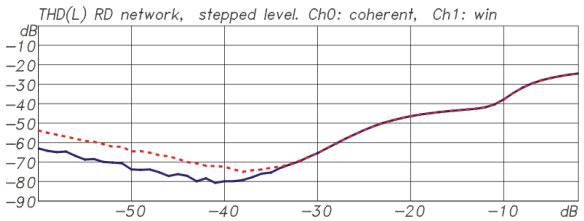


Fig. 8: THD vs. level measured with traditional method. Solid: coherent analysis, dotted: windowed analysis.

Below -35 to -40 dB, the curves in the figures start to rise again. This occurs because the noise floor is rising relative to the fundamental as the level of the excitation signal decreases. The THD values taken in this region do not originate from true distortion, but simply from noise residing at the frequency bins of the harmonic.

Notably, the curve from windowed analysis takes a higher course in this region than the one obtained from coherent sampling. The Albrecht 5-term window [2] used throughout this work has its first spectral minimum positioned at ±5 bins. So, the capture range for each component must comprise 9 spectral bins for windowed analysis. Almost 10 dB more noise energy is picked up

compared to coherent sampling (where just the central bin of each component is evaluated).

The minor differences between the results of the two methods are offset by the considerable speed-up possible with the level sweep method. With the traditional method, each level step is sent out twice the FFT-interval time T_{FFT} in order to establish steady state (with the AD data acquisition only beginning in the second half). Being so, the traditional method takes 11.3 seconds, compared to the 1.4 seconds chosen and proven feasible in this example (on modern PCs, the processing time, including the FFTs, can be almost neglected).

Using a hybrid approach, it is possible to retain the higher noise immunity of the non-windowed analysis at low levels while avoiding leakage errors occurring at higher levels. Non-windowed analysis is used as long as no leakage is detected by the analysis routine. As soon as considerable leakage (indicating rising distortion) starts to spill into the spectrum, windowing is activated.

4. A SIMULATION

The undue time variance encountered in the experimental verification of course makes a comparison of the two methods somewhat unfair. For this reason, a simulation with a memory-less distortion algorithm based on the hyperbolic tangent (Fig. 9) followed the practical trial.

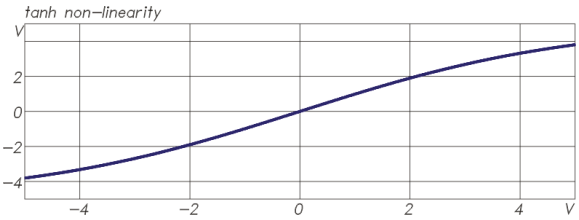


Fig. 9: A piece of the hyperbolic tangent used as characteristic input-output curve for distortion simulation,

The odd order distortion introduced by this characteristic curve was chosen to reach approximately the same end value as in the practical trial. The simulation was performed with 24 bit samples. Fade speed, dB step width and total length were maintained as before.

Fig. 10 shows the results for stepped and continuous level rise, both in conjunction with windowed FFT. The THD difference is now less than 0.1 dB in the level range from -35 dB upwards (Fig. 11).

One may conclude that the level sweep is a viable (and still improvable) approach to performing level-dependant distortion measurements in a fraction of the time required by the traditional method.

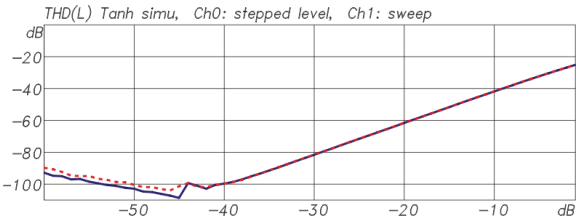


Fig. 10: THD measured with stepped levels (solid) and with level sweep (dotted), both deploying windowed analysis

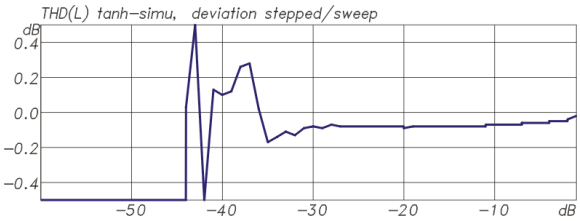


Fig. 11: Difference of results from Fig. 10

5. CONCLUSION

In a similar way that continuous *frequency sweeps* provide benefits for transfer function measurements, continuous *level sweeps* allow one to circumvent the annoying stabilizing period mandatory in steady-state measurements. The gliding level method reduces the overall measurement time by 50% or more compared to the traditional stepped single tone approach with same spectral and level resolution.

If an exact FFT bin frequency is used and the DUT is only mildly non-linear, it is acceptable to perform the subsequent FFT-analysis without windowing, retaining all advantages of coherent sampling, especially higher selectivity in the presence of noise.

Even strongly non-linear devices can be measured with surprisingly good concordance. Moreover, the method can reveal signal processing artifacts which don't manifest themselves at steady levels.

On the other hand, rapidly varying linearity will always create a certain amount of leakage which can make windowing indispensable.

When testing DA converters, a small level range below full scale can't be evaluated and has to be complemented with steady state measurements.

Finally, some distortion tests explicitly demand steady state conditions – in this case, there is of course no alternative to the traditional method.

The level sweep method is equally well suited for *intermodulation measurements* with two or more tones faded in simultaneously. However, treatment of this subject is beyond the scope of this paper.

6. ACKNOWLEDGMENTS

The work has been supported by the Brazilian National Council for Scientific and Technological Development (CNPq).

7. REFERENCES

[1] Fredric J.Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform", *Proc. IEEE*, January 1978, p. 51

[2] Hans-Helge Albrecht, "A Family of Cosine-sum Windows for High-Resolution Measurements", Proceedings. (ICASSP '01) 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing

[3] R.A.Wannamaker, Stanley Lipshitz, John Vanderkooy, J.N.Wright, "A Theory of Nun-Subtractive Dither", IEEE Transactions on Signal Processing, vol 48, p.499-516, (2000 Feb).

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Correção de *Clicks* Através de Métodos Bayesianos Computacionalmente Eficientes

Flávio R. Ávila[1] e Luiz W. P. Biscainho[1]

[1] Universidade Federal do Rio de Janeiro, DEL/Poli, LPS - PEE/COPPE
Caixa Postal 68504, Rio de Janeiro, RJ, 21941-972, Brasil

flavio@lps.ufrj.br, wagner@lps.ufrj.br

RESUMO

Defeitos comuns em gravações analógicas, *clicks* são pulsos de pequena duração e localização aleatória que podem ser causados por deposições ou arranhões na mídia. Eliminá-los do sinal contaminado envolve localizar as amostras corrompidas e substituí-las por estimativas das originais. Os métodos de restauração da literatura vão desde heurísticos eficientes a bayesianos sofisticados que modelam sinal original e ruído estatisticamente. Sendo a alta complexidade computacional exigida a maior restrição destes últimos, este trabalho propõe um modelo estatístico alternativo para o ruído, que usa um número significativamente menor de parâmetros. São apresentados resultados para gravações reais e sintéticas, e comparações com os de outro método.

0 INTRODUÇÃO

Os *clicks* são um dos defeitos mais comuns em gravações de áudio antigas. Esse tipo de defeito é principalmente causado por partículas de poeira ou por pequenos arranhões presentes na superfície do disco. Caracteriza-se por ter tanto localização quanto duração aleatórias. É comum que os *clicks* estejam presentes com grande freqüência ao longo do sinal: uma gravação antiga digitalizada a 44.1 kHz pode ter até 20% de suas amostras corrompidas por *clicks*.

A primeira técnica eficaz proposta para restauração digital de *clicks* descreve o sinal por um modelo autorregressivo (AR) [1], tradicionalmente usado para sinais de áudio, e trata o ruído (*click*) como uma perturbação no sinal que não pode ser descrita por esse modelo. O chamado filtro inverso permite determinar as amostras corrompidas, e a restauração do sinal pode ser feita por interpolação. Essa técnica é computacionalmente eficiente, porém não explora as propriedades desse tipo particular de ruído, que poderiam ser usadas para melhorar o resultado.

Em [1] é descrito um método de restauração em que tanto o sinal original (não-corrompido) quanto o ruído são modelados estatisticamente, e o problema é tratado de acordo com o paradigma bayesiano. As amostras corrompidas são descritas pela soma do sinal original

com o ruído, que é modelado como branco gaussiano de média zero e variância constante. Essa técnica exige que se determine para cada amostra se o ruído está presente ou ausente, o que nos casos práticos gera um custo computacional elevado.

Para tentar reduzir o custo computacional, é proposto um modelo para o ruído em que cada *click* é descrito por apenas três variáveis, e uma característica particular do *click* é explorada.

Nas próximas seções, apresentamos a modelagem matemática do sinal e do ruído, alguns métodos bayesianos para correção de *clicks* que serão usados como padrão de comparação em relação à proposta aqui apresentada, uma descrição do método proposto e finalmente a apresentação dos resultados de simulações.

1 MODELAGEM DO SINAL

Trabalhos anteriores nesta área [1] utilizam um modelo para o sinal em que o *click* é considerado como ruído aditivo presente em apenas algumas amostras do sinal. Considerando $x(n)$ o sinal original (que foi gravado) e $v(n)$ o ruído, o sinal corrompido é descrito como:

$$y(n) = x(n) + i(n)v(n), \tag{1}$$

onde $i(n)$ determina se a n-ésima amostra está corrompida ou correta. Isto é, se $i(n) = 0$, a amostra está correta; se $i(n) = 1$ amostra está corrompida.

O objetivo do sistema de restauração é estimar o sinal $x(n)$ sendo conhecido o sinal $y(n)$. Usualmente, isso é feito em duas etapas: detecção e interpolação. Na etapa de detecção, o objetivo é decidir quais amostras estão corrompidas, ou seja, estimar o sinal $i(n)$. A tarefa da etapa de interpolação é estimar o sinal $x(n)$, sendo conhecidos os sinais $y(n)$, que é diretamente observado, e $i(n)$, obtido pelo procedimento de detecção.

Nos trabalhos considerados neste artigo, o sinal original, $x(n)$, é descrito por um modelo autorregressivo (AR) [1]. Em alguns trabalhos, atribui-se um modelo estatístico para o ruído, enquanto outros usam apenas o fato de que o ruído é uma perturbação que não pode ser descrita pelo modelo especificado para o sinal. Nesse caso, a parte do sinal considerada corrompida é tratada como perdida e a correlação existente no sinal de áudio é explorada para se determinar essas porções do sinal. Na próxima seção serão relatados os trabalhos anteriores relevantes para este artigo, e mais detalhes sobre os modelos usados.

2 RESTAURAÇÃO POR MÉTODOS BAYESIANOS

Através do paradigma bayesiano, é possível incorporar um modelo estatístico ao processo de geração de *clicks*, e assim tratar o problema pela teoria de decisão bayesiana, para a detecção, e usar a teoria de estimação bayesiana para restauração do sinal corrompido. Nas próximas seções, essa abordagem será detalhada.

2.1 Restauração em Blocos

Um bloco de sinal corrompido por *clicks* pode ser descrito vetorialmente da seguinte forma:

$$\mathbf{y} = \mathbf{x} + \mathbf{i}\mathbf{v} \tag{2}$$

onde $\mathbf{y}$, $\mathbf{x}$, $\mathbf{i}$ e $\mathbf{v}$ contêm, respectivamente, N amostras das seqüências $y(n)$, $x(n)$, $i(n)$ e $v(n)$, descritas na equação (1).

Apenas o vetor $\mathbf{y}$ é conhecido. Inicialmente queremos detectar as amostras corrompidas, isto é, estimar o vetor $\mathbf{i}$. Pela regra de Bayes, obtém-se a probabilidade *a posteriori* de $\mathbf{i}$:

$$P(\mathbf{i}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{i})P(\mathbf{i})}{p(\mathbf{y})}. \tag{3}$$

onde $p(\mathbf{y}|\mathbf{i})$ é a verossimilhança de $\mathbf{y}$ dado $\mathbf{i}$; $p(\mathbf{y})$ é a densidade de probabilidade de $\mathbf{y}$; e $P(\mathbf{i})$ é a probabilidade *a priori* de $\mathbf{i}$, que é determinada considerando-se a distribuição típica de amostras corrompidas e corretas nos sinais práticos.

Como o sinal $y(n)$ é conhecido, o denominador da expressão acima é independente de $\mathbf{i}$, e pode ser desconsiderado no processo de otimização.

Para a descrição matemática do método, é conveniente escrever os sinais em vetores contendo amostras nas posições correspondentes ao sinal corrompido e nas posições em que o sinal está correto. Para isso usa-se o sub-índice $-i$ para indicar as amostras intactas, e o sub-índice i indica as amostras corrompidas. Por exemplo, se $\mathbf{x}$ denota o vetor com as amostras do sinal original, os vetores $\mathbf{x}_{-i}$ e $\mathbf{x}_i$ são partições de $\mathbf{x}$ que contêm as amostras corretas e corrompidas do sinal, respectivamente.

Considerando o sinal $\mathbf{x}$ descrito por um modelo AR e o ruído de média zero e matriz de covariância $\mathbf{R}_{\mathbf{v}i}$, mostra-se que a verossimilhança é dada pelo formulário abaixo:

$$p(\mathbf{y}|\mathbf{i}) = \frac{\mu^l e^{\left(-\frac{1}{2\sigma e^2}E_{\mathrm{MIN}}\right)}}{(2\pi\sigma_e^2)^{\frac{N-Q}{2}}|\mathbf{R}_{\mathbf{v}_\mathbf{i}}|^{1/2}|\mathbf{\Phi}|^{1/2}}. \tag{4}$$

onde

$$E_{\mathrm{MIN}} = E_0 - \mathbf{\Theta}^T\mathbf{x}_i^{\mathrm{MAP}}, \tag{5}$$

$$\mathbf{x}_i^{\mathrm{MAP}} = \mathbf{\Phi}^{-1}\mathbf{\Theta}, \tag{6}$$

$$\mathbf{\Phi} = \mathbf{A}_i^T\mathbf{A}_\mathbf{i} + \sigma_e^2\mathbf{R}_{\mathbf{v}i}{}^{-1}, \tag{7}$$

$$\mathbf{\Theta} = -(\mathbf{A}_i^T\mathbf{A}_{-i}\mathbf{y}_i - \sigma_e^2\mathbf{R}_{\mathbf{v}_i}{}^{-1}\mathbf{y}_\mathbf{i}), \tag{8}$$

$$E_0 = \mathbf{y}_{-i}^T\mathbf{A}_{-i}^T\mathbf{A}_{-i}\mathbf{y}_{-i} + \sigma_e^2\mathbf{y}_i^T\mathbf{R}_{\mathbf{v}_i}{}^{-1}\mathbf{y}_i. \tag{9}$$

Nas equações acima, σ_e^2 é a variância da excitação do modelo AR do sinal, contida no vetor $\mathbf{e}$, $\mathbf{A}$ é a matriz do modelo AR, tal que $\mathbf{e} = \mathbf{A}\mathbf{x}$, l é o número de amostras corrompidas, Q é a ordem do modelo AR, e μ é a razão entre σ_e^2 e a variância do ruído σ_v^2 .

O vetor $\mathbf{x}^{\mathrm{MAP}}$ que aparece na equação (6) contém os valores estimados para as amostras corrompidas (MAP = *maximum a-posteriori*). Assim, o sinal restaurado é obtido inserindo-se os valores de $\mathbf{x}^{\mathrm{MAP}}$ nas posições correspondentes (determinadas pelo vetor $\mathbf{i}$) no sinal original $\mathbf{y}$.

Pela teoria de detecção bayesiana, precisa-se calcular as probabilidades *a posteriori* para todos os possíveis vetores $\mathbf{i}$ e selecionar o que gera o maior valor. Como cada elemento de $\mathbf{i}$ pode ser 0 ou 1, para um bloco de sinal de tamanho N, existem 2^N candidatos a serem testados, o que torna o procedimento inviável em casos práticos. Por exemplo, para um bloco de sinal de 20 ms de duração, amostrado numa taxa de 44.1 kHz, tem-se $N = 882$, e o número de candidatos seria de 2^{882}.

2.2 Restauração seqüencial

Uma maneira de se tornar a detecção mais eficiente é fazê-la na forma seqüencial, que ainda tem a vantagem de poder ser realizada em tempo real, à medida que novas amostras são geradas.

O procedimento consiste em se estimar a probabilidade *a posteriori* para o vetor contendo os n primeiros elementos de $\mathbf{i}$, a partir da probabilidade *a posteriori* do vetor contendo os $n-1$ primeiros elementos de $\mathbf{i}$. Imaginando-se que, no instante n, tenha-se um conjunto de P candidatos para o vetor de deteção $\mathbf{i}$, para cada um dos candidatos, haverá no instante $n+1$ dois novos candidatos: um com o novo elemento igual a 1, e outro com o novo elemento igual a 0. Isso faz com que a cada iteração, o número de candidatos dobre, o que geraria um crescimento exponencial indesejável. Para se reduzir a complexidade, adotam-se técnicas de eliminação de candidatos menos prováveis.

Os testes realizados mostram que é possível obter bons resultados usando-se um conjunto de apenas 10 candidatos. Mais detalhes sobre essa técnica, incluindo-se os critérios sub-ótimos que podem ser usados para eliminação de candidatos, são apresentados em [1].

3 MODELO MULTIPLICATIVO

A estimação seqüencial, apesar de muito mais rápida que a estimação em blocos, ainda tem um custo elevado, por necessitar da análise amostra-por-amostra do sinal. O modelo multiplicativo aqui proposto descreve cada *click* separadamente, usando um número significativamente menor de variáveis, tornando a estimação menos custosa.

Esse modelo se baseia no fato observado de que os *clicks* possuem um comportamento característico que pode ser satisfatoriamente descrito por pulsos com decaimento exponencial com taxa de decaimento, amplitude e localização variáveis, na forma:

$$v(n) = r(n) V e^{-\lambda(n-n_0)} u(n-n_0), \qquad (10)$$

onde $r(n)$ é ruído branco de variância unitária, V, λ e n_0 são respectivamente a amplitude, a taxa de decaimento e a amostra inicial do *click*; e $u(n)$ é a função degrau unitário discreta.

As três variáveis do modelo (λ, V e n_0) são supostas desconhecidas *a priori*, e devem ser estimadas a partir do sinal corrompido.

Para avaliar as potencialidades e deficiências decorrentes desse modelo, será considerado o problema de correção de *clicks* em um cenário simplificado: uma porção de sinal de áudio corrompido por um único *click*. O modelo deverá ser posteriormente generalizado para o caso de um número indeterminado de *clicks* no sinal.

3.1 Probabilidade *a priori*

A probabilidade *a priori* representa qualquer conhecimento sobre a variável antes de os dados estarem disponíveis. A seguir apresentamos os modelos *a priori* para cada uma das variáveis que descrevem o ruído.

3.1.1 Localização (n_0)

Se nenhum conhecimento *a priori* sobre a localização do *click* estiver disponível, é razoável assumir que todos os possíveis valores para n_0 são equiprováveis. Sendo assim, a densidade de probabilidade *a priori* da variável discreta n_0 é:

$$p(n_0) = \frac{1}{N} \sum_{i=0}^{N-1} \delta(n_0 - i). \qquad (11)$$

3.1.2 Amplitude (V)

Uma escolha aceitável para a PDF de V é uma distribuição gaussiana de média zero e variância σ_V^2.

$$p(V) = N(V|0, \sigma_V^2). \qquad (12)$$

Essa distribuição tem a vantagem de ter propriedades que simplificam a análise matemática. Entretanto, é mostrado em [2] que essa variável é melhor representada por distribuições *heavy-tailed*, ou seja, que comparadas à gaussiana, associam probabilidades mais altas a valores maiores.

O valor de σ_V^2 pode ser escolhido inicialmente levando-se em conta algum conhecimento subjetivo sobre o comportamento da variável. Outra possibilidade é tratar esse parâmetro como variável aleatória e associar a ela uma distribuição *a priori* não-informativa. Neste trabalho, escolheu-se a primeira opção.

3.1.3 Taxa de decaimento (λ)

A variável λ representa a taxa de decaimento do pulso exponencial, sendo, portanto, positiva. Uma distribuição que atende essa exigência e tem propriedades algébricas interessantes é a distribuição gama inversa, definida por:

$$\mathrm{IG}(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\tau(\alpha)} \lambda^{-(\alpha+1)} e^{-\beta/\lambda}. \qquad (13)$$

A média e a variância dessa distribuição são obtidas a partir dos parâmetros α e β da seguinte forma:

$$m_\lambda = \frac{\beta}{\alpha - 1}, \qquad (14)$$

$$\sigma_\lambda^2 = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}. \quad (15)$$

A partir dessas relações, podemos escolher os parâmetros que representam adequadamente a distribuição da taxa de decaimento dos *clicks*. Há ainda a possibilidade de se tratar esses parâmetros como variáveis aleatórias e incorporá-las ao modelo bayesiano. Neste artigo, optou-se por tratá-los como constantes conhecidas.

3.2 Maximização da probabilidade *a posteriori*

Define-se um vetor contendo os parâmetros do modelo do ruído (*click*):

$$\theta = \begin{bmatrix} V & \lambda & n_0 \end{bmatrix}^{\mathbf{T}}. \quad (16)$$

Para o cálculo de $p(\mathbf{y}|\theta)$ pode-se usar a equação (4) com:

$$\mathbf{i} = \begin{bmatrix} \mathbf{0}_{1X(n_0-1)} & \mathbf{1}_{1XN_p} & \mathbf{0}_{1X(N-n_0-N_p} \end{bmatrix}^T, \quad (17)$$

onde $\mathbf{0}_{1Xn}$ e $\mathbf{1}_{1Xn}$ são vetores-linhas de n elementos contendo apenas 0 e 1, respectivamente; e N_p é uma amostra escolhida de tal forma que a partir dela o ruído possa ser considerado desprezível. A matriz de covariância do ruído é:

$$\mathbf{R}_{v_i} = V^2 \text{diag}\left\{ \begin{matrix} 1 & e^{-2\lambda} & \dots & e^{2N_p\lambda} \end{matrix} \right\}. \quad (18)$$

Pela regra de Bayes, obtemos a probabilidade *a posteriori*:

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta), \quad (19)$$

onde $p(\theta)$ é a densidade de probabilidade *a priori* do vetor de parâmetros θ, cujos parâmetros são supostos independentes, isto é:

$$p(\theta) = p(V)p(\lambda)p(n_0). \quad (20)$$

A maximização da densidade de probabilidade *a posteriori* se mostrou difícil de ser feita de forma analítica. O uso de métodos de otimização multivariável (como o método do gradiente, o método de Newton etc) requer o cálculo das derivadas em relação a cada uma das variáveis, o que é aparentemente impossível analiticamente. Optou-se pelo uso de um método de otimização univariável (o método de Fibonacci [3]) usado repetidas vezes, partindo-se de um valor inicial para as variáveis de interesse. Esse valor inicial é obtido por amostragem da distribuição *a priori* de cada variável. A seguir detalhamos o procedimento para cada variável.

3.2.1 Estimação de n_0

O valor de n_0 determina quais amostras do sinal estão corrompidas, e portanto permite obter o vetor $\mathbf{y}_i$, que aparece no argumento da exponencial da equação (4). Como conseqüência, essa equação não tem expressão analítica em função de n_0. Como a função é discreta, o método de Fibonacci levou a resultados incorretos. Todavia, é possível estimar n_0 de forma trivial, calculando-se a probabilidade condicional $P(n_0|y)$ para todos os possíveis valores de n_0 e escolhendo-se aquele que gera o maior valor.

Para reduzir a complexidade computacional desse procedimento, pode ser feito algum pré-processamento simples (por exemplo, através de filtro inverso) que estabeleça um intervalo menor para os possíveis valores de n_0.

3.2.2 Estimação de V e λ

A maximazação dessas duas variáveis pelo método de Fibonacci foi bem sucedida. O intervalo inicial escolhido foi $I_i = \begin{bmatrix} 0 & 1 \end{bmatrix}$, tanto para V quanto para λ. Em ambos os casos, com apenas 10 iterações do algoritmo obtém-se a convergência com um erro desprezível, considerando que o erro de inferência estatística é muito maior.

4 RESULTADOS

Esta seção é destinada à apresentação dos resultados obtidos com o modelo proposto, e sua comparação com os obtidos pelo método bayesiano seqüencial. Foram feitos testes com sinais de áudio reais, em duas situações distintas:

(A) *click* artificial, gerados a partir do modelo multiplicativo (equação (10)), e

(B) *click* real, retirado de um outro sinal corrompido, e inserido num sinal não corrompido.

Em ambos os casos, analisou-se um bloco de sinal com apenas um *click* presente. O sinal escolhido, nas duas situações, foi uma nota de violão, bastante tonal, que torna os *clicks* facilmente audíveis. Foi selecionado um bloco do sinal de 500 amostras, com o único *click* localizado próximo à amostra 250, como mostram as figuras 1 e 3.

Para o teste com o método seqüencial, foi usado um modelo AR de ordem 10, cujos coeficientes foram determinados a partir do sinal corrompido pela minização do erro médio quadrático. A variância do ruído σ_V^2 é calculada diretamente a partir do ruído que foi inserido no sinal.

Usou-se o mesmo modelo AR para o sinal na implementação do método proposto. Os parâmetros das distribuições *a priori* das variáveis V e λ foram os seguintes: $\sigma_V^2 = 1$, $\alpha = 5$ e $\beta = 0.8$.

4.1 Caso A: *Click* artificial

O *click* foi gerado artificialmente, através do modelo da equação (10), com os parâmetros: $n_0 = 250$, $\lambda = 0.2$ e $V = 0.3$.

As estimativas obtidas para esses parâmetros foram: $\hat{n}_0 = 250$, $\hat{\lambda} = 0.1493$ e $\hat{V} = 0.2975$. Os sinais original e estimados pelos métodos seqüencial e proposto estão mostrados na Figura 2. Vemos que o método proposto fornece uma estimativa mais suave e mais parecida com o sinal original do que o método seqüencial. Além disso, é significativamente mais rápido.

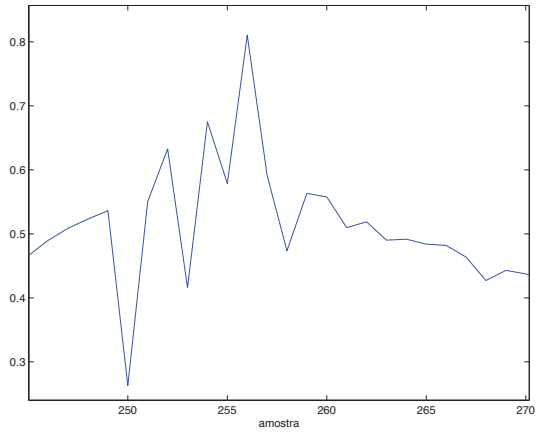


Figura 1: Sinal corrompido com *click* artificial

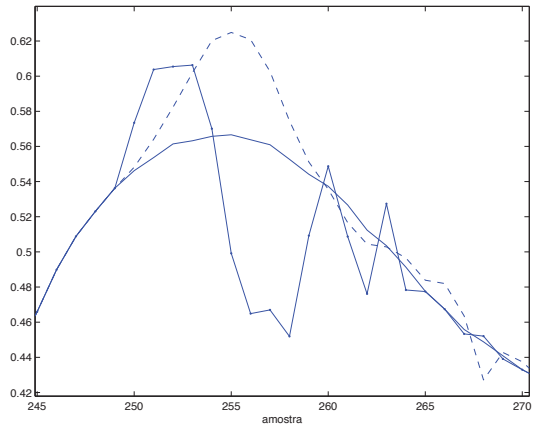


Figura 2: Sinal original (linha contínua), sinal estimado pelo método seqüencial (linha pontilhada) e sinal estimado pelo método baseado no modelo proposto (linha tracejada)

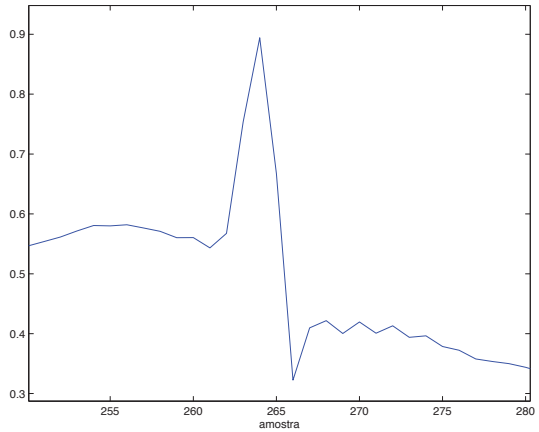


Figura 3: Sinal corrompido com *real*

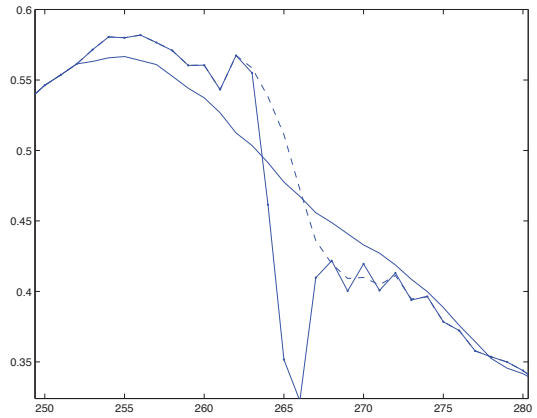


Figura 4: Sinal original (linha contínua), sinal estimado pelo método seqüencial (linha pontilhada) e sinal estimado pelo método baseado no modelo proposto (linha tracejada)

4.2 Caso B: *Click* Real

A Figura 4 mostra o sinal original e os sinais estimados pelo método seqüencial e estimado pelo método baseado no modelo proposto. Vemos que a curva obtido pelo método proposto é mais suave, e representa melhor o sinal original. Ouvindo os sinais estimados, percebemos um leve *click* no sinal restaurado pelo método seqüencial, mas nenhuma degradação no sinal obtido pelo método proposto.

5 CONCLUSÕES

No contexto dos métodos bayesianos para restauração de sinais de áudio corrompidos por *clicks*, este artigo apresentou um modelo de *click* que usa um número de variáveis significativamente menor em relação aos encontrados na literatura. Com isso, consegue-se obter um método com custo computacional menor, além de um resultado subjetivamente melhor. A limitação desse método, no momento, é a necessidade de separar o sinal em blocos e garantir que cada bloco possua apenas um *click*. A continuação do trabalho se concentra na generalização desse modelo para um número qualquer de *clicks*, para o que o método *Reversible-Jump* MCMC [4], que emprega o algoritmo de Metropolis-Hastings, parece bastante promissor.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration - A Statistical Model-based Approach*, Springer, 1998.

[2] S. Godsill and E. Kuruoglu, *Bayesian Inference for Time Series with Heavy-tailed Symmetric α-stable Noise Process*, University of Cambridge Technical Report, 1999.

[3] J. H. Mathews and K. D. Fink, *Numerical Methods Using Matlab*, Prentice Hall, 2004.

[4] C. Andrieu and A. Doucet, "Joint bayesian model selection and estimation of noisy sinusoids via reversible jump mcmc," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2667–2676, October 1999.

Sessão 5

Equalização, gerenciamento e mixagem

(Audio equalization, management and mixing)

Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Harmonic Equalizer

Andre Luiz Luvizotto,[1] César Rennó Costa,[1,2] Jonatas Manzolli,[1]
Rafael Santos Mendes,[1,3] and Fernando Von Zuben[2,3]

[1] Núcleo Interdisiplinar de Comunicação Sonora - NICS, Caixa Postal: 6166, Campinas, São Paulo, Brasil

[2] UNICAMP, Laboratório de Bioinformática e Computação Bio-Inspirada - LBiC
Caixa Postal: 6166, Campinas, São Paulo, Brasil

[3] UNICAMP, Faculdade de Engenharia Elétrica e Computacional, Caixa Postal: 6101, Campinas, São Paulo, Brasil

andre@nics.unicamp.br, cesar@nics.unicamp.br, jonatas@nics.unicamp.br,
rafael@dca.fee.unicamp.br, vonzuben@dca.fee.unicamp.br

ABSTRACT

We present here a new-fashioned graphic equalizer which acts on signal partials. The process is based on a wavelet-based filter-bank that is dynamically tuned to the note's fundamental pitch. Instead of working on fixed frequencies as traditional graphic equalizers, it matches - down to every single note - its frequency bands to harmonic partials. Furthermore, It can also be seen as a real-time subtractive synthesizer which could be used as a compositional tool as well.

0 INTRODUCTION

In this paper we present a new-fashioned graphic equalizer which works dynamically on the sound's partials based on the pitch detector and on the wavelet filter bank proposed in [1] by Béltran where a new algorithm to compute an additive synthesis model of a signal, which is based on the Continuous Wavelet Transform analysis, is outlined.

The analysis and resynthesis of musical sounds has been addressed by many authors usually with Fourier Transform related techniques. These techniques involve a representation of signals in terms of contributions equally spaced on the linear frequency axis. A method that would combine a decomposition in elementary contributions and hearing like properties would be very attractive. The Wavelet Transform is such a method [2].

A second point that we will focus is on the pitch detection algorithms. Finding the pitch of musically interesting sounds can be extremely difficult taking into

account its complexity. Musical instruments, unlike speech, have extremely large frequency range, so the pitch-tracking algorithm has to deal with a very large bandwidth and sounds that are harmonically rich and can contain inharmonic partials [3].

In short, in the first section we shall treat the main aspects about wavelets analysis and the mathematical background used in this paper. In the second section we will present the formulation of the used scale and shift parameters as well as the chosen mother wavelet. In the third section the pitch detection algorithm will be explored with an overview of the autocorrelation method. And in the last two sections there will be an explanation of the implementation and also a discussion about the results and performance of our model.

1 WAVELET ANALYSIS

Usually the process of partial extraction is based on *Short Time Fourier Transform*, but as we know by the Heisenberg uncertainty principle an ideal time localization gives rise to a non-ideal frequency localization and vice versa[4][5].

Wavelet Analysis can be seen as a powerful tool to fix this problem that provides a flexible time-frequency window and that is the main reason to choose wavelets to implement the filter-bank. In formulating a Continuous Wavelet Transform (CWT) a scale parameter is introduced to adjust the width of the sliding time window process. The mathematical basis of the Fourier Transform are sine waves whereas the basis of the CWT is a wavelet family generated by the mother wavelet ψ(t). The CWT is defined as:

$$W_f(a,b) = \int_{-\infty}^{+\infty} f(t)\psi^*_{a,b}(t)dt \tag{1}$$

where * is the complex conjugate and $\psi_{a,b}$(t) is the mother wavelet scaled by a factor a and shifted by a factor b.

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\Big(\frac{t-b}{a}\Big) \tag{2}$$

Hence, the CWT consists of breaking up a signal into scaled and shifted versions of the mother wavelet that defines a wavelet basis. We can also define:

$$\psi_a(x) = \frac{1}{\sqrt{a}}\psi\Big(\frac{x}{a}\Big) \tag{3}$$

If the $\psi_a(t-b)$ forms an orthogonal basis, equation 1 can be viewed as a inner product of the original signal $f(t)$ with :

$$W_f(a,b) = \langle f, \psi_a(t-b)\rangle \tag{4}$$

The wavelet transform is the convolution of the function f with the filter which the impulse response is $\tilde{\psi}(x)$:

$$W_f(a,b) = f * \tilde{\psi}_a(b) \tag{5}$$

where

$$\tilde{\psi}_a(x) = \psi_a(-x) \tag{6}$$

The filter frequency response of the scaled mother wavelet ψ_a(x) is:

$$\hat{\psi}_a(\omega) = \sqrt{a}\hat{\psi}(a\omega) \tag{7}$$

For each scaled version of the mother wavelet we have a new filter with a different bandwidth. To create a desire filter bank, we need to define a parameter s to shift the central frequency response of each filter generated by setting the time scale factor a. In our case we want to define a parameter s that could sweep all the harmonic bands we want of a given signal.

2 WAVELET FILTER BANK

As proposed by Beltrán in [1] we are considering a complex generalization of the Morlet wavelet given by:

$$\psi(t) = C' e^{\frac{-t^2}{2}}\left(e^{j\omega_0 t} - e^{\frac{-\omega_0}{2}}\right) \tag{8}$$

for the mother wavelet.

This function as proposed by Kronland-Martinet et al [2] requires small corrections to ensure that the formal conditions for an analyzing wavelet is satisfied. However, in practice taking $w_0 > 5$ is enough [1]. In this case the Fourier transform of the Complex Morlet's wavelet is:

$$\hat{\psi}(\omega) = Ce^{\frac{-(\omega-\omega_0)^2}{2}} \tag{9}$$

C' e C are normalization constants in the time and frequency domain, respectively.

As we are interested on a filter bank we need to have how to control the frequency resolution of each band. This way we can tune all the bands of our equalizer to the sound's partials of a given note. We should included a parameter k in the equation 9 to control the bandwidth of the filter. So we have in the frequency domain:

$$\hat{\psi}(\omega) = Ce^{\frac{-(\omega-\omega_0)^2}{2k}} \tag{10}$$

The set of scale is employed to provide a logarithm-resolution on the frequency axis. We will also insert, in equation 10, a discrete factor s_i to slide the central frequency of the first band, i.e, changing this factor we will be sliding the first band and generating all the others. To prevent aliasing problems the filter bank generates first the bands with highest frequencies. That is, it gets the fundamental pitch f_f of the played note, calculates the last frequency band that will be used and then calculates the central frequencies ω_c of the remaining bands changing s_i as follow:

$$\omega_{c_1} = f_f N \tag{11}$$

and

$$\omega_0 = s_{min}\omega_{c_1} \tag{12}$$

We will call s_1 of s_{min} cause it can be considered a tuning factor related to the Nyquist criterion which fine tunes ω_0 to the desire first (highest) band's central frequency ω_{c_1}.

And for $2 \leq i \leq N$ we have:

$$s_i = \frac{N}{n_p} \tag{13}$$

$$\omega_c = \frac{\omega_0}{s_i} \tag{14}$$

N is the number of bands and n_p is the band's number of the *i-th* frequency that is being calculated. We can see in the equation 14 the relation between the central frequency ω_c and the frequency ω_0 for all bands. Then we have in the frequency domain the follow equation:

$$\hat{\psi}(\omega) = C_s e^{\frac{-(s_i\omega-\omega_0)^2}{2k}} \tag{15}$$

where C_s is also a normalization constant. We have in equation 15 the main expression for the filter bank in the frequency domain. We need now to find its *time* representation to be able to use the convolution representation of the equation 5 to filtrate the input signal. The follow expression was obtained from the equation 8, taking into account properties of Fourier transforms and by changing the continuous factor a for the discrete one s:

$$\psi(t,s) = \frac{\sqrt{k}}{s} C' e^{\frac{-kt^2}{2s^2}} \left(e^{\frac{j\omega_0 t}{s}} - e^{\frac{-\omega_0}{2\sqrt{k}}} \right) \tag{16}$$

The equation 16 presents the mother wavelet that will be used with the frequency slide parameter s and bandwidth parameter k.

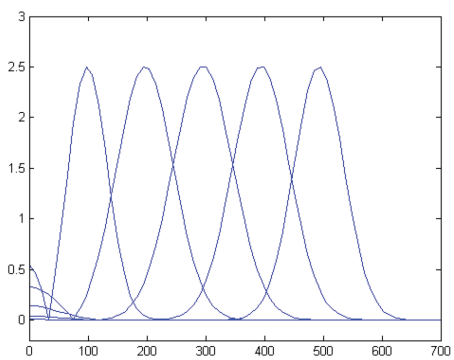


Figure 1: Filter bank generated by the mother wavelet presented in equation 16 with $\omega_c = 100$Hz, $k = 22000$ and s_{min} = 6.91 and s_i given by equations 13.

Now we'll expose the pitch detector algorithm.

3 PITCH DETECTION

Pitch detection algorithms can be classified in two separate categories, spectral-domain and time-domain based on period detection. We are interested on the second one due to its low computational cost. This kind of detection seems to be the most straightforward idea, which consists in looking to the input signal as an amplitude fluctuations in the time domain and try to find repeating patterns from the waveform that could lead us to its periodicity. Autocorrelation was the chosen method. The goal of the autocorrelation routines is to find the "similarity" between the signal and a shifted version of itself.

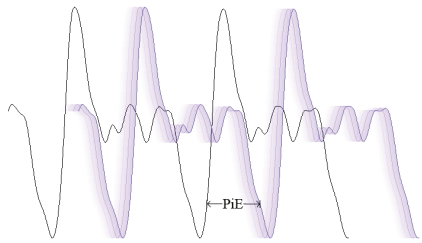


Figure 2: The strong line is the original signal and the blurred signal is a copy of the original, shifted left by an amount PiE nearing the fundamental period. Notice how the signals begin to align with each other as the shift amount nears the fundamental period.

If the signal is periodic, the autocorrelation function $y(n)$ also will be, and if the signal is harmonic the autocorrelation function will have peaks in multiples of the fundamental frequency [6]. The mathematical definitions is:

$$y(n) = \sum_{k=1}^{m} u(k)u(k+n) \tag{17}$$

4 IMPLEMENTATION

The algorithm was implemented with *Matlab* according to the diagram of the figure 3:

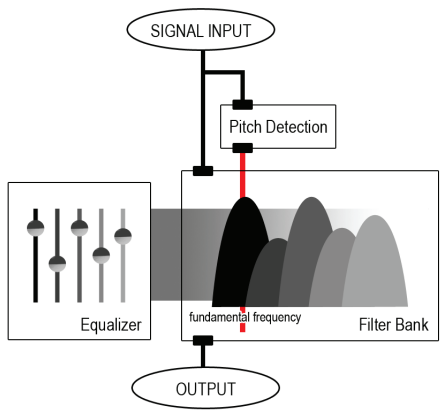


Figure 3: Implementation Block Diagram

The input signal first reaches the pitch detector (that we call by Dr. Pitch) which outputs the fundamental frequency's value f_f to the filtering stage. Then the

filter parameters ω_0 and s_i are calculated and inserted into the equation 16. As we know each band of the filter is related to a wavelet with bandwidth factor k and central frequency ω_c which is sampled with a sampling frequency f_s. As a results we have a matrix $M_{m\times n}$ where m is the number of sampled points and n the number of filter bands N.

In practice M is left-side multiplied by a weight's matrix $P_{1\times n}$ that has the informations about the gain of each band, i.e, values between 0 and 1 came from the equalizer' slides. At last this resulted matrix is time convoluted with the input sound vector providing the filtered output. When the next signal comes Dr. Pitch sends the new f_f to the filter stage. The parameters are actualized to generate the up-to-date wavelets and go on like before.

We used a Hanning window with 1024 and also 2048 samples with overlap of 50%. The window size depends on the lowest frequency (large period) of the input signal. We are looking for other methods of pitch detection that could be more efficiency and computationally better.

5 RESULTS

We applied the *Harmonic Equalizer* with ten and twelve bands to many different sounds. Guitar, bass, percussion instruments, some different guitar tunings, for example, with the 6th string tuned to C1 since we were interested in evaluate how Dr. Pitch would work at extreme conditions. These examples can be find at www.nics.unicamp.br/∼andre/equalizer.

We can see on the figure 4 that Dr. Pitch worked very well, tracking both of notes correctly. We can also observe the difference between the spectrograms of figures 4 and 5 considering the number of partials presented. Only the first five partials were letting pass in figure 5 while all others five were filtered.

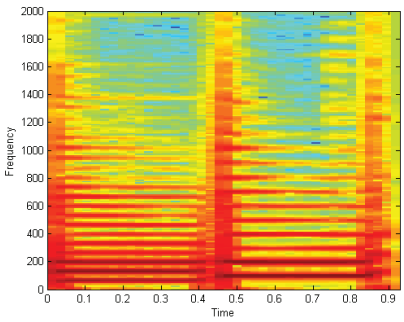


Figure 4: Sonogram of the input signal.

As a second example we have a sawtooth wave on figure 6 where its first four even harmonics were filtered. We can see on figure 7 that the resulted waveform is closed to a square wave, as expected. The more we increase the number of the filtered partials the more waveform matches to the square waveform.

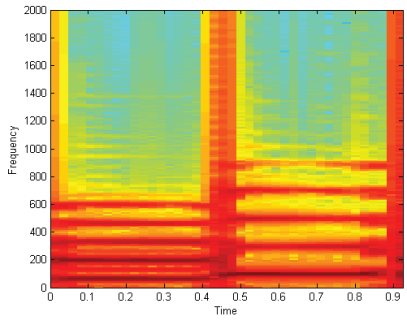


Figure 5: Sonogram of the output signal.

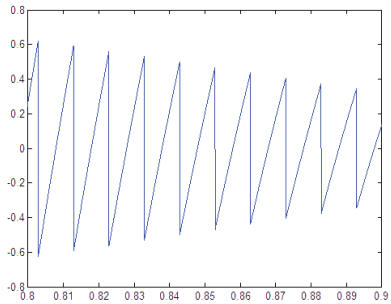


Figure 6: Original sawtooth wave.

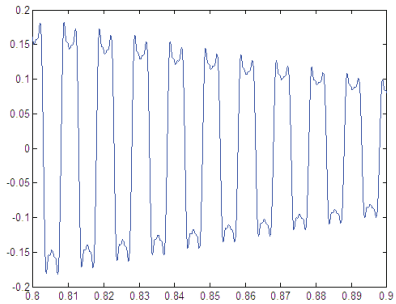


Figure 7: Filtered sawtooth wave. Notice that it is closed to a square waveform.

Finally a square wave, figure 8, were filtered letting pass just the fundamental frequency. As we can observe on figure 9 the resulted waveform is closed to a simple sine.

6 CONCLUSION AND FURTHER IMPLEMENTATIONS

We presented the first results of a new-fashioned graphic equalizer. A method of partials extraction was proposed as well as a pitch detector based on the autocorrelation function. The essential mathematical tools was revised in the first sections and a briefly discussion of the pitch detection in the section three.

This first prototype made with *Matlab* could give us a good idea of the sonic potential of the model. The

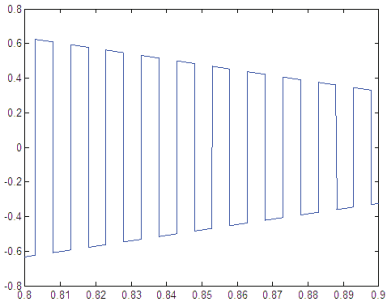


Figure 8: Original square wave.

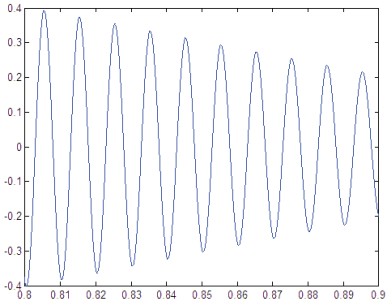


Figure 9: Filtered square wave. Notice that it is closed to a simple sine.

best results were obtained in low frequencies sounds for example with bass drums, bass and lower guitar sounds.

We are currently developing a low latency C++ implementation to be used as a cross-platform plugin. We are also verifying the possibilities of working with other pitch detection methods, for example a wavelet based algorithm, and use it as another parameter on the graphical interface that could be selected by the user depending which kind of sound it wants to filter for example bass, brass, clarinet, piano etc.

7 ACKNOWLEDGMENT

We would like to thank Professor Adolfo Maia Jr. for helping us with the mathematical revision, formalization and kindly tips.

REFERENCES

[1] J. R. Beltran and F Beltran, "Additive synthesis based on the continuous wavelet transform: A sinusoidal plus transient model," in *Proc. of the 6th Int. Conference on Digital Audio Effects*, 2003.

[2] R. Kroland-Martinet, "The wavelet transform for analysis, synthesis, and processing of speech and music sounds," *Computer Music Jornal*, vol. 12, no. 4, pp. 11–20, Win 1988.

[3] Tristan Jehan, *Musical Signal Parameter Estimation*, Ph.D. thesis, Université de Rennes 1, 1997.

[4] C. K. Chui, *Wavelets: a mathematical tool for signal processing*, Society for Industrial and Applied Mathematics, 1997.

[5] Stéphane Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[6] M. Bosi, "Pitch Detection Methods Review," *http://www-ccrma.stanford.edu/ pdelac/154/m154paper.htm*, April 1, 2007.

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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.

Gerenciador Digital com Alta Faixa Dinâmica

Swen Müller
INMETRO – Instituto Nacional de Metrologia, Normalização e Qualidade Industrial
Duque de Caxias, RJ, Brasil

smuller@inmetro.gov.br

RESUMO

Este artigo apresenta os conceitos de um novo gerenciador digital de sistemas com faixa dinâmica inédita e possibilidade de combinar livremente filtros do tipo IIR e FIR para a equalização de cada via. Para garantir a proteção tanto contra distorção quanto contra a destruição dos alto-falantes, o gerenciador possui um limitador RMS e um limitador de picos em cada via. A técnica "*look-ahead*" aplicada no limitador de picos permite reagir com agilidade e precisão sem introduzir distorções perceptíveis.

0. INTRODUÇÃO

Os gerenciadores de alto-falantes são uma peça-chave em sistemas de sonorização ativos de médio e grande porte. Eles basicamente reúnem as funções de *crossover*, equalizador e limitador. Antigamente, essas tarefas eram cumpridas por circuitos analógicos dedicados. Para cada tipo de alto-falante, filtros ativos individuais tiveram que ser projetados para viabilizar a separação e equalização de cada via da caixa acústica, a fim de obter uma resposta *overall* equilibrada. Os limitadores obrigatórios eram compostos por VCAs (*voltage controlled amplifiers*), uma fonte permanente de dores de cabeça para os engenheiros porque eles inerentemente aumentam o ruído de fundo e a distorção do áudio.

Hoje em dia, prevalecem as soluções digitais para os gerenciadores de alto-falantes. Mediante conversores AD de alta qualidade, o áudio é transformado para o mundo digital logo no primeiro estágio. Todo o processamento do sinal é realizado por processadores digitais, e o sinal só volta ao domínio analógico no último estágio, fornecendo o áudio processado aos amplificadores.

As vantagens dessa tecnologia são múltiplas:

1) O processamento digital dos sinais a princípio pode ser executado com precisão quase ilimitada. Na praxe, existem limitações devidas à arquitetura dos DSPs (*digital signal processors*), mas num software minuciosamente desenvolvido, a contribuição do processamento digital para o ruído de fundo e as distorções é desprezível. O que define então a qualidade do áudio de um gerenciador digital são essencialmente os estágios analógicos com os conversores AD e DA nas entradas e saídas. Esses por muito tempo (que felizmente já passaram) de fato constituíam um gargalo para a faixa dinâmica.

Contudo, é importante salientar que assim como desenhar um circuito analógico de qualidade inferior, é igualmente muito fácil degradar o áudio com processamento digital impróprio. Processamento digital oferece, mas não garante uma qualidade impecável de áudio!

2) O *routing*, os filtros e os limiares dos limitadores podem ser adaptados facilmente a qualquer conjunto de alto-falantes, desde pequenas caixas ativas para entusiastas de Hifi até os maiores sistemas de sonorização, com número de vias, freqüências de cortes e potência arbitrários. De fato, além da qualidade superior do processamento dos sinais, a facilidade de transferir, armazenar, organizar e alterar os *setups* contribuiu muito para a grande popularidade dos gerenciadores digitais.

3) As funções do processamento digital têm estabilidade incondicional, elas não mudam ao longo dos anos devido ao envelhecimento de componentes nem sofrem *drift* devido a mudanças das condições ambientais.

4) Alguns tipos de processamento de realização difícil ou até impossível no domínio analógico podem ser integrados com facilidade no domínio digital. O mais notório recurso desse tipo, de implementação trivial, é o *delay*, ferramenta indispensável para o ajuste temporal entre as vias de uma caixa de som e entre os *stacks* de sistemas de sonorização distribuídos. Outro exemplo de exclusividade do processamento digital são os filtros FIR *(finite impulse response)* que permitem influenciar independentemente a amplitude e a fase do sinal, recurso que só agora gradualmente ganha popularidade em produtos comerciais. Por final, limitadores de pico com capacidade de "prever o futuro" podem começar a reduzir o nível do sinal antes mesmo de um transiente provocar ceifa do amplificador ou até dano mecânico de um alto-falante.

Este artigo é baseado num novo gerenciador que reúne todas essas vantagens da tecnologia digital num produto de uma altura de unidade de *rack* (44 mm), 4 entradas e 8 saídas analógicas, uma entrada e saída no formato AES/EBU, e conexões RS232 e LAN para o controle remoto. Um cuidado especial foi dado à qualidade do áudio, particularmente à faixa dinâmica.

Infelizmente, é impossível perscrutar todos os pormenores de um equipamento tão complexo num artigo conciso de só 8 páginas. Por isso, só alguns aspectos de destaque serão abordados, a começar com a questão qual a taxa de amostragem mais adequada.

1. 48 OU 96 KHZ ?

O pró e contra do dobramento da taxa de amostragem para 96 kHz já foi amplamente discutido na década passada. Vamos rever alguns argumentos, com enfoque nos aspectos e requerimentos de sonorização ao vivo.

Quanto à audibilidade de som acima de aproximadamente 20 kHz, cuja reprodução requeria a taxa elevada de amostragem, tudo indica o seguinte: Nós não ouvimos ultra-som *de jeito nenhum*, nem de forma direta, nem de forma indireta. Cientistas japoneses [2] conduziram uma série de testes bastante reveladores para desempatar esse quesito. Neles, eles misturaram componentes ultra-sônicos ao sinal de áudio e perguntaram aos ouvintes se esses perceberiam uma diferença. E muitos de fato perceberam. No entanto, nesses ensaios, o ultra-som foi reproduzido pelo mesmo amplificador e pelo mesmo *tweeter* que a música, o que levanta a suspeita que produtos de *intermodulação* caindo na faixa audível abaixo dos 20 kHz foram na verdade o que foi percebido pelos ouvintes. Para esclarecer essa dúvida, os experimentadores repetiram os testes, desta vez reproduzindo os componentes ultra-sônicos por *tweeters* separados. Resultado surpreendente: *nenhum* dos participantes dos testes foi capaz de detectar a presença do ultra-som, o que corrobora a tese que o acréscimo de ultra-som só é perceptível quando ele gera intermodulação com componentes incidindo na faixa audível.

Na área do som ao vivo, outros aspectos têm que ser levantados em consideração . Primeiro, nenhum sistema de sonorização de alta eficiência é construído de maneira que permita a efusão de ultra-som. Os *drivers* de compressão raramente têm uma faixa útil de freqüências que se estende muito além de 12 ou 13 kHz. O que se vê na faixa acima na resposta em freqüência são meramente ressonâncias singulares em algumas freqüências. Os *wave-guides* dos *line-arrays*, responsáveis para a formação de uma onda quase cilíndrica, igualmente só funcionam até no máximo 15 kHz. Além disso, a absorção de freqüências altas no ar, particularmente em países tropicais como o nosso com alta umidade do ar, iria malograr qualquer tentativa de

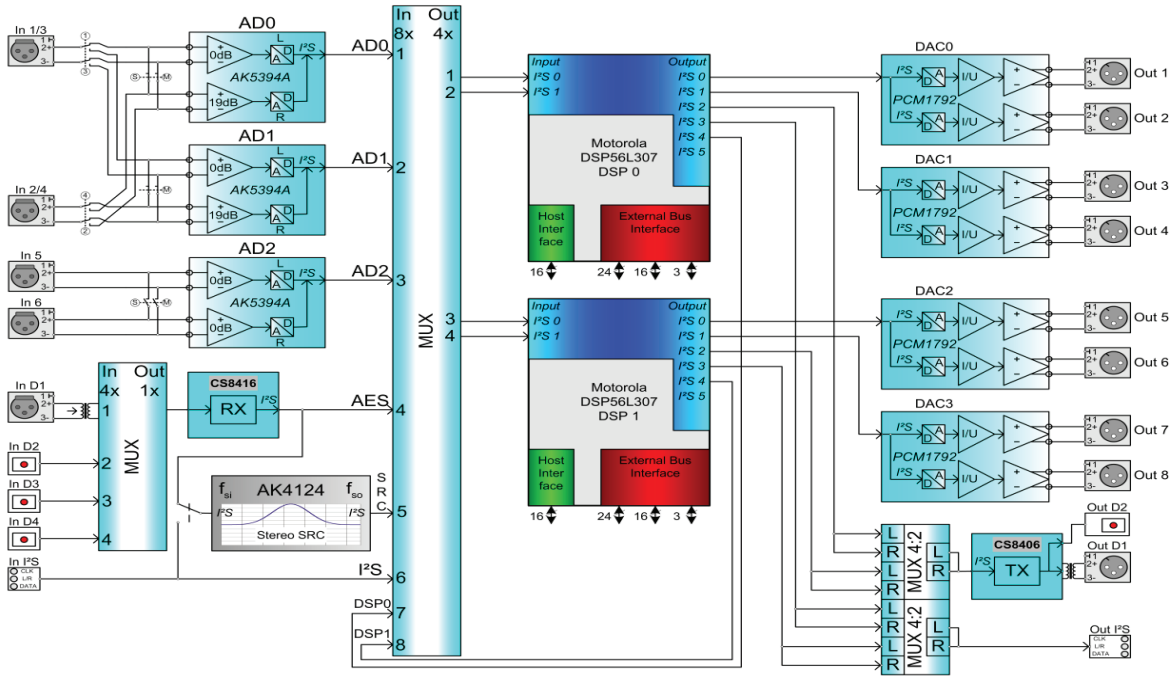


Figura 1: Diagrama de blocos do hardware para o processamento de áudio

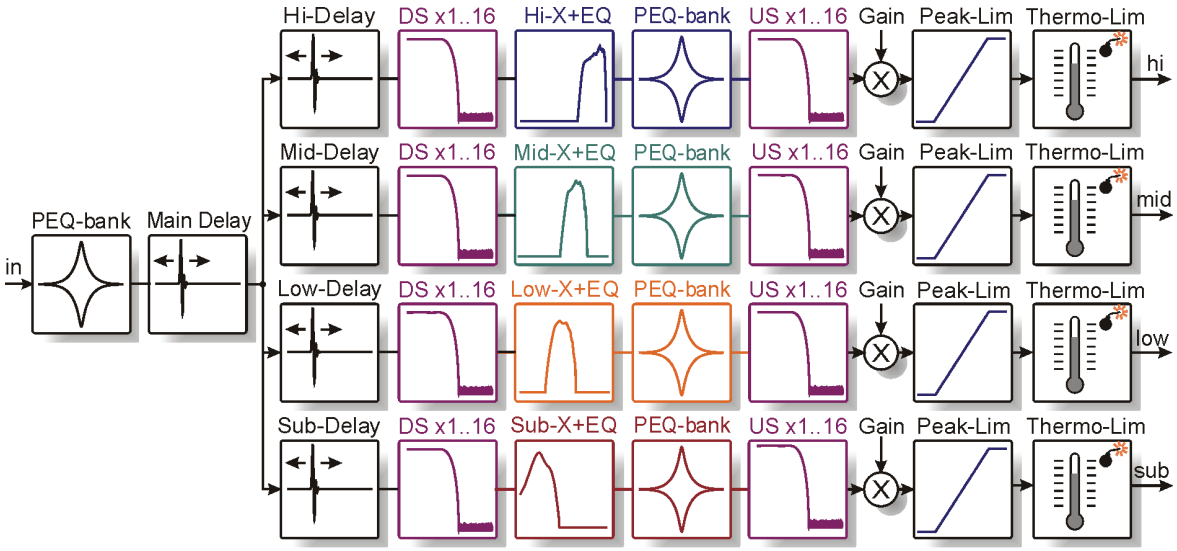


Figura 2: Processamento para 4 vias, com combinação livre de filtros FIR/IIR e redução opcional da taxa de amostragem

transmitir ultra-som através de distâncias médias e longas.

Do ponto de vista da praticidade, parece então fazer pouco sentido levantar a taxa de amostragem acima de 48 kHz num aparelho que serve para a sonorização de grande porte, e conseqüentemente, a maioria das mesas de som digital para som ao vivo trabalha com 48 kHz.

Porém, existem vários pontos em favor dos 96 kHz:

Primeiro, além de aumentar a relação S/R do processamento digital em 3 dB na banda audível (por espalhar o mesmo ruído de quantização agora até a nova freqüência Nyquist de 48 kHz), o emprego de *error feed back* ou até de *noise shapers* de ordem maior após cada requantização interna permite jogar a maior parte do ruído resultante dessa operação à banda acima de 20 kHz, onde ele não perturba mais.

Segundo, porque a resposta em freqüência de filtros IIR se assemelha mais com os gabaritos analógicos, se bem que essa questão já deixou de ser uma preocupação com a devida correção do fator Q de qualidade e o método de geração dos coeficientes segundo Orfanidis [6].

No entanto, a maior vantagem para o som ao vivo é a redução da latência, particularmente causada pelos filtros *anti-aliasing* nos conversores. Esses comumente trabalham com fase linear e resposta impulsiva simétrica, o que acarreta um atraso de tipicamente 1,4 a 2 ms para a conversão AD-DA com 48 kHz. Dobrar a taxa de amostragem logo faz despencar esse valor para a metade. Porém, uma redução bem maior da latência é possível empregando filtros com decaimento menos inclinado, o que torna a resposta impulsiva bem mais curta. Alguns conversores, como o PCM1792A, possibilitam escolher entre o filtro tradicional e outro do modo *"slow"* com resposta em freqüência bem menos íngreme. Com esse filtro, a latência cai para um terço comparado ao filtro tradicional, chegando a meros 187 µs @ 96 kHz. Aplicar o filtro *"slow"* com taxa de amostragem de 48 kHz não seria viável porque o decaimento da resposta em freqüência começaria bem antes de 20 kHz.

Devido a essas vantagens, optou se pelo emprego de 96 kHz, não obstante o fato que isso significa dobramento da carga computacional para os DSPs. Essa taxa também é mantida quando um sinal digital de 48 kHz (oriundo de uma mesa digital) é a fonte do áudio. Nesse caso, um conversor de taxa de amostragem assíncrono se empenha para adaptar a taxa do sinal à interna de 96 kHz.

2. CONVERSÃO AD COM 2 FAIXAS DE GANHO

A faixa dinâmica é um critério-chave de qualquer produto de áudio. Enquanto outros critérios igualmente importantes para a qualidade de áudio, tais como distorções harmônicas e desvios da resposta linear em freqüência, deixaram de ser uma preocupação simplesmente pelo fato que os valores atingidos hoje em dia são muito aquém da audibilidade, qualquer aumento da faixa dinâmica sempre pode ser verificado não somente pelos aparelhos de medição, como também com as nossas orelhas.

Como já mencionado anteriormente, os conversores AD e DA são peças cruciais para a qualidade total de um gerenciador. As primeiras gerações desses dispositivos não chegaram nem perto da faixa dinâmica de uma boa solução analógica e contribuíram para a má fama do áudio digital que alguns saudosistas alimentam até hoje. Desde então, a tecnologia evoluiu muito e atualmente, conversores AD e DA com uma relação S/R (sinal/ruído) entre 110 e 120 dB (calculada na banda até 20 kHz sem ponderação) são disponíveis em grande variedade por preços amigáveis.

Uma faixa dinâmica de 110 dB pode parecer exagero, tendo em vista que numa mixagem ao vivo com vários microfones no palco abertos, a relação S/R do *mix* não deve ultrapassar uns 60 ou 70 dB. Porém, musica ao vivo não é a única aplicação dos gerenciadores. Existem situações que demandam uma faixa dinâmica bem maior, por exemplo na sonorização em teatros e cinemas. Em trechos silenciosos da apresentação, o ruído de fundo do sistema de sonorização deve ser imperceptível até nos lugares mais próximos dos alto-falantes para não atrapalhar. Porém, efeitos naturais e especiais, tais como o estrondo em uma tempestade ou explosões, costumam ser reproduzidos com altíssimo nível, tornando necessário o aparelhamento do estabelecimento com alto-falantes de alta eficiência e amplificadores poderosos (e potencialmente ruidosos).

Para a reprodução de música, Fielder [1] chega à conclusão que uma faixa dinâmica de 120 dB é necessária para eliminar qualquer possibilidade de o ruído de fundo afetar a apreciação. Isso poderia nos levar a crer que simplesmente empregando os melhores conversores AD do mercado, que de fato alcançam uma faixa dinâmica de 120 dB, o ruído de fundo não deveria mais preocupar.

No entanto, não é bem assim. O motivo disso é que ainda precisamos de *headroom* para os limitadores embutidos no gerenciador. Para manter o som não-distorcido, o sinal da entrada precisa permanecer limpo mesmo com os limitadores já em ação, de preferência até a própria fonte do sinal (ou seja, a mesa de som) chegar ao limite de ceifa. Da mesma maneira, o ideal seria que o gerenciador não contribuísse significativamente com o próprio ruído de fundo quando os *master faders* da mesa de som estão fechados.

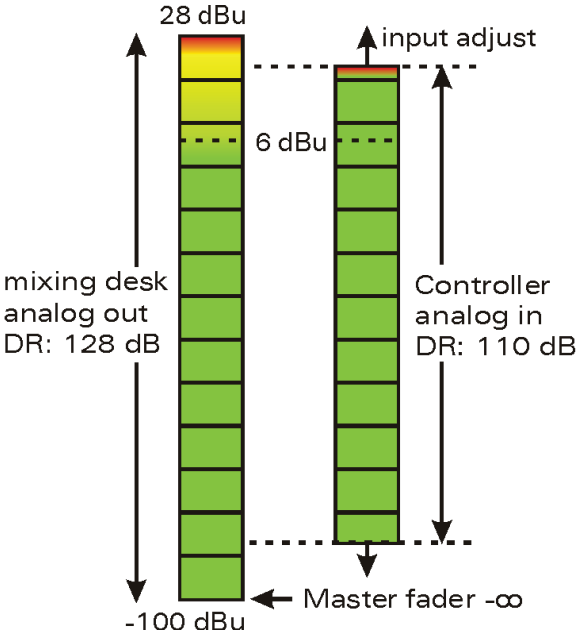


Figura 3: Gerenciador comum: Gargalo de faixa dinâmica

A Figura 3 demonstra o dilema. Como praticamente todas as mesas de som são equipadas com amplificadores operacionais alimentados com ±15 V a ±18 V, o nível de ceifa nas saídas simétricas gira em torno de 28 dBu (0 dBu = v 0.6 V RMS). Com os *faders* fechados, porém, só sobra o ruído de fundo do último estágio da mesa de som. Esse pode facilmente cair abaixo dos -100 dBu. A diferença entre esses dois valores equivale então a aproximadamente 130 dB, e portanto é esse valor da faixa dinâmica que seria desejável para um gerenciador de sistemas. Como os gerenciadores comuns do mercado chegam nem perto a esse valor, muitos deles são equipados com potenciômetros ou chaves de sensibilidade nas entradas. Sobra então para o engenheiro do som a tarefa de ajustar esses controles, avaliando criteriosamente se um amplo *headroom* ou um baixíssimo ruído de fundo é mais importante em determinada situação de sonorização.

Para tirar essa dor de cabeça do técnico, a técnica de conversão em duas faixas diferentes de ganho permite estender generosamente a faixa dinâmica da conversão AD. Funciona essencialmente da seguinte maneira: O sinal de entrada é passado a dois pré-amplificadores com ganhos diferentes, que por sua vez alimentam os dois canais de um conversor ADC estéreo (Figura 4). Os sinais digitalizados são passados ao DSP , que compensa a diferença do ganho e decide qual dos dois sinais é válido. Quando o canal com *maior* ganho começar a ceifar (ou melhor, pouco antes disso acontecer), o DSP chaveia para o outro canal com *menos* ganho.

Para que isso possa acontecer de forma inconsútil e totalmente inaudível, algumas precauções têm que ser tomadas. Primeiro, o circuito analógico tem que ser desenhado criteriosamente para que não haja *crosstalk* do canal *clipado* para o canal de *menor* ganho, dessa forma introduzindo distorção harmônica indesejada. Segundo, a diferença dos ganhos entre os dois canais é passível a pequenas variações devidas ao *drift* térmico do circuito. O mesmo acontece com o *DC offset*. Para evitar uma descontinuidade no sinal no momento da transição, ambos os valores têm que ser traçados de forma adaptativa. A velocidade da adaptação ainda é controlada pelo nível do sinal. Quando esse for baixo, a adaptação do *DC offset* é rápida e a da diferença dos ganhos mais devagar, e vice versa. Quando o canal com *maior* ganho ceifar, as adaptações param temporariamente com os dois valores congelados.

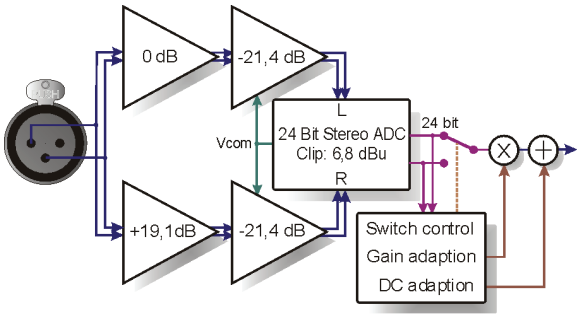


Figura 4: Princípio da conversão AD em duas faixas de ganho"

Graças a essas adaptações, o erro da estimação da diferença do ganho não ultrapassa ±0.003 dB e o da diferença dos *DC offsets* é de no máximo ±10 µV (com o valor *fullscale* sendo aproximadamente 20 Volts), já possibilitando uma transição inaudível.

Mesmo assim, algumas medidas adicionais servem para aperfeiçoar ainda mais a técnica. Para reduzir drasticamente a taxa de mudanças de um canal para o outro, o mascaramento temporal da nossa audição é explorado. Quando o canal com *maior* ganho deixar de ceifar, um período de repouso de 20 ms é esperado até escolher esse canal como fonte do sinal. Nesses 20 ms, o ruído de fundo é mascarado pelo som altíssimo presente antes da queda do nível, e portanto é inaudível. Se o sinal subir novamente dentro do intervalo de espera, duas transições (para o canal com *maior* ganho e de volta) desnecessárias são economizadas.

Adicionalmente, os limiares para o chaveamento de um canal para o outro não são iguais: Quando o nível aumentar, o limiar para chavear para o canal com *menor* ganho é de -2 dBFS (dB abaixo de *fullscale* do canal com *maior* ganho), enquanto o para chavear de volta para o canal com *maior* ganho é de -5 dBFS. Essa histerese evita que sinais com níveis estáveis provocam repetidos chaveamentos.

Mais um recurso é uma função "*look-ahead*", que faz com que um aumento do nível acima do limiar seja detectado com antecedência de 16 valores de amostragem. Esse detalhe é necessário porque os filtros *anti-aliasing* de fase linear do conversor AD já trazem reações (ondulações) antes de o sinal ceifado chegar à saída.

Com uma diferença dos ganhos de 19 dB e o emprego do conversor AK5394A, alcançando uma relação S/R de 118 dB por canal, a faixa dinâmica total ultrapassa a barreira almejada de 130 dB com folga.

A técnica descrita obviamente não melhora a relação S/R *instantânea*. Essa não pode ser melhor do que a dos canais individuais do conversor AD. Na praxe, porém, o aumento da faixa dinâmica, definida pelo nível de ceifa em relação

ao ruído de fundo sem sinal presente, é tão útil quanto seria um super-conversor que realmente atingisse essa relação S/R "por força própria". A diferença dos ganhos de 19 dB pode ser vista como uma extensão da faixa útil de níveis. Até um nível na entrada de aproximadamente 6dBu, o canal mais sensível garante uma digitalização com contribuição minúscula de ruído de fundo (< -105 dBu). Se o sinal ultrapassar esse limiar, a transição para o canal com ganho *menor* faz com que o ruído de fundo dê um salto para cima, fazendo a relação S/R *instantânea* cair de aproximadamente 115 dB para pouco menos de 100 dB. Parece dramático, mas devido ao mascaramento espectral do nosso sentido de audição, esse pulo do ruído de fundo é completamente inaudível em todas as circunstâncias. Isso até vale para os sinais mais críticos, que nesse caso seriam tons puros de baixa freqüência.

Para o técnico do som, a extensão da faixa dinâmica proporcionada pela conversão AD em duas faixas de ganho significa que ela não precisa mais se preocupar com o nível de áudio indo da mesa de som ao gerenciador de sistemas. O nível de ceifa dos dois equipamentos coincide, enquanto o ruído de fundo do gerenciador é até mais baixo do que o da mesa de som com os *master faders* fechados.

O ideal, contudo, é trabalhar com uma mesa de som digital e passar o *master mix* de forma digital (no formato AES/EBU) ao gerenciador de sistemas, evitando dessa forma perda de qualidade através de uma conversão dupla desnecessária.

3. CONVERSÃO DA

Uma alta faixa dinâmica também é desejável para as saídas de um gerenciador que fornecem os sinais processados aos amplificadores. Em contraste às mesas de som, a faixa dinâmica de amplificadores de potência normalmente é significativamente mais baixa. Exemplares excelentes atingem valores de 110 a 120 dB, mas a maioria alcança meramente de 100 até 110 dB.

Porém, a faixa dinâmica das saídas do gerenciador deveria ser notavelmente superior a esse valor. Primeiro, vale lembrar que conectar um gerenciador a um amplificador com exatamente a mesma faixa dinâmica já faria o valor total do sistema de sonorização cair em 3 dB. Uma perda ainda bem maior acontece se o amplificador possuir ganho alto, fazendo com que o nível de ceifa dele seja alcançado bem abaixo do nível máximo que as saídas do gerenciador são capazes de produzir. Nesse caso, a faixa acima desse nível se torna inutilizável. Por outro lado, como o gerenciador deve ser capaz de tirar a potência máxima de qualquer amplificador, o nível máximo das saídas do gerenciador não pode ser inferior a aproximadamente 18 dBu, garantindo que um amplificador com ganho de 26 dB (que constitui o mínimo encontrado no mercado) ainda seria capaz de produzir 7600 Watts em 2 Ohms.

Para contornar o problema de perda de faixa dinâmica no casamento com os amplificadores, muitos gerenciadores possuem potenciômetros nas saídas, alguns motorizados para permitir o controle e *recall* pelo software. Porém, esses controles aumentam os custos, podem quebrar e ainda diminuem a área disponível no painel dianteiro, comumente com altura de meramente uma unidade de *rack*. Outra possibilidade é de acrescentar potenciômetros digitalmente controlados atrás dos conversores DA. No entanto, essa solução é pouco convincente porque a faixa dinâmica desses potenciômetros não é maior do que a de bons conversores DA e eles ainda acrescentam distorção (e também aumentam o preço final do produto).

A solução mais versátil é empregar conversores DA com a faixa dinâmica mais alta possível. Optou-se pelo melhor conversor de áudio atualmente disponível no mercado, o PCM1792A da Burr-Brown. No entanto, a relação S/R espetacular de 126 dB (20 Hz - 20 kHz sem ponderação) prometida na ficha técnica é difícil de alcançar. O conversor é muito sensível a *jitter* (pequenas variações dos tempos das comutações de um *clock)*, o que não surpreende porque os *noise shapers* requerem um *master clock* bem limpo para remover a parte grossa do ruído de quantização para fora da faixa audível [3]. Quanto menor o ruído de fundo desejado, maior é a sensibilidade para *jitter*.

No gerenciador descrito neste artigo, a relação S/R alcançado com os PCM1792A é de 121 dB (lin. 20 Hz - 20 kHz). Isso ainda é bem acima do valor de qualquer outro conversor DA de áudio e reduz efetivamente o perigo de perda de faixa dinâmica na conexão aos amplificadores. Na grande maioria dos casos, os amplificadores ligados ao gerenciador serão a fonte predominante de ruído.

4. FILTROS IIR

Para realizar a funcionalidade de *crossover* e equalização, existem duas opções básicas: Os filtros com (IIR) e sem (FIR) realimentação interna. Os filtros IIR são os mais utilizados na área do áudio digital. Com uma estrutura de meramente 5 coeficientes e dois *delays*, o chamado *biquad*, eles são capazes de emular um filtro analógico de segunda ordem. Assim, eles constituem o elemento básico de filtragem e podem assumir quaisquer características também conhecidas no domínio analógico, tais como passa-baixa, passa-alto, *shelf*, *notch* e *bell*, além de outros. Para realizar funções de *crossover* com ordem superior a 2, vários *biquads* com parâmetros ligeiramente diferentes para gerar a característica desejada (*Butterworth, Bessel, Linkwitz-Riley*, etc.) são concatenados. Para a equalização dentro da faixa passante de cada via, são comumente empregados filtros paramétricos (*bell*) com freqüência central, largura de banda e ganho variáveis, além de filtros *low-shelf* e *high-shelf*.

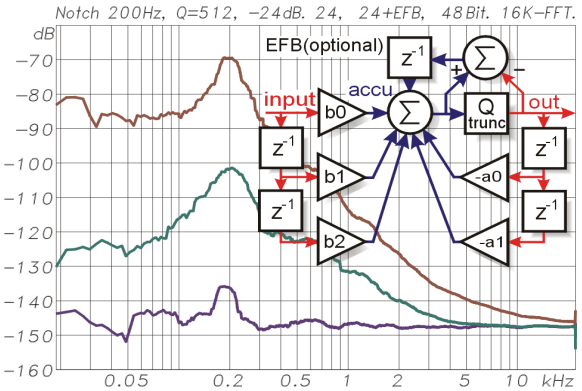


Figura 5: Topologia de um *biquad* na forma direta I com *error feed back* opcional, e ruído auto-gerado por um filtro *notch* com processamento interno de 24 bits, 24 bits com EFB, e 48 bits

O caminho do sinal para cada via é então sempre uma cadeia de múltiplos estágios conectados em série, o que infelizmente resulta em acumulo de ruído e distorção gerados nos estágios individuais, como também acontece em realizações analógicas. Isso significa que a qualidade de processamento de cada estágio deve ser bem superior ao resultado final desejado.

De fato, filtros IIR são bastante sensíveis à precisão restrita dos cálculos de um DSP. Um cálculo com meros 24 bits daria resultados deploráveis, com ruído, distorção e tons contínuos (*limit cycles*) mesmo na ausência do sinal de entrada. A introdução da técnica da realimentação do erro da requantização (*error feed back)* mostrada na Figura 5 já diminui os erros em 30 dB (curva do meio na Figura 5). Porém, para suprir as demandas do áudio profissional, só um cálculo com 48 bits integer (curva de baixo na figura5) ou 40 bits ponto flutuante garante resultados aceitáveis.

5. FILTROS FIR

Os filtros FIR seguem um conceito de filtragem diferente. Eles executam uma *convolução* do sinal de áudio com a *resposta impulsiva* do filtro desejado. Sendo assim, eles são capazes de emular quaisquer sistemas lineares e invariantes no tempo. Na área de efeitos, por exemplo, a *convolução* é utilizada para gerar reverberação natural com a ajuda de uma resposta impulsiva captada no ambiente emulado (salas de concerto famosas, por exemplo). Nos gerenciadores digitais, um só filtro em cada via executa ao mesmo tempo a função de *crossover* e toda a equalização do alto-falante. Em contrapartida aos filtros IIR, a fase pode ser tratada independentemente da amplitude com os filtros FIR. Isso possibilita basicamente três tipos de equalização:

5.1. Fase mínima

Na verdade, deveria ser chamado de "atraso mínimo". Um determinado módulo da resposta em freqüência sempre é acompanhado por uma fase específica em cada freqüência, que representa a atraso mínimo fisicamente possível para esse filtro. Quanto mais radicais e íngremes as mudanças no módulo do espectro, maior será a distorção da fase e do atraso de grupo (que é proporcional à derivada negativa da fase). O filtro FIR de fase mínima pode facilmente ser gerado dos casos seguintes através da transformada Hilbert.

5.2. Fase linear

Poderia também ser chamado de "atraso constante" ou "resposta impulsiva simétrica", pois essas são características intrínsecas dos filtros de fase linear. Significa que não importa quão radicalmente o percurso de uma resposta em freqüência possa mudar, ela não causa distorções da fase nem do atraso de grupo. Esse tipo de filtro é muito interessante para a função do *crossover* de uma caixa acústica, pois ele permite escolher uma inclinação muito alta (acima de 100 dB/oitava) para a separação das vias sem complicar o alinhamento temporal. Como maior benefício, a faixa de interferência por volta da freqüência de corte de duas vias da caixa acústica será muito estreita, praticamente eliminado problemas com *phasing* sob diversos ângulos. A grande desvantagem do filtro de fase linear é o atraso constante que ele introduz. Ele equivale à metade do comprimento de filtro FIR. Esse comprimento, por sua vez, é proporcional à seletividade do filtro (quanto mais íngreme a inclinação desejada na faixa transitória, mais comprida a resposta impulsiva armazenada no filtro FIR) e inversamente proporcional à freqüência de corte. Isso traz duas desvantagens desagradáveis: Primeiro, para atuar em baixas freqüências (por exemplo, para a separação dos *subwoofers* das torres principais), o filtro de fase linear acarreta um atraso entre aproximadamente 40 ms a 100 ms, o que inviabiliza o uso para som ao vivo. Segundo, um filtro FIR tão comprido extrapola o poder de cálculo dos DSPs se operado com a taxa de amostragem original. A conseqüência incontornável é que as vias para médios e baixos têm que ser operadas com taxa de amostragem reduzida. O *downsampler* no início do processamento e o *oversampler* no final da cadeia (Figura 2) compõem filtros passa-baixo que por sua vez acarretam um pequeno atraso, na soma algo entre 2 ou 3 ms para a via de subgrave, não impedindo o uso para som ao vivo.

5.3. Fase equalizada

Esse tipo de equalização vai a um passo além do último tipo apresentado: Ele equaliza não somente a amplitude do alto-falante, como também a fase, fazendo com que a cadeia completa do filtro mais o alto-falante exiba um atraso de grupo constante e perfeitamente independente da freqüência. Em outras palavras, ele elimina a dispersão do alto-falante. Com essa propriedade, a caixa acústica pode reproduzir de maneira perfeita um pulso curto, que vai aparecer como a resposta impulsiva simétrica de um passa-baixo ideal. Uma onda retangular também é reproduzida fielmente. Porém, isso só vale no ponto que serviu como base para as medições das respostas complexas em freqüência (compostas por módulo e fase), o que expõe uma dificuldade no uso dos filtros FIR: Enquanto a geração dos coeficientes do filtro FIR a partir das respostas é uma tarefa bastante *straight-forward*, o resultado depende da qualidade e representatividade da medição que serviu como gabarito para a equalização.

No novo gerenciador, todos os tipos de filtro podem ser livremente combinados: Em cada via, podem ser posicionados filtros IIR e/ou um filtro FIR com fase mínima, linear, ou equalização complexa, com redução opcional da taxa de amostragem entre 2 a 16 (Figura 2). Além disso, existe um banco de filtros IIR opcionais que agem logo na entrada, antes da separação para as várias vias, substituindo um equalizador de sistema externo. A alocação dos filtros IIR é dinâmica, quer dizer o número total de *biquads* aplicáveis é fixo (em função do poder computacional do DSP disponível).

O esquema mais popular de equalização numa geração anterior do novo gerenciador fez uso dos filtros FIR com equalização complexa nos agudos é médios, fase mínima nos graves, e pequenas correções de banda larga mediante o banco de filtros IIR antes da ramificação das vias. Esse método é o preferido por usuários que possuem e sabem utilizar um sistema de medição com base em FFT.

A operação exclusiva com filtros IIR é especialmente indicada para usuários que não dispõe de um sistema de medição, mas têm as recomendações do fabricante da caixa acústica relativas às freqüências de corte e equalização na mão.

6. LIMITADOR VIDENTE DE PICOS

Limitadores analógicos de pico (Figura 6) sofrem um dilema: Quando um transiente ultrapassar o limiar de acionamento, já é tarde para reduzir o volume – o pico do sinal vai crescer acima do limiar e, portanto, ainda pode sofrer ceifa. Para combater esse problema, a velocidade de regulagem (*attack*) pode ser aumentada. Mas isso, a partir de um certo ponto, vai surtir distorções bem audíveis e bastante desagradáveis. Outra possibilidade é posicionar o limiar de acionamento abaixo do nível que vai causar ceifa do amplificador ou danos mecânicos do alto-falante. Porém, isso vai ocasionar acionamento do limitador mesmo se o pico nem alcançar o limite de ceifa.

De qualquer maneira, o circuito RC para suavizar a voltagem que controla o VCA (*voltage controlled amplifier*) constitui um filtro de ordem 1 com inclinação de

meramente 6dB/oitava. Significa que o sinal de controle do VCA, oriundo do comparador que gera um sinal binário de alta amplitude (sinal do topo na Figura 7), ainda contém alto teor de componentes de alta freqüência que ocasionam distorções feias. Utilizar um filtro de ordem maior não é remédio viável num limitador analógico porque iria aumentar a lentidão de reação.

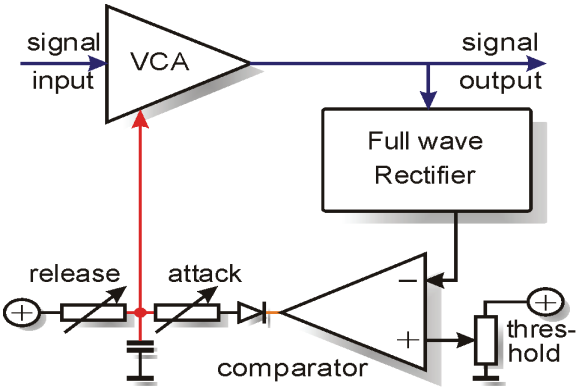


Figura 6: princípio de um limitador analógico de picos

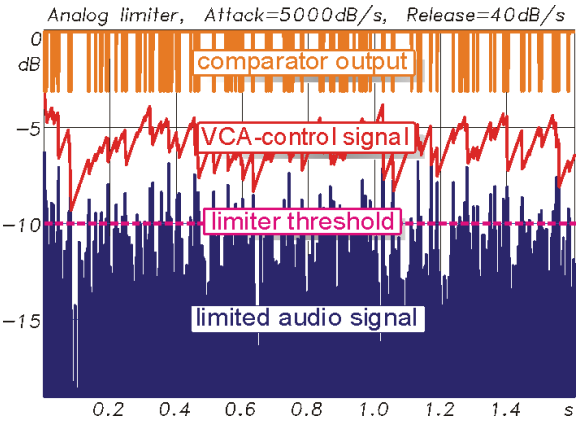


Figura 7: Sinal saindo do comparador, sinal de controle do VCA, e áudio saindo do VCA no limitador analógico

Para completar a lista das desvantagens, um VCA analógico aumenta o ruído de fundo e as distorções mesmo se o limitador nem está em ação.

O processamento digital permite contornar de forma elegante os problemas mencionados. Primeiro, se o limiar de acionamento não é alcançado, o limitador digital deixa passar o sinal sem alterações, não acrescentando nenhum ruído e nenhuma distorção.

Para driblar o problema que o limitador só percebe um nível alto quando já é tarde, seria útil se ele já soubesse de antemão como será o percurso do sinal no futuro próximo. Essa comodidade pode ser facilmente realizada passando o sinal por uma pequena *delay line* de apenas 1ms. Ao detectar um transiente que iria ultrapassar o limiar, o limitador pode dispor desse tempo para reduzir gradualmente o ganho do sinal, com inclinação constante em dB/s, até que o topo do pico incide exatamente com o limiar ao sair da *delay line*.

O sinal liso do "controle do VCA" (Figura 10) resultante dessa operação já diminui muito a contaminação com componentes de alta freqüência (Figura 11). Um aperfeiçoamento do método é passar o sinal de controle por um passa-baixo Bessel (na implementação, com freqüência de corte de 500 Hz e ordem 6). Para comportar o atraso de grupo constante de aproximadamente 1ms desse filtro, a *delay line* é simplesmente aumentada por esse valor.

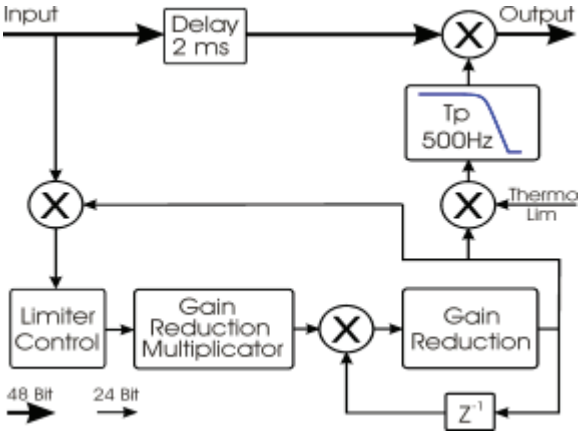


Figura 8: Princípio do limitador digital com *look-ahead*

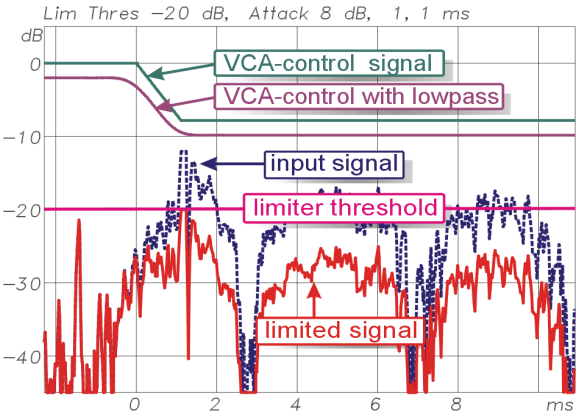


Figura 9: sinal de controle de VCA do limitador digital com *look-ahead*. Em baixo: sinal de entrada (pontilhado) e de saída

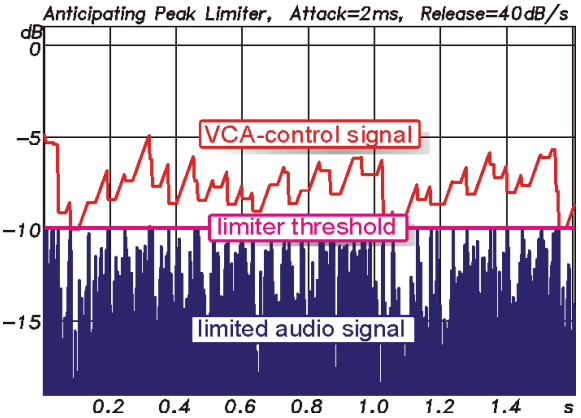


Figura 10: Mesmo trecho que na Figura 7, agora tratado pelo limitador digital: sinal de controle do VCA e sinal de saída limitado

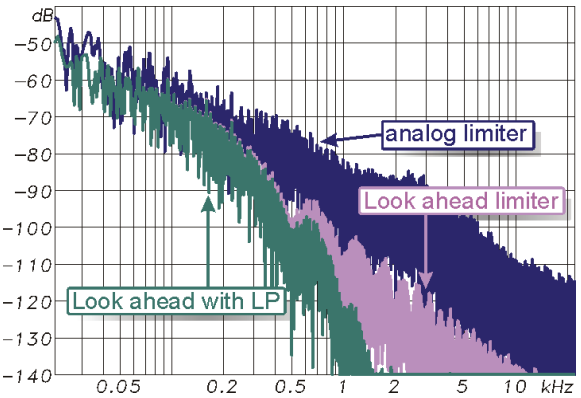


Figura 11: Espectros do sinal de controle do VCA nos limitadores analógico e digital, sem e com filtragem

O limitador é completado com lógica para realizar a fase *hold* (50 ms) e o *release*, em dB/s constante ajustável, como em limitadores tradicionais.

O procedimento descrito não significa que a ação do limitador não seja mais audível. Pelo contrário, ao apresentá-lhe um sinal muitos dBs acima do limiar e escolhendo uma taxa alta do *release* (acima de 100 dB/s), o fator de crista será reduzido consideravelmente e o áudio vai ganhar densidade (uma técnica que infelizmente hoje em dia é utilizada em excesso na masterização de CDs e transmissão de rádio, ultrapassando os limites do bom gosto). Porém, o que o método assegura é que o áudio deixa de ser modulado por resquícios da onda quadrada saindo do comparador, eliminado a aspereza agressiva do som resultante desse tipo de distorção em limitadores convencionais.

O limitador de picos serve para evitar tanto ceifa do amplificador quanto danos mecânicos dos *drivers* (excursão além do limite no caso de *woofers*, aceleração excessiva no caso de *tweeters*). Muitas vezes, por exemplo, no caso de *subwoofers* potentes, é o amplificador que de fato perde primeiro o fôlego. Para evitar ceifa dele em todas as circunstâncias, o limiar do limitador teria que ser ajustado em concordância com a potência contínua que o amplificador é capaz de produzir. Porém, tal procedimento não exploraria a capacidade da maioria dos amplificadores de reproduzir *bursts* de curta duração com nível 2 ou 3 dB (ou até mais) acima da potência contínua.

Para não desperdiçar essa reserva, o limitador de picos foi evoluído a fim de traçar a descarga dos capacitores eletrolíticos na fonte dos amplificadores, responsável pela diminuição da potência do sinal quando aplicado continuamente. Essa extensão, chamada de *"controlled overshoot"*, introduz dois novos parâmetros para a programação do limitador: O valor *surge*, que informa a potência RMS inicial de curta duração, e o valor *surge duration*, definido pela interseção da reta inicial do decaimento da potência com a linha horizontal da potência contínua (Figura 12). Esses valores podem ser avaliados facilmente ligando a carga e um osciloscópio ao amplificador (o segundo canal precisa ser carregado também) e tocando um *burst* de 100 ms ou mais com nível que causa ceifa desde o início. Se esses recursos não estiverem ao alcance do usuário, um valor *surge* 2 dB acima da potência contínua e uma *surge duration* de 30 ms combinam bem com a maioria dos amplificadores profissionais de alta potência.

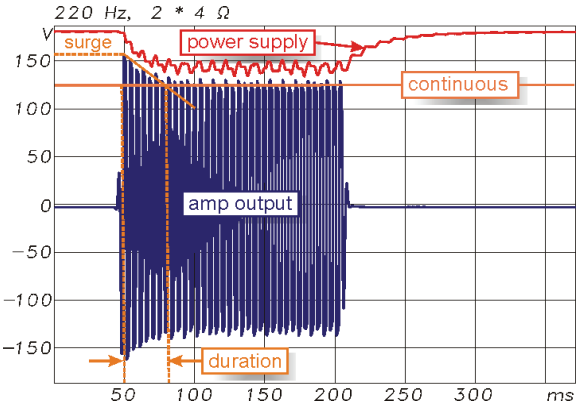


Figura 12: Medição dos parâmetros para o "*controlled overshoot*"

A descarga dos capacitores é modelada a partir do quadrado do sinal e assumindo uma carga ôhmica independente da freqüência. É claro que a impedância de caixas acústicas reais difere bastante desse ideal, mas na praxe, a aproximação do modelo funciona surpreendentemente bem.

Além do limitador do pico, existe um segundo limitador RMS para evitar superaquecimento em cada via. Esse limitador funciona independentemente do limitador de picos e somente compartilha o estágio "VCA" (a multiplicação do sinal com o fator de atenuação) com este. O limitador térmico precisa de dois valores: A potência RMS contínua que o alto-falante agüenta com duração infinita, e o *time constant*, que desta vez é a interseção da inclinação inicial do aquecimento (quando aplicar a potência contínua) com a temperatura máxima.

Como no caso do limitador de picos, a realidade é bem mais complexa porque o aquecimento é dependente da impedância e da excursão da bobina móvel, ambos fatores com forte dependência da freqüência, sobretudo em *woofers*. Porém, contando com as propriedades estatísticas do sinal musical, uma boa proteção do alto-falante sem desperdiçar muitas reservas é factível.

De qualquer maneira, um técnico de som sensato e responsável não se aventura para exprimir as últimas frações de dB NPS do sistema de som. É tolice comemorar um som altíssimo durante o show, mas depois ter que trocar *drivers* queimados aos montes.

7. CONCLUSÕES

A tecnologia digital de áudio avançou o suficiente para tornar uma solução digital de gerenciadores de sistema superior em praticamente todos os aspectos de qualidade, comparados com os antecessores analógicos. Isso vale sobretudo para critérios chaves como distorções, faixa dinâmica e precisão dos filtros de *crossover* e equalização.

Com a livre escolha de filtros FIR e IIR, o usuário pode optar para a equalização automática a partir de respostas em freqüências medidas de cada via, com total flexibilidade na escolha do comportamento da fase, ou enveredar pelo caminho tradicional de programar filtros de segunda ordem passo-a-passo.

Resta comemorar que com os avanços consideráveis que os fabricantes de caixas acústicas profissionais obtiveram sobretudo no controle direcional, com o advento dos *line-arrays* que possibilitam um som bem mais homogêneo, e com as ferramentas de medição interagindo com os gerenciadores via controle remoto, o som ao vivo alcançou um patamar de qualidade que compete com os melhores sistemas de som em casa e em estúdios.

8. REFERÊNCIAS BIBLIOGRÁFICAS

[1] Louis D. Fielder, "Dynamic-Range Issues in the Modern Digital Audio Environment", J.AES 1995, p. 322-339

[2] Ashihara Kaoru, Kiryu Shogo, "Detection Threshold for tones above 22 kHz", 110th AES Convention, Amsterdã, paper 5401

[3] Steven Harris, "The Effects of Sampling Clock Jitter on Nyquist Sampling Analog-to-Digital Converters and on Oversampling Delta-Sigma ADCs", J.AES 1990, S.537-542

[4] John Vanderkooy, Stanley P. Lipshitz, "Resolution below the least Significant Bit in Digital Systems with Dither", J.AES 1984, S.106-113

[5] John Vanderkooy, Stanley P. Lipshitz, "Dither in Digital Audio", J.AES 1987, S.966-975

[6] Sophocles J. Orfanidis, "Digital Parametric Equalizer Design With Prescribed Nyquist-Frequency Gain", J.AES, p. 444, junho 1997

Sociedade de Engenharia de Áudio
Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, 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 the delay out of digital mixing systems

Dr Robin Clark
Allen & Heath Limited, Cornwall, United Kingdom
robin.clark@allen-heath.com

ABSTRACT

Digital mixing systems are susceptible to signal propagation delay and internal delay paths, causing intelligibility problems for musicians and colouration when mixing correlated signals. Reducing latency is difficult due to high channel counts, processing/configuration, audio I/O and networking requirements. Latency minimization in hardware, DSP architecture, signal processing and the practical mixing system set-up is discussed. The work aims to generate dialogue and inform the audio engineer of the problems associated with latency in large scale mixing systems at component and system level.

0 INTRODUCTION

Signal propagation delay or 'latency' is the signal delay through a system from input to output, independent of frequency. Digital mixing systems are susceptible to higher latency than in analogue systems. This increase in latency is reported to produce intelligibility problems – predominant in musician in ear monitor applications.

Digital mixing systems are also susceptible to delays in internal signal paths causing frequency response colouration when correlated sources, with different delay paths, are mixed. (Correlated sources are two signals with similar audio content, for example a stereo pair of channels or microphones in close proximity with similar pickup).

Reducing latency and delay paths in large scale mixing systems can be problematic. Problems are accentuated through the requirement of high input, bus and mix counts, processing at multiple points in the signal path and multiple internal processing fx engines. Furthermore, providing flexible system configuration, digital I/O interface, audio networking and different sampling rates is also problematic in reducing latency and delay paths.

Initially the work explores the various effects of latency and delay paths. System signal delays at hardware component level are discussed, including A/D, D/A, sample-rate converters, digital I/O interfaces, giving explanation into the reason for these latencies and effects from minimising them. DSP architectures are also discussed along with their implications on latency and issues arising from distributed processing and reconfiguration. Techniques are discussed to minimise latency in key processing algorithms – mixing, dynamics and equalisation. The side effects of minimising latency in these algorithms is also detailed. The work concludes with latency and delay path analysis in a practical digital mixing system set up.

1 LATENCY/TIMING ISSUES

1.1 Instruments

Figure 1 shows the musicians direct and monitor mix signal paths. The musician's own instrument direct path is either acoustic or a local amplifier/speaker. The latency of the direct path is simply based on distance from source and in critical cases the musician can position themselves close to the direct path to reduce latency.

It is when the musician uses the monitor mix that latency timing problems start. If the musician monitors their own instrument through their direct path and the rest of the instruments through a monitor mix, their own direct path is

a pre-event in time to the rest of the instruments. The musician can feel they are playing out of time.

If the musician has no direct local feed and monitors all instruments, including their own, purely through the system monitor mix, their instrument is aligned to the rest of the monitor mix. For some musicians this still upsets their timing through the monitor latency delay between playing the note and hearing it. Typically musicians report timing problems with monitor mix latency delays above 5 ms, feeling disconnected from their instrument, whilst some percussive musicians claim to be affected with much lower latencies, around 1 ms (typical snare drum attack envelope is 0.5ms). In these scenarios in-ear monitors do help reduce latency by eliminating the acoustic path from monitor to ear. However there is still latency incurred through the digital monitor mixer, digital mixers typically have a basic input to output latency of around 2ms.

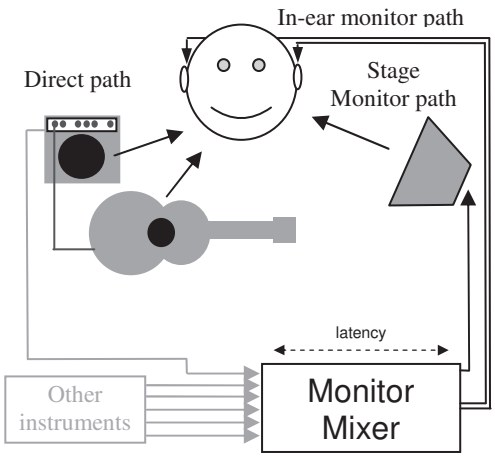


Figure 1 Musician with direct and monitor signal paths

1.2 Vocalists

Vocalists have an 'in-head' direct reception of their own voice (directly through their skull). The 'in-head' path has negligible latency. In-ear monitors provide vocalists with high quality audio direct to the ear. It is when the delayed in-ear monitor feed sums with the 'in-head' reception that intelligibility problems occur for the vocalist.

Vocalists notice effects during slow and continuous pitch change. Pitch change produces an effect such that the delayed monitor feed is at a pitch different to the current 'in-head' pitch. The vocalist hears their current in-head direct path pitch and also hears the latent pitch – producing colouration or 'micro-flanging', reducing intelligibilty. This pitch change effect seems to be noticeable for monitor latencies of above 4 ms (informal tests at Allen & Heath).

Vocalists report other colouration effects using in-ear monitors at lower latencies. These effects are described as lack of vocal power, 'phasy' sound, 'dull' sound or 'less intelligible'. This is due to comb filtering caused by the time difference between the 'in-head' signal with the in-ear monitor mix (due to the latency in the in-ear monitor mix). The change in frequency response characteristic or basic comb filter is easy to model (1), where d is the latency time through the in-ear path.

$$\mathrm{Vocal}_{\mathrm{head_path}}(t) + \mathrm{Vocal}_{\mathrm{inear_path}}(t-d) \qquad (1)$$

Figure 2 shows the resultant amplitude-frequency response of mixing two identical sources together with a latency difference of 2ms (comb filtering in the vocal range, 250 and 750 Hz). However this is too simplistic, the in-head feed is not identical to the in-ear feed in level or frequency response. If the in-head signal is attenuated by 10dB relative to the in-ear signal the comb filtering is much less severe, figure 3. This demonstrates comb filtering in the percieved frequency response (in the vocal region) of the summed vocal for a typical latency of 2ms. Reducing the latency to 0.5ms, the comb filter response dips at 1 and 3 kHz, also in the vocal frequency range.

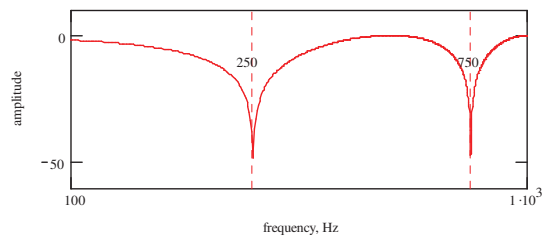


Figure 2 Comb filter amplitude-frequency response, two identical sources, latency delay of 2ms.

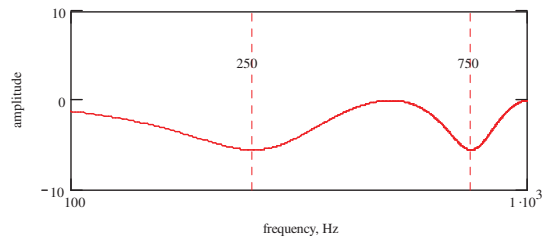


Figure 3 Comb filter amplitude-frequency response, 10dB relative difference in level, latency delay of 2ms.

1.3 Delay differences in signal paths

Digital mixing systems are susceptible to delays in internal signal paths. This causes problems when mixing signals together that have any relative delay and are correlated signals. If two correlated audio sources with a delay difference are mixed together a comb filter effect occurs. Figure 4 shows the comb filtering effect for two signals mixed together with a delay path difference of only 20 us (1 sample at 48kHz) with a 10dB level difference. Noticeable high frequency roll-off, above 2 dB at 12kHz. As the delay path increases any comb filtering effect becomes more noticeable in the audio frequency band. Figure 5 shows the frequency response for a delay path difference of 4 samples at 48kHz sampling rate. Level difference of 20 dB.

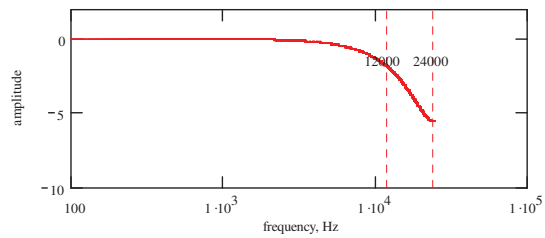


Figure 4 Comb filter effect, two signals mixed with a delay difference of 20 microseconds 10dB difference in level.

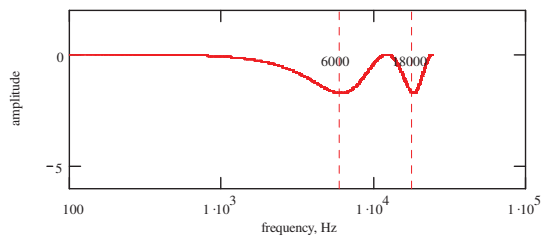


Figure 5 Comb filter effect, two signals mixed with a delay difference of 83 microseconds (4 samples) with a 20dB difference in level.

2 LATENCY IN HARDWARE

This section discusses the latency associated with the hardware components in a digital mixer, ADC/DAC converters, Digital IO and sample rate converters, audio networks and DSP archiectures, as shown in Figure 6.

2.1 Analogue to Digital Converters (ADC)

ADC's require anti-aliasing filters to reject signal bandwidth above the Nyquist frequency. Originally these were high order analogue filters with good stop band rejection and flat pass band, but suffered high audio band phase distortion. Delta-sigma oversampling techniques shifted the anti-aliasing filter into the digital domain as a post modulator decimation filter. These were implemented as FIR filters for linear phase response, unfortunately increasing latency, typically by 30 to 50 samples. If the FIR filter length is reduced, latency is reduced but at the expense of stop band attenuation and passband ripple. Compromising stop band attenuation results in out of band signals reflecting into the audio band. Compromising pass band ripple results in audio colouration. Passband ripple is intuitively an amplitude frequency response issue, but is directly connected to pre-echo - transient smearing in the impulse response of linear phase filters.

One technique to reduce latency is to employ IIR filters in the decimation stage [1]. IIR filters can be designed to provide sufficient stopband rejection and good passband with low latency. However the IIR decimation filter has a non-linear phase response. Phase compensator techniques can be employed to approximate a linear phase response. Since IIR filters are not linear phase they do not suffer from pre-echo but their impulse response is still of importance for audio transparency. IIR filter based decimators can result in the overall latency typically reducing to 10 samples.

Doubling sampling rate reduces IIR and halves FIR filter latency. At higher sampling rates the decimation filter can roll-off earlier relative to the Nyquist frequency (since the passband is higher relative to the human ear's passband). This relaxes the filter roll-off characteristic reducing latency, especially in FIR designs.

2.2 Digital to Analogue Converters (DAC)

DAC filtering stages are employed for analogue reconstruction and image rejection (rejecting bandwidth of multiples of the sampling rate). Higher order analogue filters can also be used externally (post DAC) and can relax the requirements of the digital filter stage. These digital filters can be implemented as FIR, typically resulting in latencies of 20 to 40 samples.

IIR filter implementation again reduces latency. Typical latency figures for a DAC IIR filter design is 10 samples, with the same design issues as for ADC. At higher sampling rates latency is reduced in the FIR and IIR filter topologies. A doubling of sampling rate can half the latency and shifts the phase distortion and rejection band to the higher less audible frequency range again also facilitating simpler filters, which could reduce latency further.

2.3 Digital IO and Sample Rate Converters

AES3, SPDIF, 2 channel optical, ADAT and MADI digital audio formats are low latency digital audio transports, typically with a 2 sample delay for a transmitter or receiver stage. These formats rely on synchronous clocked encoded audio data streams passing between transmitters and receivers. Complex systems are kept synchronous using a house sync, where one device is clock master and all other devices slave from this sync, the slave receiver clock or a dedicated sync, e.g AES11.

If multiple sampling rates exist, or digital equipment cannot slave to an external clock then sample rate conversion (SRC) is employed at the receiver. This de-couples the sampling domains with a penalty in latency.

SRC's employ interpolation and decimation stages to adjust effective sampling rate, typically producing latencies of 40 samples each - around 1.6ms total at 48kHz. If the SRC is always downsampling the decimator can be disabled, decreasing latency by 40 samples. However, this is not an easy scenario to predict for live applications. Some SRC's provide interpolator buffer options to reduce distortion for large changes in incoming sampling rates, increasing latency by 40 samples. However if the application does not need to accomodate large sampling rate changes the buffering option should be disabled.

2.4 Digital audio networks

There are many audio network technologies, all of which have advantages and disadvantages and have suitability to different applications. For the purposes of this work we are solely interested in latency.

EtherSound is an Ethernet based audio network which dedicates the 100Mbits bandwidth to a synchronous framework for audio channel delivery, embedding some control bandwidth into the synchronous framework. This provides a fixed latency system, 5 samples audio input to audio output, with a small latency per hop for cable propagation delay. EtherSound does not attempt to phase lock audio across nodes. Delay paths of a few samples can exist in Ethersound systems across nodes due to the propagation delay and frame clock phase difference between nodes. However Ethersound nodes can be up to 64 channels so it is easy to keep correlated sources on one node and thus in phase.

AES50 is loosely Ethernet based and implemented on a synchronous framework to deliver fixed low latency of around 5 samples. AES50 incorporates a transmitted clock to provide sync across devices (to phase lock devices giving fixed delay paths on all inputs). However specialised routers are required to implement beyond point to point increasing latency by typically 5 samples.

OptoCore is a synchronous bus aggregating audio streams, USB video and Ethernet traffic onto a 1Gbit time division multiplex bus optical network. The bus presentation time on and off is 2 samples at 48kHz, 41.6 us. Each node of up to 8, introduces a transmission delay of 0.2us. This is an extremely low latency bus, relying on

relatively expensive optical links to achieve a high bandwidth long haul interconnect.

Cobranet is a fully Ethernet compliant audio network working freely alongside Ethernet asynchonrous traffic. Cobranet achieves this through buffering and transmitting audio as multi-sample bundles over standard Ethernet packets - merging the audio with standard traffic packet sizes. This audio buffering increases latency considerably, originally 5ms audio input to output. However Cobranet offers low latency modes by reducing the buffer length at the cost of bandwidth. Typical figures for latency range from 5 ms down to 1.3 ms.

IEEE1394 (Firewire) is a multimedia based bus network capable of streaming audio and video - designed as a multimedia PC interface. Latencies vary across implementations, slightly higher than synchronous Ethernet, for example, 400us, 20 samples, input to output. It is often used for computer audio interfacing and is dependent on driver latency at the PC end.

USB is a popular point to point bus interface for computer audio based systems. Latency is strongly reliant on software drivers at the computer end. Specialised low latency drivers can achieve 3ms per hop. Standard ASIO drivers often achieve 10ms latency per hop.

2.5 DSP architectures

Large scale digital mixing systems generally employ several, often dozens of DSPs in a signal processing network. These DSP's share and pass audio through either point to point multi-channel streaming and/or a shared common bus. From a DSP hardware interconnection perspective there is little difference in latency. All audio interconnects have to transfer data within a sampling frame and require I/O data buffering inside the DSP. DSP latency from input to output is typically 3 samples. It is the actual DSP architecture that has more influence on latency and delay paths.

- **Fixed DSP architectures**

Fixed DSP architectures have a fixed I/O structure, DSP's are permanently configured to perform a fixed set of audio processing tasks. This system has the ability to be very low latency, since the network can be designed to provide minimal signal path through the DSP architecture. However these architectures are usually very fixed in configuration, and to change the number of mixes or change the way the mixes operate can often be impossible. They also tend to suit simple smaller mixers, where the number of channels and mixes is not bound by I/O or processing capacity at each DSP node. Owing to the small nature of these architectures latency, input to output, could typically be the path through 2 to 4 DSPs, for example 6 to 14 samples (120us to 300us at 48kHz).

- **Flexible DSP architectures - 'DSP farm'**

A DSP 'farm' contains multiple DSP's in mesh or cluster arrangements, employing processor nodes and/or shared memory for audio distribution across the farm. These architectures are employed for very large mixing systems, for reconfigurability of the entire DSP architecture across large channel counts. DSP farms require multiple routing stages and large amounts of signal traffic across processor nodes, creating high latency.

Latency in a DSP farm is governed by process scheduling. This is a complex design process, based on what flexibility is required in signal routing, mixing and processing at certain points of the audio signal path [2]. The mixing and processing requirements dictate DSP utilisation and routing - latency and delay paths change based on the configuration. To compensate for the delay path differences, further delays are inserted into audio channels to equalise the delay paths - increasing latency further. DSP farms can typically involve 30 to 70 DSP's. Given high amounts of processing concurrency, a signal path may still involve 8 DSP's, including DSP routing and delay balancing a latency approaching 50 samples (1ms at 48kHz) could be typical.

- **Hybrid DSP/FPGA systems**

Low latency and optimal delay path control is best achieved with fixed concurrent or parallel DSP structures where possible. Flexibility and reconfigurability is best achieved through an efficient centralised router that can route to key positions in the DSP system. This can be used to provide reconfiguration and flexibility within a fixed framework. Using recent advances in FPGA technology it is feasible to provide centralised routing to all large clusters of DSP. These clustered DSP's can operate concurrently, reducing latency, using the centralised router to sum the bus contributions and re-route mixes to mix processing blocks. Such hybrid routers provide DSP architectures that are flexible with low latency. For example, a hybrid DSP structure may contain 30 to 40 DSP's with a signal path route from input to output of several DSP's, however using minimal routing this may result in a typical latency of 15 to 20 samples.

FPGA technology is now also being used in performing concurrent DSP processing. Integrating multiple DSP tasks onto one FPGA. This is similar to concurrent DSP operation and yields low latency mixing solutions.

3 LATENCY IN DSP ALGORITHM DESIGN

The key DSP algorithms in mixing systems fundamental to latency are mixing, inserts, equalisation and dynamics processing.

3.1 Mixing

Mix algorithms are largely dependent on DSP and routing architecture and how signals get to and from the mix bus. Concurrency is heavily used in DSP mixing applications reducing latency. For example, inputs maybe processed in parallel and then mixed to partial bus in parallel. These are then summed to form a mix bus.

Mixing systems send mixes to mixes. Careful ordering and structure has to be applied to this process to avoid delay path problems. It should ideally incurr zero delay so that any input and mix can be mixed without any colouration effects. Often a delay path has to exist somewhere, it is then best to partition the process such that mixes that never mix to each other are partitioned sensibly.

3.2 Processing Inserts

Inserts provide a channel send to an additional processor (external or internal to the mixer) and return the signal into the channel just after from where it was sent. This requires additional DSP routing, a latency of approximately 4 to 8

samples. Inserting a traditional analogue channel processor incurrs additional ADC and DAC latency in the insert path, figure 6. This additional latency would be typically between 20 to 70 samples, depending on ADC/DAC latency.

An external digital insert has a lower latency if no SRC's are used in the inserted device or insert return. This can be achieved using a house sync or selecting the inserted device to sync from its input, forcing the return path to be in sync. Then all SRC's in the chain can be disabled. Typical latency for a non SRC'd insert would be 6 to 8 samples. Latency in an SRC stage is typically 40 samples.

Internal inserts or plug-ins can be used in some mixing systems, avoiding the external interface latency of 4 samples and the complexity of SRC's.

Inserts can also be patched over networks. The latency overhead of the network means a low latency network is essential. However a low latency network is considerably lower latency than a SRC based digital interconnect.

Although latency is very small for synchronous digital inserts there are still delay path differences between inserted and non inserted channels. If these channels have correlated programme and are mixed comb filtering will occur. Applying delay compensation on the non inserted channel avoids this. Delay compensation requires the application of the exact number of samples to work, the engineer needs access to this information. If the inserted channel has no correlated programme on other channels delay compensation is not neccessary.

3.3 Equalisation

There are 2 types of digital equaliser filter section, IIR filters (infinite impulse response - uses feedback paths) and FIR filters (finite impulse response - purely feedforward). IIR, if implemented as minimal phase filters, produce the same frequency phase response and group delay as the analogue counterpart. FIR filters allow more flexibility in filter design and are not minimal phase and have no analogue counterpart. FIR filters offer filter responses not possible in the analogue domain for example steep slope cross-overs with linear phase characterisitics. However FIR filters incurr high latency, often 100's samples. These are useful in loudspeaker management, where this may correspond to a few meters in speaker placement. But as a processing equaliser in a mixer channel would be problematic for latency sensitive signals.

3.4 Dynamics

Dynamics processing such as Gates, Compressors and Limiters have a finite attack response. This can result in transient material not being processed by the correct gain reduction. For example, a Gate may not open quickly enough, responding to a fast transient, or a Limiter may not attenuate a fast transient.

A solution is to place a 'lookahead' delay in the signal path, compensating for this finite transient response, this results in the dynamics processor responding instantaneously. However this lookahead delay is adding latency in the signal path. Some Limiter lookahead delay lengths for LF speaker protection applications can be up to 50 samples. This is a considerable channel delay. The relevance of lookahead delay for most Compressors and Limiters in channel processing is questionable since Limiters in channels are not used as transient clip protectors with very fast attack times.

Gates typically benefit from lookahead delay, providing an instantaneous attack response. As an alternative to lookahead delay, very fast attack can be manually set up but this can cause a popping sound. There are other complex techniques employed as alternatives to lookahead delay that use attack and depth shaping to reduce popping. These techniques can produce very fast attack without popping and without lookahead delay.

If lookahead delay is available and can be bypassed to reduce latency on a per channel basis, then delay path differences can occur between channels.

4 DELAY PATH/LATENCY ANALYSIS IN A PRACTICAL SYSTEM

Figure 7 is a diagram of a typical digital mixing system setup with some of the interface and processing applications discussed.

4.1 Analog and Digital Inputs

The signal path starts with local analog inputs, containing an ADC, where the latency is dependant on sampling rate and decimation filter with considerable variation typically between 0.2 and 0.9 ms. It is important all analog inputs within the main mixer have equal delay paths to avoid colouration when mixing correlated sources. Using the same ADC type across inputs is naturally sensible and using a common sampling clock (phase locking the sampling process across converters).

Digital inputs are either decoupled from the system sampling rate through the use of one SRC or purely synchronous through the use of system sync'ing. Inclusion of the SRC could increase latency by typically 1ms. If the digital input is purely a playback music source latency is not important and the SRC is commonly employed to reduce sync issues and convert from 44.1kHz up to 48 or 96kHz sampling rate. If the digital input is from an instrument source, the monitor path latency has to be considered. It is then important for the engineer to know whether an SRC is in circuit and whether it is possible to sync the device to the mixing system and disable the SRC.

The latency through the analogue and digital input stages will not be the same, producing delay path differences between analogue and digital sources. This is only important if correlated sources exist between analogue and digital inputs. An example of this could be an instrument feeds the mixer with a direct digital output and an analogue mic send. The engineer has to be aware and should experiment with applying delay correction on these correlated inputs.

4.2 Audio Networks

Audio networks can be used to provide remote I/O across a stage, feeding instrument sources into the mixer and monitor mix outputs from the mixer to the musician. Therefore low latency networks are essential for this application.

Delay path differences between inputs at different network nodes and relative to the local inputs also need consideration. For example, delay compensation in the input channel can be employed to compensate for delay path differences in input sources with correlated pick up. A more practical solution is to keep correlated sources

within a node on the network and not spread them across the network and local inputs.

Networked inputs are also used to share inputs between two mixing systems, for example the monitor mixer and front of house mixer. Relative delay differences between these types of inputs are not important since both systems are seperate and do not mix to the same point. However sharing inputs may make the network topology more complex and increase latency - and low latency is critical for musician monitors. One solution is to provide two network ports, one for system expansion to other mixers and the other for local networking, remote input/outputs, inserts etc. This keeps things simple and structured, an ingredient for lower latency.

Recorder sends and playback feeds are not latency critical but are delay path critical, all channels should be coherent. This is ensured for some networks, USB, Firewire and OptoCore. Using EtherSound one network node should be used to avoid delay path differences.

4.3 Insert processing

In Figure 7 various insert interfaces are shown. Inserting an analogue device into a channel requires DSP routing, ADC and DAC.

Inserting a digital processor but using analogue I/O increases latency with a further ADC and DAC stage inside the external inserted device, at this point latency can escalate. With poor latency converters this type of insert could add over 3ms of latency to the channel.

An external digital processor should naturally be inserted through digital interconnect. The insert device can be configured to slave from its input stream, slaving to the mixer, syncing the return stream to the mixer too, alleviating the need for SRC at any point in the insert path keeping latency low.

Internal inserts or plug-ins still require DSP routing and do not for practical issues offer much reduction in latency than an external sync'd digital device.

All inserted channels have some change in delay path relative to an uninserted channel – causing delay path effects. All channels can be delay compensated. However if there are various different delay paths due to various different plug-ins and external inserted devices, this becomes complex and often unneccessary unless they are correlated sources.

Mix inserts are essentially the same, only sensitive to latency issues if they are in the monitor path (Groups, Aux and sometimes matrix). Inserting processors onto Main outputs does not typically cause latency problems. Delay path differences between mixes of the same type is critical, for example if a stereo aux mix used for in-ear monitoring has a sample delay relative between the pair, the musician will find this objectionable (often described as a spatial or phasy sound).

Minimising the requirement for inserts avoids additional latency and delay path effects. The mixer should contain high quality equalisers, gates, compressors, limiters and de-essers to facilitate this. Specified outboard processors do not sound good with latency and delay path problems.

4.4 Delay path effects in Mixes and Outputs

There are some practical issues concerning delay path differences in mixes. Mixing inputs or mixes with other mixes can produce comb filtering since mixes are typically correlated with each other. Bus structures can be designed to accomodate this but each mix is passed through additional processing (this in itself produces colouration filtering effects). However non processed mixes are fed to matrix mixes these source mixes should be taken from points with no delay path differences to avoid comb filtering. This should be a user option in the matrix source feeds. Small delay path differences between a stereo pair in-ear mix are audible. Delay path differences feeding acoustic drivers are less audible, given that a 41.6us (2 samples at 48kHz) delay path difference is approximately 12mm positional error for a speaker placement.

5 CONCLUSIONS

Latency in musician monitors produce various effects that degrade intelligibility (vocal pitch change effects and timing problems for percussive musicians). These are not problems in low latency mixing systems, where the signal path input to output is typically less than 2 ms. To maintain this the engineer has to use insert points sensibly, and not involve SRC digital IO on latency critical monitor channels, using sync'd (non SRC) digital inserts. Also to be mindful of analogue inserts on percussive monitor feeds and especially digital processors with analog IO (doubling the ADC/DAC count in the insert path). Low latency audio networks are also critical for low latency monitor systems.

It is also shown that comb filtering effects occur in very low latency systems where the 'in-head' feed interferes with the 'in-ear' monitor feed and when correlated audio is mixed through delay paths. This produces colouration effects in the monitor mix in the vocal region. The use of EQ lift at these frequencies in the vocal region may improve intelligibility for vocal in-ear mixes. The exact frequencies naturally depend on the latency and channel insert topology.

There are large differences in latency for various ADC and DAC's. Increasing the sampling rate at the converter stage does reduce latency, but once low latency converters are employed halving latency is small and can be outweighed by the complexity and increase in latency in the DSP processing requirements for higher sampling rates.

As mentionned syncing of digital IO to avoid the use of SRC helps minimise latency considerably. Furthermore the development of low latency SRC's would be of benefit to this application. Digital IO and plug-in inserts do simplify the latency model but still increase latency through delay compensation. The ultimate solution is excellent processing on the channel to alleviate the need for inserts in these applications.

Mixing system manufacturers could provide more latency path information in system specification. However with the use of inserts this becomes a complex subject for the engineer to distill.

The paper has attempted to highlight problems associated with latency and delay paths in large scale mixing systems at component, software and system level. Ultimately the work aims to provoke discussion of latency issues across the disciplines of audio networking, mixer configuration and signal processing for live monitor applications.

Acknowledgements

The author would like to thank Yves Ansade, Steve Green, Morten Lave and Bob Martin for discussions concerning latency in converters and audio networks. Also Carey Davis, Chris Panks, Leon Philips and Glenn Rogers for discussions concerning latency in mixing systems.

6 REFERENCES

[1] Steve Green, 'A New Perspective on Decimation and Interpolation Filters'. Cirrus Logic Publication, July 2004.

[2] Matthias Rosenthal, 'Concept and Design of a Reconfigurable Parallel Processing System for Digital Audio'. Published PhD Thesis, Hartung-Gorre Series in Micro-electronics Vol65. ISBN 0936-5364.

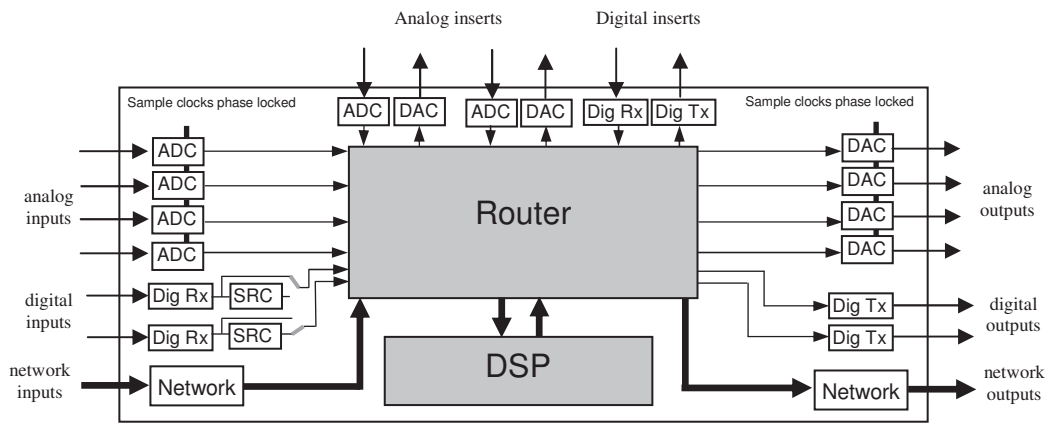


Figure 6 Digital mixer hardware component latency analysis

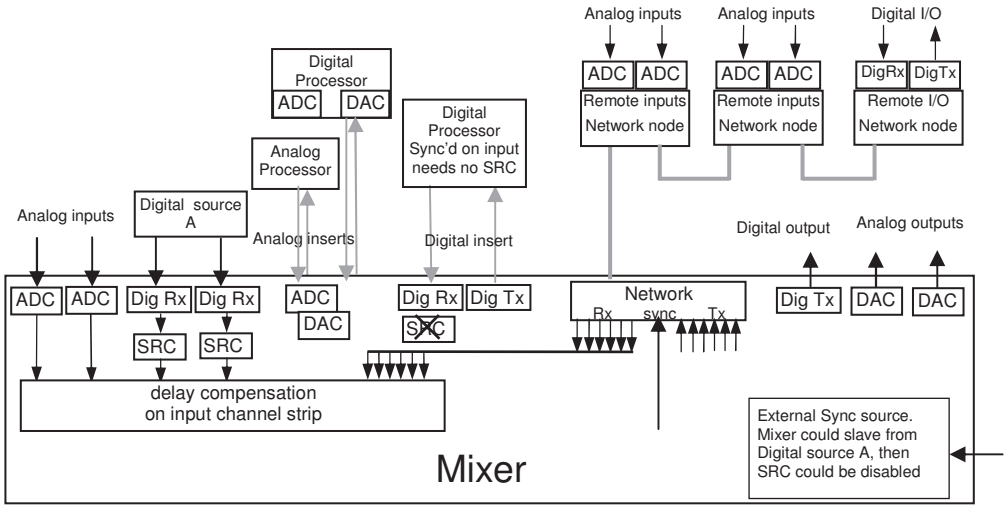


Figure 7 Practical mixing system set up - system latency analysis

Índice de Autores

Author Index

Avila, F. R. 132
Barbedo, J. G. A. 47
Barbon Jr, S. 17, 28, 33
Batalheiro, P. B. 20
Biscainho, L. W. P. 11, 58, 66, 74, 82, 105, 113, 132
Clark, R. 151
Costa, C. R. 138
Dedecca, J. G. 97
Diniz, F. C. C. B. 58
Diniz, P. S. R 82
Esquef, P. A. A. 66
Faria, R. R. A. 89
Freeland, F. P. 82
Furlanete, F. 51
Goldemberg, R. 51
Gomes, M. H. A. 97
Guido, R. C. 17, 28, 33
Gunzi, A. S. 82
Haddad, D. B. 20
Jesus, R. A. 113
Lopes, A. 47
Lovisolo, L. 74
Luvizotto, A. L. 138
Manzolli, J. 51, 138
Masiero, B. S. 119
Mendes, R. S. 138
Müller, S. 127, 143
Nascimento, R. G. 97
Nascimento, V. H. 119

Netto, S. L. 11, 58
Nunes, L. O. 66
Oliveira, L. C. 51
Paiva, R. C. D. 11
Petraglia, M. R. 20, 39
Ribeiro, C. E. P. 39
Sanchez, F. L. 17, 28, 33
Santo, N. M. E. 127
Silva, E. A. B. 74
Silva, J. R. M. 39
Tavares, T. F. 47
Tcheou, M. P. 74
Tygel, A. F. 105
Vieira, L. S. 17, 28, 33
Vital, R. S. 89
Zuben, F. 138

Audio Engineering Society - Seção Brasil

Anais do 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil

Proceedings of th 5th AES Brazil Conference
11th AES Brazil National Convention

Patrocinadores:

Expositores:

Acoustic	Pride	Sabra Som
Attack	Oversound	Selenium
Audiocare	Roland	Sennheiser
Beyma	Royal	SLM Brazil
Decomac	Revista Backstage	Snake Pro
Hinor	Revista Igreja	Sotex
Hot Sound	Revista Música & Mercado	Staner
Libor	Revista Música e	Studio R
Meteoro	Tecnologia	Yamaha