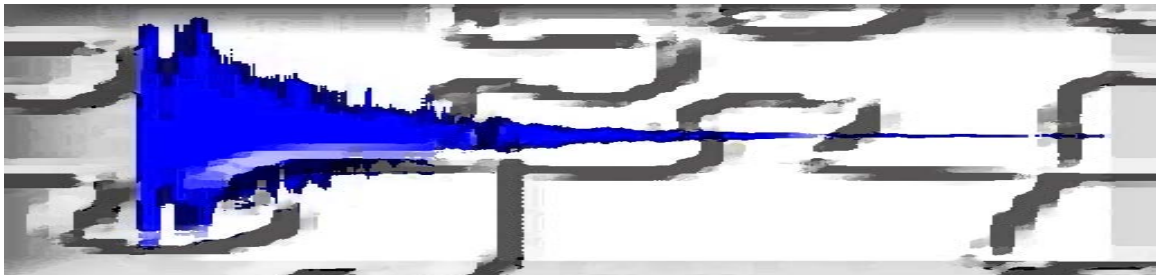
INSTITUTO SUPERIOR TÉCNICO

UNIVERSIDADE TÉCNICA DE LISBOA

INSTITUTO SUPERIOR TÉCNICO



New Techniques for Estimation of Acoustical Parameters

Joel Vera Cruz Preto Paulo

Supervisor: *Doctor José Luís Bento Coelho*

Thesis approved in public session to obtain the PhD Degree in Electrical and Computer Engineering

Jury final classification: *Pass with Merit*

Jury

Chairman: *Chairman of the IST Scientific Board*

Members of the Committee:

Doctor John Vanderkooy

Doctor Moisés Simões Piedade

Doctor Mário Alexandre Telles de Figueiredo

Doctor José Luís Bento Coelho

Doctor Luís Manuel de Jesus Correia

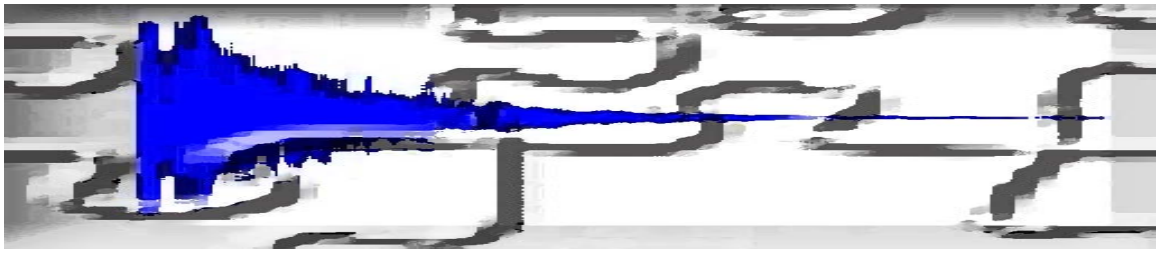
Doctor Diamantino Rui da Silva Freitas

2012

INSTITUTO SUPERIOR TÉCNICO

UNIVERSIDADE TÉCNICA DE LISBOA

INSTITUTO SUPERIOR TÉCNICO



New Techniques for Estimation of Acoustical Parameters

Joel Vera Cruz Preto Paulo

Supervisor: *Doctor José Luís Bento Coelho*

Thesis approved in public session to obtain the PhD Degree in Electrical and Computer Engineering

Jury final classification: *Pass with Merit*

Jury

Chairman: *Chairman of the IST Scientific Board*

Members of the Committee:

Doctor John Vanderkooy, Distinguished Professor Emeritus, University of Waterloo, Canada

Doctor Moisés Simões Piedade, Professor Catedrático do Instituto Superior Técnico, da Universidade Técnica de Lisboa

Doctor Mário Alexandre Telles de Figueiredo, Professor Catedrático do Instituto Superior Técnico, da Universidade Técnica de Lisboa

Doctor José Luís Bento Coelho, Professor Associado (com agregação) do Instituto Superior Técnico, da Universidade Técnica de Lisboa

Doctor Luís Manuel de Jesus Correia, Professor Associado (com agregação) do Instituto Superior Técnico, da Universidade Técnica de Lisboa

Doctor Diamantino Rui da Silva Freitas, Professor Associado da Faculdade de Engenharia da Universidade do Porto

Funding Institutions

FCT - Fundação para a Ciência e Tecnologia

2012

Título: *Novas Técnicas para Estimação de Parâmetros Acústicos*
Nome: Joel Vera Cruz Preto Paulo
Doutoramento em Engenharia Electrotécnica e de Computadores
Orientador: Doutor José Luís Bento Coelho

Resumo

As medições da resposta ao impulso de salas (RIR), em situações de níveis elevados de ruído de fundo, conduzem frequentemente a relações sinal-ruído (SNR) baixas. Nestes casos, os resultados das medições podem não ser aceitáveis, mesmo aplicando a técnica de médias temporais aos sinais de teste. Mais ainda, se a cadeia electroacústica apresentar não-linearidades ou se o sistema não for invariante no tempo, a SNR final pode ser severamente degradada. Por outro lado, a correcta estimação dos parâmetros acústicos de recintos utilizados em espectáculos ou espaços públicos requer que as medições acústicas sejam realizadas com a presença do público. Contudo, por questões de incomodidade ao ruído, as medições são geralmente realizadas com a sala vazia, aplicando-se apenas um factor de correcção que contempla o efeito da absorção sonora da audiência. Este procedimento não considera a variabilidade de determinados parâmetros ambientais, tais como a humidade relativa e gradientes de temperatura, que ocorrem durante o espectáculo e que poderão influenciar em grande escala a resposta em frequência do sistema. Neste sentido, desenvolveram-se novas técnicas para serem utilizadas em situações de ruído não-estacionário baseadas na minimização da energia nos domínios do tempo e da frequência da resposta da sala aos sinais de teste e é proposto um método que utiliza modelos perceptuais auditivos de forma a mascarar os sinais de teste por forma a não serem percebidos pela audiência, permitindo caracterizar o espaço acústico convenientemente. Adicionalmente, é apresentado um método para a identificação e minimização do fenómeno da variância temporal no espaço acústico utilizando uma sonda que emite um sinal de teste na banda dos ultrasons. São apresentadas e discutidas as vantagens e fraquezas das técnicas propostas, juntamente com casos de estudo e aplicações.

Palavras-chave: técnicas de medição acústica, parâmetros acústicos, resposta ao impulso de salas, baixa SNR, acústica de salas, modelos perceptuais, incomodidade ao ruído, ultrasons, efeitos da variância temporal

Title: New Techniques for Estimation of Acoustical Parameters
Name: Joel Vera Cruz Preto Paulo
PhD Degree in Electrical and Computers Engineering
Supervisor: Doctor José Luís Bento Coelho

Abstract

The measurement of the Room Impulse Response (RIR) under high background noise levels frequently means one must deal with low Signal to Noise Ratios (SNR). If such is the case, the measurement might yield unreliable results, even when synchronous averaging techniques are used. Furthermore, if there are non-linearities in the apparatus or system time-variances, the final SNR can be severely degraded. On the other hand, the accurate estimation of the acoustic parameters of enclosures used for live performances, such as theatres and other public areas requires the measurements to be carried out in the presence of an audience. However, in order to avoid annoyance of the audience, people are usually kept out and a correction factor related to the equivalent sound absorption of the audience area is applied. This procedure does not consider the variability of some relevant parameters such as the relative humidity and the temperature gradients that occurs in a venue during a live show, which may have a considerable influence on the room frequency response. Therefore, new techniques are developed for situations of non-stationary ambient noise based on the minimization of the energy of the ambient noise both in the time and in the frequency domain and a new method is proposed where filtered swept sine signals are used in agreement with a perceptual auditory model in order to minimize the annoyance, thereby allowing the estimation of the acoustic parameters in the presence of the audience. Additionally, a measurement technique was set-up with the purpose of monitoring the acoustical media, using ultrasonic probe test signals, to minimize the effects of time variance phenomena. Advantages and weaknesses of the proposed techniques are discussed and study cases and applications are presented.

Keywords: Acoustic measurement techniques, acoustic parameters, room impulse response, SNR, room acoustics, perceptual models, sound nuisance, ultrasound signals, time variance effects.

to

Paula Silva

Separo os pensamentos em mil pedaços, grânulos inertes desprovidos de aparente significado
E, sem hesitação, espalho-os de mão aberta pelo espaço sónico em redor
tentado escondê-los da percepção de quem por ali passa,
para que mais tarde,
quando se não sentir mais o cheiro da tempestade,
depois de devidamente encadeados e somados,
possam representar a essência de algo de teor acabado.

I separate thoughts into a thousand pieces, inert granules devoid of apparent meaning, and,
without hesitation, I spread them with my open hands through the sonic space surrounding,
trying to hide them from the perception of the humans passing by,
in a way that, so much later,
after being overlapped and added,
when the scent of the storm is impossible to feel,
they may represent the essence of something over content.

Table of Contents

List of Figures

List of Tables

List of Symbols

$\overline{\alpha}$ – mean sound absorption coefficient of the room boundaries;

ρ – density of air (kg/m^3);

$\mathrm{IACF}_t(\tau)$ – Inter-Aural Cross-correlation Function;

IACC – Inter-Aural Cross-correlation Coefficient;

$\phi_{SS'}$ – Cross-correlation function;

ϕ_{SS} – Auto-correlation function;

$\circledast$ – circular convolution operation;

$\otimes$ – circular cross-correlation operation;

$\langle\oplus\rangle$ – segments concatenation operator;

A_k – amplitude of the *k* partial;

f_k – frequency of the *k* partial;

φ_k – phase of the *k* partial;

pvr – peak-to-valley ratio;

Δf_w – window function bandwidth;

Δf_b – bin separation frequency;

l – normalizing factor applied to the weighting average technique;

$N_k(f)$ – power spectral density for the frequency band *k* (Watt/Hz);

A_w – overlap factor;

d_w – amplitude deviation of the overlap factor;

$Res(n)$ – room output response with additive noise;

$Res_{av}(n)$ – room output response with additive noise using the average technique;

$\overline{p_n^2}$ – mean square value of the noise;

$\overline{p_s^2}$ – mean square value of the speech signal;

$\tilde{p}_0$ –rms value of the reference sound pressure level, 2x10^{-5} N/m^2;

$\tilde{p}$ – rms value of the instantaneous sound pressure (N/m^2);

$\Phi_{S_{MLS}y}$ – vector of entries of $\phi_{S_{MLS}y}(n)$;

f_S – sampling frequency (Hz);

T_S – sampling period (s);

Δ_G – processing gain of the MLS technique against the impulse signal;

$\overline{I_i}$ – white noise intensity (Watt/m^2);

$w(n)$ – time analysis window;

$T_{S_{MLS}}$ – length of the MLS signal (s);

RT60 – reverberation time (s);

T20 – reverberation time based on a 20 dB evaluation range (s);

RIR – room impulse response;

IR – impulse response;

$\mathrm{SNR}_{HíbrEspct}$ – signal-to-noise ratio of the Hybrid spectral technique;

$\mathrm{SNR}_{HíbrTemp}$ – signal-to-noise ratio of the Hybrid temporal technique;

SNR_{av} – signal-to-noise ratio obtained from the averaging technique (dB);

$h_{filtr}(n)$ – impulse response of the filter;

W_s – sum of the weights;

τ – time delay (s);

λ – wavelength (m);

$\underline{\otimes}$ – overlap-add;

$\delta(t)$ – Dirac impulse;

AVAC – conditioned air, heating and ventilation systems;

A – maximum amplitude of the auto-correlation function (amplitude of IR);

A_0 – room reference effective absorption area of the adjacent room (m^2);

AGC – Automatic gain controller;

BM – Basilar membrane;

BW_c – bandwidth of each critical band;

B_w – window bandwidth;

c – sound speed in the air (m/s);

C80 – Clarity 80 (dB);

CF – crest factor (dB);

D – decay rate of the linear regression;

d – distance from the sound source, (m);

D50 – Definition 50 (%);

Dc – critical distance (m);

E – energy density for a particular location in the room;

E_0 – initial energy density;

E_1 – energy corresponding to the time where the decay curve crosses the background noise;

E50 – energy of the first 50 ms of the impulse response;

E80 – energy of the first 80 ms of the impulse response;

E_d – energy density of the direct sound wave;

E_N – energy of noise;

E_0 – initial energy of noise;E_e – energy density of the reverberant field (steady-state);

E_{rev} – energy of the reverberant field;

E_{Total} – total energy of the impulse response;

EDT – Early decay time;

ERB – Effective Rectangular Band;

F – modulation frequency (Hz);

$f(x)$ – characteristic polynomial;

$f_{cm}(n)$ – characteristic of the transfer function of the measurement chain;

$f_{pre}(n)$ – pre-emphasis filter, inverse filter of the $f_{cm}(n)$;

f_{nxnynz} – natural oscillation frequencies or room resonances;

f_{sch} – Critical Frequency or Schroeder Frequency;

f_{s} – sampling frequency;

$H(f)$ – Fourier transform of the room impulse response;

$h(t)$ – room impulse response;

$h'(t)$ – room impulse response affected by noise;

h_L – impulse response measured of the left ear;

h_R – impulse response measured of the right ear;

I – acoustic intensity (Watt/m^2);

k – propagation constant (rad/m);

K – MLS sequence order;

K_d – arbitrary delay;

$k_{\mathrm{i}}(t)$ – i Volterra Kernel component;

L – MLS sequence length (samples);

$L(t)$ – energy decay of the impulse response (dB);

L_1 – noise level in the room (dB);

L_2 – noise level in the adjacent room of the auditorium (dB);

L_d – maximum deviation of the IR from the true value (dB);

L_I – sound intensity level (dB);

L_{seg} – length of the segments in the hybrid sequences approach (samples);

L_p – sound pressure level (dB);

L_W – sound power level (dB);

m_{net} – Combined masking threshold;

M – number of sequences;

MS – mean square value;

N – number of segments of the sequence;

$n(n)$ – equivalent noise;

$N(n)$ – external noise;

NMR – noise-to-mask ratio;

SMR – signal-to-mask ratio;

N_B – number of frequency bands in the filter bank;

N_F – constant of proportionality relative to the power spectral density of the test signal;

p – sound pressure (N/m^2);

PR – perfect reconstruction;

NPR – nearly perfect reconstruction;

Q – directivity factor;

R – room constant;

$r(t)$ – residual part of the sound;

R [n] – random variable with normal distribution, with zero mean value and unity variance.

STI – Speech Transmission Index (dB);

RaSTI – Rapid Speech Transmission Index;

E_B – critical band noise energy level;

E_T – tone masker energy level;

S_A – total area of the room boundary surfaces (m^2);

S – sound power of the speech;

$S(f)$ – Fourier transform of the system input signal;

$s(t)$ – system input signal;

Seq_{MLS_RES} – output segmented sequence;

$S_{MLS}(t)$ –MLS sequence;

SNR– signal-to-noise ratio (dB);

SNR_{BA} – initial signal-to-noise ratio (dB);

SF – spread of masking function;

T_q – Threshold in quite;

T – Duration of the swept sine test signal;

t_o – time when the sound source is switched off (s ou ms);

t_1 – end of the analysis (s or ms);

TDS – Time-Delay Spectrometry;

TI – Transmission Index for the speech (dB);

T_{seg} – duration of the segments for the hybrid sequences (s);

TH_N – noise thresholds;

TH_T – tone masking thresholds;

UFB – uniform filter banks;

NUFB – non-uniform filter banks;

V – volume of the room (m^3);

W – sound power, (watt);

W_k – weighting coefficient of the TI;

Y – array of the y(n) entries;

$Y(f)$ –frequency response of the system;

$y(t)$ – system output;

ΔSNR – processing gain of the SNR (dB);

ω – angular frequency (rad/s);

ω_1 – initial angular frequency (rad/s);

ω_2 – final angular frequency (rad/s);

$z(f)$ – Critical band number, after Zwicker;

QMF – Quadrature mirror filters;

PQMF – Pseudo-quadrature mirror filters;

DFT – Discrete Fourier transform;

DCT – Discrete cosine transform;

MDCT – Modified discrete cosine transform;

DWT – Discrete Wavelet Transform;

PR – Perfect reconstruction;

UFB – Uniform filter bank;

NUFB– Non-uniform filter bank;

NPR – Nearly perfect reconstruction systems;

BM – Basilar membrane;

IHC – Inner Hair Cells;

OHC – Outer Hair Cells;

CF_{BM} – Characteristic frequency on the basilar membrane;

Chapter 1

INTRODUCTION

1.1 General considerations

The accurate and objective characterization of an acoustic system is a complex task. The main difficulty usually lays on the fact that the acoustic parameters have to adequately describe not only the quantities involved but also the subjective interpretation of the acoustical phenomena by the human auditory system. Therefore, the assessment of the sound properties should be based on objective parameters that are able to describe both the physical phenomena and subjective issues related to the human hearing and perception mechanisms.

Effective techniques for sound measurements in enclosed spaces are then necessary to correctly characterize the acoustic conditions the enclosure offers to the sound field to be perceived by a listener in the audience.

Objective and subjective parameters are established to describe the sound energy spatial distribution and time behavior, such as energy decay or reverberation, and the speech intelligibility or music clarity, respectively.

The acoustic measurement provides knowledge and insight into the key factors responsible for the subjectivity of the acoustical phenomena in a room and allows for a diagnostic to define guidelines at the design stage. For instance, if for some reason an auditorium has to be frequently used for performances different to those it was originally designed for, or if deficiencies are detected in a room, such as change of geometry or finishing materials, after completion of building, acoustic measurements should be made so that appropriate corrections could be designed and applied.

In the different stages of building of special auditoria (such as theatre rooms, concert halls, live performance rooms) it is a good practice to carry out acoustic measurements to check the design objectives, and if necessary, to design corrective solutions. This is a responsible methodology having the obvious benefits of anticipating problems, thereby reducing the building costs and ensuring the desired results.

In some cases, the measurement of the impulse response (IR) should be performed keeping in mind that the system under test is to be used with people in its final usage. In fact, the correct estimation of the acoustic parameters of concert halls, listening rooms, sport halls, transport stations, etc., should be made with the people inside the enclosure due to the excess of sound absorption offered by the audience and to temperature and humidity variations of the air medium with the consequence of changes in the reverberation time, the frequency response, etc. Another scenario concerns broadcast audio chains. Usually, the characterization of the audio channels, terrestrial or satellite is made by using a “dead” channel, secondary channel, used in off-line mode, and only after its correct configuration is the broadcasting program moved there. However, if some anomalies occur during the broadcast operation, affecting the audio contents, detecting and solving the problem is a difficult task without cutting the transmission.

The system theory states that the properties of a linear time invariant system (LTI) are included in its IR, or alternatively, in its frequency response, in the sense of its Fourier transform. However, in real situations the assumption of linearity and time-invariance for the room under test is not ensured, thereby imposing deviations in the results of the acoustic measurements.

The subjective interpretation of the sound signal (speech or music) in a room is usually described (modeled) by a number of acoustic parameters and indicators related to both the enclosure and the human auditory system and can be studied by using IR derived analysis. Therefore, the measurement of the IR is, no doubt, one of the most important procedures in the acoustic or electroacoustic fields.

1.2 Motivation

A number of issues related to the sound signal analysis worked as the motivation to develop the techniques presented in this thesis:

- In room acoustic measurements, the disturbing sounds are often characterized by non-stationary noise, with pronounced fluctuations in the time and in the frequency domains, such as the generality of musical material and speech for significant time periods.
- Considering the perceptual mechanisms of the human hearing system and the characteristics of some testing signals, such as the swept sine, the acoustic tests could be performed by considering the masking effects of the auditory system.
- Spectral analysis of various common audio signals usually present during live event performances show insignificant acoustic energy above 40 kHz. This fact suggests the possibility of using an ultrasound test signal probe simultaneously with the acoustic measurements for the purpose of comparing the deviations imposed to both test signals, audio and ultrasonic, due to time variance phenomena and thereby assess their influence in the results.

Measurement techniques robust to noise must be investigated if sufficiently high values of overall SNR are necessary to achieve accurate results.

1.3 Objectives

The main goal of the work described herein was to develop new techniques to measure the IR able of dealing with situations of background noise with characteristics of (i) time and frequency varying signals (non-stationary case), (ii) showing tonal-like spectra for reasonable periods of time (segment stationary case) allowing the use of the auditory masking approaches, and (iii) showing frequency bandwidth lower than 40 kHz to allow the use of ultrasound test

signal probe. These techniques should be able to allow reasonable values of SNR to estimate the IR accurately and should represent an improvement of the performance of the existing acoustic measurement techniques.

Room acoustic measurements are carried out in two situations:

- In empty rooms – the enclosure is measured without people inside. The maximum sound pressure levels are limited only by the electroacoustic apparatus and the SNR is maximized at the reception;
- With people inside the room – the audience play an important role in the correct characterization of a venue due to the change of the mean absorption coefficients observed in the room during the musical performance. The sound levels used for the test signals should be low, not causing significant nuisance to the people. Usually, the SNR shows low values at the reception and, therefore, supplemental post-processing techniques are applied.

Each technique was developed to be adapted to the specific situation of noise. Nevertheless, they can be used simultaneously, as a complement.

To highlight particular issues related to the acoustic characteristics of the systems, the study has focused on the performance of the techniques in the time and frequency domains for a number of selected parameters.

The performance of the techniques was compared with the standard techniques in order to enhance the relative advantages and disadvantages as well as to study their weak and strong points.

1.4 Contribution of the thesis

This thesis encompasses the development of new measurement techniques or the enhancement of others for the estimation of the impulse response and a number of parameters usually used to characterize the system's acoustic conditions. The techniques are intended mostly to be used in situations of low SNR, with particular noise characteristics of the types: (i) very non-stationary

both in the time and frequency, and (ii) stationary tonal-like. Additionally, a system based on an ultrasonic array device and modules of statistical classification was developed to detect and attenuate the effects of the time variance phenomena in the acoustical results.

New techniques based on maximum length sequences, MLS, and on logarithmic sine sweeps were proposed.

Swept sines and MLS are deterministic signals, therefore, the synchronous time averaging technique can be applied in order to improve the overall SNR. Two different approaches regarding the synchronous time averaging technique were proposed, namely;

- Segment Weighted Average, SWAve – This method consists of splitting each captured frame into *N* segments. The average technique is applied to the weighted version of each segment. In order to improve the SNR, the segment with the larger energy will be more attenuated and the segment with the lower MS value will be more valued.
- Spectral Segment Weighted Average, SSWAve – This method is similar to the SWAve where each segment is filtered by a filter bank. The averaging technique is applied to the weighted version of each frequency band among all the segments. The frequency band with the larger energy, for the same frequency band, will be more attenuated and the frequency band with the lower energy will be more valued.

These techniques can be used with advantage in situations of moderate to high level non-stationary disturbances.

A new method was developed that uses filtered swept sine signals following a perceptual auditory model, taking the simultaneously masking effect into account, in order to minimize the annoyance. In the Segmented Test Signal plus Noise with Modification procedure, STSNM, the swept sine frame is split into several segments being implanted only in the frequency bands masked by the audio material.

This technique allows the estimation of the acoustic parameters in the presence of an audience since the test signals are not perceived by people.

A measurement technique consisting of an ultrasound loudspeaker array was set-up with the purpose of monitoring the acoustical medium, searching for time variance phenomena, for low

SNR situations. A probe test signal in the ultrasonic band is sent to the room by using an ultrasound loudspeaker array simultaneous with the test signal frames. The relevant parameters for establishing time-variance and associated thresholds are then estimated from the acquired ultrasonic sound. The valid test signal frames, which pass the thresholds test, are labeled with a weighting factor depending of its significance. Otherwise, the frames are discarded, and do not enter in the averaging process.

A method to identify and classify different types of road pavements was developed. This work led to a Portuguese patent of invention (in Portuguese), [Paulo09P]. This method uses statistical learning techniques applied to a set of features extracted from near field road noise. The main concepts developed in this study worked as background knowledge for other parts of the thesis.

Finally, a didactic framework based on Matlab programming language was developed to test the acoustic techniques developed throughout this thesis. This framework provides a user-friendly interface based on the GUI of Matlab, allowing *in situ* or in post-processing mode, the characterization and study of a number of room acoustic parameters extracted from the RIR. Figure 1.1 shows these techniques in a schematic fashion.

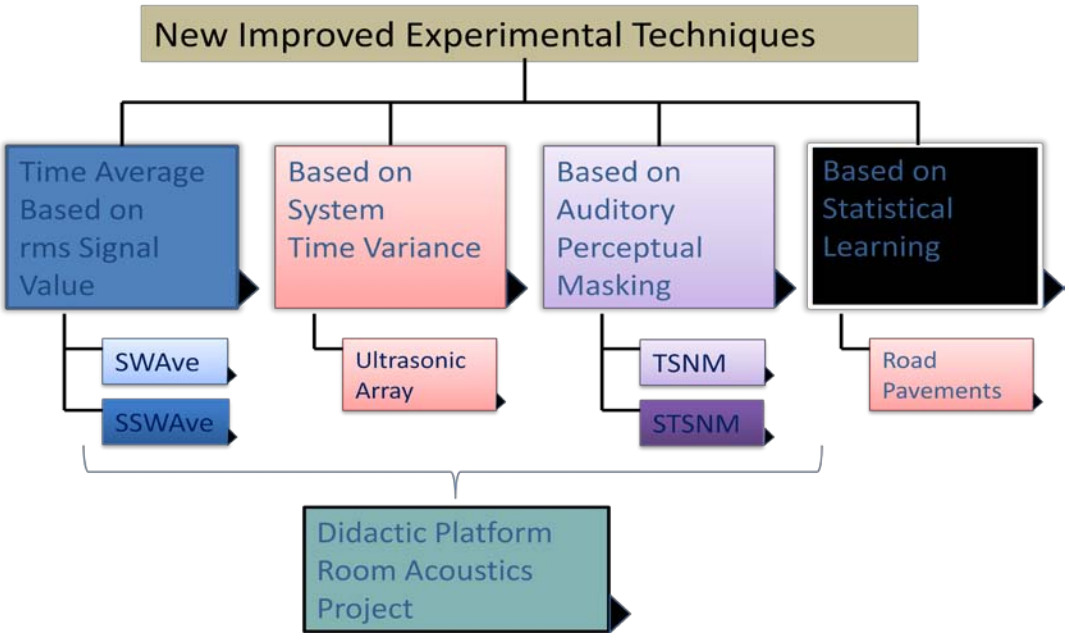


Figure 1.1 – Schematic presentation of the developed work at the aim of this thesis.

1.5 Thesis development

This document consists of seven chapters where the concepts and results of the work, together with discussions and conclusions, are described.

In Chapter 1 – INTRODUCTION, the foundations for the study are established. A general perspective of the room impulse response concept, its importance and applicability in the room acoustics field is presented, with an emphasis on issues related with the SNR and the time variance phenomena. The objectives and the contribution of the thesis for the room acoustics fields are laid out.

Chapter 2 – SOUND BASICS is split into two main parts. The first part describes the basic sound phenomenology, from sound generation to sound wave propagation in open air and inside enclosures. A brief description of the acoustic parameters estimated from the RIR, related to the perception of the human auditory system, is presented. The second part deals with fundamentals of the human auditory system. The human hearing mechanism is described, emphasizing aspects related to auditory masking effects, within the purposes of this study.

Most concepts and experimental techniques currently used for room acoustic measurements, including auditory perceptual models, digital signal processing techniques and statistical learning are presented in Chapter 3 – SOUND MEASUREMENT AND ANALYSIS. Acoustic measurement methods using MLS and swept sine signals in which the developed techniques are based on, are also described in detail in this chapter, with their advantages and deficiencies highlighted.

Chapter 4 – SIGNAL-TO-NOISE RATIO IMPROVEMENT consists essentially of two parts. The first part describes two new techniques, the Segment Weighted Average and the Spectral Segment Weighted Average, developed with the aim of improving the synchronous averaging method. The second part details the use of ultrasound probe signals for detection and attenuation of the time-variance phenomena occurring during the acoustic measurements. An ultrasound sensor array based testing system is presented.

A perceptual model developed toward the use of the simultaneous masking effects of the human auditory system to minimize the audience annoyance during tests, is described in Chapter 5 – TECHNIQUES FOR ANNOYANCE MINIMIZATION. A brief description of

perceptual models commonly used in perceptual audio coding algorithms is made as an introduction to the perceptual model concept adopted with the techniques developed for strict use in the presence of an audience.

Case studies, results and discussions are presented in Chapter 6 – PRACTICAL APPLICATIONS. The techniques based on the synchronously averaging method, the Segment Weighted Average and the Spectral Segment Weighted Average, are compared with the MLS and swept sine methods highlighting their advantages and weaknesses. Examples of application of the techniques based on perceptual models, namely Tone or Noise probes plus Noise with Modification and the use of ultrasound transducer arrays are shown. The swept sine frame is used to build a tempered musical scale. Therefore, music tracks for acoustic tests are defined by conveniently selecting midi files. Experimental results of application of a Bayesian classifier for the statistical classification of road pavements using near field vehicle rolling noise measurements are shown and discussed.

Chapter 6 – CONCLUSIONS briefly describes the main results and presents the conclusions of the work. The relevant results are emphasized and discussed stating advantages and deficiencies of the new techniques and the boundaries of their applicability in the real world. A picture of further applicability and development regarding future commercial implementation of the techniques is also presented.

Some of the results and concepts presented in this thesis are also reported in the selected publications presented in the References section. The basic procedure to mask the test signals perceptually is presented in Appendix F1. Appendix G1 describes a demonstration tool/didactic framework, showing a user-friendly interface to implement the new techniques, allowing the characterization of a number of room acoustic parameters. The remaining appendices cover some additional information data needed to the implementation and explanation of the work presented in the thesis.

Chapter 2

SOUND BASICS

2.1 The concept of sound

The concept of sound lies on the physical disturbance of the ambient medium (gas, liquid or solid) which is detected (pressure) by the human auditory system and perceived by the human brain as a sensation. In the air, the pressure variations are transmitted as sound waves, a series of mechanical compressions and rarefactions around the mean atmospheric pressure,

$$p_{acoust} = p_{amb} - p_{atm(norm)} \tag{2.1}$$

The physical phenomenon responsible for the generation and radiation of sound is related to the vibration of the air particles from their equilibrium position (the air has the properties of an elastic system). Therefore, the total contribution of the energetic changes between the particles, due to the collision of each particle with the adjacent, produces a movement from the source position, following a specific direction.

The origin of this phenomenon, for the purposes of this study, is related to the vibration of a rigid surface which applies movements to the medium producing compression (zones of higher pressure-higher density of air molecules) and expansion (zones of rarefaction) to the surrounding air particles. A sketch is shown in Figure 2.1.

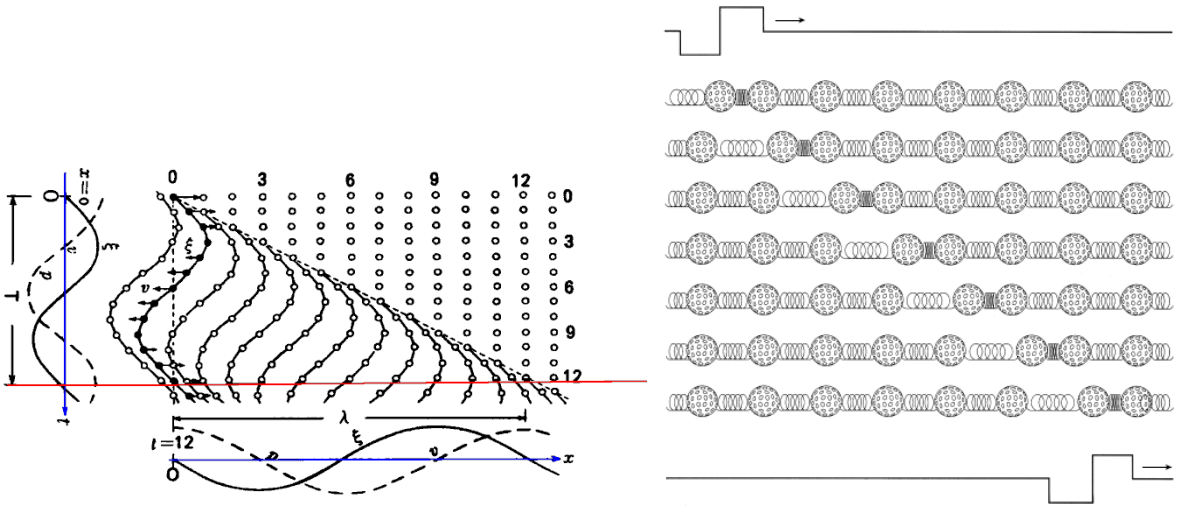


Figure 2.1 – Generation of an acoustic wave from the sinusoidal oscillation of a planar rigid surface; each particle is defined by: ξ – displacement, v – velocity and p – pressure (left side); example of a sound pulse propagating in a material using a golf ball spring model (right side), (from [Howard09]).

The motion of the medium particles vibrating about their mean position in sound waves can be described by the displacement of the particles, or more usually of their velocity.

Figure 2.2 depicts snapshots of the sound propagation due to the vibration of a pulsating spherical rigid surface (left image) and an infinite rigid plane (right image).

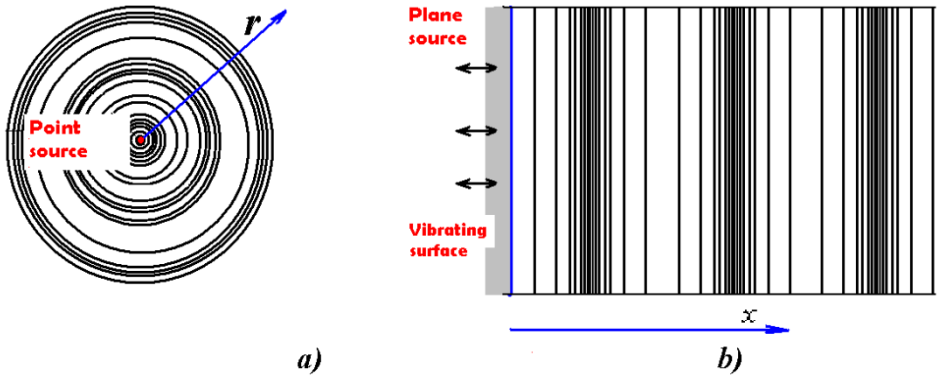


Figure 2.2 – Generation of an acoustic wave produced by pulsating a spherical rigid surface (left side) and by the vibration of an infinite plane (right side).

Assuming the medium is homogeneous and at rest, the speed of sound in the air is given by:

$$c = 331.5\sqrt{1+\frac{t}{273}} \tag{2.2}$$

where c is the speed of sound in the air (m/s), t stands for the medium temperature (ºC).
The wavelength is given by

$$\lambda = \frac{c}{f} \tag{2.3}$$

where f is the oscillation frequency in Hz.
The propagation of sound is governed by the principles of conservation of momentum and of conservation of mass expressed, respectively, by:

$$\text{grad } p = -\rho_0 \frac{\partial \mathbf{v}}{\partial t}$$

$$\rho_0 \text{ div } \mathbf{v} = -\frac{\partial \rho}{\partial t} \tag{2.4}$$

where p stands for the sound pressure, **v** is the vector particle velocity, t is the time, ρ is the total density including its variable part, $\rho = \rho_0 + \delta\rho$, ρ_0 is the static value of the gas density. It is assumed that variations of p and ρ are small when compared with their static values p_0 and ρ_0. Moreover, the magnitude of the particle velocity **v** should be much smaller than the sound velocity c.

Considering ideal gases and eliminating the particle velocity **v** and the variable part $\delta\rho$ from eq(2.4), the wave equation is obtained

$$c^2 \Delta p = \frac{\partial^2 p}{\partial t^2} \tag{2.5}$$

where $c^2 = \kappa \frac{p_0}{\rho_0}$ and κ is the adiabatic exponent or the coefficient of thermal conduction (for air κ = 1.4) and where the Laplace operator $\Delta \equiv$ div grad has been introduced as a kind of a shorthand notation.

Because this equation holds information of the propagation of sound waves in any lossless fluid, it is of capital importance for almost all acoustical phenomena. In fact, all acoustical relevant quantities, sound pressure, density and temperature variations are included in the equation.

2.2 Sound in free-field

The three dimensional propagation of an acoustic wave can be modeled by the wave equation, which is a linear relationship relating pressure p at position $d = (x, y, z)$ and time t, as

$$\nabla^2 p = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{2.6}$$

where $\nabla^2 p = \partial^2 p/\partial x^2 + \partial^2 p/\partial y^2 + \partial^2 p/\partial z^2$. The general solution to the wave equation in one dimension, which describes the sound pressure, p, is given in terms of two arbitrary functions, F(·) and G(·), as

$$p(t, x) = \mathrm{F}(c\ t - x) + \mathrm{G}(c\ t + x) \tag{2.7}$$

where F(·) and G(·) describe the waves traveling in the positive and negative x-direction, respectively, with a velocity c, which corresponds to the speed of sound, and assuming their second derivatives exist.

By considering a decomposition of the pressure function $p(t,x)$ in harmonic waves, one can use complex exponential functions for F and G, resulting:

$$p(t, x) = p_0 e^{jk(ct-x)} = p_0 e^{j(\omega t - kx)} \tag{2.8}$$

where $k = \omega/c = 2\pi/\lambda$ is the wavenumber or propagation constant corresponding to the angular frequency ω and p_0 is the amplitude of the sound pressure at the origin.

If the radiating surface is planar and rigid, then the sound pressure in a point in the space develops in plane sound waves, and at the distance d from the source $kd>>1$ and considering a lossless fluid, no absorption of energy by the air particles,

$$p(t, x) = p_0 \cos(\omega\ t - kx) \tag{2.9}$$

.

Small size sources (where all dimensions are << λ/4 and much smaller than the distance from the source) may be assumed to be point sources and spherical waves are radiated:

$$p(t,d) = \frac{1}{d} p_0 \cos[\omega(t - \frac{d}{c})] \quad . \tag{2.10}$$

To cater for sound sources radiating in non-plane or non-spherical fashions, such as loudspeakers, Equation 2.10 can be generalized, assuming spherical phase, as

$$p(t,d,\phi,\theta) = \frac{A}{d}\Gamma(\phi,\theta)\cos\left[\omega(t - \frac{d}{c})\right] \tag{2.11}$$

where $\Gamma(\phi,\theta)$ stands for the directivity function of the source, altitude and azimuth, normalized to unity.

The sound pressure level at a distance *d*, with $\tilde{p}_{rms} = \sqrt{1/T\int_T p^2 dt}$ representing the root mean square pressure of *p*, is expressed by:

$$L_p = 20\log_{10}\frac{\tilde{p}_{rms}}{\tilde{p}_{ref}} \quad (\text{dB}) \tag{2.12}$$

where $\tilde{p}_{ref}$ is the reference root mean square pressure with a value of 2×10^{-5} N/m^2.

The acoustical Intensity (or flux of energy vector), *I*, can be defined as the product of the sound pressure, *p*, by the particle velocity, *v*, embodying the magnitude of the sound field and the direction of propagation at each point of the space (it is a vector) and is expressed by:

$$I = \frac{1}{T}\int_0^T pv\, dt \quad . \tag{2.13}$$

The acoustical Power, *W*, is a measure of the energy passing through the radiating surface (if the surface *S* closes the sound source, then *W* corresponds to the radiated sound source energy), is obtained from:

$$W = \int_S I\, dS \quad . \tag{2.14}$$

For spherical wave propagation of a point source in open space at a distance d from the source and disregarding air absorption:

$$I = \frac{W}{4\pi d^2} \quad (\text{watt}/\text{m}^2) \,. \tag{2.15}$$

The sound intensity level can be defined as

$$L_I = L_W + 10\log_{10}\frac{1}{4\pi d^2} \quad (\text{dB}) \tag{2.16}$$

where L_W is the acoustical power level defined as:

$$L_W = 10\log_{10}\frac{W}{10^{-12}} \quad (\text{dB}) \,. \tag{2.17}$$

2.3 Sound in enclosures

The distribution of sound in rooms is different from open spaces, making its study a much more complex task. In fact, the sound propagation in enclosures is conditioned by its physical boundaries (floor, ceiling and walls).

The sound field is evaluated by solving the wave equation (deterministic method) and applying the appropriate boundary conditions or by considering a diffuse sound field distribution (statistical approach).

2.3.1 Solution of the wave equation

The wave equation is solved by applying the room boundary conditions. However, in the real world, due to the room geometry and the variety of the acoustical properties of the materials, very complex mathematical calculi are associated with this procedure. Therefore, in the following it will be assumed that the room boundaries are perfectly reflecting and rigid, i.e. the normal acoustic surface impedance is infinite (the sound absorption of the materials is zero, absorption coefficient $\alpha = 0$, the particle velocity components, v, normal to the surfaces cancel). This assumption allows, in a first approach, to study a number of acoustical phenomena occurring in real closed spaces. For these purposes, consider a room of physical dimensions L_x, L_y and L_z. Moreover, it can be shown that the wave equation yields non-zero solutions taking account the boundary condition only for particular discrete values of k_n called ‘eigenvalues’,

where the subscript *n* stands for each direction of the three-dimension space. The wave equation, the Helmholtz equation, that governs the propagation of sound in the three-dimensional space, considering harmonic excitation, is given by:

$$\frac{\partial^2 p}{\partial x^2}+\frac{\partial^2 p}{\partial y^2}+\frac{\partial^2 p}{\partial z^2}+k^2 p=0 \tag{2.18}$$

where p is the sound pressure, k is the propagation constant, ω is the angular frequency, and c is the speed of sound in the air.

The variables can be separated yielding the wave equation solution for a plane wave that is propagating in the positive x, y, and z-direction across this space:

$$p(x,y,z)=p_1(x)p_2(y)p_3(z) \tag{2.19}$$

Each term depends only of one space coordinate. The replacement of Equation 2.18 into Equation 2.19 corresponds to a linear system of three differential equations. For instance, for p_1:

$$\frac{\partial^2 p_1}{\partial x^2}+k_x^2 p_1=0 \tag{2.20}$$

and from the boundary conditions corresponding to rigid walls

$$\frac{\partial p_1}{\partial x}=0 \quad \text{for } x=0 \text{ and } x=L_x \quad . \tag{2.21}$$

Similar considerations are applied to $p_2(y)$ e $p_3(z)$. The constant terms are related as:

$$k_x^2+k_y^2+k_z^2=k^2 \quad . \tag{2.22}$$

The general solution of the Equation 2.21, considering harmonic time law sound pressure excitation, is given by

$$p_1(x)=\mathrm{A}_1\cos(k_x x)+\mathrm{B}_1\,\mathrm{sen}(k_x x) \tag{2.23}$$

where A_1 and B_1 are constants that depend of the boundary conditions. Hence, since the particle velocity is zero and the pressure is an extremum at x=0 and x=L_x, B_1=0 and $\cos(k_x L_x) = \pm 1$

imposing that k_xL_x be an integral multiple n of π. Then, the constants for each dimension related with the wavenumber are given by:

$$k_x = \frac{n_x\pi}{L_x}, \quad k_y = \frac{n_y\pi}{L_y}, \quad k_z = \frac{n_z\pi}{L_z} \quad . \tag{2.24}$$

Replacing these results in Equation 2.22 yields

$$k_{n_xn_yn_z} = \pi\left[\left(\frac{n_x}{L_x}\right)^2 + \left(\frac{n_y}{L_y}\right)^2 + \left(\frac{n_z}{L_z}\right)^2\right]^{1/2} \quad . \tag{2.25}$$

This result corresponds to the eigenvalues of the wave equation. The eigenfunctions or the eigenmodes of the room (resonances or normal modes) related with these eigenvalues are given by

$$p_{n_xn_yn_z}(x,y,z) = \mathrm{C}\cos\left(\frac{n_x\pi x}{L_x}\right)\cos\left(\frac{n_y\pi y}{L_y}\right)\cos\left(\frac{n_z\pi z}{L_z}\right) \tag{2.26}$$

where C is a general constant. This expression represents the tri-dimensional standing wave (excluding the harmonic dependence of $e^{j\omega t}$). This result shows an infinite number of sound pressure distributions for different combination of n_x, n_y and n_z which satisfy Equation 2.26. The numbers n_x, n_y and n_z indicate the numbers of nodal planes perpendicular to the x-axis, the y-axis and the $z$axis, respectively. Figure 2.3 depicts, as an example, the sound pressure distribution on the plane $z = 0$ to the eigenmode $n_x = 1$, $n_y = 1$ and $n_z = 0$.

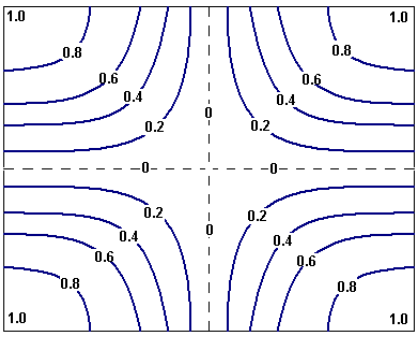


Figure 2.3 – Sound pressure distribution in a rectangular room of rigid walls for the eigenmode (1,1,0).

More complex sound pressure distributions are obtained for different combinations of n_x, n_y and n_z. The natural oscillation frequencies or room resonances of the enclosure are obtained from Equation 2.26 yielding:

$$f_{n_x n_y n_z} = \left[\left(\frac{c\, n_x}{2 L_x} \right)^2 + \left(\frac{c\, n_y}{2 L_y} \right)^2 + \left(\frac{c\, n_z}{2 L_z} \right)^2 \right]^{1/2} . \tag{2.27}$$

The distribution of the resonance frequencies, corresponding to the eigenmodes, is obtained experimentally by measuring the transmission characteristics of the room. In practice, this can be made by applying a sinusoidal sweep to the room that covers the frequency range of interest placing the sound source and the microphone in opposite corners. Figure 2.4 shows the transmission characteristics of a room in terms of the sound pressure level L_p for each mode.

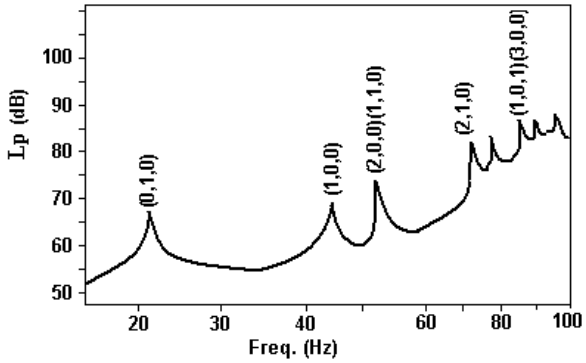


Figure 2.4 – Transmission characteristics of a rectangular room of dimensions 4.8 m x 7.7 m x 2.8 m.

As observed, for the low frequency bands, the resonances are located far away from each other. However, the increase in frequency produces an increase of the density of the resonances showing a more uniform spectrum.

That is the reason why the wave equation method is used in the study of rooms only in the low frequency range and for small/medium rooms. This frequency limit is the so-called Critical Frequency or Schroeder frequency and is simply estimated by

$$f_{sch} >= 2000 \sqrt{\frac{R60}{V}} \tag{2.28}$$

where *R60* is the reverberation time, in seconds (a measure of sound energy decay rate in the room; this concept will be developed in more detail in the next section) and *V* stands for room volume, in cubic meters. The operation above this frequency only assures high "modal overlap", ensuring that time-averaged fluctuations of reverberant sound over space or frequency obey to certain statistical laws. Above the Schroeder frequency, the statistical behavior of reverberant sound follows a well defined law. However, the assumption of diffuse field conditions is not accomplished yet. The Schroeder frequency is the frequency at which the average modal overlap index (the average modal bandwidth divided by the average frequency spacing between modes) is 3.

In general, the room under test is far from a rectangular room with rigid walls. Therefore, the sound absorption coefficients and the current room geometry should be considered, increasing the mathematical complexity of calculation in an exponential fashion. Actually, the explicit evaluation of the eigenvalues and eigenfunctions requires generally the application of numerical methods such as the Finite Element Method (FEM) or Boundary Element Method (BEM).

Nevertheless, this is an exact method, involving all the physical and acoustical quantities and with the advent of fast computers, running with parallel processing tasks using multi-core processors and large memory capabilities this is currently very often used.

2.3.2 Determination of the sound field by the assuming diffuse or reverberant field (statistical method)

The sound field distribution in a room created by a radiating sound source consists of the sum of the contributions of the direct wave plus the reflected waves on the room boundaries. Figure 2.5 presents a tridimensional model of an auditorium where the multiple paths of acoustical waves, from the stage to a location in the audience area, assuming specular radiation, are represented. Only up to three reflections for each ray were considered in this model.

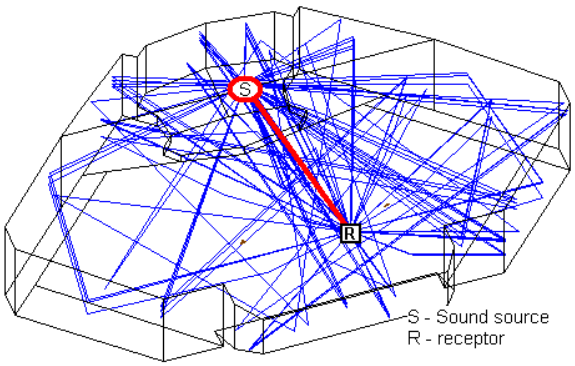


Figure 2.5 – Virtual 3D model of the sound field distribution in an auditorium; red line: direct sound wave path.

The multiple sound wave reflections on the boundaries result in the decay of the sound energy after the sound source is turned off. The main processes related to this loss of energy with time are (i) losses in the propagation medium (molecular sound absorption due to collisions of air particles), and (ii) energy absorption by the surrounding material elements.

By considering these aspects, especially those referred in (ii) and (iii), a time analysis of the reflections showing the attenuation and delay of each propagation path from the source to the target is possible. In this approach, the sound source radiates acoustic rays (echogram, following laws identical to the classical optics) [Bodiund77, Lochner64]. Figure 2.6 depicts the echogram related to the analysis in the Figure 2.5 assuming auditorium full seating capacity (occupancy of 100%).

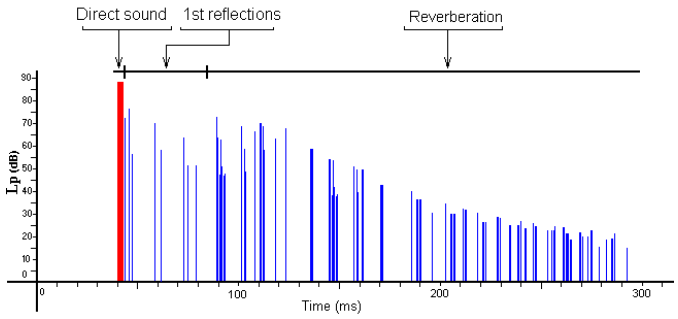


Figure 2.6 – Echogram corresponding to the case in Figure 2.5.

The echogram is split into different time regions related directly with the human hearing perception. In a first approach, two different regions are considered for the reflected waves: (i) first reflections – temporal section starting at the time of appearance of the direct wave and

ending at 30 to 80ms, depending on the required study. The reflections occurring within this period of time are usually well separated in time and are psycho-acoustically related with the volume of the room (spaciousness), differences of absorption coefficients of the materials and the relative location of the sound source and reception point (ii) Reverberation zone – the density of the reflections is large and therefore, the study can be based on a statistical process approach. In this situation, the directionality of the sound is lost, making difficult the recognition of the physical characteristics of the room near the spectator. Nevertheless, the notion of spaciousness is kept.

Echograms can be very helpful in the investigation of important acoustic issues, such as:

- Distance from sound source to the reception location, d - The time delay of the direct sound wave, referred by the red line, can be related with the sound speed c to estimate the distance d;
- Distance between reflecting surfaces, d_r – The relative delay between the direct sound wave and the first reflections (blue lines near direct sound wave) is related to the extra path tracked by each sound wave;
- Best choice of the acoustic materials – Significant differences among the amplitudes of the first reflections, usually, indicate very different sound absorption coefficients of the material used for the boundary elements near the receptor. This is a situation to avoid due to problems of correct perception of the sound directionality;
- Incorrect project of the geometry for the room boundaries – If the amplitude of the direct sound wave is smaller than some reflected sound wave, the reason might be related to the wrong orientation of a number of reflecting panels to form some sort of a parabolic system where the reception point is placed near the focus or coincidence of reflections;
- Mean absorption of the room – after the first reflections, the energy in the room should follow an exponential decay law due to the high density of reflections. The energy decay rate is related directly with the mean absorption of the acoustical materials applied to the room [Agerkvist93, Barron73]. Considering the example in Figure 2.5,

for a room occupancy lower than 100%, a higher amplitude of the reflections in the tail of the echogram would be observed (corresponding to a slow dissipation of energy in the room with time). If decay of energy in the room does not follow an exponential law, this may indicate a mixture of modes with different reverberation times due to the use of reflecting panels with very different sound absorption.

For rooms with low to medium volumes and high energy decay rates, the reverberation phenomenon is identified by the human auditory system as a sound colored by the specific characteristics of the room (geometry and room surface materials). An interesting test consists of listening to different impulse responses. The perception of the dimension of rooms with large volumes (low energy decay rates) is associated with the very distinct time delay difference between the direct and the reflected sound waves (spaciousness) and the long reverberation times. Therefore, the reverberation concept holds important spacial information about the volume and geometry of the room.

Rooms boundaries (walls, floor and ceiling) with highly reflective materials (very low sound absorption), such as, concrete, ceramic tiles or glass panels, usually, lead to high magnitude of the frequency response in the high frequency bands. Therefore, the sound field becomes almost diffuse, not facilitating the intelligibility of the speech. On the other hand, carpets, thick curtains, and fiber based materials usually have high absorption properties in frequency bands above 1 kHz. The extensive use of these types of materials leads to deficiencies in the high frequency bands, imposing a sensation of low acoustical comfort. This sensation of discomfort is related to the perceived loss of spaciousness.

Another related issue concerning room acoustics is the sound field distribution inside an enclosure. In the area close to the sound source (depending of the characteristics of the room, as will be shown next), though far away from the near field, the sound field is made almost entirely of the direct sound waves. Therefore, the laws of free field propagation (open air conditions) are applied, meaning that the sound pressure levels drop to a half of its initial value by doubling the distance to the source. After a certain distance from the source, the reverberant field is considerably higher than the direct sound. The Critical Distance, D_c is then defined as the distance where the acoustical intensities of the direct and of the reverberant fields are equal.

After this distance the direct sound fades rapidly and the sound field tends to be homogeneous. The perception of directionality is lost; this zone is called reverberant or diffuse field.

The estimation of the sound field distribution is formally studied from the acoustical energy density [Agerkvist93]. The study based only on the energy density approach does not consider the direction associated to the acoustical quantities. Therefore, the mathematical formalism of the concepts is much simplified. The energy at a point is given by:

$$E = E_d + E_e \tag{2.29}$$

where E is the energy density at a point in the room, E_d is the direct sound wave energy density of, and E_e is the reverberant field (steady-state situation) energy density.

Direct sound wave energy.

This quantity corresponds to the energy in one particular point of the space belonging to the distant field (far from the sound source) considering free field propagation conditions and is calculated from the acoustical intensity given by Equation 2.30:

$$E_d = \frac{I}{c} = \frac{1}{c}\frac{W}{4\pi d^2}. \tag{2.30}$$

Reverberant field energy density

A diffuse sound field is assumed for the estimation of this quantity, i.e. there is no privileged propagation direction. Therefore, a statistical approach is applied [Lundeby95].

The energy on all room boundaries (acoustical intensity) is given by

$$I = \frac{c\,E\,S\overline{\alpha}}{4} \tag{2.31}$$

where S is the total area (m^2), and $\overline{\alpha}$ is the room mean sound absorption coefficient.

Considering a sound source with acoustical power W, the fundamental energy equation in the room is given by:

$$V\frac{dE}{dt} = W - \frac{c\,E\,S\overline{\alpha}}{4} \tag{2.32}$$

where V is the total room volume (m^3).

The solution of the differential equation (2.32) is obtained by applying the initial condition $E = 0\,|_{t=0}$,

$$E = \frac{4W}{cS\bar{\alpha}}\left[1 - e^{-\frac{cS\bar{\alpha}}{4V}t}\right] \tag{2.33}$$

The energy increases exponentially from the moment when the excitation starts until the equilibrium point is reached, when the energy radiated by the source equals the energy dissipated in the room due to sound absorption of the room boundaries.

The energy density in the steady-state situation, E_0, is obtained at $t \to \infty$,

$$E_0 = \frac{4W}{cS\bar{\alpha}}. \tag{2.34}$$

The energy applied to the room after the first reflection on a wall is $W(1-\bar{\alpha})$. This energy corresponds to the energy lost in the steady-state situation, then

$$E_e = \frac{4W}{cS\bar{\alpha}}(1-\bar{\alpha}) = \frac{4W}{cR} \tag{2.35}$$

where $R = \frac{S\bar{\alpha}}{(1-\bar{\alpha})}$ is the so-called room constant.

The energy density for different distances from the source is estimated by replacing E_d and E_e in Equation 2.29,

$$E = \frac{1}{c}\frac{W}{4\pi d^2} + \frac{4W}{cR} = \frac{W}{c}\left[\frac{1}{4\pi d^2} + \frac{4}{R}\right]. \tag{2.36}$$

The energy density is related with the sound pressure by:

$$E = \frac{p^2}{\rho c^2} \tag{2.37}$$

where ρ stands for the air density (kg/m3).

By replacing in Equation 2.36, the total sound field for each point in the room is defined by:

$$\overline{p^2} = \rho c W \left(\frac{1}{4\pi d^2} + \frac{4}{R} \right) \tag{2.38}$$

or

$$L_p = L_W + 10\log_{10}\left(\frac{1}{4\pi d^2} + \frac{4}{R} \right) \quad \text{(dB)} \quad . \tag{2.39}$$

When $R \to \infty$ ($\bar{\alpha} \to 1$) the second term approaches zero, corresponding to the sound propagation in free field conditions. Figure 2.7 shows a set of curves of the sound pressure distribution in an enclosure for a number of values of R.

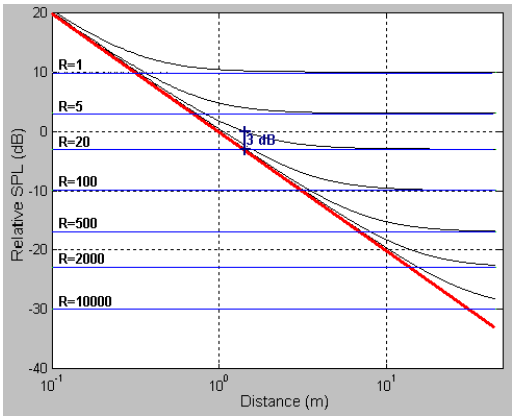


Figure 2.7 – Sound pressure distribution in an enclosure for different of R at different distances from the source.

The Critical Distance D_c is then calculated from:

$$\frac{1}{4\pi D_c^2} = \frac{4}{R} \tag{2.40}$$

yielding

$$D_c = \sqrt{\frac{R}{16\pi}} \cong \frac{\sqrt{R}}{7} \quad . \tag{2.41}$$

Figure 2.8 depicts the sound pressure levels distribution for the example in Figure 2.5 using three sound sources in the audience area.

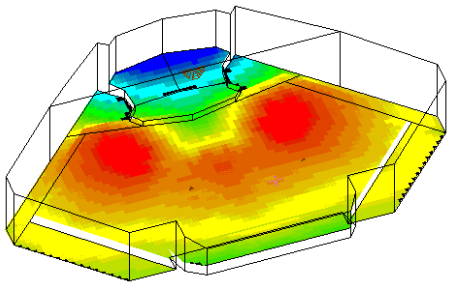


Figure 2.8 – Sound pressure levels distribution inside the auditorium.

If the sound radiation is not omnidirectional, a directivity factor Q different from unity must be considered in Equation 2.39 and 2.40 becomes:

$$L_p = L_W + 10\log_{10}\left(\frac{Q}{4\pi d^2} + \frac{4}{R}\right) \quad \text{(dB)} \tag{2.42}$$

For large venues, modeling the sound field uniquely on statistical energy distribution can be a fair approach, essentially (i) if the variation of the absorption coefficients of the surface materials is not significant when compared with the α mean value, or (ii) for frequency bands above 500 Hz.

Often, the solution of the wave equation and statistical approaches are used simultaneously given their complementarity.

2.3.3 Room acoustical quality criteria

The complete acoustical characterization of an enclosure requires the definition of some objective parameters based on the subjective human perception regarding the assessment of the acoustic quality of the room. The parameters obviously depend of the intended room function. Whereas the reverberation and/or the sound level attenuation with distance from the source may be sufficient in an industrial hall, a more comprehensive set of parameters must be used in more demanding rooms, such as concert or theatre halls. A number of other parameters that correlate well with the subjective impression are based on data calculated from measured impulse responses in the room.

Some important room quality parameters, defined to cater for the acoustic comfort are presented next.

Reverberation Time, RT60

The reverberation time is defined as the time needed for the sound pressure level or the space-averaged sound energy density in a room to decrease by 60dB (a factor of 10^{-6}) from its initial value i.e. after the sound source is turned off. Experimentally, the RT60 is estimated from the energy decay curves.

The method of obtaining decay curves can be based on the room impulse response by reverse (backward) integration of squared impulse responses, the so-called Schroeder integration [Schroeder66].

$$\left\langle S^2(t)\right\rangle = N_F \int_t^{T_i} h^2(\tau)d\tau \tag{2.43}$$

where N_F is the constant proportional to the power spectral density of the test signal, $h(t)$ is the room impulse response RIR and T_i is the time used to truncate the $h(t)$ (is infinite in absence of background noise).

As Figure 2.9 shows, care must be taken in choosing the initial time for the integration (or the truncation point of the RIR) due to the background in order to avoid significant deviations of the estimated decay curves and, therefore, errors on the RT60. An example of this situation is depicted in the right side of the figure. It is customary to compensate for the final 'lost' energy.

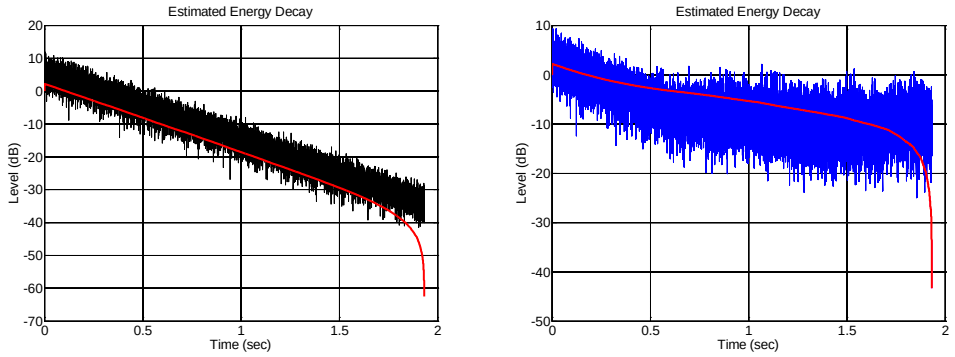


Figure 2.9 – Schroeder method to estimate the RT60; left side, clean RIR; right side, RIR corrupted by background noise.

Another method frequently used consists of directly recording the sound pressure level decay after exciting the room, the so-called interrupted noise method, and applying linear regression techniques to obtain $L(t)$,

$$L(t) = -D\ t + E_0 \tag{2.44}$$

where D is the slope of the linear regression, and E_0 is the initial energy value.
The time required for a 60dB decrease of the acoustical energy is the reverberation time

$$\mathrm{RT60} = \frac{60}{D}\ (s) \quad . \tag{2.45}$$

Figure 2.10 shows a scheme for the estimation of the RT60.

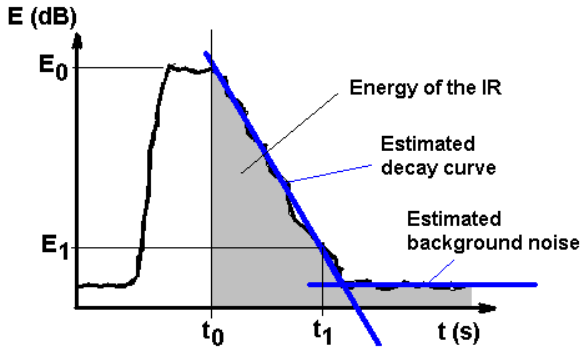


Figure 2.10 – Estimation of the RT60 by the interrupted noise method.

The concept of signal-to-noise ratio applied to the RIR is defined by the ratio of the peak value by the noise appearing at the tail of the room impulse response (the decay curve has already sunk in the background noise floor),

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\max[h(t)]}{\sqrt{\frac{1}{t_1 - t_0}\int_{t_0}^{t_1} h^2(t)\,dt}}\right) \tag{2.46}$$

where t_0 and t_1 are the initial and final time for the calculation of the noise energy of the RIR. Sometimes, the calculation of the background noise is executed for the time interval just before the starting point of the RIR. This procedure is useful when the decay curve is not completely faded for the tail of the RIR.

In practice, the SNR is usually not sufficient enough to assure a dynamic range of the energy decay curve of 60dB due to background noise. The dynamic range of the energy decay E_0 - E_1 could then be lowered assuming the RT60 is calculated by extrapolation of the decay curve considering a range of 60dB. Normally, a dynamic range of 20dB (or 30dB), between −5 and −25dB (−5 and −35dB) below the initial level, is used [Chu78, Faiget98, Vorlander94, ISO/DIS3382]. The RT60 corresponds to t_1 - t_0.

The reverberation time is an indicator of the sound environment quality perceived inside the enclosure. For very low values of RT60, the listener feels a claustrophobic sensation (a quasi-anechoic chamber is an example) i.e. the perception of spaciousness is lost, only the direct sound is heard. For very high values, reverberation dominates over the direct sound and the intelligibility becomes difficult and confusing (cathedrals and reverberant chambers are examples) [Blesser01].

The ideal RT60 corresponds to an intermediate value when the sound signal is well perceived with comfort, being much related to the type of usage of the room. Figure 2.11 shows the optimal value of RT60 versus the function of the room, for the 500 Hz frequency band.

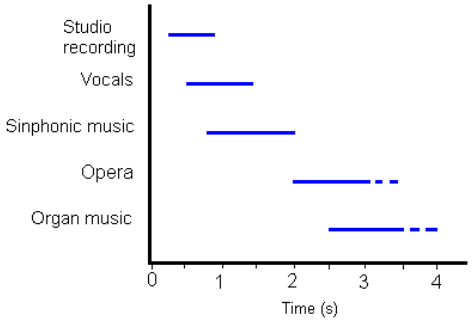


Figure 2.11 – Ideal RT60 versus the type of usage of the room.

It is acknowledged that the reverberation time plays a very important role and there is sufficient background experience on how long or short it should be depending of the size of the room and be related to the type of the performance room; theatre, room for music performance, etc.

In auralization studies for virtual simulation of room acoustics, this information is used to better characterize the real room space. Therefore, the reverberation of the sound program (music or speech) is adjusted [Bodiund77].

Early Decay time, EDT

The EDT is considered the RT60 of short duration. The energy decay of 10dB from the initial value is analyzed and the extrapolation to 60dB is applied. The main reason for analyzing the first part of the energy decay curve is related with psychoacoustic issues. Recently, it has become common to characterize the rate of sound decay in its initial portion by the EDT. Listening tests based on binaural impulse responses recorded in different concert halls confirmed that the EDT is a more reliable indicator of the subjective impression of *reverberance* [Gade94]. In fact, it is in the first part of the decay curve that the early reflections occur. Therefore, this measure is dependent of the measurement position, being, for example, useful to characterize an audience area in a venue with respect to aspects of the auditory system, such as sound brightness and speech intelligibility.

On the other hand, the human auditory system can barely process energy variations above 10dB, at least in situations of performative musical events.

This criterion also relates the reaction of the human auditory system to impulsive sounds (sharpness concepts) and its sound intensity [Lochner64].

The ratio of the EDT to RT60 is used to make an energy balance between the initial and total energy of the decay curve (room impulse response).

As mentioned above, the energy of different parts of the room impulse response is related perceptually. Usually, the RIR (or the decay curve) is split into two parts: (i) the first part from the beginning to a specific instant *t*, and (ii) the second part from that instant to the end (considering the ending instant assigned to the truncation point of RIR) [Agerkvist93].

This approach considers some psychoacoustic issues. For instance, the minimum value of 50ms is usually assigned to the instant *t* assuming the human auditory system is not able to discriminate time intervals lower than this value. Therefore, as an attempt to quantify the first reflection's effect [Barron73, Lundeby95], a number of different subjective analyses are proposed for different values of *t* (50, 80 or 95 ms).

Clarity 80, C80

This criterion consists of the ratio of the acoustical energy belonging to the first 80 ms of the RIR and the rest of the energy, as defined by Reichardt (1972). This approach accounts for the first reflections against the diffuse sound field and is evaluated from:

$$\mathrm{C80} = 10\log_{10}\frac{\int_{t=0}^{80ms} h^2(t)\,dt}{\int_{t=80ms}^{\infty} h^2(t)\,dt} = 10\log_{10}\frac{E80}{E_{Rev}} \quad \text{(dB)} \tag{2.47}$$

where E80 is the energy of the first 80 ms of the RIR, and E_{Rev} is the energy of the reverberant sound field.

The *C80* is very often used for the assessment of rooms dedicated to musical performances.

Notice that this parameter can be a negative value. In fact, it is almost consensual that the optimal value for listening to the music is about 0dB. However, the range of –2 to 2dB to symphonic music and 1 to 4dB for opera are wide accepted as a good indication of listening comfort. For speech intelligibility the analysis is often made using the energy of the first 50 ms of the RIR, corresponding to the C50 parameter. To avoid the syllable intelligibility decreasing below 80 % the C50 should be$\geq$ -2dB. Phrase intelligibility (text intelligibility or continuous speech), however, is increased to approximately 95 %, and is higher than the syllable intelligibility, due to the context of the words in the phrase information.

Definition 50, D50

Thiele (1953), assuming that the energy of the first reflections is mandatory for the reinforcement of the direct sound, related this energy to the energy of the all RIR.

This criterion is, as the *C80*, a measure of the fluctuations of the energy contained in the RIR.

$$\mathrm{D50} = \frac{\int_{t=0}^{50ms} h^2(t)\,dt}{\int_{0}^{\infty} h^2(t)\,dt} = \frac{E50}{E_{Total}} \quad (\%) \tag{2.48}$$

where E_{50} is the energy of the first 50 ms of the RIR, and E_{Total} is the total energy of the RIR.

This parameter is used for rooms dedicated essentially to speech. It is found that the higher the Definition the better the speech intelligibility.

Reverberation Index, RI

Beranek [Beranek65] extended the D50 concept relating the energy of the direct sound to the first reflections as:

$$\mathrm{RI} = 10\log_{10}\left(\frac{1-D50}{D50}\right) \ \text{(dB)} \qquad (2.49)$$

where D50 is given by Equation 2.48.

Enclosures such as conference rooms, theatre halls or, in a wide sense, rooms dedicated to speech, have specific characteristics when compared to auditoria and concert halls, which are mostly used for musical activities. This fact is related directly with differences on the statistical properties of the sound signals.

The speech signal consists essentially of syllabics holding vowels and consonants which are described by sinusoidal-like and impulsive signals, respectively. It is found that the syllabics containing consonants are those responsible for the speech intelligibility.

A correct understanding of the speech can be much degraded due to the multiple reflections of the sound waves in a room. The syllabics tend to overlay between them causing difficulty on the auditory processing. On the other hand, it is common sense that the intelligibility of the speech is degraded, or even impossible, if the speech sound levels speech are too low when compared with the background noise levels, i.e. temporal masking effects are prominent for low values of the SNR.

A number of criteria were stated in order to define measures for the intelligibility of speech taking account the background noise energy and reverberation time. Some of these measures are described next.

Speech Transmission Index, STI

The *STI* accounts for a number of issues regarding the intelligibility of speech but does not cater for possible non-linearities of the room system.

Speech is a non-stationary signal, showing pronounced fluctuations at different instants, nevertheless, if short time periods are assumed, less than 20-30ms, the speech signal is considered stationary. However, the time domain variations are slow allowing, for experimental purposes, the analysis to be carried out by using a signal $S(t)$ synthesized by modulating white noise signals.

$$S(t) = \overline{I_i}[1+\cos(2\pi Ft)] \tag{2.50}$$

where F is the modulation frequency, between 0.4 and 20 Hz, and $\overline{I_i}$ is the white noise intensity level.

The STI is estimated by using the *Modulation Transfer Function*, defined by Schroeder [Blesser01, Schroeder66],

$$m(F) = \frac{\int_{t=0}^{\infty} h^2(t) e^{-2j\pi Ft}\, dt}{\int_{t=0}^{\infty} h^2(t)\, dt} \tag{2.51}$$

where $\overline{p_s^2}$ is the mean square value of the speech signal and $\overline{p_n^2}$ is the variance of the noise. The numerator of Equation 2.51 is the Fourier transform of the squared RIR (spectrum of the decay curve) and the denominator stands for the total energy of the RIR, or can be interpreted as being the DC component of the numerator. The noise effects are accounted for by applying a correction factor to Equation 2.51,

$$m(F) = \frac{\int_{t=0}^{\infty} h^2(t) e^{-2j\pi Ft}\, dt}{\int_{t=0}^{\infty} h^2(t)\, dt} \frac{\overline{p_s^2}}{\overline{p_s^2} + \overline{p_n^2}} . \tag{2.52}$$

The SNR is calculated for each frequency band, usually, octave bands.

$$\frac{S}{N_k(F)} = 10 \log_{10}\left(\frac{m(F)}{1-m(F)}\right) \tag{2.53}$$

where S is the power of the speech signal, and $N_k(F)$ is the power of the noise for the frequency band k.

The mean value of the SNR is limited to the interval of $\pm$15dB

$$\left\langle \frac{S}{N_k} \right\rangle = \frac{1}{n} \sum_k \frac{S}{N_k(F)} \tag{2.54}$$

where n is the number of the modulation frequency.

The term Transmission Index, TI, for each frequency band [Chu78] is estimated by

$$\langle TI_k \rangle = \frac{\left\langle \frac{S}{N_k(F)} \right\rangle + 15}{30}. \tag{2.55}$$

The Speech transmission Index, STI, is obtained by weighting TI,

$$STI = \frac{\sum_k W_k TI_k}{\sum_k W_k} \tag{2.56}$$

where W_k is the weighting coefficient of TI.

The parameter STI allows the estimation of the maximum distance from the source for which the speech intelligibility keeps acceptable values. The STI decreases rapidly until reaching that distance maintaining approximately the same value above due to the reverberant sound field in the room.

Rapid Speech Transmission Index, RaSTI

The RaSTI is a simplified version of the STI with some differences:

- The octave frequency bands centered in 125 and 2 kHz are not accounted for;
- All the SNR, X_i, are limited to $\pm$15dB;
- The arithmetic mean $\overline{X_i}$ is calculated for all X_i.

$$\text{RaSTI} = \frac{\overline{X_i} + 15}{30}. \tag{2.57}$$

InterAural CrossCorrelation Coefficient, IACC

The IACC is an objective parameter used to quantify the subjective issues related to the perception of space in rooms (spaciousness). Therefore, it is based on the binaural RIR analyzed in octave frequency bands, recorded for several locations in the audience area.

$$\mathrm{IACC}_t = \left| \mathrm{IACF}_t(\tau) \right|_{\max}, \qquad -1 < \tau < +1 \tag{2.58}$$

where $\mathrm{IACF}_t(\tau)$ is the InterAural Cross-Correlation Function, τ is the sliding parameter, and t is the time duration of the analysis.

The $\mathrm{IACF}_t(\tau)$ is defined by

$$\mathrm{IACF}_t(\tau) = \frac{\int_0^t h_L^2(t)\, h_R^2(t+\tau)\, dt}{\sqrt{\int_0^t h_L^2(t)\, dt \int_0^t h_R^2(t)\, dt}} \tag{2.59}$$

where h_R is the RIR measured for the right ear, and h_L is the RIR measured for the left ear.
The IACC_{E3}, corresponding to the IAAC for an elapsed time t = 80ms, estimated for the octave frequency bands of 500 Hz, 1 kHz and 2 kHz, is usually applied for rooms dedicated to music performances. This index is closely associated to the perception of spaciousness experienced by a listener being a good indicator of the sound quality of the venue [Beranek88].

These criteria are based on the RIR measured for several locations in the room. Therefore, different results are obtained for each location in the audience area, and the final results should be presented as an average.

The parameters can be evaluated on frequency bands where the type of the filters is normalized for each criterion.

2.4 Hearing mechanism

Some significant aspects concerning the composition and behavior of the human auditory system are discussed in this section. Perceptual models which are representative of the main characteristics and tolerances of the human auditory system use this information. Some of these properties are closely related to the structure of the peripheral region of the human auditory system, lying between the pinna and the auditory nerves.

2.4.1 The human auditory system

The human auditory system consists of the ear, auditory nerve fibers, and a section of the brain responsible for processing issues. The sound waves are converted into sensations perceived by the auditory cortex.

The ear is the outer peripheral system responsible for the conversion of acoustic energy (sound stimulus) into electrical impulses by firing the auditory nerve. The ear is divided into three distinct parts: the outer, the middle, and the inner ear, as depicted in Figure 2.12.

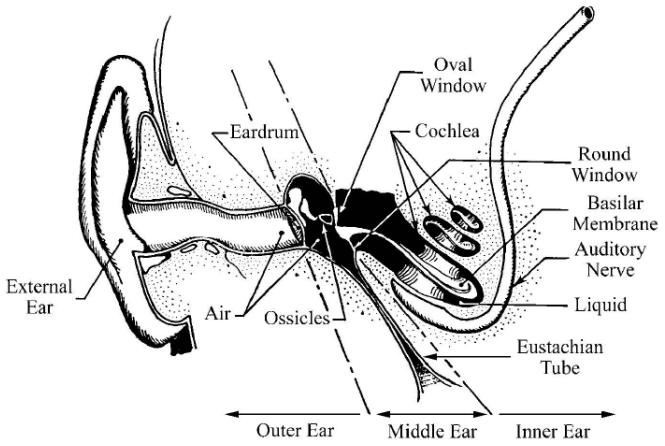


Figure 2.12 –Overall view of the main structure of the human ear, showing the outer, middle and inner ear.

2.4.1.1 The Outer Ear

The outer ear consists of the external ear, the pinna and concha (the visible part of the ear), the meatus (ear canal), and the tympanic membrane (eardrum). Sound waves incident from different angular positions are spectrally shaped by the pinna and the concha and delivered to the ear canal. The concha acts as an acoustic resonant cavity. Acoustic effects of the pinna and the concha help the interpretation of sound localization. In fact, the combined acoustic effects

are particularly useful for determining whether a sound source is in front or behind, and to a lesser extent whether it is above or below, giving more sensitivity to sounds coming from the front of the listener [Shaughnessy00].

The ear canal is a 2.5-3.5cm long tube, approximately, which drives the sound to the tympanic membrane, the ear drum. A cavity with one end open and the other closed by the tympanic membrane, the meatus acts as a quarter-wave resonator with a center frequency around 4000 Hz. It further acts as a filter with amplitude frequency response as shown in Figure 2.13.

The tympanic membrane consists of three layers: the outside layer which is a continuation of the skin lining of the auditory canal, the inside layer which is continuous with the mucous lining of the middle ear, and the layer in between these which is a fibrous structure which gives the tympanic membrane its strength and elasticity. An oscillatory motion is generated on the tympanic membrane by the external acoustic stimuli. It converts acoustic pressure variations from the outside world into mechanical vibrations in the middle ear.

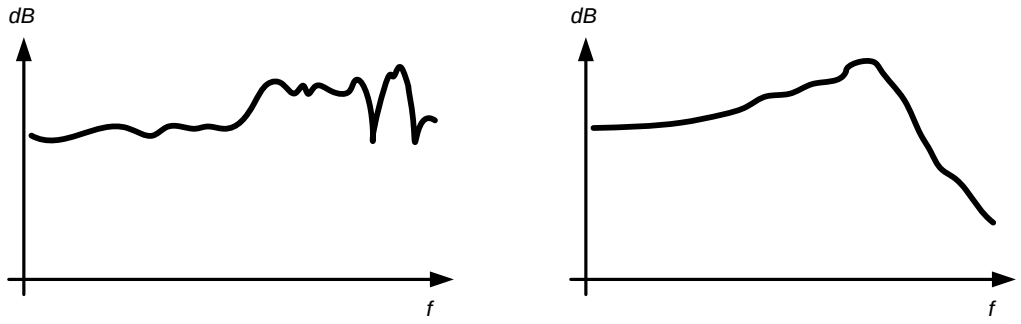


Figure 2.13 – Equalization process in the outer ear. Left side: frequency response of the pinna system for a particular direction of an incident sound wave. Right side: ear canal resonance centered at 5000Hz, approximately.

2.4.1.2 The Middle Ear

The middle ear is considered to start at the tympanic membrane, a light, thin, highly elastic structure which forms the boundary between the outer and middle ears, and ending at the oval window, a membrane at the entrance of the fluid-filled inner ear, and is constituted of the ossicles, the three smallest bones in the human body named malleus (hammer), incus (anvil), and stapes (stirrup). This part of the ear has two purposes: (i) acting as levers performing the impedance matching between the air outside the eardrum and the fluid in the cochlea, and (ii) protecting the hearing system to some extent from the effects of loud sounds.

The efficient transfer of energy from the tympanic membrane to the oval window is achieved by acting on the oval window by an effective pressure. The pressure, adjusted by mechanical means, is made to be greater than that acting on the tympanic membrane in order to overcome the higher resistance to movement of the cochlear fluid compared with that of air at the input to the ear. The pressure ratio between the stapes footplate and the tympanic membrane, in decibels, is about 30dB [Howard09].

The acoustic "reflex effect" consists of the activation of the middle ear muscles, to change the type of motion of the ossicles when loud low-frequency sounds, approximately above 75dB, are applied to the eardrum, attenuating the pressure transmission by up to 14dB. This effect is also activated during voicing in the speaker's own vocal tract [Blesser01]. Due to their mass, the ossicles act as a low-pass filter with a cutoff frequency around 1000 Hz. The reaction time for the muscles to contract in response to a loud sound is about 60–120 ms. Therefore, in the case of a loud impulsive sound, such as the firing of a gun, it has been suggested that the acoustic reflex is too slow to protect the hearing system. In gunnery situations, a sound loud enough to trigger the acoustic reflex, but not so loud as to damage the hearing systems, is often played at least 120 ms before the gun is fired.

2.4.1.3 The Inner Ear

The inner ear has a great role in both the hearing and the body balance. The hearing organ is a bony structure comprised of the semicircular canals of the vestibula and the cochlea. In a simple approach, the sound stimuli enter the cochlea, through the movement of the stapes footplate at the oval window, setting the fluid within in motion. It is the most complex part of the ear, wherein the mechanical pressure waves are converted into nerve firings to be processed eventually by the brain.

The vestibula is the organ that helps balancing the body and has no apparent role in the hearing process [Gold99]. The cochlea is a cone-shaped spiral, of approximately 2.75 turns, filled with a gelatinous incompressible fluid (endolymph or "perilymph) in which the auditory nerve terminates. At its base this tube has a cross section of about 4 mm^2, and two membrane covered openings, the Oval Window and the Round Window. The Oval Window is connected to the ossicles. The Round Window is free to move to better equalize the pressure since the endolymph is incompressible.

The cochlea has two membranes running along its length, the Basilar Membrane (BM) and Reissner's Membrane, which divide the cochlea into three channels, as seen in Figure 2.14. These channels are called the Scala Vestibuli, the Scala Media, and the Scala Tympani. Pressure waves travel from the Oval window through the Scala Vestibuli to the apex of the cochlea. A small opening (helicotrema) connects the Scala Vestibuli to the Scala Tympani.

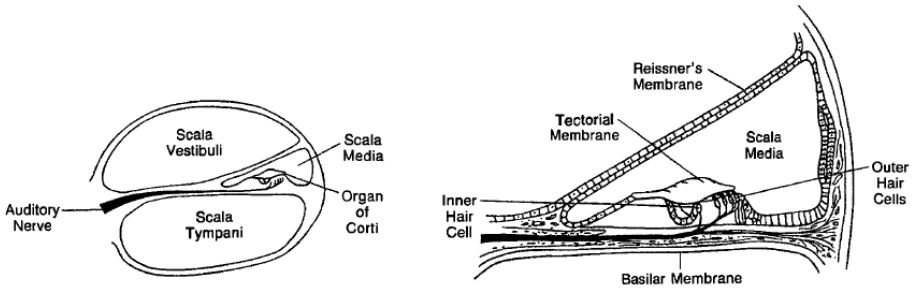


Figure 2.14 – Cross-section of the cochlea [Shaughnessy00].

The sound pressure waves then travel back to the base through the Scala Tympani, terminating at the Round Window. The velocity of sound in the cochlea is about 1600 m/s, thus, no appreciable phase delay is observed.

2.4.1.4 The Basilar Membrane and the Hair Cells

The mechanics of the Basilar Membrane (BM) can explain many effects of masking (described below). Within the BM, mechanical movements are transformed into nerve stimuli transmitted to the brain. The BM performs a crucial part of sound perception. It is narrow and thin at the base of the cochlea, gradually becoming both wider and thicker along its length to the apex. The basilar membrane therefore responds best to high frequencies where it is narrow and thin (at the base) and to low frequencies where it is wide and thick (at the apex). Each point on the cochlea can be viewed as a mass-spring system with a resonant frequency that decreases from base to apex. The BM thus acts as a spectrum analyzer. A frequency to place transformation is performed such that if a pure tone is applied to the Oval Window a section of the BM will vibrate [Békésy60]. The amplitude of the BM vibration is dependent of the distance from the oval window and the frequency of the stimulus. The BM vertical displacement is small near the oval window. Growing slowly, the vertical displacement reaches a maximum at a certain distance from the oval window. The amplitude of the vertical displacement then rapidly dies

out in the direction of the helicotrema. The frequency of a signal that causes maximum displacement at a given point of the BM is called the Characteristic Frequency (CF_{BM}).

The vibration of the BM is picked up by the hair cells of the Organ of Corti. There are two classes of hair cells, the Inner Hair Cells (IHC) and Outer Hair Cells (OHC). About 90% of afferent (ascending) nerve fibers that carry information from the cochlea to the brain terminate at the IHC. Most of the efferent (descending) nerve fibers terminate at the OHC, which greatly outnumber the IHC. Empirical observations suggest that the OHC, with direct connection to the tectorial membrane, can change the vibration pattern of the BM, improving the frequency selectivity of the auditory system [Gold99, Moore97].

Measurements from afferent auditory nerves have shown further nonlinearities in the auditory system. All IHC show a spontaneous rate of firings in the absence of stimuli. As a stimulus (such as a tone burst at the CF for the IHC) is applied, the neuron responds with a high rate of firings, which after approximately 20 ms decreases to a steady rate. Once the stimulus is stopped, the rate falls below the spontaneous rate for a short time before returning to the spontaneous rate [Gold99].

More recently, James Fulton [Fulton08], hearing researcher, describes a new theory on how the inner ear converts sound into neural impulses. In cross-section, the rigidly attached supports for the Basilar Membrane to the sides of the Cochlea, and the membrane rigidly braced longitudinally by the Organ of Corti constrain the Basilar Membrane from vibrating as traditionally described. Instead, it is likely to provide a structurally rigid base on which the hearing transducers are mounted.

Fulton describes a gel-like substance on the underside of the Tectorial Membrane, another rigid structure parallel to the Basilar Membrane, which carries a surface acoustic wave and acts like an acoustical waveguide. This gel SAW waveguide called “Henson’s Stripe” is in contact with the Inner Hair Cells. Henson’s stripe is also adjacent to a flatter plane surface of the gel in contact with the Outer Hair Cells.

The curvature of the cochlea provides the frequency selectivity observed in the Cochlea. Marcatili modelled the dispersive waveguide in optical fibres: Fulton argues that a similar mechanism to the dispersive Marcatili filter is how Henson’s stripe performs the frequency

selectivity. As the frequency reaches a “critical” frequency, the wave at that narrowly defined frequency leaves the waveguide to travel across the nearby membrane and intersects the chevrons of Outer Hair Cells.

The Inner Hair Cells in contact with Henson’s stripe sense a low-pass filtered temporal loudness signal component, while the very narrow band signals from the dispersive filter feed the Outer Hair Cells.

The hair cells transducer the mechanical vibrations in the gel into an electrical signal, which is then amplified by biological “transistors” that Fulton calls “Activas” formed at the junction between cells. Fulton’s “Electrolytic Theory” of the neuron is different from the traditional biochemical model, and may be controversial in the traditional context.

Although this new concept could change our interpretation on how the human hearing mechanisms work, as a first approach don’t change significantly the techniques proposed in this thesis. Nevertheless, could give important guidelines to improve the performance of the techniques based on perceptual models.

2.5 Sound perception

The human perception of sound is a complex process that involves not only the human auditory system but also the brain that processes the neurological stimuli to produce an interpretation of what reaches the ear. It is thus relevant to consider the basic and most relevant phenomena that may set the boundaries to define the measurement procedures in auditoria with the presence of the public. The listener attends a live performance, speech or music, for the announced program and should not be disturbed by test signals. The design of the testing method should then integrate the sound perception mechanisms. Basic considerations on these are presented next.

2.5.1 Pitch

Pitch is the human perception of how high or low a tone sounds, based on its relative position on a scale. Musical pitch is defined in terms of notes; however, there are psychoacoustic experiments to measure human perception of relative pitch as well. Absolute pitch discrimination is rather rare, occurring in only 0.01 percent of the population [Rossing90]. Relative pitch discrimination can be measured by asking subjects to respond when one tone sounds twice as high as another. Like loudness experiments, the results are complex; they

depend primarily of the frequency and can also vary with intensity and waveform. For example, if a 100 Hz tone is reproduced at 60dB and then at 80dB, the louder sound will be perceived as having the lower pitch. This phenomenon is primarily one that occurs at frequencies below 300 Hz. At the mid frequencies (500 to 3000 Hz), pitch is relatively independent of intensity, whereas at frequencies above 4000 Hz, pitch increases with level.

The spacing of frequency resonances along the BM is not linear with frequency. The scale that relates the resonant frequency to position on the BM is called the Bark scale, or critical band scale [Scharf70]. In engineering terms, it approximates to a log scale. This matches our perception of pitch, following a logarithmic law. Musical instruments are also tuned logarithmically.

2.5.2 Loudness

Due to the way the ear transduces and interprets the sound, an amplitude dependence on sensitivity occurs. In fact, since the outer hair cells are distributed along the length of the BM, they react to feedback from the brainstem, altering their length to change the resonant properties of the BM.

The inner and outer hair cells interact together in an active feedback system, which increases the movement of the BM for quiet sounds, and suppresses it for loud ones. In effect, the hearing system behaves as a sophisticated Automatic Gain Control mechanism, AGC [Howard09]. This causes the frequency response of the BM to be amplitude dependent.

In terms of amplitude sensitivity, the human perception of loudness follows a logarithmic law. This should come as no surprise to audio engineers – they usually measure levels and responses in decibel (dB) which is a logarithmic scale $\{20*\log_{10}(\text{amplitude})\}$ – it is the AGC function of our ears which causes us to hear this way.

For the lowest frequencies (up to 200 Hz) and for the highest frequencies (above 15 kHz), the sensitivity is much lower than for the medium frequency bands (range of 500 Hz to 6 kHz). Some humps and bumps are reported in the contours above 1 kHz due to the resonances of the outer ear. This will have a first resonance at about 3.4 kHz and, due to its non-uniform shape, a second resonance at approximately 13 kHz. Therefore, the hearing system frequency response is a function of both amplitude and frequency.

The absolute minimum hearing threshold in quiet results from the facts that (i) the inner hair cells fire at random, even in the absence of any incoming sound, and (ii), in silence, the blood flowing around the regions of the inner ear becomes audible, as shown in Figure 2.15. This set of curves, called equal-loudness contour or "Fletcher-Munson" [Fletcher33] curves, is a measure of sound pressure level, over the frequency spectrum, for which a listener perceives a constant loudness when presented with pure steady tones. The unit of measurement for loudness levels is the phon. Each curve is designated by its phon value corresponding to the L_p at 1 kHz.

Nevertheless, listeners usually show different thresholds in quiet and therefore an average is taken as the threshold of hearing.

The human hearing system is able to indentify sounds of pressure levels on a dynamic range of about 140dB.

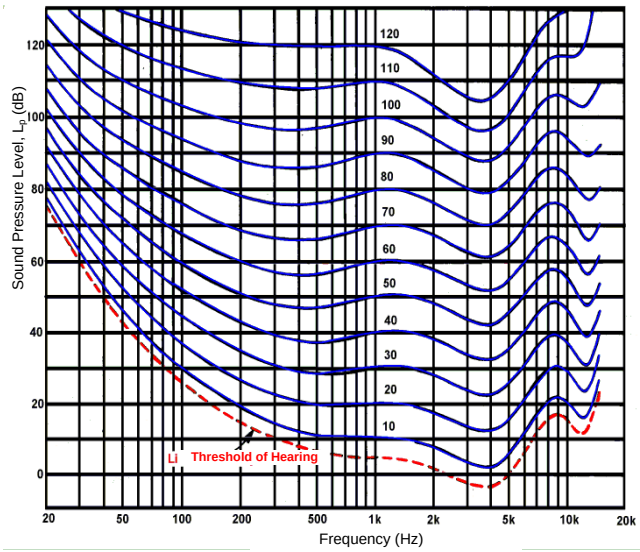


Figure 2.15 – Loudness curves; the red curve is the threshold of hearing in quiet.

The following practical expression is an approximation of the threshold in quiet at frequency f (in Hz) [Terhardt82],

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \text{ (dB)} \quad (2.60)$$

Although the loudness curves provided an accurate measure of the relative loudness of tones, their shape was too complicated for use with an analog sound level meter. To overcome this

problem, electrical weighting networks or filters were developed, which approximate the Fletcher Munson curves. These frequency weighting schemes were designated by letters of the alphabet, A, B, C, and so forth. The A, B, and C-weighting networks were designed to roughly follow the 40, 60, and 80 phon curves (turned upside down), respectively, as shown in Appendix A1. However, only the A and C curves are still in general use. The A-weighted level (dBA) is the most common single number measure of loudness. The C-weighting network is used mainly as a measure of the broadband sound pressure level.

2.5.3 Critical Bands

Auditory perception is based on a critical band analysis on the inner ear. A critical band is the bandwidth around a center frequency beyond which subjective responses of the hearing system abruptly change [Scharf70]. The importance of the critical bands comes from the fact that the hearing system discriminates between energy in and out of a critical band. Additionally, the simultaneous masking property of the hearing system, which will be discussed later, is related to the critical bands.

The physiological basis of the critical bands is not clear. However, the existence of the critical bands is certainly related to the function of the basilar membrane. Each point on the basilar membrane is tuned to a frequency called the characteristic frequency; that is the place at which the travelling wave caused by a stimulus reaches its maximum amplitude [Scharf70]. We can assume a non-ideal bandpass filter, which is referred to as an auditory filter, centered at each characteristic frequency [Moore96]. The effective bandwidth of the bandpass auditory filter is defined as the critical bandwidth. The length of each critical band corresponds to 1.3 mm or 1.5 mm on the basilar membrane [Hartmann97]. Moore [Moore96] defines a critical band as the Effective Rectangular Band (ERB) which is the bandwidth of an ideal bandpass filter centered at any frequency (the area under the squared-magnitude of the ideal filter equals that of the auditory filter centered at that frequency). According to this author, each ERB covers 0.9 mm on the basilar membrane. Note that there is no border between the critical bands and a band can be specified for any point on the basilar membrane.

The bandwidth of the critical bands was firstly measured by Fletcher [Fletcher40]. According to his experiment, in order to measure the bandwidth of a critical band centered at any frequency, a tonal signal is made inaudible by a narrowband noise centered at that frequency. If

the bandwidth of the noise is increased, the level of the inaudible sinusoid can be increased too. When the bandwidth of the noise exceeds a certain value, i.e., the critical bandwidth, the level of the sinusoid input remains almost constant.

This experiment is based on a few assumptions: (i) the hearing system contains a bank of overlapping bandpass linear filters, (ii) the listener is assumed to perceive only the output signal of the auditory filter centered at the tonal signal frequency, and (iii) the threshold of the signal is only determined by the power of the noise at the output of the auditory filter [Moore96]. The power of the noise at the output of the bandpass filter determines the threshold of the test tone. Since the bandpass filter is tuned to the frequency of the sinusoid, the tonal signal will be passed but a great part of the noise will be removed (when the noise bandwidth is greater than the critical bandwidth). The part of the noise which passes through the filter has a remarkable effect on making the tonal signal inaudible. In this experiment, increases in noise bandwidth result in higher noise power at the output of the bandpass filter as long as the noise bandwidth is less than the filter bandwidth. However, when the noise bandwidth exceeds the bandpass filter bandwidth, further increase in noise bandwidth will have a little effect on the output noise power. This bandwidth, at which the signal threshold stops increasing, is the definition of the critical bandwidth.

Different experiments are used to demonstrate the concept of critical band, as shown in Figure 2.16. First, the detection threshold for a narrowband noise signal between two masking tonal components remains constant as long as the frequency separation between the tones remains within a critical bandwidth. Beyond this bandwidth, the threshold rapidly decreases (Figure 2.16c). Second, a notched-noise experiment can be constructed with masker and maskee playing reversed roles (Figure 2.16b,c). Third, in this experiment the beating effect of two tones is used. For frequency differences of less than 12.5 Hz, approximately, a beating sensation occurs. When the frequency difference is increased above 15 Hz, the test signals sound “rough”. There is a frequency difference for which the two tonal components can be heard as separated (critical bandwidth).

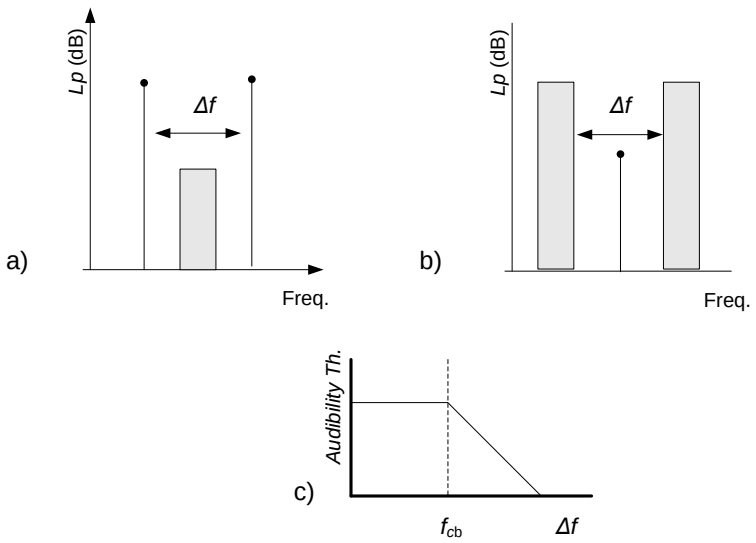


Figure 2.16 – Critical Band measurement Methods (from [Painter00]).

Experiments have shown that the width of the critical bands is relatively wider at low frequencies. In fact, the signal is processed in the inner ear on a nonlinear scale called the Bark scale (Bark is the unit of perceptual frequency and a critical band has a width of one Bark). Therefore, as shown in Figure 2.17, the peripheral section of the hearing system can be modeled as a non-uniform filterbank consisting of bandpass filters with bandwidths equal to critical bandwidths. About 75% of the critical bands are below 5 kHz and hence the hearing system receives more information from low frequencies. The critical bandwidth is almost constant from 100 Hz up to 500 Hz, approximately. Above this upper frequency, increases of about 20% of the center frequency are observed [Zwicker91].

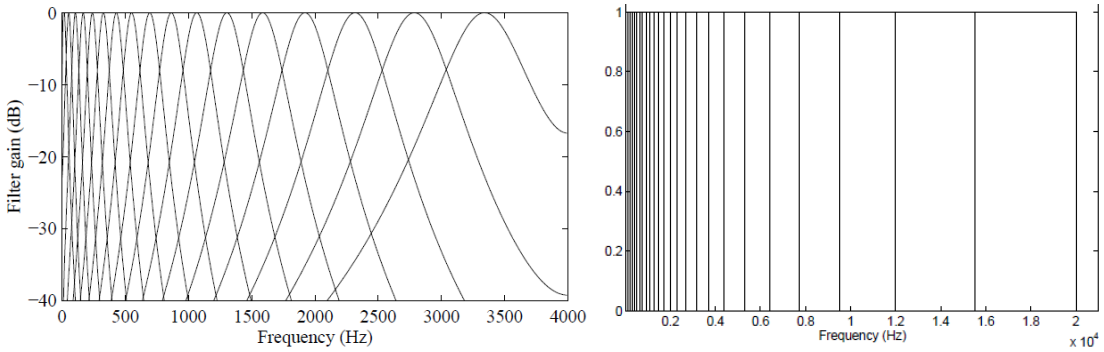


Figure 2.17 – Example of the non-uniform auditory filterbank (approximated to the auditory filters, left side, and rectangular, right side); notice the different range of frequency axis.

There is a relation between the distance along the basilar membrane and its frequency resolution. Considering the fact that a length of 1.5 mm on the basilar membrane represents approximately 1 Bark, near the top (far from the oval window) a length of 0.1 mm represents a frequency difference of about 7 Hz whereas near the base 0.1 mm represents 450 Hz.

Many analytical expressions have been proposed in the literature to relate the critical band number z (in Bark) to frequency f (in Hz). Zwicker [Zwicker99] proposes the following

$$z(f) = 13 \operatorname{atan}\left(\frac{f}{1315.5}\right) + 3.5 \operatorname{atan}\left[\left(\frac{f}{7500}\right)^2\right]. \tag{2.61}$$

The bandwidth of each critical band as a function of center frequency can be approximated by

$$BW_c(f) = 25 + 75\left[1 + 1.4\left(\frac{f}{1000}\right)^2\right]^{0.69} \tag{2.62}$$

Glasberg and Moore [Moore97] proposed the following relation number of

$$ERB = 24.7 + \left(\frac{f}{228.8} + 1\right). \tag{2.63}$$

An example set of critical bands is listed in Appendix B1.

Another model used to represent the critical bands is based on the Mel scale. This approach gained popularity on the ASR, Automatic Speech Recognition, community with the computation of the MFCC, Mel Frequency Cepstral Coefficient (this issue will be treated in Chapter 3). Nevertheless, the Bark scale is a better approximation to model the human auditory system.

The *Mel* frequency (in Mel) scale is related with the linear frequency, f (in Hz), by

- $Mel(f) = f$, for $f<700$ Hz;
- $Mel(f) = 2595 \log_{10}(1+f/700)$, for $f>700$ Hz.

2.5.4 Auditory Masking

The mechanism of auditory masking and the human hearing system in general is not well understood. Nevertheless, we experience masking effects in everyday life. The American Standards Association (ASA) defines masking as the process or the amount (customarily measured in decibels) "by which the threshold of audibility is raised by the presence of another (masking) sound". Or, in other words, masking is a characteristic of the hearing system by which a weaker audio signal becomes inaudible by a louder signal occurring simultaneously or close in time [Moore97]. Masking can be thought of as the filter's effectiveness in analyzing a component at its center frequency (maskee) being reduced to some degree by the presence of another component (masker) whose frequency falls within the filter's response curve. The degree to which the filter's effectiveness is reduced is usually measured as a shift in hearing threshold, or "masking level. In daily life we observe many examples of the masking effects. For example during a conversation in a very noisy environment, one needs to raise the voice as an attempt to be understood.

The masking phenomena reflect the limited frequency and temporal resolutions of the hearing system. Although there are physiological explanations for the masking phenomena of the hearing system, it is almost impossible to develop comprehensive theories only based on physiological data. In order to obtain enough reliability data, one needs to have access to the inner parts of the human hearing system such as the cochlea, basilar membrane, nerve fibers - the brain without causing damage. This constraint is overcome by psychoacoustic data collected through subjective tests.

There are different masking effects related to different mechanisms. Simultaneous masking occurs when the masker and the maskee (masked signal) are presented to the hearing system at the same time. In addition to simultaneous masking, time domain masking phenomenon, referred to as temporal masking, occurs when the masker and the maskee are presented close in time, but not simultaneously. The phenomenon of masking a signal occurring before the beginning of the masker is called pre-masking or backward masking. Another form of the temporal masking, which is referred to as post-masking or forward masking, happens if the masked signal occurs after the end of the masker.

2.5.4.1 Simultaneous Masking

Simultaneous masking occurs when two stimuli are simultaneously presented to the auditory system and one of them is made inaudible by the other. Physiological evidence reveals that the simultaneous masking is due to the operation of the basilar membrane and the hair cells.

There are two theories for the mechanism of simultaneous masking [Scharf70]. One theory suggests that the masker produces a great amount of activity on the basilar membrane such that any activity caused by a weaker signal may become undetectable. In fact, the hair cells detect the strongest vibration in any local region (critical band) along the basilar membrane. The simultaneous masking pattern is well predicted by this theory which models the hearing system as a bank of linear filters. The second theory, which is highly nonlinear, suggests that the masker suppresses the activity which the masked signal would produce if there is no masker. Based on this theory, the neural response to a tone at the characteristic frequency of that neuron might be suppressed by a tone which does not produce any neural activity in that neuron. Many researchers believe that the first theory plays the dominant role in the simultaneous masking mechanism.

Although the physiologically-based theories mentioned above explain the simultaneous masking phenomenon, the analytical masking models are developed, as mentioned above, using psychoacoustic data. Briefly, some psychoacoustic findings about the simultaneous masking properties of hearing are presented. Figure 2.18 illustrates this principle by helping in understanding the findings of experiments associated with masking. Tones that are close in frequency mask each other more than those that are widely separated.

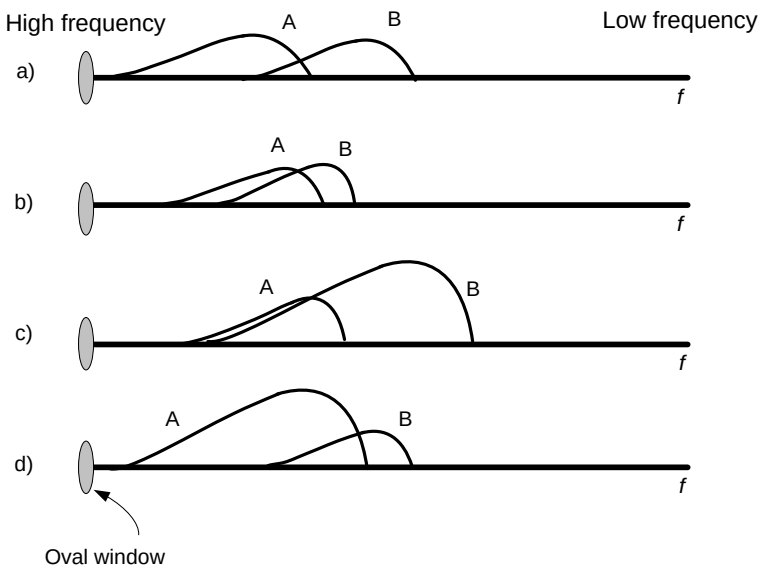


Figure 2.18 – Overlap of Regions of the Cochlea (Rossing, 1990); simplified response of the basilar membrane for two pure tones A and B: (a) the excitations barely overlap; little masking is noticed, (b) there is an appreciable overlap; tone B mask tone A more intensively, (c) the more intense tone B almost completely masks the higher-frequency zone of tone A, (d) the more intense tone A does not completely masks the lower-frequency zone of tone B.

To determine the masking pattern (curve) of a simple stimulus, the test signal (maskee) is varied keeping the masker fixed. The masking threshold at any frequency is the level of the test signal when it is just inaudible. Figure 2.19 shows the approximate masking curves due to narrowband noises centered at 200, 400, 800, 1200, 2400 and 3500 Hz, for different levels. As it is observed, the maximum of the masking curves depends of the center frequency. The peak of the masking curve is 2–6 dB below the excitation level [Zwicker99].

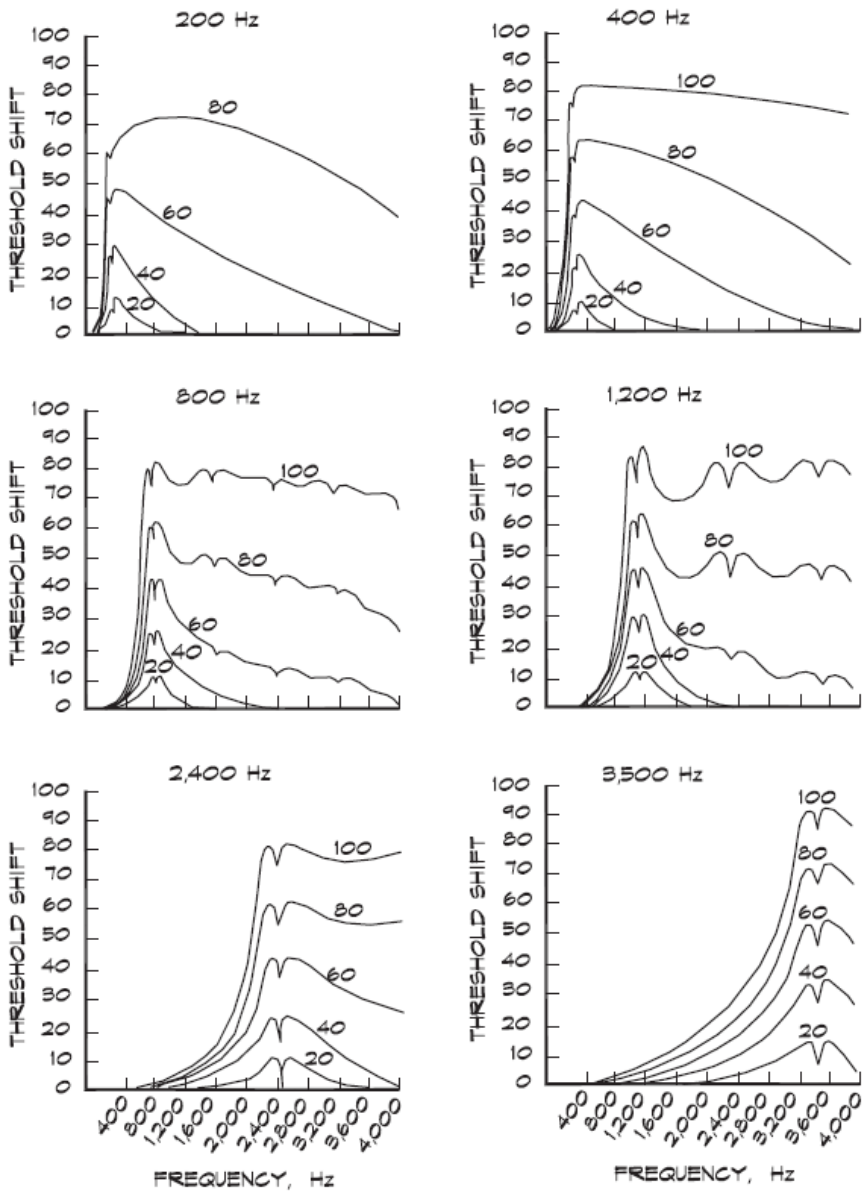


Figure 2.19 – Pure Tone Masking Curves (from [Howard09]).

Based on psychoacoustic experiments, although the lower slope of the masking curve is almost independent of the masker level, the upper slope (towards the higher frequencies) depends of the level of the masker. As Figure 2.19 shows, the higher the excitation level the lower the upper slope is. This nonlinear behavior of the hearing system is attributed to the saturation of the outer hair cells in the cochlea at high levels [Scharf70].

Note that the nature of the masker as being noise-like or tonal-like has an impact on the masking curve. For instance, the maximum of the masking curve due to a single tone is sharper

(peaky) [Zwicker99]. Additionally, the distance between the masker level and the masking threshold is greater for tonal signals.

2.5.4.2 *Temporal Masking*

Temporal masking occurs when the masker and the maskee are not presented to the hearing system at the same time. The temporal masking characteristic of the hearing system is asymmetric, meaning that the backward masking effect is much less than the forward masking. Backward masking is effective about 5 ms before the occurrence of a strong stimulus, whereas forward masking lasts up to 200 ms [Zwicker99]. This is demonstrated experimentally by using a burst of white noise followed by a short tone. Usually, this tone would be audible above a level of 20dB.

An effect of the postmasking is less audibility of low energy consonants following a high energy vowel. Figure 2.20 shows the temporal masking pattern due to a short burst of a tonal signal [Zwicker99].

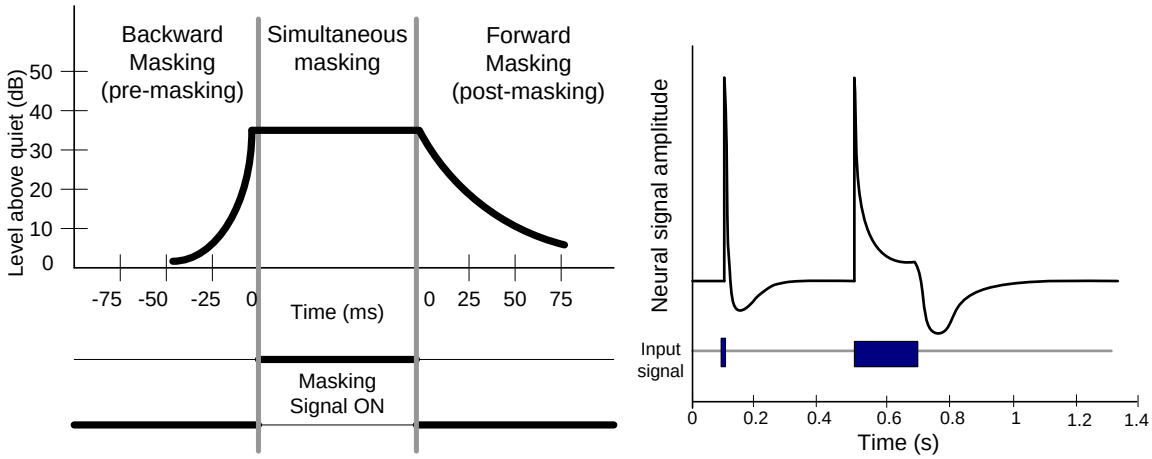


Figure 2.20 – Temporal masking pattern (dashed curve, left figure) due to a short burst of a tonal signal starting at 100 ms and ending at 300 ms; overall response of a group of inner hair cells to a short and long tone burst of 1kHz (5ms then 200ms).

Three important features are apparent. Firstly, the hair cells are most sensitive to the onset of sounds. Secondly, the hair cells take a finite time to recover their full sensitivity after the end of a sound. Thirdly, the magnitude of this recovery depends of the duration of the sound. This gives rise to a process known as temporal masking. If a sound (of a similar frequency) occurs

during the period of recovery, the hair cells may be unable to register it, hence it may be inaudible.

Although psychoacoustic experiments reveal the temporal masking effects, this phenomenon is not well understood. Temporal masking effects suggest that the brain might integrate sound over a short time interval or perhaps the brain simply processes loud sounds faster than soft sounds [Zwicker99]. Moore suggests that the following different phenomena contribute to the forward masking effects which occur after the end of a masker [Scharf70]:

• Temporal overlap of the basilar membrane responses to different stimuli might play a role in the temporal masking. This phenomenon contributes to temporal masking for about 10 ms after the end of the masker.

• Short term neural fatigue at higher neural levels might reduce the perception of the activity of the maskee which occurs after the masker.

• The neural activity as a response to the masker persists at higher levels after the end of the masker. This activity masks the activity produced by the maskee. This effect is also suggested to occur at stages higher than the auditory nerve.

2.5.4.3 Combined Masking Threshold

Both temporal and simultaneous masking effects are used to compute the combined masking threshold. Some research has been done to find how to combine these two different phenomena. A simple way to deal with this problem is to linearly sum the masking levels. According to some experiments this model is not appropriate; and therefore another model called power-law has been proposed in the literature [Soulodre98, Beaton96, Veldhuis92] as follows

$$m_{net} = (m_{p1} + m_{p2})(1/p), \tag{2.64}$$

where m_{net} is the net masking level due to two masking levels m_{p1} and m_{p2}. The value of 0.3 for parameter p is found to be the best match to experimental data [Soulodre98].

Chapter 3

SOUND MEASUREMENT AND ANALYSIS

3.1 Measurements techniques

Sound field measurements evaluate the characteristics of the sound signal in the major domains: energy, frequency content, and time history. Sound levels, both overall and in frequency bands, statistical indexes or other are determined.

In enclosures, the sound fields are complex due to the interaction of the different reflections, the acoustical behavior of the boundary surfaces, and the different contributing sound sources both inside and outside. A number of acoustical parameters are usually evaluated to characterize the sound field inside the room and the acoustical properties of the enclosure.

The acoustic parameters can be estimated from the room impulse response, IR, (or from its transfer function which is the Fourier transform of the impulse response) measured in the room considered as a linear and time-invariant system, LTI. Therefore, the experimental determination of the room impulse response is a fundamental task in room acoustics.

3.1.1 General

The system under test, either enclosure or open air facility for music performances or electroacoustic system used for sound reinforcement, can often be considered to be linear and time-invariant. However, in practice, this assumption is not fulfilled and non-linearities in the

measurement chain, time-variance phenomena and insufficient SNR are constraints that must be dealt with [Ciric02].

The *Crest factor*, *CF*, can be defined as a measure of the energy effectively applied to the system under test estimated by the ratio of the peak value of the signal to its total energy, in dB, as shown in

$$\mathrm{CF} = 10\log_{10}\frac{\max\left|s(t)\right|}{\frac{1}{T}\int_{T} s^2(t)\,dt} \tag{3.1}$$

where T is the period of the signal.

The maximum peak value allowed in each point of the measurement chain should not be exceeded in order not to increase the distortion produced by the saturation of the electronic and electroacoustic devices making the measurement chain. Therefore, the test signal should be normalized to the "weakest" element of the chain as an attempt to inject the maximum possible energy to the system avoiding non-linear distortion. This parameter is important for comparing the performance of different measurement techniques as long as the more energy is applied to the system the higher the SNR is expected to be observed at the reception. Therefore, the *CF* should have a low value, ideally close to zero.

The measurement techniques usually employed in architectural acoustics are more or less suitable to particular situations showing advantages and deficiencies due to their intrinsic characteristics such as immunity to time variance phenomena and non-linearities and issues of nuisance to the in-occupants (when the measurements are carried out with the audience in the room). Therefore, the criteria for selecting the measurement apparatus and respective method of estimation the IR, depends on the purpose of the use of the acoustic system [Stan02]. Hence, it is useful to establish comparison terms emphasizing advantages and deficiencies among the different types of the common acoustical measurement approaches:

- *Impulsive sound source* – follows the basic concept of the impulse response (pistol shots, balloon pops).
  - Advantages: simplicity; mains power is not necessary; low sensitivity to time-variance.

  - Deficiencies: low immunity against distortion (non-linearities); low noise rejection; low skills of repeatability; high crest factor.

- *Time-Delay Spectrometry, TDS* –
  - Advantages: reasonable SNR (low crest factor); low sensitivity to time-variance; robustness to harmonic distortion.
  - Deficiencies: weak energy and ripple at the low frequency bands.
- *Stepped Sin*e
  - Advantages: high SNR (low crest factor);
  - Deficiencies: time consuming; low noise rejection; excitation of eigenmodes of the room; only the magnitude and phase of the frequency response to each sampled frequency are estimated.
- *Linear Sine Sweep*
  - Advantages: high SNR (low crest factor); good noise rejection for mid/high frequency bands.
  - Deficiencies: time consuming; excitation of eigenmodes of the room; poor rejection against distortion; low noise rejection for low frequency bands (constant SNR across all frequencies).

These measurement techniques are not intended to be used in live situations with an audience, at least without additional procedures to mask the signals conveniently. Instead, the following approach is an example of the use of music or speech signals as test signals in order to achieve total or quasi-total test signal measurement transparency for the audience. In fact, since the Dual-Channel FFT Analysis assumes the system under test (the room) is a LTI system, according to the systems theory, the transfer function of the system is merely the ratio of the spectrum of the system output to the input spectrum as shown by

$$y(n) = x(n) * h(n) \longleftrightarrow Y(e^{j\omega}) = H(e^{j\omega}).X(e^{j\omega}), \tag{3.2}$$

$$H(e^{j\omega}) = \frac{Y(e^{j\omega})}{X(e^{j\omega})} = \frac{H(e^{j\omega}).X(e^{j\omega})}{X(e^{j\omega})} . \tag{3.3}$$

Therefore, assuming the input signal has sufficient energy for all spectrum range of interest, for a period of time, the impulse response and the frequency response are estimated.

- Advantages: possibility of measuring the magnitude and phase of the frequency response using the musical program and in real time (only for the case of *Real Time Deconvolution* method, *RTD*, approach);

- Deficiencies: deviations in the magnitude response; low noise rejection; low immunity against distortion; results are not shown in real time due to the time need in the averaging methods used to improve the SNR and the impulse responses lengths limited (excepting for the RTD method). Accurate results are only obtained for short room reverberation times.

Examples of commercial implementations based on this approach are – SIM® Audio Analyzer trademark of Meyer Sound, SIA SmaartLive® trademark of EAW and Real Time Deconvolution, RTD, which uses analysis windows with different lengths, namely Time-Frequency-Constant Window, TFC, increasing the time-frequency resolution and therefore, the accuracy. The time length of the TFC window changes as a function of frequency. The window duration is inversely proportional to frequency so that multiplying the window duration at a given frequency by that frequency yields a constant value. It allows for the calculation and display of impulse response (IR) data of up to 10 seconds in real-time and with fast on-screen display refresh rates. However, some deficiencies occur in case of the synchronization of several sound sources in live performances, such as in sound reinforcement systems using several sound delay channels to cover a larger audience area with less variability of the sound pressure levels.

The weaknesses of the methods mentioned before do not make them best choices for room acoustics measurements in many situations. Maximum Length Sequences (MLS) and logarithmic swept sine techniques have proved to be more suitable to estimate the impulse response accurately, as described in detail in the next sections.

3.1.2 Maximum Length Sequences technique, MLS

Pseudo-random sequences are used due to their broadband spectral characteristics and the autocorrelation function corresponds to an impulse [Scharf70, Hellman72]. In fact, a binary periodic sequence $x(n)$, with values of ± 1, are designated by pseudo-random if the following requisites are verified [Davies66, MacWilliams76]:

- The mean value or the DC component, $\bar{x}(n) = 1/L \sum_{n=0}^{L-1} x(n)$, equals $1/L$, where L is the length of the sequence. In short, the excess of symbols “+1” to “-1” is unity.

- Half of the number of consecutive identical symbols has length 1, a quarter has length 2, an eighth has length 3, and this rule is repeated until the number of consecutive symbols reaches the order of the sequence.

The normalized auto-correlation function,

$$\phi_{ss}(n) = \frac{1}{L}\sum_{k=0}^{L-1} x(n)x(n+k) \tag{3.4}$$

shows two values only $\begin{vmatrix} 1 & n=0 \\ -1/L & 0<n\le L \end{vmatrix}$

Therefore, maximum length sequences are characterized by having a unit impulse for the auto-correlation function, except for a residual DC component, which is insignificant for most real situations, when the sequence length is large [Rife89]. The amplitude of the impulse response for the time n=0 is equal to the length of the sequence, L.

As a consequence of the periodicity, this type of signal shows a discrete spectrum. The power spectral density is flat (excepting for the DC component) with random phase.

The acoustic measurements are performed by exciting the system under test by a MLS frame, $S_{MLS}(t)$, and capturing the system response, $y(t)$, simultaneously. The filtered signal is correlated with the original sequence as follows:

$$\begin{aligned}\phi_{SS'}(t) &= [S_{MLS}(t) * h(t)] * S_{MLS}(-t) \\ &= [S_{MLS}(t) * S_{MLS}(-t)] * h(t) = \int_{-\infty}^{+\infty} h(\tau)\,\phi_{SS}(t-\tau)\,d\tau \approx \delta(t) * h(t) \approx h(t),\end{aligned} \tag{3.5}$$

where $\phi_{SS'}$ stands for the cross-correlation function, ϕ_{SS} stands for the auto-correlation function, and * represents the circular convolution operation. As long as the MLS is a periodic signal, the convolution operations are replaced by the circular convolution.

The main difference between the spectra of the unit impulse and of MLS signals lays in the phase characteristics. The amplitude frequency responses are identical with very different crest factors ($\approx L$ for the pure impulse and 0dB for the MLS).

An interesting finding is that the result of the correlation between input and output signals is independent of the MLS test frame. Hence, the RIR is simply determined by the cross-correlation between the MLS frame and the filtered MLS frame.

A simplified block diagram is illustrated in Figure 3.1 for a typical RIR measurement setup.

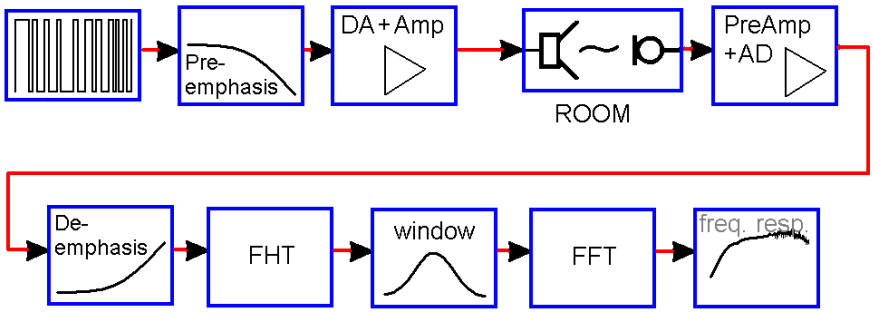


Figure 3.1 – Scheme of measurement setup using the MLS frame for IR estimation.

This technique has the ability of spreading the spectral energy uniformly over the entire length of the sequence. That is the reason why the MLS technique shows good SNR characteristics.

The pre-emphasis and de-emphasis modules are optional. They are used mostly in situations of background noise with non-flat spectral characteristics. Often, in architectural acoustics where machinery, such as the air conditioning and renovating air systems, share the same space or are inside the system under test, the low frequency bands are severely affected by noise. Therefore, the pre-emphasis filter is used to boost the part of the spectrum more affected by noise. At the reception point, the inverse filter, de-emphasis filter, is applied for compensation. A number of advantages and weaknesses of the technique can be identified:

- Advantages: The type of test sound is not unpleasant (it is similar to some nature sounds, i.e. sea waves and water cascades). Possibility of usage in the presence of people; high SNR (low crest factor, 0-11dB, depending of the filter used). The SNR can be increased by augmenting the sequence length.
- Weakness: very sensitive to time-variance; poor robustness to harmonic distortion (the non-linearities issues can be attenuated by applying windowing techniques to the sequence previously enlarged, since the tail of the IR is more affected. It is assumed insignificant time variance phenomenon); time consuming on setting up the measurement chain (adjusting the levels of each stage) to minimize the effects of the non-linearities produced by the electroacoustic equipments used on the measurements. The use of dedicated

measurement methods (such as the Inverse Repeated Sequence technique [Dunn93, Ream70]).

3.1.3 Logaritmic swept sine technique

The swept sine technique, using exponential chirps, consists of exciting the system under test by a logarithmic sinusoidal sweep with continuous varying frequency, $x(t)$, explored in detail by Farina [Farina00], expressed by:

$$x(t) = \sin\left\{\frac{T\omega_1}{\ln(\omega_2/\omega_1)}[e^{(t/T)\ln(\omega_2/\omega_1)} - 1]\right\}, \tag{3.6}$$

where T is the time duration of the sine sweep $x(t)$, ω_1 is the initial angular frequency, and ω_2 is the final angular frequency.

The RIR is obtained by performing the linear convolution between the captured signal from the system (by the microphone) and a filtered version of the original swept sine, the inverse filter. The test signal, swept sine, is filtered to compensate for the non-flat spectral characteristic. In fact, the magnitude frequency response shows a fall of 3dB/octave, the same as the pink noise. The inverse filter $f(t)$, built from the swept sine signal, is designed to satisfy the relation:

$$x(t) * f(t) = \delta(t - K_d), \tag{3.7}$$

where * denotes the linear convolution operation and K_d is the arbitrary delay.

Exponential chirps spend equal time for each octave band providing an SNR gain inversely proportional to the frequency.

The spectrum of the swept sine, $S(\omega) = |S(\omega)|e^{j\phi(\omega)}$, is approximated by

$$|S(\omega)| \approx \frac{1}{\sqrt{\frac{1}{2}\left|\frac{d\gamma(\omega)}{d\omega}\right|}} \tag{3.8}$$

and the phase is given by

$$\phi(\omega) = -\int_0^{\omega} \gamma(v)dv \tag{3.9}$$

where the group delay, $\gamma(v)$, is the functional inverse of $\omega(t)$ [Muller01, Stan02].

The compact representation of the magnitude in Equation (3.8), by considering the signal energy in the frequency domain between two nearby frequencies [$\omega-$, $\omega+$] is approximated by the signal energy in the time domain over the corresponding time interval [$\gamma(\omega-)$, $\gamma(\omega+)$]. In essence, the sinusoidal sweep power at any given frequency is proportional to the "time spent" on that frequency - the inverse frequency trajectory derivative. The original derivation and more complete explanation are found in [Stan02]. Figure 3.2 shows a simplified block diagram describing this technique.

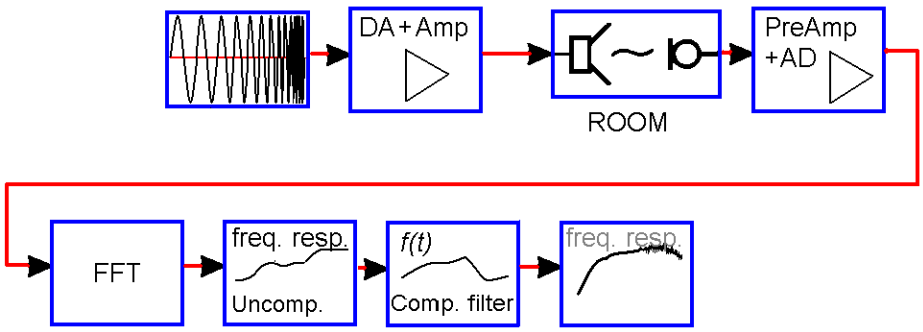


Figure 3.2 – Basic block diagram showing the processing modules of the swept sine technique.

One possible procedure for designing the inverse filter $f(t)$ can be:

1 – The test signal is time reversed and delayed to assure causality, $x_1(t)$. This time reversal causes an inversion in the phase spectrum. Therefore, the resultant signal from the convolution $x_2(t) = x(t)*x_1(t)$, shows linear phase response, corresponding to a pure delay only; however, its spectrum is squared.

2 – The magnitude of the spectrum of $x_2(t)$ is divided by the squared magnitude of the spectrum of the swept sine $x(t)$ to find $f(t)$.

The use of the linear convolution avoids problems of temporal aliasing. Nevertheless, one should be aware of the transient parts generated at the beginning and at the end of the sweep.

The non-linearities are easily removed from the RIR by correctly choosing the measurement time: its effect appears instants before the impulse response and can be cut out. The main advantages and weaknesses of the technique are:

- Advantages: high SNR (most of the energy is concentrated at a simple frequency, which has an associated low crest factor, 3dB. Although the swept sine is a kind of frequency modulated signal, this is a valid approach since the frequency varying rate is relatively low). The SNR can be increased by enlarging the sequence; Good robustness to harmonic distortion (the effects regarding the non-linearities can be removed by using sequences larger than the necessary); low sensitivity against time-variance; Magnitude frequency response shows a fall of 3dB/octave. This fact permits increasing the energy for the low frequency bands and has the ability to protect the high frequency transducers against overheating. Usually, the use of pre-emphasis techniques is not necessary.
- Weakness: excitation of the eigenmodes of the room. The narrowband test signals are susceptible of exciting the eigenmodes of the system, essentially for the low frequencies, originating some fluctuation in the system response. The reverberation time is incorrect for some situations. Possibility of saturation of the reception equipment; It is not a pleasant sound to be used in the presence of people; however, as it will be explained in Chapter 5, it is suitable for use in the perceptual models approach.

Table 3.1 shows examples of *CF* values for the MLS and swept for different types of filters.

Table 3.1 – Crest factor for the MLS and swept signals with different filters applied.

Filtering type	Crest factor (dB)	
	MLS	swept sine
Broadband (no filter applied)	0.0	3.0
Broadband (output of electroacoustic system)	~10.0	~4.1
Equal energy bands (pink noise)	11.3	~3.0
A Weighting	12.0	7.2
Octave bandpass	14.5	14.5

The pros and cons balance makes the swept sine-based signals currently the preferred excitation for electroacoustic and room acoustic measurements.

3.2 Distortion

3.2.1 Non-linear distortion

The main importance of the linearity property in systems theory lies on the use of the Fourier transform, and thus, on the correspondence between the IR (transient response) and the frequency response (steady state). For example, the harmonic distortion due to loudspeaker non-linearities is described by [Farina00]:

$$w(t) = s(t) * k_1(t) + s^2(t) * k_2(t) + s^3(t) * k_3(t) + \ldots + s^N(t) * k_N(t), \tag{3.10}$$

where $k_i(t)$ is the i_{th} Volterra Kernel component.

In practice, some measurement techniques are able to nearly separate the linear (reverberation component) and the non-linear (distortions – harmonic, inter-modulation etc.) components. However, it is recommended to limit or even cancel the non-linearities at the origin by using a pre-distortion filter at the input of each acoustic transducer.

3.2.2 Time variance

Time invariance simply means that the environment parameters are kept constant over the duration of the measurement.

The time variance phenomenon occurs essentially in three distinct situations during the acoustic measurements: (i) in open space, for long distances due to air mass movements (stadiums) or inside enclosures due to movements of people and extraction/injection and convection of air by the air conditioning systems, (ii) during long term measurements in enclosures due to temperature variations in between measurements or to temperature gradients over the test system volume, and (iii) by excessive heating of the transducer moving voice coils and accessories of sound transduction, clock jitter and sampling frequency drift on the AD/DA converter devices.

The problem of time variance has been addressed by a number of authors. Significant deviations of the sound pressure levels and the parameters estimated from the RIR, such as the reverberation times and the energy ratios, were reported when the MLS method is applied in buildings and or in free field, or in the presence of air conditioning systems, rotating microphones and temperature changes [Chu95, Bradley96, Vorlander97a] and in occupied halls due to air convection, people movement, and temperature gradients [Griesinger96]. Vorlander [Vorlander97a] also discussed the concepts of inter-periodic and intra-periodic time variance. The inter-periodic term refers to time variance effects occurring between different test signal frames, even when the system under test is considered to be time invariant over each measuring period. Examples are sampling frequency drifts of the measuring equipment or small amounts of temperature changes since this variation is considered neglected during the time period of each test frame. The intra-periodic term stands for the time variance effects over the measuring period of each test frame, such as the jitter in the measuring hardware, wind and air convection effects.

A detailed research of inter-periodic and intra-periodic effects produced by the time variance for MLS based tests was addressed in [Svensson99]. A model based on the time stretching of signals, resampling each frame with a slight change in the sampling frequency, was proposed for the inter-periodic effects, and delay modulation was used to deal with the intra-periodic time variance. It was found that inter-periodic and intra-periodic time variance effects cause similar errors in MLS results, showing an apparent level loss and a high-frequency slope of +12dB/octave, in room acoustics or electroacoustics measurements.

A simple procedure to detect the presence of time variance phenomena in the IR, independent of the type of the test signal was suggested in [Satoh07]. In this procedure, the detection of time variance is based on the short-term running cross-correlation function between each impulse response before synchronously averaging the IR. It is calculated for 8 kHz octave band (or to high frequency bands) to estimate the so-called time-lag parameter, which is the difference between the maxima of the cross-correlation function. This parameter provides a measure of how low (short time-lag) or how strong (large time-lag) the time variance is affecting the IR. The relative error between single frame and averaged frames is also computed.

Occasionally, IR measurements in large or in constrained architectural dimensions makes use of separate playback and recording devices. Therefore, problems of synchronism due to inexact digital clocks (clock drift) might arise [Vanderkooy93]. This problem can be seen as a kind of inter-periodic time variance, as addressed above, since the difference in clock speed is small and constant. The accumulated errors over time become problematic for longer duration recordings used for robust convolution-based IR measurement methods such as MLS and swept sine. Bryan [Bryan10] shows that the unwanted clock drift imposes allpass filtering on the results and proposes methods of compensation based on resampling and compensation filtering. For proper compensation, accurate clock drift rate is made via implicit or explicit estimation.

These issues are critical when synchronous time averaging is used as an attempt to increase the SNR.

3.2.3 Practical considerations

Measuring chain

The overall SNR is estimated by the difference between the maximum value of the energy decay curve and the background noise level. The factors affecting directly this parameter are the SNR inside the room, the excessive distortion of the electroacoustic devices, the length of the test signal or the number of the averages [Nielsen97]. The curves shown in Figure 3.3 illustrate these issues. This figure shows SNR values, or energy decay range, versus output test signal levels for different background noise levels, for MLS signals of order k=16 and 10 averages. The maximum SNR values are bounded by the situation of absence of noise (red curve).

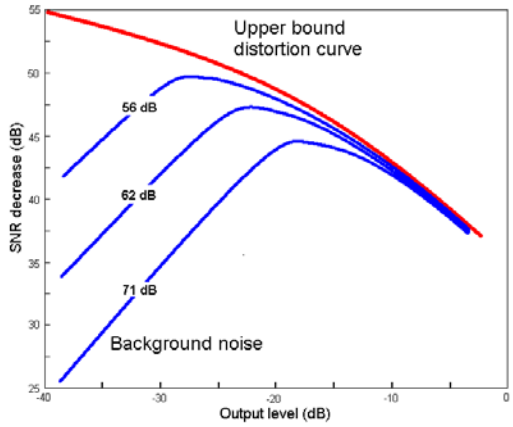


Figure 3.3 – SNR values, or energy decay range, versus output test signal levels for different background noise levels, for MLS signals of order k=16 and 10 averages.

As the figure shows, the admissible values of SNR are those in the area located below the Distortion Limit curve. This curve can be built for each acoustic measurement setup and decays by increasing the excitation sound levels due to the increase of distortion produced, essentially, by the loudspeakers. Accounting for the background noise, the remaining curves show an optimal value for the test signal sound level which maximizes the SNR. Therefore, the optimal value acts as a boundary for the SNR in the room: below, it is affected by the background noise, whereas above it will suffer from the distortion produced by the loudspeakers.

In the case of time invariant systems, the increase of the SNR is achieved by increasing the length of the test signal or by applying averaging techniques.

Synchronization

The synchronism between the input and output signals is an important issue to assure the best results in terms of SNR and frequency response, assuming that the system under test is LTI . In fact, the total delay between input/output has different causes due essentially to (i) the delay between playing and recording on the sound boards (due to internal buffering issues techniques), and to (ii) the propagation delay of sound waves. The internal latency inherent to transferring data from and to the central processing module is not an issue since real time requirements are not needed, in the sense that the results of the analysis haven't to be shown immediately. Different types of degradation occur for the MLS and swept sine signals when the captured frame, which corresponds to the system response, is truncated, removing the beginning or the last part, due to initial synchronism problems. The first part of the captured frame is removed when the window of length *L*, used to select the system response, starts after the first sample of the captured frame, the first samples are lost. The last part of the frame is removed when the window of length *L* begins before the first sample of the captured frame, implying the last samples are lost.

For the case of the MLS signal, as long as the energy is equally distributed along the spectrum, it is expected that the effect of truncation will decrease the SNR uniformly for the whole spectrum. In fact, a reduction equal to the time delay, the number of the truncated samples, is verified in the IR amplitude. If this delay is assumed to be much lower than the sequence length, the results are not much affected. For the swept sine tests, since the instantaneous frequency evolve over time, only the first or the last parts of the sweep (the first or the last frequency bands) is affected. This situation is avoided by making the sweep range larger than necessary (the starting and ending frequencies are extended). This procedure has also the advantage of reducing the distortion produced on the edges of the frequency range set to the sweep due to frequency leakage effects on the impulse response processing.

Considering periodic test signals, the length of the frames, *L*, is just the necessary to avoid time aliasing problems. The first captured frame does not contain all the information of the system response (the initial part did not reach the steady state and the tail of the captured signal, above *L*, is truncated, decreasing the SNR for the MLS signals and erasing the high frequency

information for the swept sine signals). Therefore, an initial frame is usually inserted in each measurement, with no time delay in between the ensemble frames. At reception, this initial frame is discarded. This procedure can be overcome if the length of the frames is several times the reverberation time.

As described above, a number of recommendations can be applied to improve the accuracy of the acoustic measurement results, increasing the SNR, such as correctly calibrating the measurement chain, using pre-emphasis techniques, synchronously time averaging and correctly choosing the measurement technique for each particular case. However, in situations of very low SNR, these procedures prove to be insufficient, thus imposing the use a large number of test signal frames. As an example, consider that an increase of 20dB is necessary. Then, 100 frames are needed, corresponding to a measurement time of about 3.5 minutes, assuming each frame has the duration of 2 seconds. However, if an additional increment of the SNR is required, the measurement time duration could become prohibitive and time variance problems might arise.

3.3 Optimization of the decay curves

The decay curves are a measure of how the acoustical energy fades with time in an enclosure after switching off the sound source. Considering a diffuse acoustical field model, the energy decay follows an exponential law. Although the energy decay in a decibel scale is a linear curve, in practice, this statement is not completely fulfilled due to (i) the natural room modes (essentially for low frequencies) – they may cause the decay curve to warble, especially at low frequencies due to the low mode density, (ii) the background noise – it limits the decay curve at some point, reducing the dynamic range, and (iii) the non-diffuse sound fields - the decay curve may bend resulting in a decay curve contributed of multiple energy decay rates, in case of an uneven distribution of the sound absorption of the room boundary surfaces. These constraints occur in almost all acoustic measurement situations. Expedient methods for use in automatic acoustical analysis procedures, in order to optimize the decay range, are described next.

Decay curve / background noise crosspoint estimation

Consider the case of the exponential decay corrupted by additive noise. The noisy squared impulse response is modeled by

$$h_N^2(t) = E_N + E_0 e^{-6\ln(10)t/RT} \tag{3.11}$$

where E_N is the noise energy, E_0 is the IR energy at the instant t=0 and RT is the reverberation time. The term -6ln(10)/RT is the necessary damping constant to achieve the correct RT.

The decay curve analysis for different values of the noise energy, computed by the backward integration (Schroeder integral), shows that for high background noise levels, the linear decay range could be small, making the correct RT estimation difficult or even impossible (the corresponding decay curve could never reach the true slope). Therefore, the optimal time instant to make the truncation is a vital issue when using systematic tools.

An intuitive approach consists of applying least squares line fitting techniques to estimate the slope of the curve, optimal decay rate, as given by

$$\min_k \int_{t1}^{t2} \left\{ h_e(t) - \left[k(t) + a \right] \right\}^2 dt \tag{3.12}$$

where $h_e(t) = 10\log_{10}\left\{h_N^2(t)\right\}$ and the a coefficient is the ordinate at the origin, fixed at the beginning.

Assuming the response acquisition is sufficiently large to avoid time aliasing and the noise floor was entirely reached in order to be estimated accurately, an estimate of the noise floor is usually possible. Faiget [Faiget98] suggested a procedure consisting of two regression lines, attached to the sloped part and to the noise floor. The decay starting point is determined by avoiding in the computations the direct sound and the first relevant early reflections corresponding to a multiple of the mean free path between the source and the receiver, according to Kutruff [Kutruff00]. The decay ending point is chosen at 5dB above the noise floor. The regression for the noise is basically computed by manually selecting an interval at the end of the response. The truncation point is the crosspoint of the two curves. This method, however, needs visual selection of some points of the curve by the user. In practice, the

regression for the noise part is computed to assure the noise floor energy is almost constant, the reverberant field (possible reasons for regression lines not to be horizontal are that the acquired response is shorter than the real duration of the room decay or that noise characteristics are not stationary).

This method can be simplified by using only the regression line applied to the decay slope, and making an estimate of the noise floor as the energy of the last part of the response (usually 10% of the end of the tail). Nevertheless, care must be taken on the selection of the interval [t_1, t_2] for the regression, since problems can arise related to situations of measured responses inherently nonlinear as a two-stage decay (initial decay rate or early reverberation and late decay rate or reverberation) or beating (fluctuation) of the envelope because of two modes very close in frequency. Other possible situation refers to acoustics in coupled spaces. Here, the Schroeder decay function decomposition, for several decays, based on a Bayesian analysis is applied in order to find the slopes [Xiang06].

A method based on nonlinear regression where a measured and Schroeder-integrated energy-time curve is fitted to a parametric model of a linear decay plus a constant noise floor was proposed by Xiang [Xiang95]. This is computed in an iterative fashion, starting from a set of initial values for the model parameters and applying a gradient descent procedure to search for a least squares optimum. Nonlinear optimization is mathematically more complex than linear fitting and care should be taken to guarantee convergence of the algorithm. Following the idea of using nonlinear optimization for improved robustness and accuracy, Karjalainen proposed a Nonlinear Optimization of a Decay-plus-Noise Model, a parametric model, as an attempt to improve the accuracy of the estimated decay curves [Karjalainen02]. This method proved to outperform traditional methods especially under very low SNR conditions.

Lundeby [Lundeby95] presented an iterative algorithm for automatically determining the background noise level, the decay-noise truncation point, and the late decay slope of an impulse response. This algorithm is used within the scope of this thesis, thus, a more detailed description of the algorithm is given next.

1. The squared impulse response is averaged in local time intervals in the range of 10–50 ms, to yield a smooth curve without losing short decays.

2. A first estimate for the background noise level is determined from a time segment containing the last 10% of the impulse response. This gives a reasonable statistical selection without a large systematic error, if the decay continues to the end of the response.
3. The decay slope is estimated using linear regression between the time interval containing the response 0dB peak, and the first interval 5–10dB above the background noise level.
4. A preliminary crosspoint is determined at the intersection of the decay slope and the background noise level.
5. A new time interval length is calculated according to the calculated slope, so that there are 3–10 intervals per 10dB of decay.
6. The squared impulse is averaged into the new local time intervals.
7. The background noise level is determined again. The evaluated noise segment should start from a point corresponding to 5–10dB of decay after the crosspoint, or a minimum of 10% of the total response length.
8. The late decay slope is estimated for a dynamic range of 10–20dB, starting from a point 5–10dB above the noise level.
9. A new crosspoint is found.

Steps 7–9 are iterated until the crosspoint is found to converge (max. 5 iterations).

Response analysis may be further enhanced by estimating the amount of energy under the decay curve after the truncation point. The measured decay curve is artificially extended beyond the point of truncation by extrapolating the regression line on the late decay curve to infinity. The total compensation energy is formed as an ideal exponential decay process, the parameters of which are calculated from the late decay slope.

A simple procedure to reduce the noise floor from the impulse response consisting of subtracting the mean square value of the background noise from the original squared impulse response, before the backward integration, was proposed by Chu [Chu78]. For stationary noise, the resulting backward integrated curve is an approximation of the ideal decay curve.

The SNR can be increased by simply replacing the squared single impulse response with the product of two impulse responses measured separately at the same position [Hirata82].

Nevertheless, under adverse noise conditions, a direct determination of the decay curve from the squared and time-averaged impulse response has been noted to be more robust than the backward integration method [Satoh98].

Alternative methods to estimate the decay curves of rooms

In the daily acoustic measurement affairs, the estimation of the sound decay curves is based on the excitation of a sound source by a test signal and capturing the room response by a microphone. The common procedures are the Interrupted Noise Method [ISO3382-2], a burst of broad-band or narrow-band noise is radiated into the test enclosure (in this case, the decay curve is merely the sound pressure levels captured by the microphone), or cross-correlation techniques, such as the MLS or swept sine (the decay curve is estimated from the impulse response). In each situation, the combination of the sound source and microphone positions are known.

However, a significant effort has been dedicated to estimate the decay curves from passively received microphone signals, i.e. the sound source is unknown. The objective is to establish blind methods for determining RT when the room geometry and sound absorption properties are unknown, or when a controlled test sound cannot be employed. These methods are feasible with speech sounds and are particularly important for incorporating in hearing-aids or hands-free telephony devices or to help controlling the reinforcement sound system in live conference venues. There are essentially two blind approaches, (i) the blind dereverberation, aiming at recovering a sound source by deconvolving the room output with the unknown room impulse response; this technique faces the difficulty of minimum phase issues, limiting the applicability of the method, and (ii) semi-blind methods based on the characteristics of the enclosures by learning approaches using neural networks [Cox01], or truly blind method using a maximum-likelihood procedure [Ratnam03] or statistical estimators based on a time-frequency room decay model related to Polack's statistical reverberation model [Wen08].

3.4 Analysis/synthesis filter banks for audio applications

Given the general assumption that audio signal spectra vary slowly with time, i.e. they are statistically stationary, and they can be described by short-term analysis, it is both practical and efficient to represent the signal in the frequency domain. In addition, it is meaningful from the human-perception point of view to be able to separately manipulate spectral components of the signal according to the critical bands filters.

The filter bank framework provides the best medium for the removal of redundancy, i.e. information that is not necessary to uniquely identify the signal, and irrelevancies, i.e. information that perceptually is not important. Another significant interest is related to noise removal of spectral components of the signal.

3.4.1 General overview

A filter bank, which is a parallel bank of bandpass filters covering the entire spectrum of interest, is an efficient approach to analyze the frequency contents of a signal. The analysis filter bank divides the signal spectrum into frequency sub-bands and generates a time-indexed series of coefficients representing the power located in each band. Depending of the scheme used, they are designated as uniform, *UFB*, or non-uniform, *NUFB*, filter banks. In order to reconstruct the signal, after spectral modification, it is necessary to synthesize all the frequency bands. These procedures are described in terms of an analysis/synthesis framework.

The selection of the appropriate filter bank is an important task for the algorithm designer. This is related to the trade-off between time and frequency resolution, or more precisely, analysis of stationary and impulsive signals. The filter bank might have some of the following properties, depending on the proposal of the project [Vaidyanathan90]. The requisites for the filter banks regarding the specificities of the techniques developed within the scope of this thesis can be listed as :

- *Good channel separation* – Each bandpass filter should have good skills for the rejection of the energy of the adjacent bands (the filter order or selectivity, nevertheless, the phase characteristic is an issue). The filters overlapping response should be low, not compromising the group delay.
- *Strong stopband attenuation* – The ratio of the main to the secondary lobes should be high, ideally above 90dB.
- *Perfect reconstruction, PR* – It should assure that the output signal is a replica of the input, except for a delay and perhaps a scale factor, if no modification to the filtered signal is applied. This is an important issue for applications where the test signal frame has to be analyzed and synthesized before applying correlation techniques to estimate the RIR (see, as an example, the acoustical techniques developed in Chapter IV). Often, this property is only partly achieved due to aliasing issues and these filtering structures are termed nearly perfect reconstruction systems, *NPR*. This design goal in perceptual audio coding has the advantage that the noise which is added in the coding/decoding

process is generated all by the quantizer. Therefore, since the noise source is known, it can be controlled/masked by the signal.

- *Good Frequency and Temporal Resolution* – In perceptual audio applications, the bandwidth of each bandpass filter (in the filter bank) should be equal to or narrower than the width of the narrowest critical band, i.e., 100 Hz. In general, most uniformly spaced filter banks cannot meet both these constraints given the large variation in width of the critical bands with frequency.
- *Availability of fast algorithms* – This statement has to be accomplished mostly for real time applications.

A number of time to frequency mapping algorithms are available, differing on the degree to which they allow for source component separation and source redundancy extraction. Just to mentioned a few, discrete Fourier transform (DFT), discrete cosine transform (DCT), quadrature mirror filters (QMF), pseudo QMF (PQMF), modified DCT (MDCT), hybrid filter banks, Discrete Wavelet Transform, DWT, are time to frequency mapping techniques mentioned in the literature for audio analysis/synthesis and in particular for perceptual audio coding [Jayant93, Krasner79, Princen87, Vettefii95, Sinha93, Princen95, Tsutsui92, ATSC95, ISO/IEC11172-3].

The next section briefly describes some of the filter bank structure implementation used or tested in this research.

3.4.2 Implementation

The most popular filter bank approaches used in audio signal analysis and synthesis are (i) Butterworth constant power, (ii) Cosine Modulated Pseudo-QMF M-band banks, PQMF, (iii) Discrete Wavelet Transform, DWT and (iv) FFT filtering techniques.

The Butterworth constant power filter bank (the sum of the absolute squared transfer function of the bandpass filters equals unity) is used mostly for analysis purposes only. In fact, although the property of PR is not met, the channel separation and the stopband attenuation are in general better controllable and higher than in the others approaches. The digital implementation of this filter bank has some drawbacks due to the strong discrepancy between the filter coefficients of the different band frequencies and the low frequency bands shows filter coefficients near zero. Therefore, it is convenient to group the frequency bands in different sets

and use different sampling frequencies for each set. This is the first approach in the application of multirate filter banks.

The PQMF, also known as the polyphase filter bank, performs a cosine modulation of a lowpass prototype filter to realize parallel *M*-channel filter banks with nearly PR characteristics and has a wide range of applications in the audio analysis. That is the meaning of the prefix "*Pseudo*". This filter bank splits the audio spectrum into *M* channels of uniform bandwidth as depicted in

Figure 3.4. The *M* analysis filters are characterized by individual impulse responses $h_k(n)$, frequency responses $H_k(\theta)$, for $0 \leq k < M$ and have normalized center frequencies $\frac{(2k+1)\pi}{2M}$.

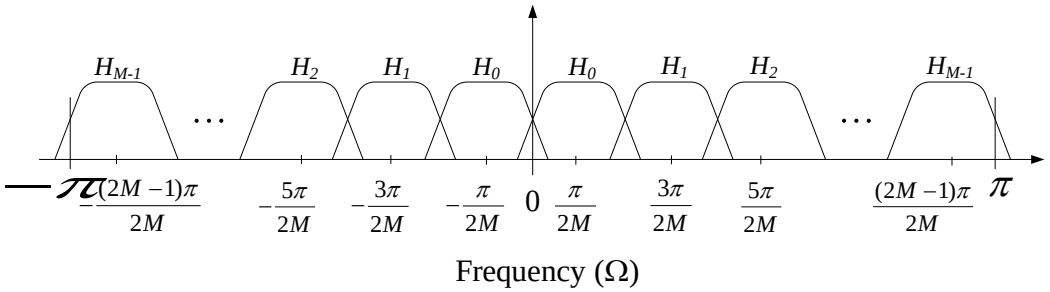


Figure 3.4 – Magnitude Response, Oddly-Stacked Uniform M-Band Filter Bank.

The interpretation of this filter band is more suitably described by the analysis/synthesis framework depicted in Figure 3.5. The input signal, $x(n)$, is first filtered by a set of FIR bandpass filters of order $(L\text{-}1)$, $H_k(z)$. To achieve a critically sampled or maximally decimated signal representation, i.e, the number of subband samples is equal to the number of input samples, the signals are decimated by a factor of *M*.

$$v_k(n) = x_k(n) * h_k(n) = \sum_{m=0}^{L-1} x(n-m)h_k(m), \quad k = 0,1,...,M-1 \tag{3.13}$$

yielding

$$y_k(n) = x_k(n) * v_k(Mn) = \sum_{m=0}^{L-1} x(n \quad M\text{-}m)h_k(m), \quad k = 0,1,...,M-1 \tag{3.14}$$

.

After applying spectral modification according to our needs, the intermediate samples $w_k(n)$ are interpolated (upsampled) by the factor M.

$$w_k(n) = \begin{cases} \tilde{y}_k\left(\frac{n}{M}\right), & n = 0, M, 2M, 3M, \ldots \\ 0, & otherwise \end{cases} \tag{3.15}$$

The imaging spectrum distortions of each subband channel introduced by the upsampling process is eliminated by processing the sequences $w_k(n)$ with a parallel bank of synthesis filters, $G_k(z)$. The final stage consists of combining the sequences to form the output modified signal $\hat{x}(n)$.

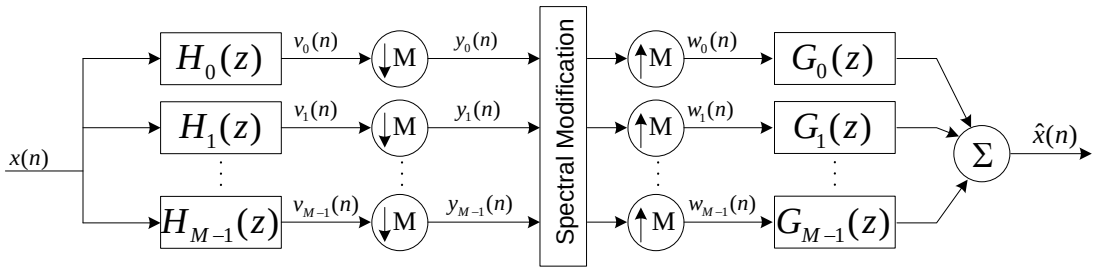


Figure 3.5 – Uniform M-band maximally decimated analysis/synthesis filter bank.

The analysis and synthesis filters are carefully designed to cancel aliasing and imaging distortions to attain the *PR* property.

In some situations, it is more convenient to apply a filter bank that firstly divides the audio spectrum in octave or fractional octave bands to best fit the behavior of the human auditory system and/or to make the analysis of some other signals conveniently, such as the swept sine signal. This arrangement is very useful when the signal energy exhibits bandwidth dependent distribution among frequency bands. Therefore, efficient structures and design procedures for general nonuniform filter banks (NUFBs) are often studied. This has the advantage of relating the time and frequency scales correctly. With regard to NUFBs, a number of design methods have been proposed in the literature over the years [Hoang89, Cox86, Nayebi93, Kovacevic93, Liu93, Watanabe98, Princen95, Argenti98, Lee95, Li97]. However, of the various available methods, only a few have linear phase (LP) property which is crucial in some applications such

as image coding. This is because the analysis/synthesis filters with an LP property can avoid artifacts in the reconstructed image. It should be pointed out that even though some methods resulti in LP NUFBs [Nagai97, Wada95] they usually have a high design cost and implementation complexity and their performance may not be satisfactory.

To develop such a filter bank, one possibility consists of using the Discrete Wavelet Transform, DWT. In fact, wavelet transforms are closely related to tree structured digital filter banks, and thus to multiresolution analysis. Tree structured filter banks give rise to nonuniform filter bandwidths and nonuniform decimation ratios in the subbands. These statements hold the fundamental issues to the wavelet transform.

DWT sub-band decomposition offers increased flexibility over the filter bank of fixed bandpass filters since identical filter bank magnitude frequency responses can be obtained for many different choices of a wavelet basis, or equivalently, choices of filter coefficients.

Under certain assumptions, the DWT acts as an orthonormal linear transform $T : R^N \rightarrow R^N$. For a finite support wavelet of length K, the associated transformation matrix Q is fully determined by a set of coefficients $\{c_k\}$ for $0 \leq k \leq K\text{-}1$. The Q matrix transformation can be viewed as a filter bank implementation. This scheme is depicted in Figure 3.6.

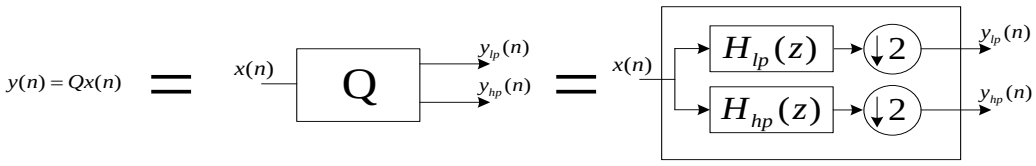


Figure 3.6 – Filter bank interpretation of the DWT.

Application of the transform matrix Q to the input signal $x(n)$ of size Nx1 generates an output signal $y(n)$ of size Nx1 in the wavelet transform domain. The signal $y(n)$ can be separated into two signals of size (N/2)x1 of approximation and detail coefficients $y_{lp}(n)$ and $y_{hp}(n)$, respectively. The spectral contents of the signal $x(n)$ transformed in $y_{lp}(n)$ and $y_{hp}(n)$ corresponds to the frequency subbands performed in 2:1 decimated output sequences from a QMF bank.

Therefore, recursive DWT applications effectively pass input signals through a tree-structured cascade of lowpass and highpass filters followed by a 2:1 decimation factor at every node. The forward/inverse transform *Q* matrix of a particular wavelet is associated with the corresponding QMF analysis/synthesis filter bank. The usual wavelet decomposition implements an octave-band filter bank structure, also called a dyadic filter bank when implemented using a binary tree structure, as illustrated in Figure 3.7. In the figure, frequency subbands associated with the coefficients from each stage are represented schematically for an audio signal sampled at 44.1 kHz.

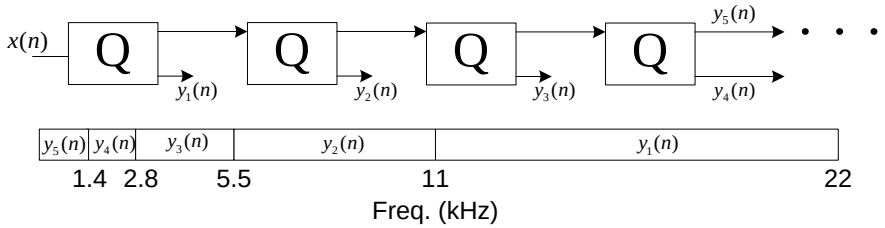


Figure 3.7 – Subband decomposition associated with a DWT transform.

Notice that the first two low frequency bands have equal bandwidth.

The octave synthesis filter bank corresponds to the transposed structure of that depicted in Figure 3.7, applying all the statements made about the analysis filter bank. In order to achieve *PR* at the end of the analysis/synthesis framework, all subbands have to exhibit the same delay. Therefore, a delay element is inserted in the high frequency channels.

For a better frequency resolution, the wavelet packet, DWPT, can be used to decompose both the approximation and the detail coefficients at each stage of the tree, as shown in Figure 3.8.

It may be noted that downsampling the highpass signal in each two-channel analysis filter bank leads to a flip in frequency of the highpass band and shifting down to the baseband. Therefore, the lowpass channel in a filter bank following a highpass output corresponds to the upper half and the highpass channel to the lower half of the preceding highpass channel.

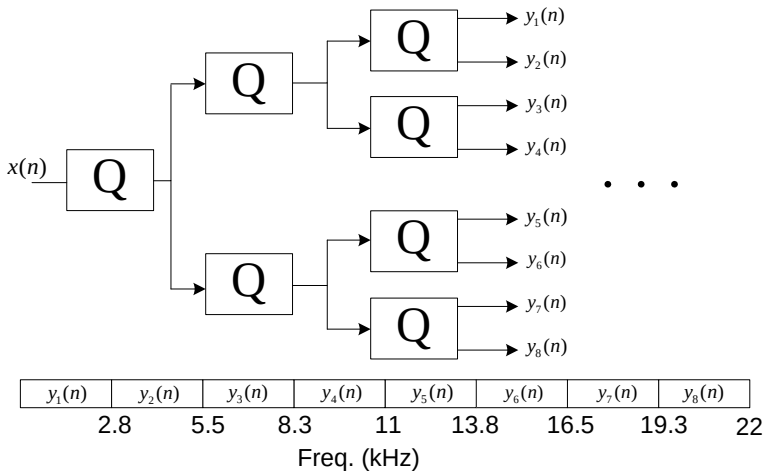


Figure 3.8 – Sub-band decomposition associated with a DWPT transform; non uniform decompositions trees are also possible.

In the figure, frequency subbands associated with the coefficients from each stage are represented schematically for the 44.1 kHz sample rate. Wavelet or wavelet packet decompositions can be tree structured as needed to decompose the input signal in a set of frequency subbands adapted to the frequency contents. Non-uniform decomposition trees can be applied to better follow the signal constraints.

The third-octave filter banks, whose cutoff frequencies are related by $1:\sqrt[3]{2}$, cannot be implemented precisely with the tree structure of two-channel filter banks. Only rough approximations are possible.

Another structure used to separate audio bands is the filter bank based on the FFT, the so-called Harmonic filter bank or phase vocoder analysis. In the Short-Time Fourier Transform, STFT, which implements a uniform FIR filter bank [Duda73 and Barron73], each FFT bin can be regarded as one sample of the filter-bank output in one channel [Portnoff76], i.e. a parallel bank of N bandpass filters as illustrated in Figure 3.9., where the outputs $y_k(n)$ are computed from the modulation of the $x(n)$ and lowpass filtering.

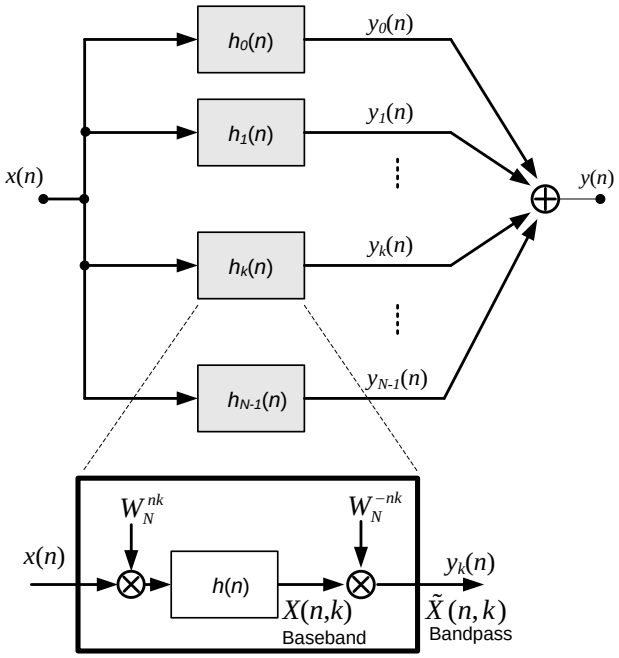


Figure 3.9 – Filter bank approach based on the STFT (from [Zölzer02]).

In fact, this structure simulates a bank of overlapping band-pass filters each one centered on an integer multiple of a base frequency Ω_a; i.e., at frequencies $\Omega_k = k\Omega_a$, for $k = 1, .., N$, where Ω_a is referred to as the analysis frequency, and N is a constant number of bands. This is a kind of Uniform M-Band filter bank.

The amplitude frequency response of each filter $|H_k(\Omega-\Omega_k)|$ has a maximum value of unity at $\Omega= k\Omega_a$. Also, each filter function is zero or very small for $\Omega \leq (k-1)\ \Omega_a$ and $\Omega \geq (k+1)\ \Omega_a$. This approach is useful for applications of periodic music signals described by harmonic spectral tonal contents. Actually, for a periodic signal with constant fundamental frequency Ω_0 exactly at Ω_a and fixed harmonic amplitudes A_k, each filter will produce a sine wave with frequency $\Omega_k = k\Omega_a$ and amplitude A_k. Additional advantages can be pointed out as facilitating manipulations of some sound parameters such as (1) time scaling—changing the duration of a sound without altering its pitch, (2) pitch transposition—changing pitch without affecting duration, or (3) time-varying filtering.

The filters are described by the impulse and frequency responses:

$$h_k(n) = h(n)e^{j\frac{2\pi nk}{N}} = h(n)W_N^{-nk}, \quad k = 0,1,...,N-1 \tag{3.16}$$

$$H_k(e^{j\Omega}) = H(e^{j(\Omega-\Omega_k)})W_N^{-nk}, \quad \Omega_k = \frac{2\pi}{N}k \tag{3.17}$$

.

The outputs $y_k(n)$ are obtained by filtering the input signal $x(n)$ by the bandpass filter $h_k(n)$ as

$$y_k(n) = x(m) * h_k(n) = \sum_{m=-\infty}^{+\infty} x(m)h_k(n-m) = \sum_{m=-\infty}^{+\infty} x(m)h(n-m)W_N^{-(n-m)k}$$

$$= W_N^{-nk} \sum_{m=-\infty}^{+\infty} x(m)W_N^{mk}h(n-m) = W_N^{-nk}X(n,k) \tag{3.18}$$

.

Since the bandpass filters are implemented from complex coefficients, the output is a complex value as well, thus

$$y_k(n) = \tilde{X}(n,k) = W_N^{-nk}X(n,k) = W_N^{-nk}\left|X(n,k)\right|e^{j\tilde{\varphi}(n,k)} \tag{3.19}$$

where $\tilde{\varphi}(n,k) = \frac{2\pi k}{N}n + \varphi(n,k)$.

Therefore, this filter bank is the so-called complex baseband implementation. The baseband signals $X(n,k)$ (short-time Fourier transform) are computed by modulation of $x(n)$ with W_N^{nk} and lowpass filtering for each channel k. The modulation of $X(n,k)$ by W_N^{-nk} yields the filtered bandpass signal $\tilde{X}(n,k)$. The output signal $y(n)$ is the sum of all bandpass signals.

A typical implementation of this filter bank is the N stage heterodyne filter consisting of a complex valued oscillator with a fixed frequency Ω_k, a multiplier and an FIR filter. The multiplication shifts the spectrum of the sound to the origin Ω=0, and the FIR filter limits the width of the frequency shifted spectrum.

The main difference from classical bandpass filtering is that in this implementation the output signal is located in the baseband. This representation leads to a slowly varying phase $\varphi(n,k)$

and the derivative of the phase is a measure of the frequency deviation from the center frequency Ω_k.

Although this type of filter is more suitable for use with periodic signals, once it assigns the harmonics and filter band one-to-one, the frequency bands can be arranged conveniently to form different filter banks structures such as octave and third octave bands, thanks to its property of perfect reconstruction. This filter bank will be used in Chapter 4.

3.5 Perceptual models

In this section, the fundamentals of psychoacoustic models used on the perceptual audio coders are addressed. This is the basic study for the implementation of some perceptual techniques developed for acoustical systems characterization, as presented in Chapter 5.

3.5.1 General

A great effort on the development of accurate algorithms that explore the human auditory masking effects, used for the lossy compression of audio signals, has been reported in the last decades. In compression of audio signals, the aim is to hide the noise introduced by the coding process below the masking threshold making the noise unperceivable by people. This will then render the coding process transparent, enabling a better compression without audible degradation of the signal. A comprehensive review of perceptual audio coding was published by Painter and Spanias in 2005 [Painter05].

More recently, the masking properties of the ear have also been taken into account for improvement of the quality of noise reduction algorithms. Specifically, instead of attempting to remove all the noise from the signal, these algorithms attempt to attenuate the noise below the audible threshold. In the context of short-time spectral magnitude (STSM) subtractive algorithms, this reduces the amount of modification to the spectral magnitude, thus reducing artifacts. This is of great importance where the resulting signal needs to be of very high quality.

Within the framework of this research, the test signals are applied to the musical material keeping its sound level below the audible threshold in order not to be perceived by an audience.

In order to take advantage of the psychoacoustic effects, it is convenient to distinguish between two types of simultaneous masking, namely tone-masking-noise [Scharf70], and noise-

masking-tone [Hellman72]. Moreover, here, the term “noise” refers to signals of random wideband or narrowband behavior as opposed to tonal component signals. In the case of tone-masking-noise, a tone occurring at the center of a critical band masks noise of any subcritical bandwidth or shape provided the noise spectrum is below a predictable threshold directly related to the strength of the masking tone. The second masking type follows the same concept with the roles of masker and maskee reversed. Both cases are examples of simultaneous masking phenomena described in Chapter 2. Nevertheless, inter-band masking has also been observed, i.e., a masker centered within one critical band has some predictable effect on detection thresholds in other critical bands. These effects, also known as the spread of masking, are often modeled in coding applications by an approximately triangular spreading function. A convenient analytical expression [Schroeder79] is given by:

$$SF_{dB}(x) = 15.81 + 7.5(x + 0.474) - 17.5\sqrt{1 + (x + 0.474)^2} \tag{3.20}$$

with x in barks and $SF_{dB}(x)$ in dB. After critical band analysis is performed and the spread of masking has been accounted for, masking thresholds in psychoacoustic coders are often established [Jayant93]:

$$TH_N = E_T - 14.5 - B \tag{3.21}$$

$$TH_T = E_B - K \tag{3.22}$$

where TH_N and TH_T, in dB, are respectively, the noise and tone masking thresholds due to tone-masking-noise and noise-masking-tone, E_B and E_T are the critical band noise and tone masker energy levels, and B is the critical band number. Depending upon the algorithm, the parameter K has typically been set between 3 and 5dB. Masking thresholds are commonly referred to in the literature as (bark scale) functions of just noticeable distortion (JND). One psychoacoustic coding scenario might involve first classifying masking signals as either noise or tone, next computing appropriate thresholds, then using this information to shape the noise spectrum beneath JND. Note that the absolute threshold (T_{ABS}) of hearing (threshold in quiet) is also considered when shaping the noise spectra, and that *max* (JND, T_{ABS}) is most often used as the permissible distortion threshold. Notions of critical bandwidth and simultaneous masking in the

audio coding context give rise to some convenient terminology illustrated in Figure 3.10, where the case of a single masking tone occurring at the center of a critical band is considered.

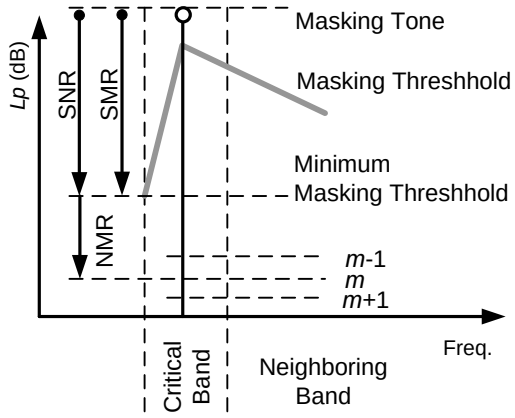


Figure 3.10 – Schematic representation of simultaneous masking effect, spreading function, in dB (after [Noll93]).

A hypothetical masking tone occurs at some masking level. This generates an excitation along the basilar membrane which is modeled by a spreading function and a corresponding masking threshold. For the band under consideration, the minimum masking threshold denotes the spreading function in-band minimum. Assuming the masker is quantized using an *m*-bit uniform scalar quantizer, noise might be introduced at the level *m*. Signal-to-mask ratio (SMR) and noise-to-mask ratio (NMR) denote the logarithmic distances from the minimum masking threshold to the masker and noise levels, respectively. Nevertheless, in an attempt to increase the accuracy of the spreading functions model to the basilar auditory mechanisms non-linear approaches are discussed and first proposed by [Baumgarte95]. Since the non-linear approaches are more complicated to implement increasing the processing load, its use in real time applications is often not possible. Therefore, the triangular spreading functions model is generally used instead.

One of the most frequently used masking models is the Johnston perceptual model [Johnston88] and given its importance in modern perceptual audio coding it will be examined in some detail. This model was adopted with some modifications in the ISO 11172-3, MPEG-1 Layer I and II (the so-called Psychoacoustic Model 1). Although a number of more sophisticated perceptual models are currently used in commercial audio coders, their benefits

are not used within the scope of this study, and therefore only the relevant characteristics will be mentioned.

The concept of psychoacoustic model was developed by Johnston [Johnston88] for coding of audio sampled at 32 kHz. This model was used by Tsoukalas et al. [Tsutsui92] for speech enhancement. Johnston's method calculates the auditory masking threshold to determine how much noise the coder can add before it becomes audible, following these steps to calculate the masking threshold:

- Analyze the signal in critical bands;
- Apply the spreading function (or spread of masking) to the critical band spectrum;
- Calculate the spread of masking threshold;
- Account for absolute thresholds;
- Relate the spread of masking threshold to the critical band masking threshold.

The complete description and implementation of the Johnston auditory model is presented in Appendix D1.

3.6 Sinusoidal synthesis methods

This section deals with audio synthesis of some segments of musical tracks as an attempt to improve the usage of some acoustical measurement techniques developed in the present research.

Audio synthesis is the name given to a number of techniques for creating artificial sounds from real world or following new music aesthetics. A number of audio synthesis approaches are available based on:

- signal modulation techniques – in amplitude, phase and frequency - based on concepts from radio transmission, the distortion frequencies (modulations) are applied to a sine wave, resulting in a number of partials derived from the modulation frequencies;
- Fourier-based techniques – these techniques use the principle that a complex waveform is made up of a sequence of harmonics and partials. Subtractive synthesis (starts with a harmonically rich waveform, such as square or sawtooth wave, and filters out unwanted spectral components), additive synthesis (adds together multiple sine waves, each one a

harmonic or partial, to produce a complex final result). May also add noise for sibilant effects;

- Wave shaping synthesis – it is a technique for distorting an input waveform using a characteristic function;
- Granular synthesis – generates a sequence of sound grains. Each grain is an extremely short (10-100 ms) burst of sound. The grains are output in quick succession to produce a rapidly changing sound, which can be further filtered and envelope shaped;
- Physical modelling – a number of different techniques which attempt to reproduce instruments by simulating (modeling) the actions of the components within them.

However, only the sinusoidal model, which is a type of additive or subtractive synthesis, is illustrated throughout this section due to its capabilities of representing well complex sounds made of tonal harmonic components and partials keeping in mind one is interested in replacing some parts of music segments to be used in the perceptual techniques presented in Chapter 5.

3.6.1 Introduction

The musical sounds can be viewed, in a first approach, as made of a tonal part, or stationary pseudo-periodic part, plus a residual part. The tonal part, often named the deterministic part of the signal (voiced speech signal, sustain and release phases of tones produced by resonant instruments) refers essentially to the harmonic or partial components of the musical instruments in its steady state and can be conveniently decomposed into partials. Each partial usually corresponds to a mode of vibration of the producing sound system and is modeled by a sinusoid with given amplitude, phase, and frequency. The residual noise part, stochastic part such as the turbulences, unvoiced speech signal is considered to have the remaining signal energy. However, with this definition, some tonal components, which cannot be parameterized accurately, belong to the noise part.

The steady state condition in this context is from a few tens of milliseconds to several seconds, depending of the purposes of the intended analysis (corresponding in the lower limit, for instance, to the audio coding usage and the upper limit to applications in room acoustics). That is the sufficient segment of signal to properly analyze its spectral components or time waveform energy features. In this sense, an attempt to produce a replica of a sound, as similar

as possible from the perceptual point of view, passes by an implementation of a system capable of analyzing the tonal and the residual parts separately. This system is usually referred to as Sinusoidal plus Residual Modeling [Serra90]. A simple scheme is illustrated in Figure 3.11.

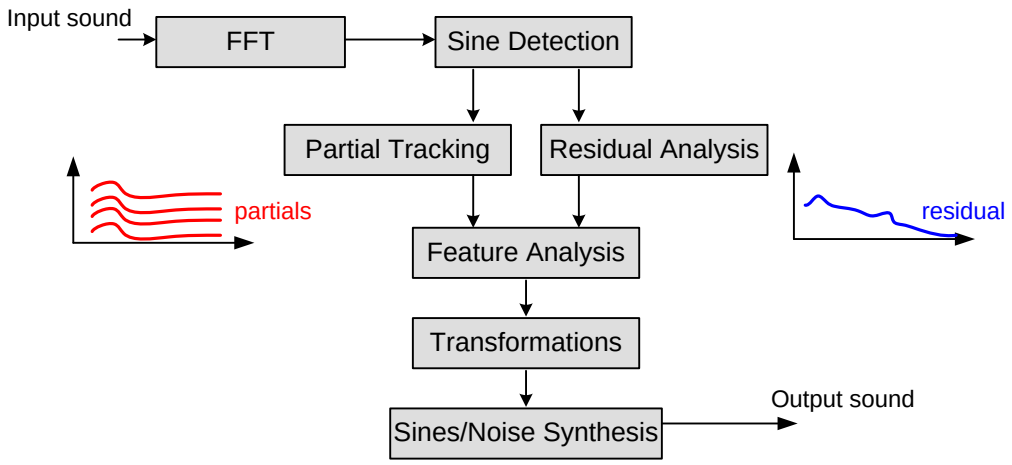


Figure 3.11 – Scheme of the Sinusoidal plus Residual Modeling approach.

The relevant topics for the implementation of this model are (a) Detection/Estimation of Sinusoids, (b) Partial Tracking, (c) Transient Modeling, (d) Multiresolution analysis, (e) Residual Analysis/Modeling, (f) Feature-based Analysis/Synthesis, and (g) Synthesis of Sinusoids/Noise parts. The most important topics for the purposes of the present study are described next.

Sinusoidal modeling can be used for deterministic component extraction, noise removal, pitch shifting, time stretching, sound morphing, audio coding, etc. However, in the literature the use of sinusoidal modeling is discussed mostly in an analysis-synthesis framework, such as the Spectral Modeling Synthesis approach, SMS, extensively discussed in [McAulay86, Serra97, Zölzer02], where new sounds are produced from the original by means of reconstruction with sinusoids, time stretching, pitch shifting, frequency warping etc.

The harmonic sinusoidal model is an alternative approach. This model uses the concept of harmonicity of musical sounds in the analysis procedures. In fact, harmonic sinusoidal modeling is an extension of sinusoidal modeling, with additional frequency constraints so that it is capable of directly modeling tonal sounds consisting of a module for finding harmonic partials at a certain frame, a module for tracking harmonic partials over time, and a module for

estimating sinusoidal parameters. This model, in a first approach, treats a sound (a musical note) as a whole, as an independent entity. There have been attempts to study pitched sound by grouping individual partials [Brown94, Ellis96] that fit in a harmonic relationship, the so called *Computational Auditory Scene Analysis* (*CASA*). Sound, in this context is referred to as a multiplexed signal - an aggregate of a number of sounds produced by a number of sound sources.

Within the framework of the present study, though, only the sinusoidal modeling will be considered.

3.6.2 Considerations on DFT computation

The control of the STFT is accomplished by using four parameters: window type, window-length, hop-size, and FFT-size.

Zero padding

The FFT of size N is usually chosen to be the lower power of two that is at least twice the window length, M. The difference of N-M samples is filled with zeros, *zero padding*, Zr. This technique has advantages since zero-padding in the time domain corresponds to interpolation in the frequency domain hence making it easier to detect the tonal frequency components. Therefore, this process does not increase the spectrum detail, and does not overcome the time/frequency trade-off. Nonetheless, by adding intermediate interpolated bins, the frequency resolution increases.

Zero-phase window

The phase detection is facilitated by using a zero-phase window, i.e. the windowed data (using an odd length window) is centered about the time origin by applying a circular shift operation of half the window length. Therefore, a zero-centered data frame appears in the length N_{FFT} input buffer. The first $(M-1)/2$ samples of the windowed data, the "negative-time" portion, are stored at the end of the buffer, from sample $N-(M-1)/2$ to $N-1$, and the remaining $(M+1)/2$ samples, the zero and "positive-time" portion, are stored starting at the beginning of the buffer, from sample 0 to $(M-1)/2$. Then, all zero padding occurs in the middle of the FFT input buffer.

Hop-size

The time advance, in samples, from one frame to the next is called the *hop-size*, *H*. Once the spectrum has been computed at a particular point in the waveform, the STFT process hops along the waveform and computes the spectrum of another section in the sound. This hop size *H*, is an important parameter. Its choice depends very much on the purpose of the analysis. In general, more overlap will provide more analysis points and therefore smoother results across time, though the computation expense is proportionately larger.

A general and valid criterion is that the successive frames should overlap in time in such a way that all data are weighted equally. For overlap-add synthesis, as described in the next section, this criterion is strictly followed. However, for other synthesis techniques or different applications, this criterion may be overly conservative when the signal is stable and too adventurous for fast-changing signals. Therefore, the choice of *hop-size* is determined by the application and the sound characteristics and with convenient choice of the window type the resulting envelop adds approximately a constant value, as shown in the expression

$$A_w(m) \triangleq \sum_{n=-\infty}^{+\infty} w(m - nH) \approx const. \tag{3.23}$$

The so-called amplitude deviation of the overlap factor, d_w, as a measure of the deviation of the A_w from a constant, is given by

$$d_w = \frac{\max_m[A_w(m)] - \min_m[A_w(m)]}{\max_m[A_w(m)]} \text{x}100 \tag{3.24}$$

.

For certain window types, overlap factors values assure the windows add perfectly to a constant value (examples: rectangular, hanning and hamming windows).

3.6.3 Sinusoid plus residual model

Consider a parametric sinusoidal representation based on the STFT. The sinusoid parameters are estimated by keep tracking of the amplitude, frequency, and phase of the most prominent peaks over time in the series of spectra returned by the STFT. Hence, the sound produced by synthesizing sine waves for each peak trajectory found is equal to the original in a perceptual sense. Nevertheless, the noise (residual) part of a sound has also to be considered in the

analysis. The standard sinusoid plus residual model is first described [McAulay86, Serra89] considering a sound signal

$$s(t) = \sum_{k=1}^{P} A_k(t) \cos(\varphi_k(t)) + r(t) \tag{3.25}$$

where P is the number of partials representing the sound, $A_k(t)$ is the amplitude of the time varying sinusoid or the k partial at instant t, $\varphi_k(t) = \varphi_k(0) + 2\pi \int_0^t f_k(\tau) d\tau$ is the angle or phase of the sinusoid k at instant t, $\varphi_k(0)$ is the initial phase of partial k (phase at time = 0), and $r(t)$ stands for the residual part of the sound with time-varying short-term spectrum. Additive noise signal (which is considered a stochastic signal) is modeled by the time-domain convolution of white noise with a time-varying frequency-shaping filter.

The model works in two main steps: first, the tonal part (sinusoidal components) is estimated from the STFT of the audio steaming for subsequent times. In fact, in the presence of concurrent sinusoids and noise, it is desirable to isolate the sinusoid of interest from other events before estimating the parameters. As long as the STFT can be viewed as a bank of bandpass filters centered for each bin, it is often used as a first approach of sinusoid analysis. Hence, the P triplets (f_k, A_k, φ_k), which are the parameters of the additive model representing points in the frequency-amplitude plane at time t, are computed. For stationary pseudo-periodic sounds, these parameters evolve slowly and continuously with time, controlling a set of pseudo-sinusoidal oscillators commonly termed tonal components or partials. The second part estimates the residual component by subtracting the tonal component from the original waveform. This procedure preserves the time shape of the waveform provided the instantaneous phases of the original sound are matched. Different windows can be used to perform the spectral analysis of the residual part allowing different time-frequency resolutions. Figure 3.12 depicts the detailed block diagram sinusoidal plus residual model.

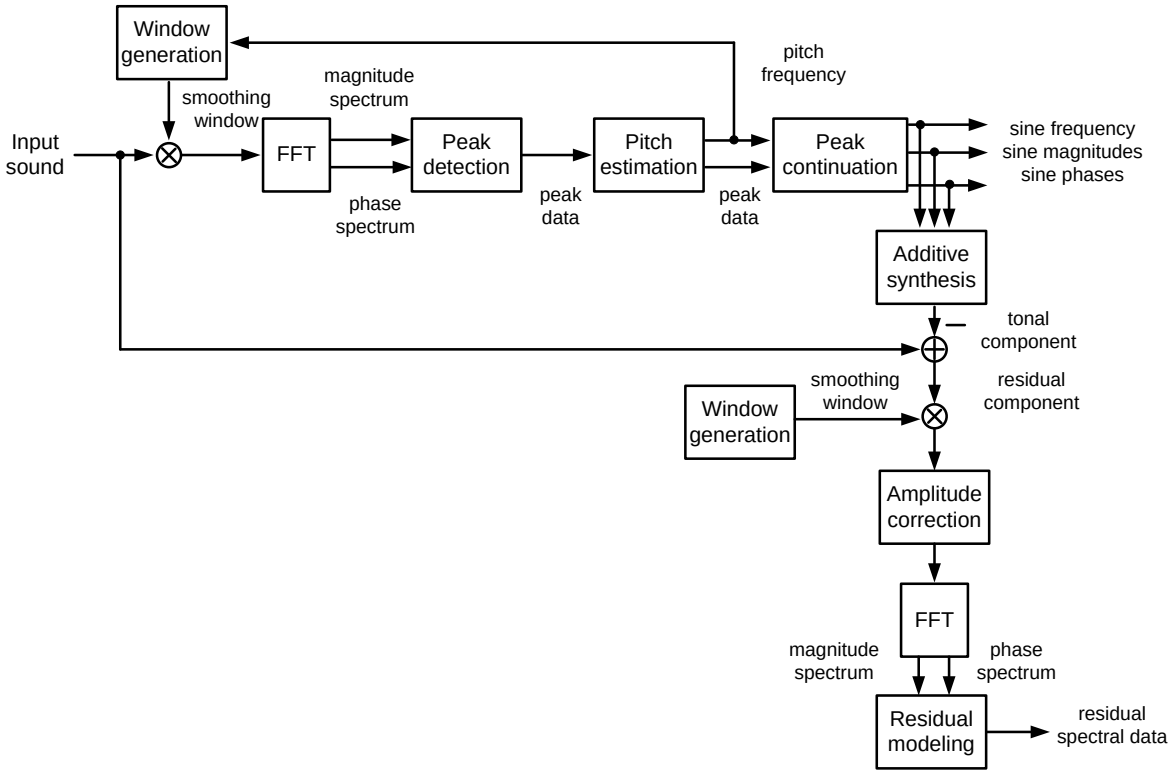


Figure 3.12 – Block diagram of the sinusoidal plus residual analysis [Zölzer02].

The full block diagrams of frequency-tracking analysis/resynthesis system for the tonal part are illustrated in Figure 3.13.

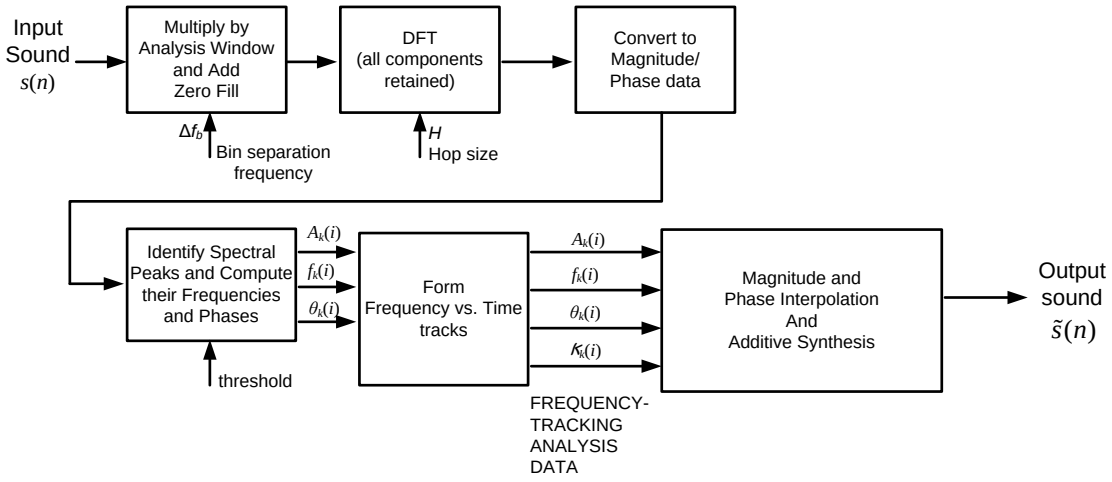


Figure 3.13 – Block diagram of frequency-tracking analysis/resynthesis system constituted by the analysis method and the resynthesis method based on frequency-tracking data.

Residual Noise Synthesis

Different parametric models have been proposed for the stochastic component, such as parametric noise spectral envelopes, parametric bark-band noise or hybrid signal models in which the residual of the sinusoidal analysis-synthesis is decomposed using a discrete wavelet packet transform (DWPT), and then represented in terms of carefully selected principal components. Nevertheless, a simple procedure consists of synthesizing the residual part of the sound in the frequency domain. When the analyzed residual component is represented by a complex quantity, with magnitude and phase, both the spectra of the residual and sinusoidal parts are added, frame by frame. Instead, if the residual magnitude has been approximated by a spectral envelope, an appropriate complex spectrum has to be generated. Therefore, simply filtering white noise with these frequency envelopes, i.e. performing a time-varying filtering of white noise, which is generally implemented using the time-domain convolution of white noise with the impulse response corresponding to the spectral envelope of a frame, will do the job. In order to avoid periodicity, this procedure is performed in the frequency domain by creating a magnitude spectrum from the approximated one, or its transformation, and generating a random phase spectrum with new values for each frame.

Integration of Sinusoidal and Residual Synthesis

Although in the process of generating the noise spectrum no windowing procedure was needed, since the data was added directly into the spectrum without any smoothing consideration, in the sinusoidal synthesis a window was applied. Its effect is canceled in the time domain after applying the IFFT. Therefore, the same window should be applied to the residual spectrum before being added to the sinusoidal spectrum. Convolving the transform of the original window with the noise spectrum accomplishes this issue, noting that there is only the need to use the main lobe of the window since it includes most of its energy. This is implemented quite efficiently since it only involves a few bins and a symmetric window was applied. A single IFFT is then used for the combined spectrum and, finally, in the time domain, the effect of the window is cancelled by imposing a triangular window. By an overlap-add process, the successive frames are combined to produce the time-varying characteristics of the sound.

The next subsections briefly describe the standard procedures for implementation of the system.

3.6.4 Frequency-Tracking analysis method

***Sinusoid detection from* STFT**

The frequency associated to each peak is estimated by analyzing the audio streaming by the successive STFT*s* of overlapped windowed segments (frames), corresponding to the spectrum $S(f)$ of the $s(t)$ waveform. A window function such as the Kaiser function with $\alpha = 6.3$ can be used for good peak separation. Usually, a zero-padding factor, Zr, of at least 2.0 is used. The windows are overlapped by a hop size of H samples or time delay of $\Delta t_{\text{frame}} = H/ f_s$, in seconds. Also, the window bin frequency spacing (after zero-padding) is $\Delta f_{\text{FFT}} = f_s / \left(Z_r N_{FFT} \right)$.

The condition for a peak detection is evaluated by considering that a local spectral peak is defined as a bin k where $|S_\gamma| \geq |S_{\gamma+1}|$, $|S_\gamma| \geq |S_{\gamma-1}|$, assuming that only one identity can occur, i.e. three equal amplitude values are not allowed. A global spectral peak is defined as a local spectral peak at bin γ so that $|S_\gamma| \geq |S_l|$ for the frame l. Nevertheless, a sinusoid should appear in the amplitude spectrum as a local peak. Moreover, a peak is considered a candidate to a sinusoid if it is apart by a defined number of bins from its neighborhood peaks.

Due to the sampled nature of the spectrum returned by the FFT, each peak is accurate only to within half a sample. Therefore, the estimated frequency and true maximum value are found by interpolation techniques applied to the three adjacent samples $S_{\gamma-1}$, S_γ and $S_{\gamma+1}$,usually by parabolic interpolation, as illustrated in Figure 3.14. The amplitude spectrum between bins γ*-1* and $\gamma+1$ is then approximated by a quadratic function, thus, the peak position of this quadratic function is used as the frequency estimate.

Zero-padding improves the interpolation accuracy, since more points are automatically inserted between the window-function bins, and the interpolation is band-limited. However, direct interpolation can be implemented by fitting a smooth curve to the three points. Ideally, a best-fit shifted version of the transformed window function should be used. In practice, it has been found that fitting a quadratic curve to the *log* of the magnitude function yields adequate results with much less computation [Serra89].

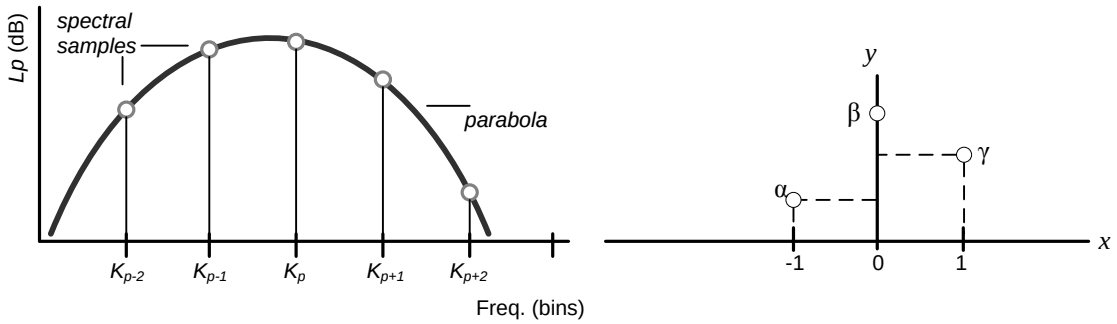


Figure 3.14 – Peak amplitude estimation by using parabolic interpolation techniques.

In order to increase the accuracy of the peak amplitude estimation, not applying a high padding-rate to all the signal due to computation cost issues, an alternative efficient spectral interpolation scheme consists of zero-padding choosing *Zr* large, such a quadratic (or other simple) spectral interpolation, using only samples in the vicinity of the maximum-magnitude sample. The peak frequency is given by

$$f_k = (\gamma + p)\Delta f_{FFT} \tag{3.26}$$

where

$$p = 0.5\frac{\log(A_{\gamma-1}) + \log(A_{\gamma+1})}{\log(A_{\gamma-1}) + \log(A_{\gamma+1}) - 2\log(A_{\gamma})}.$$

The amplitude is estimated by

$$A_k = \frac{A_\gamma}{(A_{\gamma-1} / A_{\gamma+1})^{p/4}}. \tag{3.27}$$

The corresponding phase is computed by first interpolating the values of the real and imaginary parts, a_γ and b_γ, at the peak frequency, and then computing the phase. The frequency, amplitude, and phase of each peak (A_k, f_k, θ_k) are thus computed and retained. Other FFT bin information is discarded. Usually, not every local maximum is chosen to be a peak. For example, peaks that are below a predefined threshold may be ignored. The threshold can be frequency dependent. For instance, a decreasing amplitude frequency curve, following the sound spectral energy shape, may be desirable because even though the higher-frequency

components of most musical sounds are generally weaker than the lower-frequency components, they are still very audible. In [Serra89] the use of two thresholds is discussed, one absolute and one relative to the highest peak of the spectrum of the current frame. The author also uses the log equivalent of a peak-to-valley ratio defined by

$$pvr = \frac{A_\gamma}{\sqrt{A_{\gamma-\delta} A_{\gamma+\delta}}} \tag{3.28}$$

where $\gamma - \delta_1$ and $\gamma + \delta_2$ are the bins corresponding to the first minimum below and above the peak bin γ. The peak would be rejected in case of *pvr* is above a designated threshold, *t*. Moreover, a perceptual model can be applied to the audio streaming prior to the computation of the STFT. This procedure acts as a filter of peaks not perceived by the human auditory system.

Peak Continuation

Time-frequency tracks are formed by conveniently connecting peaks of consecutive frames following peak continuation techniques. This procedure plays an important role in this subject and is the most crucial aspect of the analysis method, and there is probably no perfect way to do it. Rules for forming tracks from a heuristic method are described in the literature, though they are not optimal in any sense. For example, when the partials significantly change in frequency from one frame to the next, these algorithms easily switch from the partial that they were tracking to another one, which is closer at that point. The first and most important continuity measure is the closeness in frequency. In [McAulay86] the authors composed their sinusoid tracking using this measure alone.

The basic procedure consists of finding the best match between the peaks in frame *i* and in frame *i* + 1, organizing the peaks in time-varying trajectories. Matches are attempted between corresponding frequencies that are close together. If the number of peaks in frames *i* and *i*+1 are K_0 and K_1, respectively, and $K_0 > K_1$, some of the tracks will have to end ("death"). On the other hand, if $K_0 < K_1$, some new tracks will start ("birth"). Tracks could also start or end because the only available potential matches have excessive frequency differences. Detailed procedures for peak-tracking are given in [McAulay86, Serra97 and Serra99]. Compared to [McAulay86], in [Serra97] a dead track is revived if a successor is found within a short time.

The peak data for each track is written to a file for later use. For each frame *i*, the number of peaks K_i is given and each peak *k* is represented by amplitude $A_{k,i}$, frequency $f_{k,i}$, phase $\theta_{k,i}$, and a "link" $\kappa_{k,i}$ giving the peak number of the next frame to which it is connected. An alternative to giving the next-frame peak number is to number the tracks and give the track number of each peak, but a problem with this method is that the number of tracks usually changes continuously throughout the sound program, so that the track numbers soon get out of frequency order.

Advancements in partial tracking have mainly focused on deriving patterns for frequency and amplitude variation. Robel [Röbel02] proposed to estimate the frequency slope using the reassignment method and suggested using this slope in partial tracking. Instead of directly evaluating the frequency slope, Sterian [Sterian98] implements a linear frequency predictor as a Kalman tracker, with the frequency slope as a state variable. The use of frequency variation trends for guiding partial tracking appears in [Lagrange03] as a linear prediction model, exploring higher-order trends than a simple first-order slope. The linear prediction model is able to closely model exponential chirps and sinusoidal modulations. The predictor for a certain frame is calculated from a chosen number of previous frames. In [Lagrange05], a method for the interpolation of sinusoidal components is proposed. This method shows that autoregressive modeling of the amplitude and frequency parameters of these components allows the interpolation of missing audio data realistically, especially in the case of musical modulations such as vibrato or tremolo. The problem of phase discontinuity at the gap boundaries is also addressed.

All of the above methods are examples of pure forward tracking. One drawback relative to pure forward or backward tracking is the sensibility to local disturbance. On the contrary, forward-backward partial tracking improves the robustness by introducing a global cost function. In [Streit90] an early example of using a hidden Markov model for tracking sinusoids is described. In the context of sinusoid modeling, in [Depalle93] a method based on a hidden Markov model (*HMM*) that uses computed probabilities of peak trajectories to determine improved overall tracking is presented. Compared to forward frequency prediction models, a forward-backward tracking framework makes it possible to find optimal partials globally.

All the methods mentioned above deal with tracking individual partials. There are also methods that consider harmonic constraints in the first place. Brown [Brown92] introduces a pattern for harmonic particles that explicitly requires that the objects being tracked be harmonic. Davy [Davy03] models partial harmonicity by forcing partial frequencies close to perfect harmonicity in formulating a Bayesian model. Amplitude variation is modeled as "smooth" in [Davy03] by forcing the amplitude to be an additive combination of low-pass envelopes. Frequency variation modeled in the Bayesian framework in [Walmsley99] using single-frame harmonic models connected in a Markov chain. Dubois [Dubois07] formulates the detecting and tracking of harmonic sinusoids in a Bayesian sequential harmonic framework, which models a wider range of signals than previous Bayesian harmonic methods.

3.7 Sound descriptors

The characterization of an acoustic stimulus by a number of sound descriptors is an inevitable paradigm often used to facilitate the internal procedures of the applications by parameterization of its relevant characteristics. The main objective of this strategy was to describe the sound in different perspectives depending on the finality of the application.

These vectors of characteristics, representing a multidimensional space, are used with the main purpose to build event detection functions by selecting a number of thresholds or used on a pattern recognition approach as a mean of statistical classification of complex acoustic events. Examples of applications are indexing and browsing contents of sound metadata in databases, re-synthesis of sounds, such as in telephone transmissions (coding) or for conveniently masking test signals in an acoustical measurement approach, melody and musical genres, musical instruments identification.

An increasing number of types of audio features are available since the early research on speech recognition engines [Rabiner93], studies of classification of musical instrument sounds [Foote97, Brown99, Brown01, Jensen01, Peeters02], and psychoacoustic studies [Misdariis98, Zwicker99, Lidy05].

Although several strategies to characterize sound signals have been suggested, the interest to define audio data semantic content descriptors raised with the creation of the new standard format for indexing and transferring audio, the standard ISO/IEC MPEG 7 - Multimedia

Content Description Interface - Audio Low Level Descriptors. In [Peeters04], a systematic taxonomy on some of the functions intended to describe sound is presented for various descriptors.

3.7.1 Feature extraction

The segmentation of the sound signal is an important role in music characterization and transformation using an analysis/synthesis framework. These techniques were originally developed for speech, such as those based on pattern recognition or knowledge-based methodologies, and rapidly started to be used in music applications. Based on the segmentation of a number of sound characteristics procedure, the so-called features are extracted in order to emphasize some particularities of sound that are difficult to analyze directly from the waveform.

In fact, some acoustical techniques implemented in this research embody modules for synthesizing audio as a region-dependent approach. Therefore, a set of features that better characterize the sound according our purposes is extracted from the audio streaming segmented in overlapping frames.

A number of features used in the methods developed in this research are described in the next subsection.

3.7.1.1 Features

This group of features relies on the audio characteristics extracted directly from frame-by-frame analysis of the audio signal, over time.

The set of features (and their acronyms) used in this study are presented next.

- ***Overall sound pressure level, LAeqFull***. This is a measure of the total energy of the sound effectively perceived by people, in the range of 200 to 8000Hz.
- ***Narrow band sound pressure level***: low band – *LAeqLo* (200 - 1400 Hz), middle band – *LAeqMid* (1400 - 3000 Hz) and high band – *LAeqHi* (3000 - 8000 Hz).

  The above two types of features, after A-weighting and filtering, are calculated by using the short time average of energy according to

$$LAeqX_i = \frac{1}{W}\sum_{i=1}^{W} s_A{}^2(i)w(n-i) \quad , \tag{3.29}$$

where *X* denotes Full, Lo, Mid or Hi suffixes, *w* is the Hanning window of length *W*, and $s_A(i)$ is the noise signal (after A-weighting), in the *i*th frame.

- ***Spectral Flatness***. This is the ratio between the geometrical mean and the arithmetical mean of the spectrum magnitude.

$$SFM = 10\log_{10}\frac{\left(\prod_{k=1}^{N/2} G_p(e^{j\frac{2\pi k}{N}})\right)^{\frac{1}{N/2}}}{\frac{1}{N/2}\sum_{k=1}^{N/2} G_p(e^{j\frac{2\pi k}{N}})} \tag{3.30}$$

where $G_p(e^{j\frac{2\pi k}{N}})$ is the spectral power density calculated on the basis of an *N*-point Fast Fourier Transform.

- ***Cepstral coefficients, CC***. This is a very convenient way to model spectral energy distribution

$$c(k) = IDFT\left\{\log_{10}\left|DFT\left\{x(n)\right\}\right|\right\} \tag{3.31}$$

where DFT denotes the Fourier-transform and IDFT its inverse.

Although all coefficients model the precise spectrum, only the first *M* cepstrum coefficients are used as features (the coarse spectral shape is modeled by the first coefficients). The first coefficient, corresponding to the energy of the frame, is usually discarded. The precision (envelope) is selected by the number of coefficients taken. Usually, $M = f_s/$ (2000 Hz) is a good first guess for *M*. However, a drawback remains due to the linear frequency scale used, e.g. the frequency ranges 100–200Hz and 10–20 kHz should be approximately equally important in a perceptual sense, though the standard cepstral coefficients do not take this into account.

This feature can be further modeled by calculating the mean value and variance of each coefficient over time.

- ***Mel-frequency cepstral coefficients,* MFCC**. This is estimated by first computing the power of each Mel scale filter bank, by using a DFT, followed by the DCT calculation. It corresponds to a refinement of the cepstral coefficients algorithm to account for the logarithmic-based frequency scale. In fact, this feature approximates better the constraints of perception of human auditory system. Moreover, the MFCC are successful because, by using the Mel-frequency scale and the logarithmic of the magnitudes, a large change in MFCC vector implies large perceptual change and the use of the DCT, which drops spectral fine structure, decorrelates the features easily. Therefore, it is an improvement over the standard cepstral coefficients.

  This is another way to model the spectral energy distribution by a low-order all-pole filter.

- ***Spectral centroid, Ct***. This is the barycentre, or centre of mass, of the spectrum. It is also associated to the subjective concept of brightness of the sound. The computation of this feature is based on the magnitude of spectrum, using the STFT, performed frame-by-frame along the time axis [Li01]. The spectral centroid of the *i*th frame is defined as

$$Ct_i = \frac{\sum_{u=0}^{M} u \left| f_i(u) \right|}{\sum_{u=0}^{M} \left| f_i(u) \right|}, \tag{3.32}$$

  where M is the index of the highest frequency band and $f_i(u)$ represents the STFT of the *i*th frame.

- ***Spectral Energy Centroid, SEC***. This is the same as Ct where its computation uses the spectral energy, instead of the magnitude of the spectrum.

  The use of the squared magnitude of the spectrum, a measure of the noise energy, is aimed at a better correlation with the stimulus perceived by the human auditory system. This feature correlates with brightness subjective parameter.

- ***Band spectral energy centroid***: low band - *CtLo*, middle band - *CtMid* and high band – CtHi. These features are computed in the same way as Ct with the sound signal band filtered.

- ***Bandwidth, BW*** – the bandwidth weighted by the Spectral Centroid.

$$BW = \sqrt{\left(\sum_{u=1}^{M}(u-SEC)^2\left|f_i(u)\right|^2\right)/\left(\sum_{u=1}^{M}\left|f_i(u)\right|^2\right)} \tag{3.33}$$

- ***Modulation FM_AM-– TVATVF***. This feature accounts for the amplitude and frequency modulation produced essentially on the steady state sounds due to creative artistic expressivity. In the musical field, it is called tremolo and vibrato audio effects, respectively. *FM* and *AM* nearly always coexist in all musical instruments in a way that it is impossible to have frequency vibrato without amplitude vibrato but not vice versa. However, while the frequency modulation usually affects the full spectral contents of a musical note (harmonics or partials) as a whole, the way the amplitude modulation acts on the spectral components of sounds can be source-dependent (singing voice or musical instruments).

  Although a number of techniques are available for detection of vibrato effect the feature of Liu's [Liu05], which is based on the Time-Varying Spectrum (*TVS*) of the sound obtained from the energy density spectrum, is often used. The estimated amplitude of the current harmonic is computed from the total energy of its energy density spectrum, and the frequency is taken from the expectation value with respect to its energy density spectrum. The following adaptation of the so called Time-Varying Amplitude (*TVA*), and the Time-Varying Frequency (*TVF*) for the *j*th harmonic at the *m*th time frame is given by:

$$TVA_n^m = \sqrt{\sum_j \left|S^m(f_j)\right|^2} \tag{3.34}$$

$$TVF_n^m = \sum_j f_j \left|S^m(f_j)\right|^2 \Big/ \sum_j \left|S^m(f_j)\right|^2 \tag{3.35}$$

where j takes values between 0 and 6, $S^m(f_j)$ is the spectrum of the tonal component at frequency f_j (being f_0 the fundamental), for the frame m.

- ***Spectral Energy Maximum, FreqMax1LAeq***. This feature corresponds to the main peak of the power spectral density estimate. For example, the tire/road pavement noise spectrum profile is characterized often by a maximum sound intensity centered in the range of about 800 to 1800Hz.

- ***Secondary Spectral Energy Maximum, FreqMax2LAeq***. This feature corresponds to the secondary peak of the smoothed power spectral density of a stochastic signal. It is useful if a prominent secondary peak appears on the noise spectrum.

  These two last features are evaluated by averaging N consecutive frames (moving average method, with N=2) in the frequency domain in

$$FreqMaxXLAeq_i = \max\left\{\frac{1}{2N+1}\sum_{k=i-N}^{i+N}\left|f_k(u)\right|^2\right\}, \tag{3.36}$$

  with X = 2, FreqMax2LAeq, for the Secondary Spectral Energy Maximum, and where i is the current frame.

- ***Maximum spectral energy modulation index, FreqModIndex***. This feature is found relevant in the study of road pavement surfaces, where it was observed that, for some types of pavements, the secondary spectral energy maximum feature shows a kind of frequency modulation. This feature is applied to the *FreqMax2LAeq* feature, providing information on the frequency drift of the secondary spectral energy maximum. This feature, estimated by the derivative of the *FreqMaxLAeq2*, is relevant for thin layer pavement surfaces analyzed in near field conditions (e.g. with the close proximity method), as an indicator of the pavement quality.

- ***Third octave band sound pressure levels, Oct13B***.

- ***Delta spectrum magnitude, ΔSpecMag***. For each feature analyzed in the frequency domain, the energy difference in consecutive frequency bands in the same frame is computed.

3.7.1.2 Onset analysis

Some of the acoustical measurement techniques proposed in this thesis, such as those used with the peripheral auditory system, need to preserve the transient parts of the music signal in order not to introduce sound artifacts. Anticipating the concept behind these techniques, the music signal performs a masking mechanism so that the test signals are inaudible and therefore, care should be taken in order to not degrade the sound quality of the resulting musical piece significantly. In fact, the initial and the ending parts of a sound, the transient zones, especially the initial part, play an important role in the identification of the musical instruments and therefore, are responsible for the perception of some artifacts in the musical structure. A scheme of the transient part is illustrated in Figure 3.15.

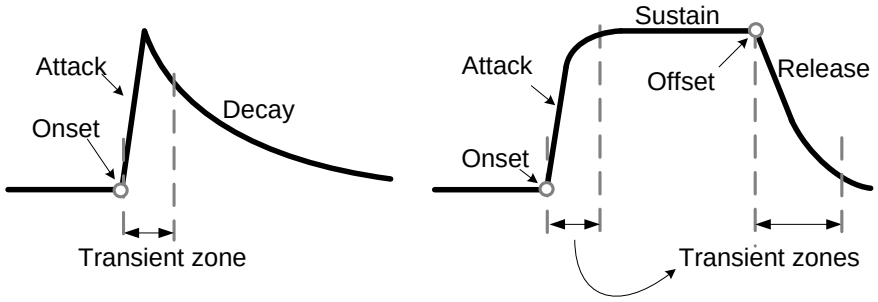


Figure 3.15 – Scheme of the transient zones of a sound including the onset, attack, release and decay of (left side) a single musical note without sustain and (right side) a prolonged single musical note.

Briefly, these notions can be differentiated as follow: attack/release - the zone of the waveform for which the amplitude envelope increases/decreases; onset and offset – the precise instant chosen to initiate the note or to finalize the sustained (steady state of the sound) part, i.e. starting point of the transient parts of the note; decay or release have similar meaning, however they are used in different situations. Release stands for sustained notes and decay for impulsive (fast) notes; transient zones are the periods of time related to the attack and release or impulsive parts of the sound. Therefore, it is considered the non stationary part of the note.

For the current purposes, the transient parts correspond only to the starting points, onsets, of the steady-state zone of the tonal sounds. The ending points, offsets, are not considered in the first place.

Onsets, sound inflections and event detection are related to a large number of signal parameter variations such as amplitude, spectral brightness, spectral standard variation, spectral slope, spectral onset slope, roll-off point, and spectral envelope. These parameters are not fully independent and some of them are highly correlated.

A good mechanism, to detect the onset making annotations in the waveform, is a helpful tool to improve the techniques used to mask the test signals. The procedure used in the majority of the onset detection algorithms consists of pre-processing the audio signal to improve the performance of the subsequent stages by applying some transformation in order to emphasize or attenuate various aspects of the signal according to their relevance to the task in hand. Some of these transformations are related to splitting the signal into multiple frequency bands, and transient/steady-state separation [Verma97]. The analysis ends with a detection function, derived at a lower sampling rate, to which a peak-picking algorithm is applied to locate the onsets.

A variety of onset detection techniques have been proposed by several authors. However, they are, in general, very dependent of the music content. Therefore, a good choice lies in the detection function suggested by Bello et al. [Bello04a] due to its ability on deal with a wider variety of signal types. The complex detection function $df_c(n)$ combines the energy and phase based on the onset detection as shown in the expression:

$$df_c(n) = \frac{1}{N}\sum_{k=0}^{N}\left(\left\|\tilde{X}_k(n) - X_k(n)\right\|\right)^2 \qquad (3.37)$$

.

It computes the complex spectral difference between the current frame $X_k(n)$ of a STFT and a predicted target frame, $\hat{X}_k(n)$. Energy change or deviations in phase acceleration produce peaks on the detection function. These deviations occur in transients or during pitch changes (often called tonal onsets, where no perceptual energy change is observed) enabling the detection of onsets in a wider range of signals. A more comprehensive description and discussion can be found in [Bello04a].

Chapter 4

SIGNAL-TO-NOISE RATIO IMPROVEMENT

4.1 Introduction

The accurate estimation of the acoustical parameters of enclosures used for live performances, such as theatres, concert halls, conference rooms, sport stadiums and other public areas requires the measurements to be carried out in the presence of an audience. However, for reasons of possible annoyance, people are usually kept out and a correction factor related to the equivalent sound absorption of the audience area is applied. Moreover, the variability of some relevant parameters, such as the relative humidity and the temperature, that occurs during the event and may have influence on the room frequency response, is not usually considered.

The accurate estimation of the Room Impulse Response (RIR) requires large signal-to-noise ratios (SNR) [Rife89, Vorlander97b, Borish83, Fausti98]. However, in the case of high background noise levels, the SNR can be very low. Therefore, the measurements for sound insulation, reverberation time (RT), and magnitude frequency response as well as for sound quality indexes such as C80, STI, EDT and IACC may yield unreliable results, even when an averaging technique is used [Farina00].

This research assumes that the enclosure to be tested is modeled by a Linear Time-Invariant System (LTI) corrupted by additive noise. This is a good approach to most *in situ* testing

situations. Music will be used as the disturbing signal here, owing to this specific approach. The SNR can be obtained from

$$\mathrm{SNR} = \frac{\mathrm{MS}\big(y(n)\big)}{\mathrm{MS}\big(d(n)\big)}, \tag{4.1}$$

where $y(n)$ is the test signal filtered by the room, $s(n)*h(n)$, the $d(n)$ stands for additive noise and the $res(n) = y(n)+d(n)$ is the captured signal.

The noise inside the enclosure under test can be characterized, in a first approach, as (i) a quasi-stationary process with quite lengthy intervals, for instance transportation noise or industrial noise, or (ii) non-stationary noise showing an intermittent behavior that changes both in time and in frequency with pronounced energy fluctuations.

The research reported here deals with situations of high sound levels and non-stationary noise and examines the acoustic energy both in time and in frequency.

In fact, music sound as background noise shows a non-stationary behavior, changing rapidly over time and frequency, as illustrated in Figure 4.1. Therefore, the acoustical testing methods should consider this issue examining the acoustic energy both in time and frequency in order to further increase the SNR.

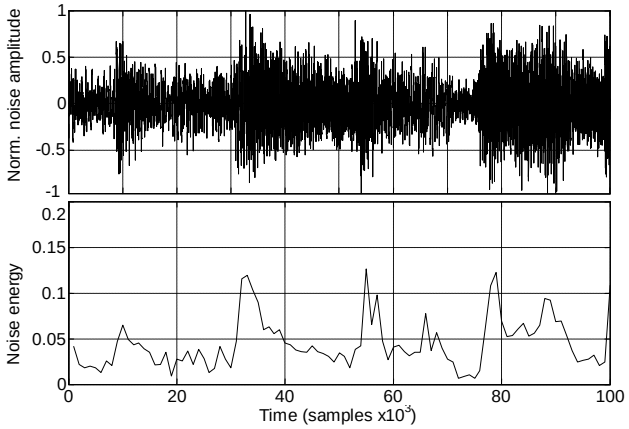


Figure 4.1 – Instantaneous noise amplitude (top image) and related MS value (bottom image).

In this sense, the analysis was conducted by considering the signal energy for different time slots [Nielsen97, Paulo03].

Since swept sine and MLS are deterministic signals, synchronous standard time averaging technique can be applied in order to improve the overall SNR_{av}

$$SNR_{av} = \frac{MS(y(n))}{MS\left(\frac{1}{M}\sum_{i=0}^{M-1} N_i(n)\right)} \tag{4.2}$$

where M stands for the number of test frames used.

When the background noise (music program) is added to the test signal, the noise source cannot be controlled by the technician. If the level of the non-stationary background noise is high, the MS of the overall frame must be minimized for the RIR measurement.

The averaging technique performs a summation, where the correlated signals (the test signals) are amplified and the uncorrelated disturbance noise components are attenuated. The increment of the SNR resulting by averaging the synchronized M sequences is given as $\Delta SNR = 10\log_{10}M$. Therefore, by doubling the number of the sequences averaged, an increase of 3dB for the SNR can be expected. This result is valid for any type of noise assuming low correlation between different frames. For music signals used as noise, where a certain correlation exists between the captured frames, the SNR increment rises asymptotically to the theoretical value when the number of averages is increased. Nevertheless, the measurement times may increase substantially when averaging techniques are used. For example, if an increase of 20dB for the SNR is required, 100 sequences need to be sent to the room, which comes to a total time of about 3.5 minutes for a 2 seconds frame. However, if this SNR increment is not sufficient and a supplementary increase of 10dB is required, the measurement time will expand to 35 minutes (1000 frames).

Moreover, time-variance issues occur frequently during acoustic measurements. Assuming validity of the LTI model, whenever subtle time variance problems occur, a balance between the number of averages required and the increase of the MLS frame length should be considered [Borish83]. In this case, the length of the test signal frame sequences should be increased (doubling the frame length leads to an increase of the SNR by 3dB) when the inter-sequence time variance (between different sequences) is more visible than the intra-sequence time variance (during the same sequence). However, for a long frame length, the intra-sequence

time variance could prevail against inter-sequence time variance. MLS signals are more affected than swept sine signals by the time-variance phenomenon. Therefore, a comprehensive optimization of the acoustic measurement set-up is recommended and a pre-emphasis procedure should be applied.

The next sections describe averaging-based approaches designed to increase the SNR in situations of non-stationary long-term high sound levels.

4.2 Weighted Average technique - WAve

Instead of applying the averaging technique directly to the frames, a weight factor which depends of the energy of each frame is applied:

$$\mathrm{SNR}_{av} = \frac{\mathrm{MS}(y(n))}{\mathrm{MS}\left(\frac{1}{W_s}\sum_{i=0}^{M-1} l_i\, N_i(n)\right)} \tag{4.3}$$

where l_i is a weight factor and W_s is the normalization factor. This procedure improves the SNR even further, for some kinds of disturbances [Nielsen97].

4.3 Time-Spectral Hybrid Technique

The room being assessed is excited by a set of M frames. An entire set of frames is sent to the room one after the other or with delays between them. The purpose of the latter procedure is to keep the ambient noise statistics as different as possible during the presence of each frame. However, in this case, each time the test signal is applied to the room at least two concatenated sequences are sent. This procedure assures that the second frame holds all of the room's acoustical information. The first sequence of each set should be discarded (this sequence does not hold the full steady-state part of the response). Obviously, this procedure is of less importance for frame lengths several times greater than the expected RT.

For high level non-stationary background noise, the mean square, MS, value of the resulting captured signal (a measure of the acoustic energy of the sequence filtered by the room plus noise) must be kept to a minimum in order to exclude further noise energy components [Paulo03, Paulo08].

A separate description in time and in frequency is presented next for a better understanding of the technique.

4.3.1.1 Segmented MLS technique

This method consists of splitting each captured test signal frame into *N* segments followed by the estimation of the MS value for each windowed segment [Paulo05a, Paulo05b]. From each transmitted sequence, the segments with the lowest energy are extracted and a new sequence is built by using the overlap-add method. Figure 4.2 describes this arrangement.

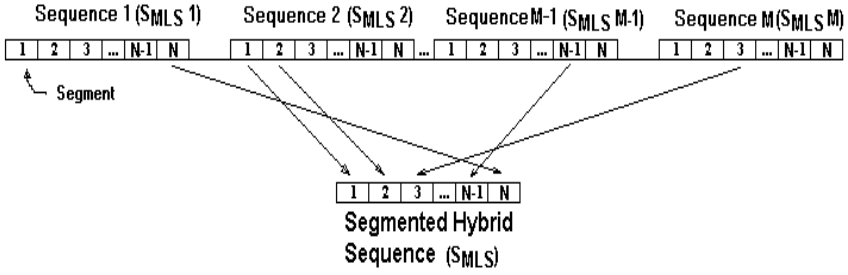


Figure 4.2 – Implementation of the Segmented MLS technique.

This procedure acts as a noise scrambler, decorrelating the noise added to each new frame due to the contribution of segments originated in different frames, at different times. The resulting sequence corresponds to the lowest energy Seq_{MLS}, thus ensuring the highest SNR value.

The ML sequence is segmented by applying the time windowing method,

$$S_{MLS}(n) = \sum_{k=1}^{N} Seg_{MLS_K}(n) = \sum_{k=1}^{N} S_{MLS}(n)\ \ w(n - kL/N + 1 \quad , \tag{4.4}$$

with $$w(n) = \begin{cases} 1 & 0 \le n \le L/N \\ 0 & n > L/N \end{cases},$$

where N is the number of segments in each sequence, L is the sequence length and $L_{seg} = (L+1)/N$ is the length of each segment, except for the last one that is equal to $[(L+1)/N]$-1, due to the odd number of samples of each ML sequence. Although a rectangular window shape is featured in Eq. (4.4) for simplicity's sake, other types of window shapes are commonly used, to attenuate the leakage effect, with a 50% overlap. The response of the room to an MLS signal is given by

$$y(n) = \sum_{k=1}^{N} y_k(n) \tag{4.5}$$

where $y_k(n) = y(n)\,w(n - kL/N + 1 \;\Rightarrow\; [S_{MLS}(n) * h(n)]\,w(n - kL/N + 1$

Considering the influence of the disturbing ambient noise, it follows that

$$Res(n) = \sum_{k=1}^{N} \hat{Res}_k(n) \tag{4.6}$$

where

$$\hat{Res}_k(n) = Res(n)\, w(n - kL/N + 1) = \left[S_{MLS}(n) * h(n) + N(n) \right] w(n - kL/N + 1) \tag{4.7}$$

The SNR is calculated after the estimation of the MS value for each segment

$$\mathrm{SNR}_{Seg} = \frac{\mathrm{MS}(y_1(n)) + \mathrm{MS}(y_2(n)) + \ldots + \mathrm{MS}(y_N(n))}{\mathrm{MS}(N_1(n)) + \mathrm{MS}(N_2(n)) + \ldots + \mathrm{MS}(N_N(n))} = \frac{\mathrm{MS}(y(n))}{\sum_{k=1}^{N} \mathrm{MS}(N_k(n))} . \tag{4.8}$$

The overall SNR_{Seg} is maximized by minimizing the denominator of Eq. (4.8). The MS value of the $y(n)$ is proportional to the acoustical power of the test signal used to excite the room. This value remains the same for the whole set of sequences. For this reason, this value was kept constant in order to compare and evaluate the performance of the Segmented MLS technique against the standard MLS technique.

Hence, the segments with the lowest MS value among the whole set of captured sequences are selected, and,

$$N_k(n) = \min\left\{ \mathrm{MS}(N_i(n)) \right\} \qquad i \in [1, M] . \tag{4.9}$$

This method shows an improvement, ΔSNR, when compared to the standard technique, given by

$$\Delta\mathrm{SNR} = \frac{\mathrm{SNR}_{Seg}}{\mathrm{SNR}_{Stand}} = \frac{\mathrm{MS}(N(n))}{\sum_{k=1}^{N} \mathrm{MS}(N_k(n))} \geq 1 \tag{4.10}$$

where $N_k(n)$ is defined in Eq. (4.9).

4.3.1.2 Spectral MLS technique

This method is an enhancement of the Segmented MLS technique, by considering the analysis by frequency bands. At the receiver point, a filter bank is applied to each sequence by splitting the audio spectrum into multiple complementary sub-bands. The energy within each frequency

sub-band is then computed. The sub-bands with the lowest spectral energy are chosen to compose a new sequence. This resulting sequence will have the highest SNR, assuming the energy of the sequences sent is kept constant. Figure 4.3 illustrates this procedure.

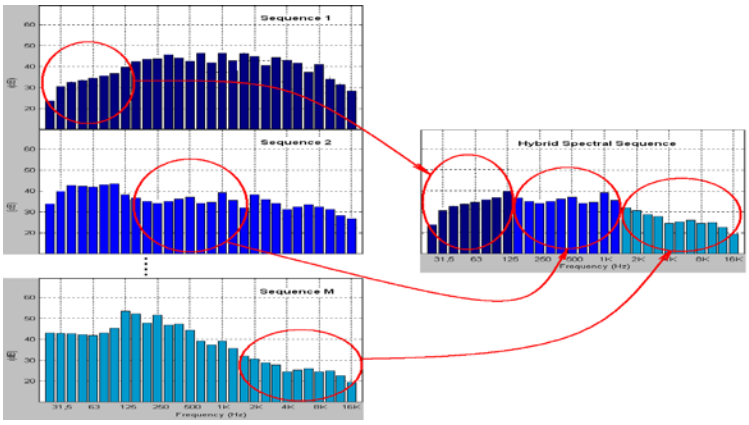


Figure 4.3 – Spectral MLS technique procedure for composing the sequence with the lowest sub-band energy.

The MS value of the system response $y(n)$ is constant for the whole set of sequences (the acoustical energy of the signal excitation is kept constant for all the ML sequences). Then, Eq. (4.5) yields

$$\mathrm{SNR}_{Spect} = \frac{\mathrm{MS}(y_1(n)) + \mathrm{MS}(y_2(n)) + \ldots + \mathrm{MS}(y_B(n))}{\mathrm{MS}(N_1(n)) + \mathrm{MS}(N_2(n)) + \ldots + \mathrm{MS}(N_B(n))} = \frac{\mathrm{MS}(y(n))}{\sum_{j=1}^{B} \mathrm{MS}(N_j(n))} \tag{4.11}$$

The maximum value of SNR_{Spect} in Eq. (4.11) is obtained by minimizing the MS value of the noise $N_k(n)$ within each sub-band,

$$N_j(n) = \min\left\{\mathrm{MS}(N_i(n))\right\} \qquad i \in [1, M]\,. \tag{4.12}$$

This new method leads to an increase in the SNR

$$\Delta\mathrm{SNR} = \frac{\mathrm{SNR}_{Spect}}{\mathrm{SNR}_{Stand}} = \frac{\mathrm{MS}(N(n))}{\sum_{j=1}^{B} \mathrm{MS}(N_j(n))} \geq 1 \tag{4.13}$$

where $N_j(n)$ is defined by Eq. (4.12), and SNR_{Spect} and SNR_{Stand} are the SNR of the Spectral MLS technique and the standard MLS technique, respectively.

The Time-Spectral Hybrid technique corresponds to the combination of the Segmented MLS and Spectral MLS methods. This technique is viewed as an acoustic energy time/frequency map split by cells corresponding to the windowed segments analyzed in frequency bands. The assembly of new MLS frames by choosing the cells belonging to deeper valleys (time/frequency segments with lower energy) yields better results than the standard MLS technique for high levels and non-stationary noise profiles. This result is closely related to the noise profile and time-variance occurrence, though to achieve a pre-defined SNR or a magnitude frequency response error, fewer MLS frames are required for the averaging process.

In this technique, from each set of *M* sequences a new one is built up provided it has the lowest energy. However, this may lead to worse results than those exhibited by the standard averaging technique (by summing up all the sequences) in situations where several frames or homologous segments (segments corresponding to the same position in different frames) share low energy noise. To improve the performance of the Hybrid MLS technique, a threshold level is applied to the energy cells as a decision-making rule for rejecting segments. All the segments showing energy lower than this threshold are used to build new Hybrid MLS frames, thus increasing the overall SNR with the application of the averaging technique. Replication of the segments in the case of incomplete frames is possible. Although the replicated segments do not contribute to increase the SNR, there are others with different noise profiles. The increase in SNR depends of the ratio of the different segments to the replicated segments. Therefore, the final SNR strictly depends on each particular situation.

The complete block diagram for the Hybrid MLS technique is depicted in Figure 4.4.

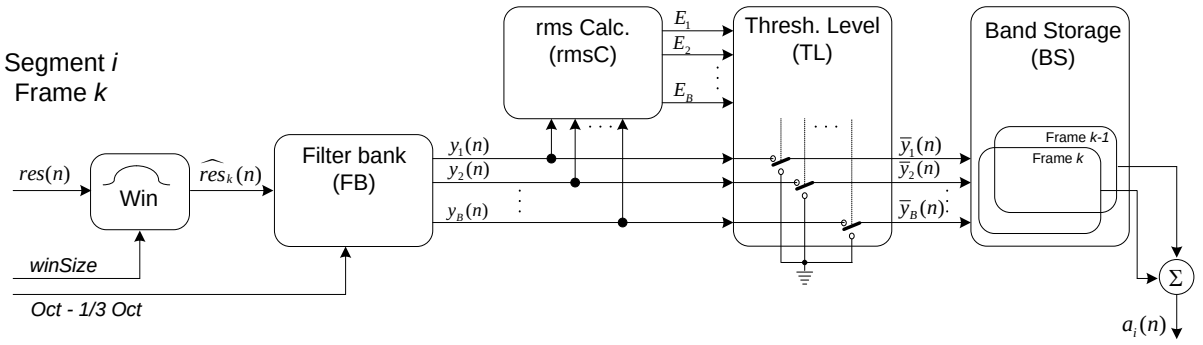


Figure 4.4 – Complete block diagram for the Hybrid MLS technique for the analysis/processing of a segment and accumulation, $a_i(n)$ with the homologous segment from the previous frame.

Although the SNR is usually increased by applying this procedure, the acoustic measurements could be very time consuming to achieve a predefined SNR value, depending of the noise statistics. In fact, since the resulting frame is built from a set of segments, in time and/or frequency, selected from different test signal frames, the performance of the technique could be low.

As an attempt to overcome the Time-Spectral Hybrid technique drawback, different approaches using weighting energy of the segments can be attempted, as described in the next sections.

Moreover, although in the early phase of research of the Time-Spectral Hybrid technique, only the MLS technique was considered on the evaluation, this concept can be applied to other measurement techniques based on test signal frames, such as the swept sine signal.

4.4 Segmented Weighted Average technique - SWAve

This technique consists of splitting each captured frame into *N* segments and applying a weight factor to each segment before averaging [Paulo05a]. Figure 4.5 depicts the procedure schematically.

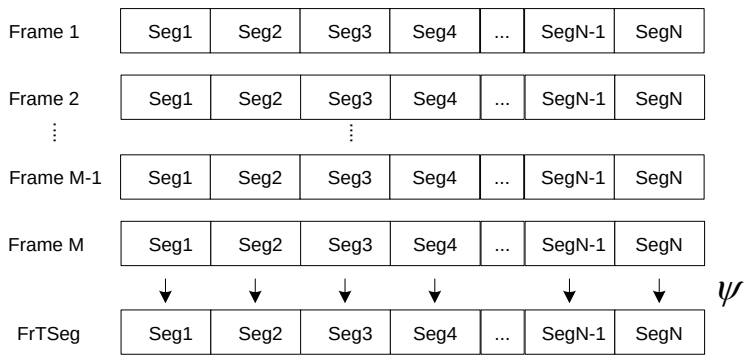


Figure 4.5 – Time Segmented technique implementation; the resulting frame FrTSeg corresponds to the weighting average,ψ, of a set of M fames.

The average technique is applied to each column, weighting each segment according to the ψ operator that depends of the output of the system, *res*(*n*). This function consists of applying a weight factor to each segment according to its MS value. In order to improve the SNR, the segments with the larger MS value will be more attenuated and the segments with the lower MS value will be more emphasized. Therefore, the segment with the lowest energy, less corrupted by the noise, will be more valued. At the end of the process, a new sequence is

generated by the overlap/add method (assuming that non-rectangular overlapping windows are used on the windowing procedure) of the weighted averaged segments. This procedure, corresponding to a segment weighting average method, ensures that the resulting frame yields the highest overall SNR.

The $\psi\{Res_k(n)\}$ operator expresses the accumulation, a_k, of the homologue segments, for the column k, defined by

$$a_k(n) = \psi\left\{Res_k(n)\right\} = \frac{1}{W_k}\sum_{i=1}^{M} l_{ik} Res_{ik}(n) = \frac{1}{W_k}\sum_{i=1}^{M} l_{ik}\left(y_{ik}(n) + N_{ik}(n)\right) \tag{4.14}$$

where $l_{ik} = \dfrac{1}{\mathrm{MS}(Res_{ik}(n))}$ is the weight factor for the segment k and frame i, is the normalization factor for the column k and $N_{ik}(n)$ is the disturbance signal belonging to the k segments. $W_k = \sum_{i=1}^{M} l_{ik}$

Assuming that the energy of $y(n)$ remains constant for the set of segments in the same column as in Figure 4.5, Eq. (4.14) can be rewritten as follows,

$$a_k(n) = y_k(n) + \frac{1}{W_k}\sum_{i=1}^{M} l_{ik} N_{ik}(n) = a(n)w[n-(k-1)L_{seg}] \quad , \tag{4.15}$$

where $a(n)$ is the complete resulting frame, $w(n)$ is the window used to analyze each segment and L_{seg} is segment length, in samples, defined by

$$a(n) = FrTS\ \mathbf{g}(n) = \sum_{k=1}^{N} a_k\left[n-(k-1)L_{seg}\right]$$

$$= \sum_{k=1}^{N}\left(y_k(n) + \frac{1}{W_k}\sum_{i=1}^{M} l_{ik} N_{ik}(n)\right) = y(n) + \sum_{k=1}^{N}\left(\frac{1}{W_k}\sum_{i=1}^{M} l_{ik} N_{ik}(n)\right)$$

$$= y(n) + \sum_{k=1}^{N}\sum_{i=1}^{M}\frac{l_{ik} N_{ik}(n)}{W_k} = \sum_{k=1}^{N}\frac{\mathbf{L}_k \mathbf{N}_k(n)}{W_k} \tag{4.16}$$

where $\mathbf{L}_k = [l_{1k},\ l_{2k},.. l_{Mk}]$ and $\mathbf{N}_k(n) = [N_{1k}(n),\ N_{2k}(n),.\ N_{Mk}(n)]^{\mathrm{T}}$

The MS value of the noise component is given by

$$\mathrm{MS}(N'(n)) = \mathrm{MS}\left(\sum_{k=1}^{N}\sum_{i=1}^{M}\frac{l_{ik}N_{ik}(n)}{W_k}\right) = \mathrm{MS}\left(\sum_{i=1}^{M}\sum_{k=1}^{N}\frac{l_{ik}N_{ik}(n)}{W_k}\right) \tag{4.17}$$

Assuming a low correlation of the noise between consecutive frames (and segments), then

$$\simeq \sum_{i=1}^{M}\mathrm{MS}\underbrace{\left(\sum_{k=1}^{N}\frac{l_{ik}N_{ik}(n)}{W_k}\right)}_{\text{frame } i} \simeq \sum_{i=1}^{M}\sum_{k=1}^{N}\mathrm{MS}\left(\frac{l_{ik}N_{ik}(n)}{W_k}\right) = \sum_{i=1}^{M}\sum_{k=1}^{N}\frac{\mathrm{MS}\left(l_{ik}N_{ik}(n)\right)}{W_k^2}$$

$$= \sum_{i=1}^{M}\frac{\mathrm{MS}\left(l_L(n)N_i(n)\right)}{W^2} = \frac{M}{W^2}\mathrm{MS}\left(l_L(n)N(n)\right) \tag{4.18}$$

with $W^2 = \sum_{k=1}^{N} W_k^2$ and $l_L(n) = l_{i[n/L_{seg}]}$. The MS value of $y(n)$ is proportional to the acoustic power of the excitation signal applied to the room which is kept constant for the set of the M frames. Hence, the MS value should be normalized in order to compare the performance of the Segment Weighted Average against the standard average technique for the swept sine or MLS methods. The SNR_{FrTSeg} is finally estimated by

$$\mathrm{SNR}_{av} = \frac{W^2}{M}\frac{\mathrm{MS}\left(y(n)\right)}{\mathrm{MS}\left(l_L(n)N(n)\right)} \tag{4.19}$$

When compared with the standard technique, the SNR improvement is expressed by

$$\Delta\mathrm{SNR} = \frac{\mathrm{SNR}_{FrTSeg}}{\mathrm{SNR}_{Stand}} = \frac{W^2}{M^2}\frac{\mathrm{MS}\left(N(n)\right)}{\mathrm{MS}\left(l_N(n)N(n)\right)} \geq 1 \tag{4.20}$$

.

A more accurate estimation of the ΔSNR can be obtained if the noise statistics are well known.

4.5 Spectral Segmented Weighted Average technique - SSWAve

This method consists of splitting each captured frame into N segments and filtering each segment by a filter bank, which splits the audio spectrum into multiple complementary sub-bands. The energy within each sub-band is then computed [Paulo05b]. A scheme of this procedure is shown in Figure 4.6.

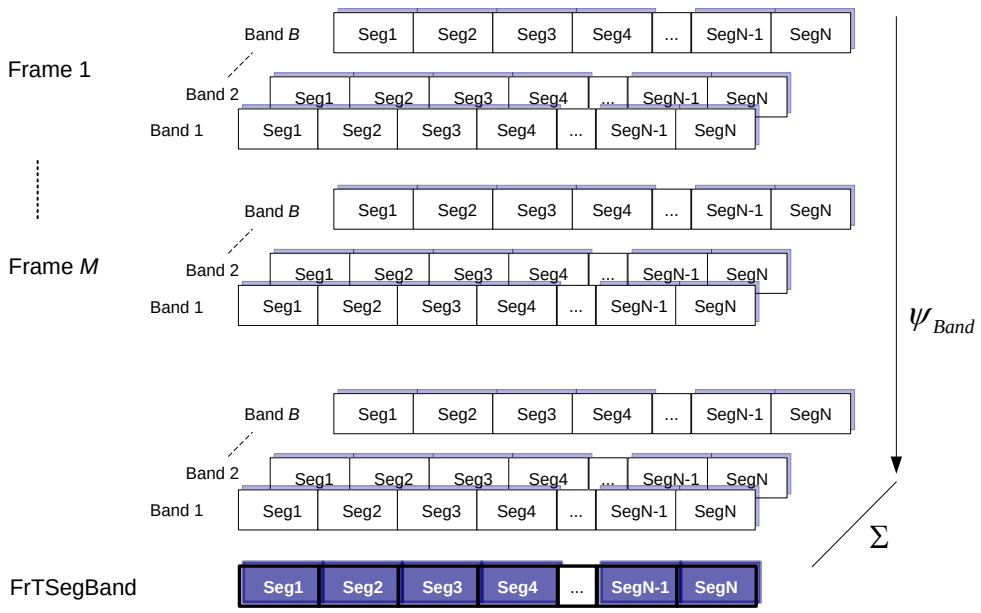


Figure 4.6 – Spectral Segmented technique implementation; the resulting frame FrTSegBand corresponds to weighting average, ψ_{Band}, of a set of *M* fames.

The average technique is applied to the weighted version of each segment band according to the ψ_{Band} operator. This operator applies a weight factor to each segment related to its MS value. In order to improve the SNR, the segment with the larger MS value, for the same frequency band, will be more attenuated and the segment band with the lower MS value will be more valued. At the end of the process, a new sequence is built up by the summation of all bands corresponding to the same segment and concatenation of the weighted averaged segments.

This procedure is almost identical to SWAve with the weighted operator ψ (energy of the overall signal levels) replaced by a weighted operator applied to each frequency band.

A $\psi_{Band}\{Res_k(n)\}$ operator, expressing the accumulation, a_{kj}, of the segment bands for the column *k* and frequency band *j* can be defined by

$$a_{kj}(n) = \psi_{band}\left\{Res_{kj}(n)\right\} = \frac{1}{W_{kj}}\sum_{i=1}^{M} l_{ikj} Res_{ikj}(n) = \frac{1}{W_{kj}}\sum_{i=1}^{M} l_{ikj}\left(y_{ikj}(n) + N_{ikj}(n)\right) \tag{4.21}$$

where $l_{ikj} = \dfrac{1}{\mathrm{MS}(Res_{ikj}(n))}$ is the weight factor, $W_{kj} = \sum_{i=1}^{M} l_{ikj}$ is the normalization factor and $N_{ikj}(n)$ is the disturbance signal, for the segment *k*, band *j* and frame *i*.

Assuming that the energy of $y(n)$ remains constant for the set of segments of the same frequency band (in the same column, see Figure 4.6), Eq. (4.21) can be rewritten as

$$a_{kj}(n) = y_{kj}(n) + \frac{1}{W_{kj}} \sum_{i=1}^{M} l_{ikj} N_{ikj}(n) = a_j(n) w[n-(k-1)L_{seg}] \quad , \tag{4.22}$$

where $a_{kj}(n)$ is the complete resulting frame for a particular frequency band, $w(n)$ is the window used to analyze each segment and L_{seg} is segment length in samples, defined by

$$a_k(n) = \sum_{j=1}^{B} a_{kj}(n) \, . \tag{4.23}$$

Finally, the complete resulting frame is given by

$$\begin{aligned} a(n) = FrTSegBand(n) &= \sum_{k=1}^{N} a_k \left(n-(k-1)L_{seg} \right) \\ &= \sum_{k=1}^{N} \left\{ \sum_{j=1}^{B} \left(y_{kj}(n) + \frac{1}{W_{kj}} \sum_{i=1}^{M} l_{ikj} N_{ikj}(n) \right) \right\} = y(n) + \sum_{k=1}^{N} \sum_{j=1}^{B} \left(\frac{1}{W_{kj}} \sum_{i=1}^{M} l_{ikj} N_{ikj}(n) \right) \\ &= y(n) + \sum_{j=1}^{B} \sum_{k=1}^{N} \sum_{i=1}^{M} \frac{l_{ikj} N_{ikj}(n)}{W_{kj}} \end{aligned} \tag{4.24}$$

The MS value of the noise component is given by

$$\mathrm{MS}(N'(n)) = \mathrm{MS}\left(\sum_{j=1}^{B} \sum_{k=1}^{N} \sum_{i=1}^{M} \frac{l_{ikj} N_{ikj}(n)}{W_{kj}} \right) = \mathrm{MS}\left(\sum_{j=1}^{B} \sum_{k=1}^{M} \sum_{i=1}^{N} \frac{l_{ikj} N_{ikj}(n)}{W_{kj}} \right) \tag{4.25}$$

Assuming low correlation between noise in consecutive segments,

$$\begin{aligned} &\simeq \sum_{j=1}^{B} \sum_{i=1}^{M} \mathrm{MS} \underbrace{\left(\sum_{k=1}^{N} \frac{l_{ik} N_{ik}(n)}{W_{kj}} \right)}_{\text{Frame } i} \simeq \sum_{j=1}^{B} \sum_{i=1}^{M} \sum_{k=1}^{N} \mathrm{MS}\left(\frac{l_{ikj} N_{ikj}(n)}{W_{kj}} \right) = \sum_{j=1}^{B} \sum_{i=1}^{M} \sum_{k=1}^{N} \frac{\mathrm{MS}\left(l_{ikj} N_{ikj}(n) \right)}{W_{kj}^2} \\ &= \sum_{j=1}^{B} \sum_{i=1}^{M} \frac{\mathrm{MS}\left(l_{LB}(n) N_{ij}(n) \right)}{W_j^2} = \frac{M}{\sum_{j=1}^{B} W_j^2} \mathrm{MS}\left(l_{LB}(n) N(n) \right) \end{aligned} \tag{4.26}$$

with $l_{LB}(n) = l_{i[n/L_{seg}]j}$ for band j. The $\mathrm{SNR}_{FrTSegBand}$ is finally estimated by

$$\mathrm{SNR}_{av} = \frac{\sum_{j=1}^{B} W_j^2}{M} \frac{\mathrm{MS}(\mathrm{y}(n))}{\mathrm{MS}(l_{LB}(n)N(n))} \tag{4.27}$$

When compared to the standard technique, the SNR improvement is expressed by

$$\Delta\mathrm{SNR} = \frac{\mathrm{SNR}_{FrTSegBand}}{\mathrm{SNR}_{Stand}} = \frac{\sum_{j=1}^{B} W_j^2}{M^2} \frac{\mathrm{MS}(N(n))}{\mathrm{MS}(l_{LB}(n)N(n))} \geq 1 \text{ ,} \tag{4.28}$$

4.6 Swept sine against MLS using average methods

In order to compare the performance of the logarithmic swept sine with the MLS techniques, tests were conducted for the so-called averaging methods Ave, WAve, SWAve and SSWAve, described above, and by following different analysis procedures based on the energy decay and the power spectral density of the RIR [Paulo09b]: (i) Energy Decay Ratio, EDR – corresponding to the excess of energy, in a noisy environment, related to the exact RIR (noiseless situation), EDR = $\int_0^{Tr} \frac{h_N^{\,2}(t)}{h^2(t)} dt$ (Tr represents the truncation time of the RIR), (ii) Energy Decay Ratio Distribution, EDRD is the time distribution of the EDR, EDRD(t) = $\int_0^{t} \frac{h_N^{\,2}(\lambda)}{h^2(\lambda)} d\lambda$, (iii) Power Spectral Density Ratio Distribution, PSDRD is the spectral distribution of the excess of energy of the ratio of the PSD in a noisy environment to the exact RIR, PSDRD(f) = $\int_0^{f} \frac{G_N^{\,2}(\gamma)}{G^2(\gamma)} d\gamma$, (iv) Background Noise – the noise of the tail of the RIR, and (v) Valid Decay, VD is the maximum decay, in dB, that the energy decay curve is able to follow the exact Energy Decay, for a predefined maximum error. This parameter was estimated by using the degree of non-linearity (ξ) and the degree of curvature (c) of the energy decay curves, according to the ISO 3382 – part 2 Standard [ISO3382-2]. The truncation point of the impulse response, used to draw the decay curves with the Schroeder method, was first estimated by applying the Lundeby method and adjusted if the c and/or the ξ parameters exceed the predefined value 3.5.

4.6.1.1 Experimental tests and results

Acoustic measurements were carried out in a classroom with RT60 of about 0.9 second (for the octave bands centered in 500, 1000 and 2000 Hz). The tests were conducted for the full audio band or for three octave bands centered at 500, 1000 and 2000 Hz (which are the frequency bands most used to quantify classroom acoustical parameters such as RT60, C80 and STI). A sample rate of 44.1 kHz and 16 bits/sample was used.

Music signals with different combination of SNR (-30, -24, -18, -12, -6 and 0dB) and of number of averages, NAve (4, 8, 16, 32 and 64), were investigated.

The results of the EDR analysis are shown in Figure 4.7 for all the combinations of SNR and NAve values. For simplicity, only the results for a) SNR of -18 and NAve 32 and b) SNR of -24 and NAve 8 are emphasized.

As a first conclusion, one can state that the swept sine always has a better performance than the MLS. The excess of noise difference remaining in the Energy Decay curve, between swept sine and MLS, varies from about 4 to 7dB for the standard average method, Ave, and from 5 to 9dB for the Spectral Segmented Weighted average method, SSWAve. Another remarkable finding is the improvement of about 9dB of the SSWAve against the Ave method. Furthermore, results of measurements with the MLS technique and using the Ave method show a loss of about 13dB when compared to measurements with the swept sine technique and using the SSWAve method.

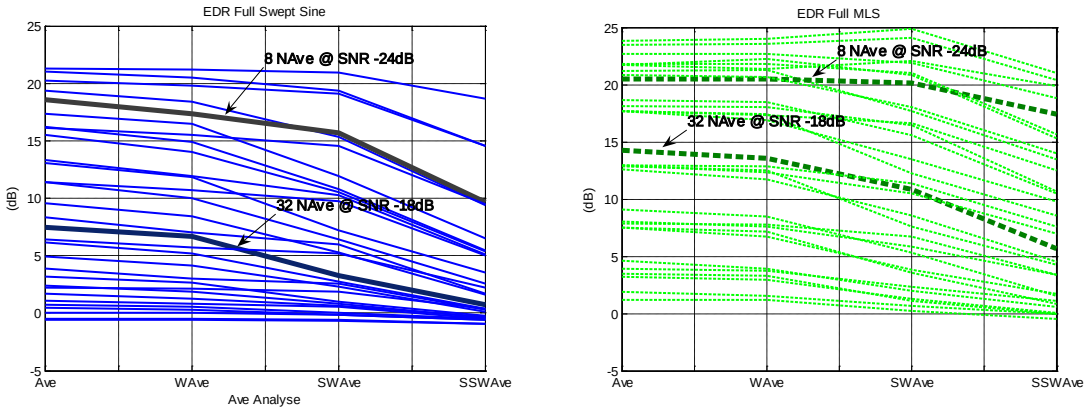


Figure 4.7 – Energy Decay Ratio and Power Spectral Density Ratio for swept sine (left) and MLS (right) techniques for the full audio band for a series of combinations of SNR (-30, -24, -18, -12, -6 and 0dB) and number of averages, NAve (4, 8, 16, 32 and 64).

The EDRD curve generalizes the EDR analysis. This approach facilitates the visualization of how a particular ED curve follows the exact ED curve with the time. The final values of the EDRD correspond to the EDR. All the combinations of SNR and NAve values are depicted in Figure 4.8.

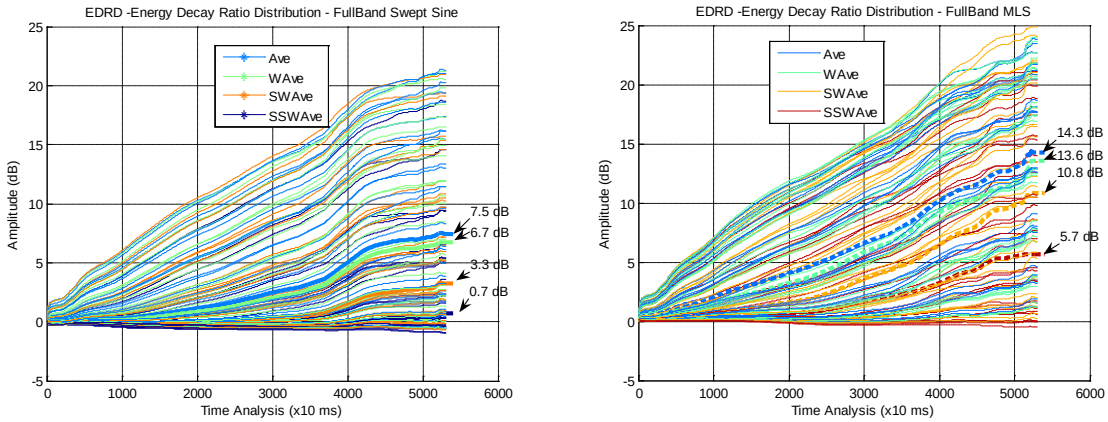


Figure 4.8 – Energy Decay Ratio Distribution for swept sine (left) and MLS (right) techniques for the full audio band; the heavy curves represent one particular analysis for SNR of -18 dB and 32 NAve.

The PSDRD analysis deals with the spectral noise accumulation vs. frequency, as depicted in Figure 4.9.

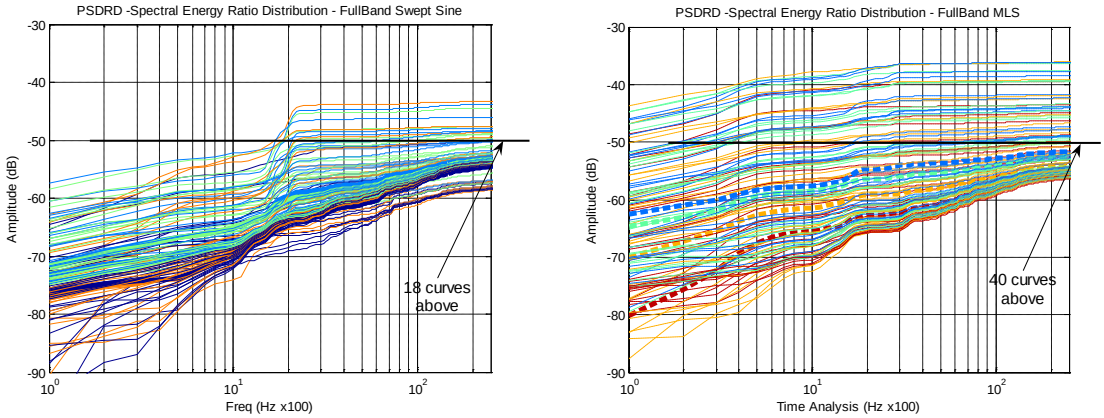
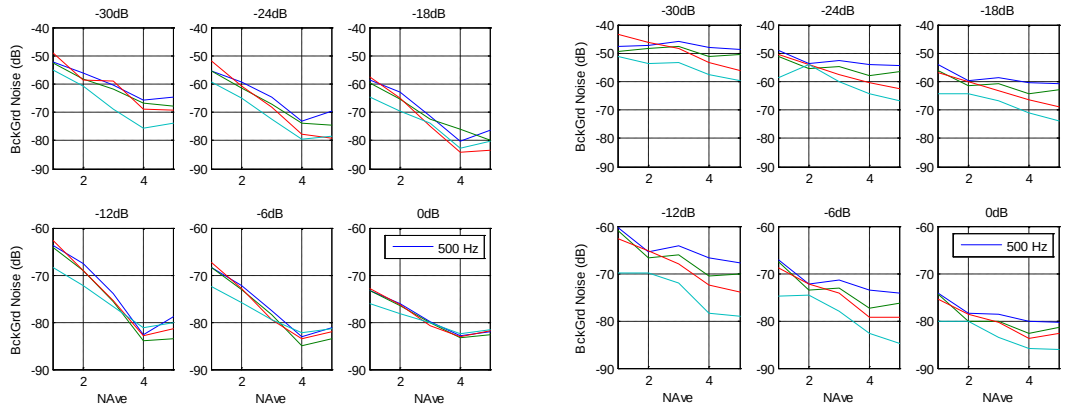
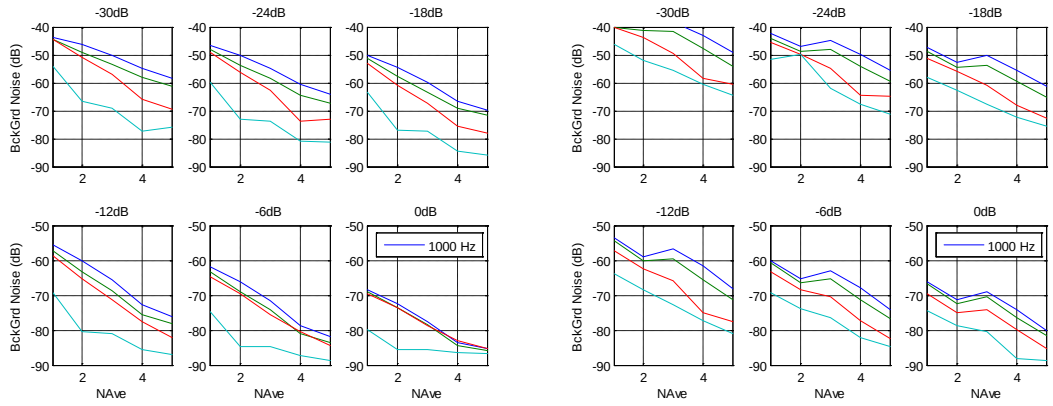


Figure 4.9 – Power Spectral Ratio Distribution for swept sine (left) and MLS (right) techniques for the full audio band.

Two main conclusions can be drawn from the results of PSDRD shown in Figure 4.9. First, the MLS technique always leads to an excess of noise higher than the swept sine. This can be inspected by defining a threshold level (a level of -50dB is shown in the figures as an example) and counting the number of analyses above this level. In this example, 18 curves for swept sine and 40 for the MLS. Second, the MLS technique shows much more low frequency noise than the swept sine, in particular for low SNR and/or low NAve. Once again, a remarkable improvement of the SSWAve against the Ave method can be observed.

A series of results of the background noise for several combinations of SNR and NAve, and for different frequency bands analyses are depicted in Figure 4.10. The parameter NAve of 1, 2 3, 4 and 5 corresponds to 4, 8, 16, 32 and 64 averages, respectively. In a general sense, the swept sine exhibits much lower background noise levels than the MLS, especially for low SNR levels. A figure of 10dB or more is achieved for SNR up to -18dB, even for NAve = 4. Again, a remarkable improvement of the SSWAve against the Ave method is observed, in most cases higher than 10dB. With the increment of SNR the results are not so obvious.





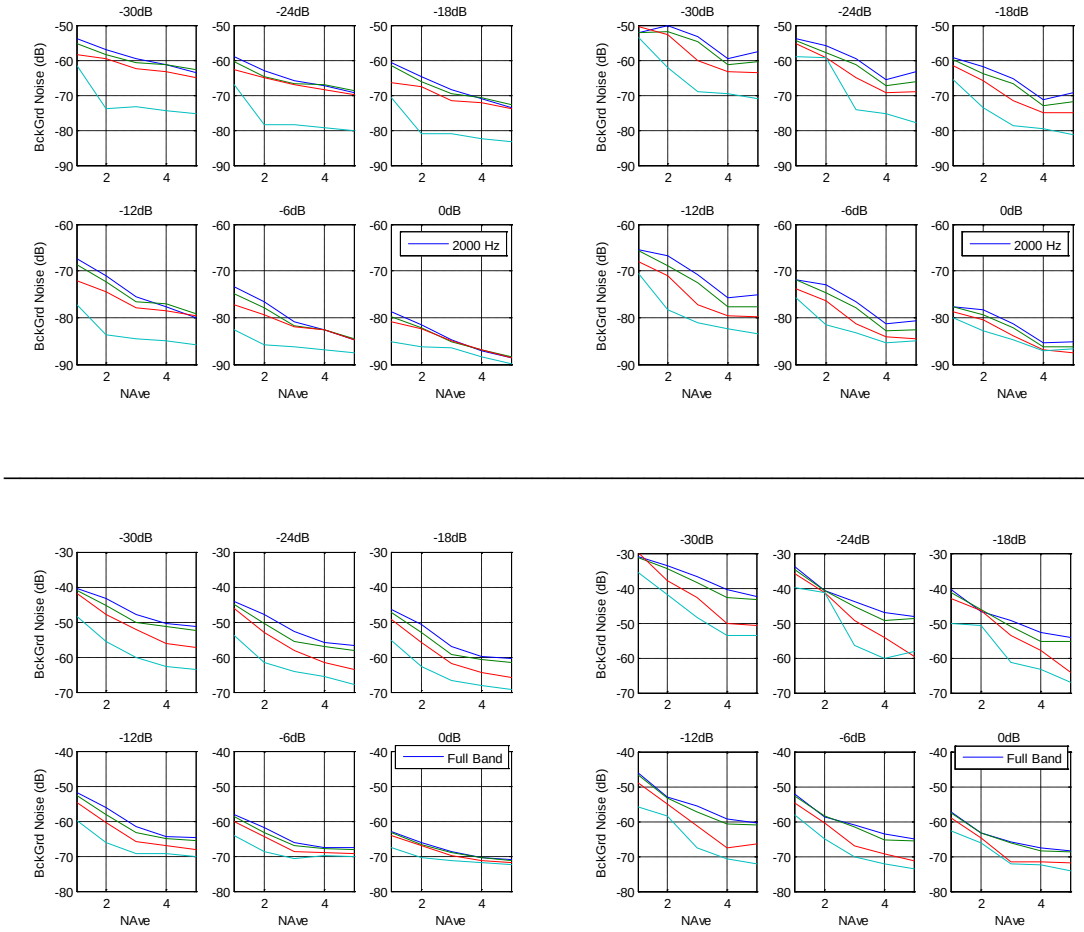
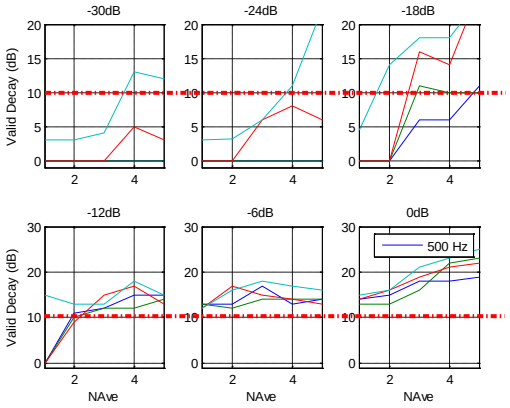
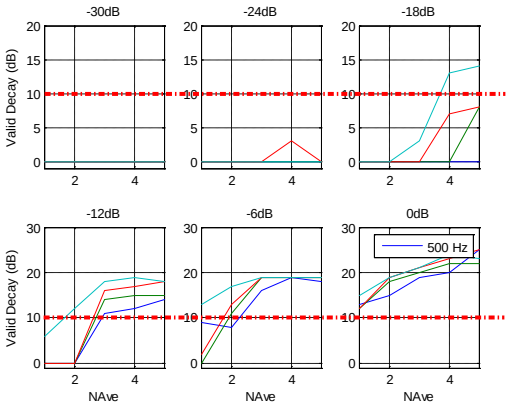
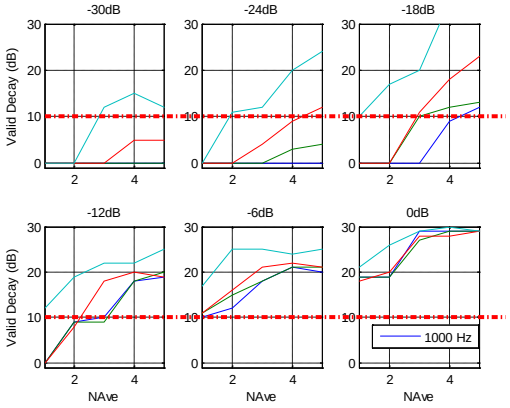


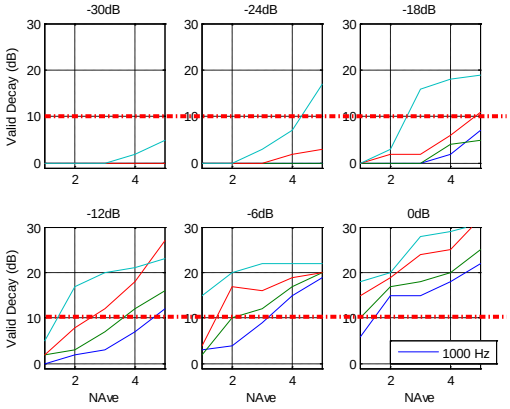
Figure 4.10 – Background noise estimated after processing the RIR for the swept sine (figures on the left) and MLS (figures on the right) techniques; SNR and NAve for 500, 1000, and 2000 Hz bands and for full band; Ave – blue, WAve – green, SWAve – red, SSWAve – cyan.

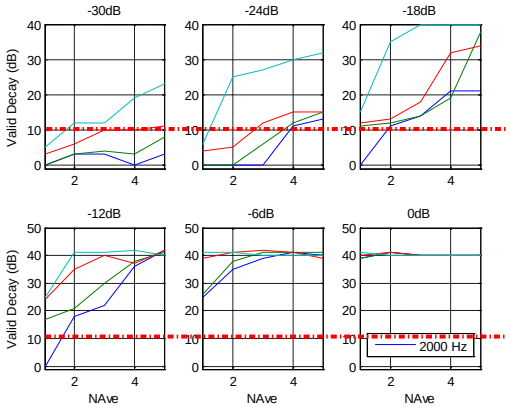
An important issue related to the accuracy of the acoustical parameters is the peak-to-noise ratio of the energy decay curve. However, in some situations, a more restrictive and precise approach for accurately estimating parameters such as the EDT and T20 is needed. Therefore, a measure of the likelihood of the energy decay curves (measured with and without noise), namely Valid Decay, depicted in Figure 4.11, was estimated. This parameter was analyzed for several combinations of SNR and NAve, for the octave bands centered at 500, 1000, 2000 Hz and for the full audio band.

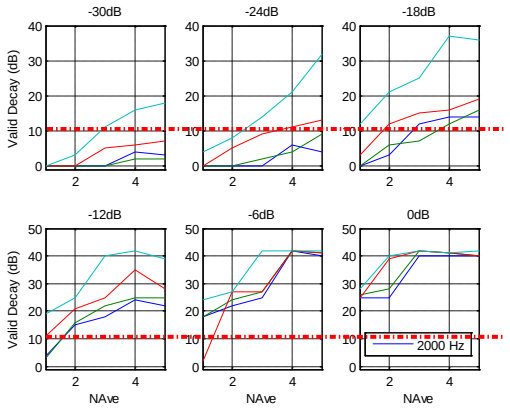

-30dB
-24dB
-18dB
-12dB
-6dB
0dB
Valid Decay (dB)
NAve
500 Hz


-30dB
-24dB
-18dB
-12dB
-6dB
0dB
Valid Decay (dB)
NAve
500 Hz


-30dB
-24dB
-18dB
-12dB
-6dB
0dB
Valid Decay (dB)
NAve
1000 Hz


-30dB
-24dB
-18dB
-12dB
-6dB
0dB
Valid Decay (dB)
NAve
1000 Hz


-30dB
-24dB
-18dB
-12dB
-6dB
0dB
Valid Decay (dB)
NAve
2000 Hz


-30dB
-24dB
-18dB
-12dB
-6dB
0dB
Valid Decay (dB)
NAve
2000 Hz

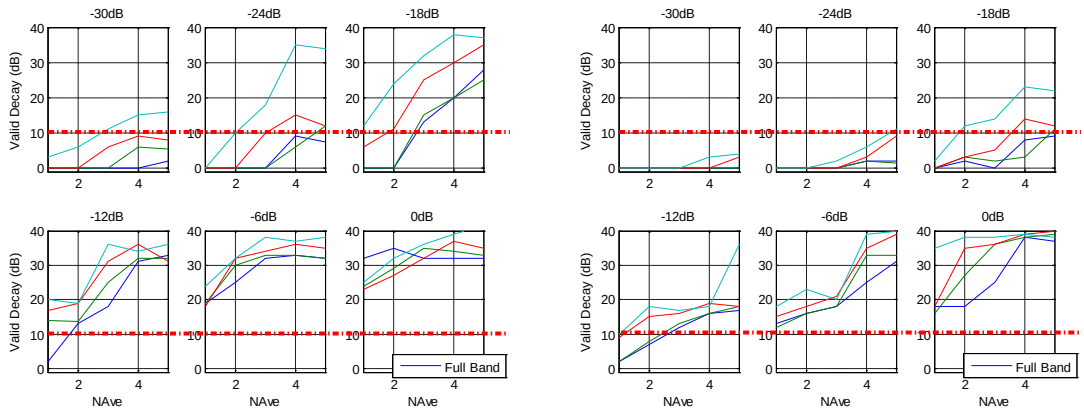


Figure 4.11 – Valid Decay variation estimated from the Energy Decay Curve for different combinations of SNR and Nave, for 500, 1000, and 2000 Hz bands and full band;Ave – blue, WAve – green, SWAve – red, SSWAve – cyan.

For a comparative assessment of the swept sine and MLS performance, the threshold level of 10dB was considered as marked on each plot in Figure 4.11. The interpretation of the plots for full band, with the MLS, leads to the interesting conclusion that for SNR up to -18dB the threshold of 10dB is only reached if the SSWAve method is used. Instead, for the case of the swept sine technique this threshold is achieved for SNR of -30dB with SSWAve or for -18dB using standard average method. This conclusion can nearly be drawn from the analyses in octave bands. Again, there is no doubt about the superiority of the swept sine over the MLS, at least for situations of low SNR. The same conclusion can be stated for the SSWAve and Ave averaging methods.

Although the results of the SNR shown by these experiments give a notable supremacy of the swept sine method against the MLS method, theoretically, assuming ideal conditions (linear systems and absence of time variance), and using the same signal levels for the two techniques, the final conclusions should be the same independent of the techniques used, since the signal and noise are treated in the same way in each frequency band. However, if peak signal levels are equal in the two techniques, swept sine measurements are superior to MLS measurements, even for linear systems. This is because in order to have the same spectral power for each frequency band, the MLS signal must be pink-noise filtered, and thus the crest factor is approximately 8 dB worse.

Nevertheless, when nonlinearities are considered the measurement systems have a different behavior. Swept sine techniques excite only harmonic distortion, while MLS techniques produce only intermodulation. Each has its place, and one does not really know which correlates better with actual perception. Since the main task here is to determine the linear room acoustic parameters, each technique should be judged for its ability to recover the best linear impulse response. Moreover, the swept sine works much well than MLS with a system slightly time variant. These could be the reasons for the experimental results stated in this section.

4.6.1.2 Preliminary conclusions

The EDR and EDRD analysis showed increments of up to 7dB for the Ave and 9dB for the SSWAve method. A remarkable increase of about 13dB between Ave and SSWAve was observed for the same SNR and number of averaged frames (NAve). The PSDRD measure revealed that the MLS technique always shows an excess of spectral noise higher than the swept sine and the MLS technique shows much more low frequency noise than the swept sine, especially for low SNR and/or low NAve. The Backgnd noise analysis showed, in most cases, a remarkable improvement of about 10dB of the SSWAve over the Ave method. The Valid Decay for full band leads to a surprising conclusion that for SNR up to -18dB the threshold of 10dB is only reached when using the SSWAve method, with the MLS. However, in the case of the swept sine technique, this threshold is achieved for SNR of -30dB with SSWAve or for -18dB using a standard averaging method. Therefore, the swept sine technique can lead to the same results with a much lower SNR or with less number of averages. If acoustic measurements with the presence of an audience require minimization of annoyance, a better masking of the test signal is achieved by music (lowering the SNR) and due to time-variance effects the number of the test frames should be reduced [Paulo09b]. Therefore, the swept sine technique is more suitable than MLS for these measurements.

As a final conclusion, for environments with very low SNR, the swept sine technique proved to estimate the acoustical parameters better than the MLS technique in all situations.

4.7 Detection and compensation methods to account for time-variance

In Chapter 3, the problem of time variance effects in acoustic measurements was addressed in a general way.

In this section, an attempt to assess the effects of time variance experimentally is presented. For this purpose, a convenient detection and, if possible, attenuation of its effects, is made by setting up a framework using sound descriptors extracted from the test signals.

4.7.1 General

The room impulse response is usually estimated by using synchronous averaging methods to enhance the SNR. However, the theoretical increment of 3dB in SNR by doubling the number of the testing signal frames is only achieved under the assumption that the system under test is time-invariant. Otherwise, the SNR will be upper-bounded and, subsequently, the expected increase will not be achieved. In real situations of auditorium acoustic measurements, time variance is due to air mass movements and temperature gradients caused by the presence/movement of people and by the air conditioning systems. The gas state parameters are interdependent, which means that changing one of the parameters will result in corresponding alterations to the others. Assuming that the closed rooms have a constant volume or a constant pressure, the influence of temperature variations on air parameters are analyzed for isobaric and constant-volume processes. During these processes, the change of the temperature has the strongest influence on the relative humidity of air. As a consequence, an increase of attenuation for the high frequency bands appears. Temperature fluctuations are important since they cause fluctuations of the speed of sound [Tatarskii61].

Nevertheless, it is not feasible to monitor the parameters responsible for the time variance phenomenon, such as temperature variations or air movements, during the acoustic measurements, for the whole volume of the room. Therefore, the task of detecting this sort of effect consists of directly analyzing the captured test signals or some variants of it such as the impulse response.

A basic approach to inspect the occurrence of time-variance phenomena during the acoustic measurement tests consists of estimating the short-term running cross-correlation function applied between each filtered impulse response, before synchronous averaging [Chu95,

Svensson99, Satoh99]. This procedure is not usually applied directly to the acquired test signal due to background noise.

Based on the cross-correlation function, a number of additional parameters are adopted to increase the accuracy of the procedure of detecting time variance phenomena.

However, in low SNR situations, due to non-controlled sound disturbances, such as machinery, or due to low levels test signals for the purpose of annoyance minimization, allowing the presence of people, the cross-correlation function applied to the RIR does not detect the time variance effects accurately. With this in mind, an alternative method to detect its consequences based on monitoring the air medium by using test signals in the ultrasonic frequency band was investigated.

The next sub-sections describe the effects of the time variance showing their consequences on the simulated and acoustic measurement results for the standard approach of its detection and the alternative method based on an ultrasonic-based device.

In particular cases, it is also possible to attenuate the time variance effects on the results by classifying the test frames as valid or not valid as a measure to not degrading the time averaging procedure.

4.7.2 Effects on the acoustic measurements

Time-variance effects on the acoustic measurements were inspected in a mid-size reverberant chamber of volume 150m^3 (~5x5x6m), approximately, with a reverberation time in the range of 2-2.5s in the mid-frequency bands. Since this enclosure is assumed to be a more controlled environment than a regular auditorium, thus assuring repeatability, the time variance occurrence can be better recognizable and quantifiable.

A number of tests covering the time variance generating mechanisms were carried out. A set of devices comprising a heater and a mid-size electromechanical fan with three rotating speeds were used. Additionally, the air was disturbed by manually waving with a pad of area 0.6m^2, as an attempt to force significant air masses movements.

The MLS and the Swept Sine methods were both used for comparison purposes. Since the reverberant chamber has a maximum reverberation time of about 3.5s, at low frequency bands,

to cover the whole frequency band, frames of length 5.8 s (18 order for the MLS) and sampling frequency of 44100 Hz were chosen. Sets of 16 frames for -42, -36, -30 and -24dBFS were used (corresponding to a minimum sound pressure level of about 50dBSPL at the microphone position). The temperature, the relative humidity and the air speed parameters were registered.

The early decay time (EDT), reverberation time (T20), SNR and the maximum valid decay of the decay curves were analyzed. In order to guarantee reliable results, the degree of non-linearity (ξ) and the degree of curvature (c) of the energy decay curves were calculated, according to ISO 3382–part2 [ISO3382-2]. The truncation point of the impulse response, used to draw the decay curves with the Schroeder method, was firstly estimated by applying the Lundeby method [Lundeby95] and adjusted if the c and/or the ξ parameters exceed the value 3.5 [Paulo11a].

Table 4.1 lists the ambient parameters registered in the different tests.

Table 4.1 – Air medium parameters registered inside the reverberant chamber

EXPERIENCE TYPE	Temperature (°C)		Rel. Humidity (%)		Air speed (m/s)	
	min	max	min	max	min	max
Ref (in quite ambient)	21.6	21.6	45.9	46.4	0	0
FanHi + Heater	23.7	26.7	40.1	47.4	0	1.7
FanLo	21.7	21.8	42.4	42.6	0	0.7
Heater	24	24	49.3	49.4	0	0
WavingPad 7x	21.8	21.9	44	44.3	0	1.3
WavingPad 20x	21.8	21.9	45	45.1	0	0.9
WavingPad20 + Heater	24.7	25.2	44.2	45.1	0	0.8

Reference tests – static air medium conditions

Figure 4.12 depicts the SNR values estimated for the air medium kept as static as possible (no mechanisms to produce time variance phenomena applied).

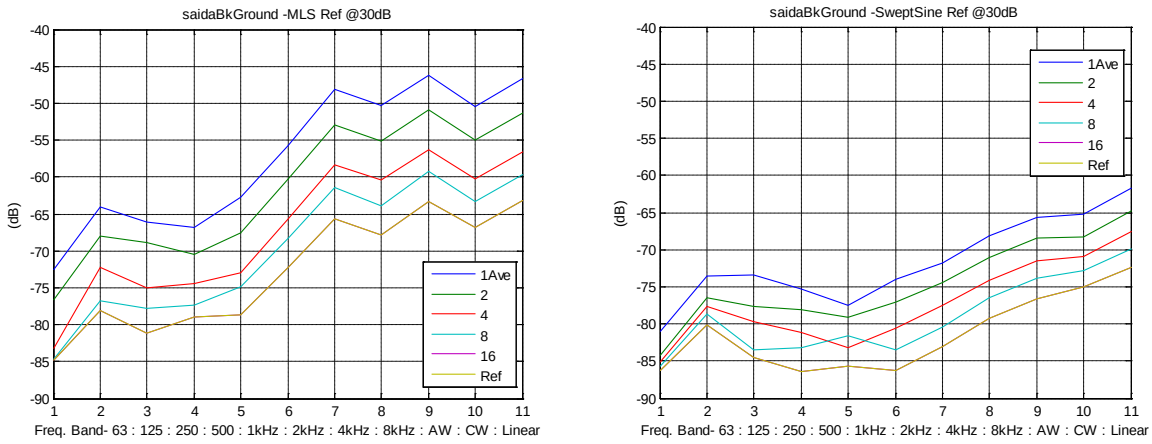


Figure 4.12 – SNR results for MLS and swept sine measurements in the reverberant chamber with no time variance; the 16 Ave is the same as Ref.

As Figure 4.12 shows, the swept sine reveals a much higher immunity to noise than the MLS technique and a tendency of decreasing the SNR with the increase of frequency is observed. This might be explained by the fact that some degree of time variance is always present in the measurement system, being more significant for high reverberant enclosures and time duration of the tests. Therefore, a higher SNR gain is expected for the high frequency bands with the swept sine technique.

The Valid Decay is defined here as the maximum decay range of the Schroeder curves, for a given error, used to estimate the reverberation time. This parameter is obviously related to the SNR. However, as shown in Figure 4.13, this relation is not always directly proportional due to the fact that, as mentioned above, sometimes the decay curves exhibit a degree of curvature (a possible reason for this is that the decay represents a mixture of decaying modes with different decay rates). This fact can be clearly observed for the full band related decay curves. Therefore, in order to attain the predefined degree of curvature parameter *c* below the defined threshold (of 3.5) the Valid Decay should be reduced. This is easily observed for the full band estimate, the Valid Decay value is the same for all the averages.

Notice the drop of the Valid Decay values at the 250 Hz octave band. This frequency band is affected by a strong room resonance, corresponding also to approximately the Schroeder frequency. This drop is more prominent for the swept sine technique proving that with this

technique the room resonances are more easily excited and thus care must be taken in the calibration of the measuring chain.

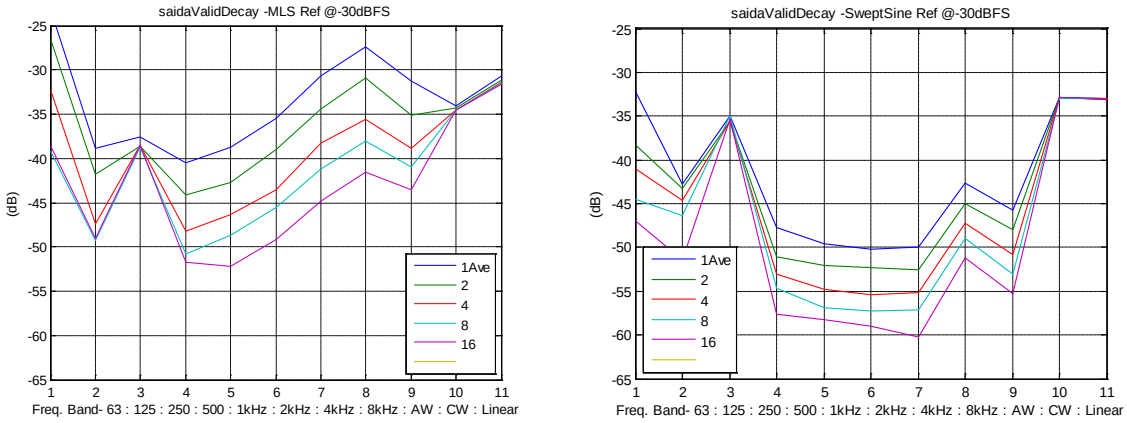


Figure 4.13 – Valid Decay for octave bands and for different average values, for MLS and for swept sine methods.

Air movement tests

Figure 4.14 depicts the values of EDT and of the reverberation time (T20) in octave bands for waving a pad 20 times during each set of 16 frames, using the MLS technique. The pad was moved slowly in order to simulate a real situation of a live event with the presence of people and air conditioning systems switched on. Similar results were obtained with the swept sine.

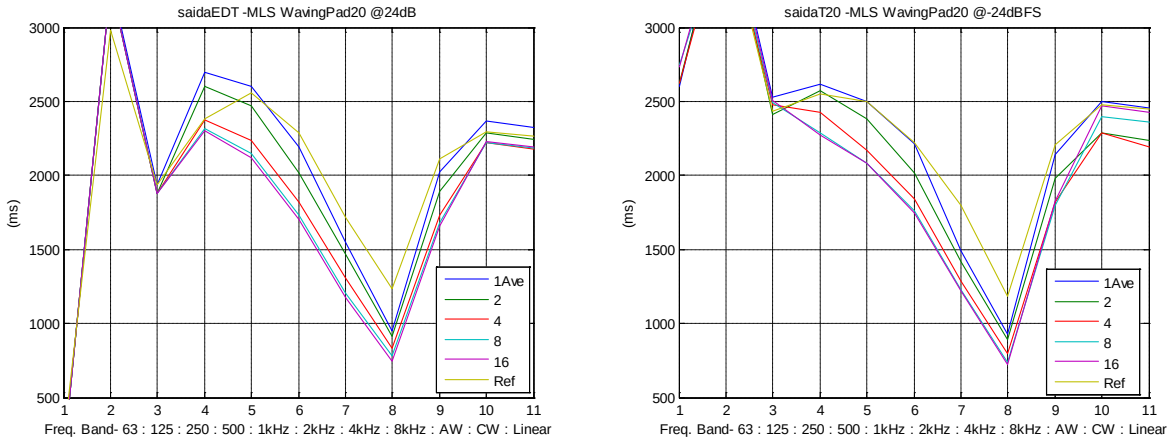


Figure 4.14 – EDT and T20 taking account the air movement by waving a pad; each curve corresponds to averaging the frames from 1 to 16; the Ref curve corresponds to the situation of a quiet environment; AW – Weighting A, CW - Weighting C.

Table 4.2 summarizes the results highlighting the error for the EDT and T20.

Table 4.2 –EDT and T20 errors (in percentage) due to air mass movements

		MLS				SweptSine			
-24dBFS									
Freq.		EDT	Err (%)	T20 (s)	Err (%)	EDT (s)	Err (%)	T20 (s)	Err (%)
500	Wav20	2,3	11	2,3	14	2,3	11	2,3	14
	Ref	2,6		2,7		2,6		2,6	
1k	Wav20	2,1	20	2,1	18	2,1	18	2,1	18
	Ref	2,6		2,5		2,6		2,6	
2k	Wav20	1,7	23	1,8	21	1,8	20	1,8	18
	Ref	2,2		2,2		2,2		2,2	
4k	Wav20	1,2	27	1,2	24	1,2	24	1,3	20
	Ref	1,6		1,6		1,6		1,6	
8k	Wav20	0,8	26	0,7	32	0,8	24	0,8	27
	Ref	1,0		1,1		1,0		1,1	
	Ave Mid Bands		18		18		16		17
	Ave Total		21		22		19		20

Deviations of about 20% at the mid-frequency bands and approximately 30% at the high frequency bands were obtained. The results in Figure 4.14 also show a tendency of T20 and EDT to reach a down boundary value as increasing the number of averages. In fact, the curves with averages of 8 and 16 are much closer than the others. This finding might be related to the particular room environment conditions and acoustical characteristics, such as the volume of the enclosure and reverberation time.

It should be noted that an error higher than 10% cannot be tolerated in the evaluation of absorption or reverberation time [kuttruff01]. Therefore, these results reveal that great care must be taken in acoustic measurements in venues with air conditioning systems actuating, and especially in very reverberant rooms.

Several authors [Vorlander97a, Svensson99] have mentioned the sensitivity of the MLS to time variance phenomena. Effectively, the tests performed with all different approaches show a remarkable higher immunity of the swept sine when compared with the MLS, achieving gains of the SNR of about 15dB. The performance of the swept sine and MLS techniques due to air mass movement (with the pad, softly waving 20 times, care was taken when waving the pad to not make significant noise) is shown in the Figure 4.15.

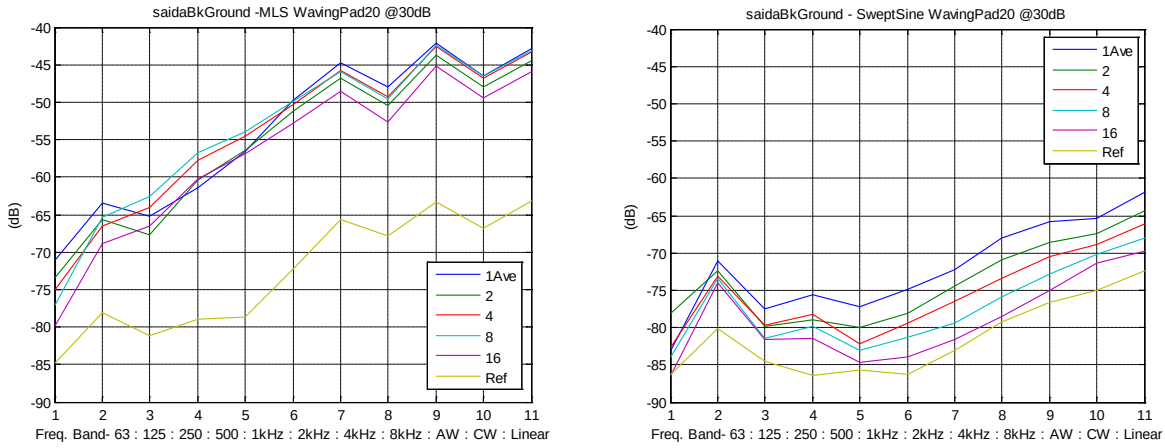


Figure 4.15 - SNR results for the MLS and swept sine measured in the reverberant chamber with air mass movement.

The swept sine technique is considerably less affected by time variance effects than the MLS technique. In fact, a gain of approximately 20dB is achieved for almost all frequency bands. By doubling the number of the averaged frames, a decrease in the noise levels of 3dB is observed with the swept sine but not with the MLS technique.

Tests with temperature gradients (space and time)

The test results with the enclosure at different constant temperatures do not reveal significant deviations of the energy decay, T20 and EDT. If the speed of sound changes, one could expect the density of sound reflections inside the enclosure would change too. Depending on whether the temperature increases or decreases, the acoustic energy will be more or less absorbed by the materials of the walls, ceiling and floor, affecting the T20 and EDT values. Two reasonable explanations for the minor deviations detected can be put forward: (i) the imposed temperature variation in the room, between tests, was about 3 degrees Celsius, which implies little variation on the speed of sound and minor changes on the reflections density, and (ii) the reverberant chamber boundaries are built mostly with concrete, with very low absorption coefficients, also causing little difference on the absorption of the travelling sound waves for the different temperature environments.

Since in large venues the temperature is not constant through the room, the tests were carried out with temperature and air mass movements by using an electric fan and a large pad. The results are shown in Figure 4.16 for the T20 using a heater and waving the pad slowly. Both

swept sine and MLS results are affected by this air disturbing situation. However, whereas the T20 for the swept sine method changes gradually, by augmenting the number of averages, the MLS results change abruptly, showing again its weakness regarding time variance effects.

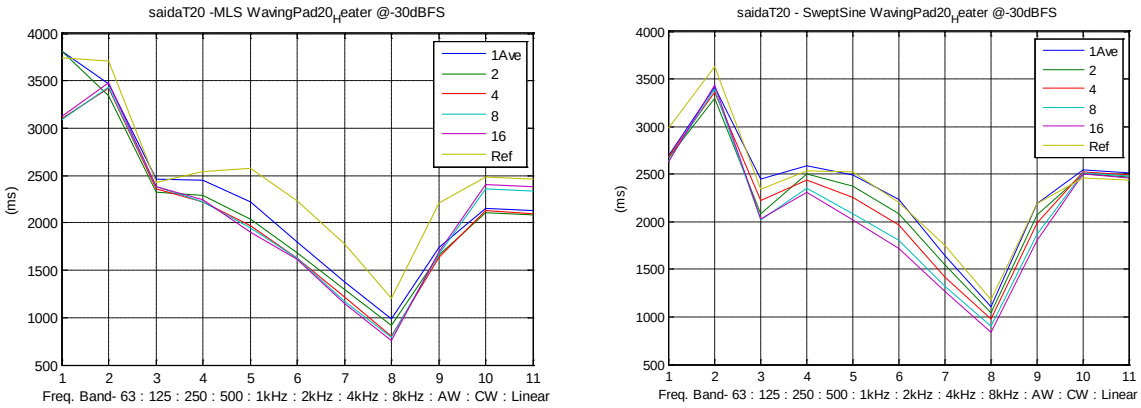


Figure 4.16 - T20 vs. octave bands, for different averages.

A sort of distortion on the estimated IR was found in the experiments with the heater, with a number of glitches appearing on the tail of the IR. These glitches are more prominent on swept sine measurements, though they are also present in the MLS technique as shown in the Figure 4.17.

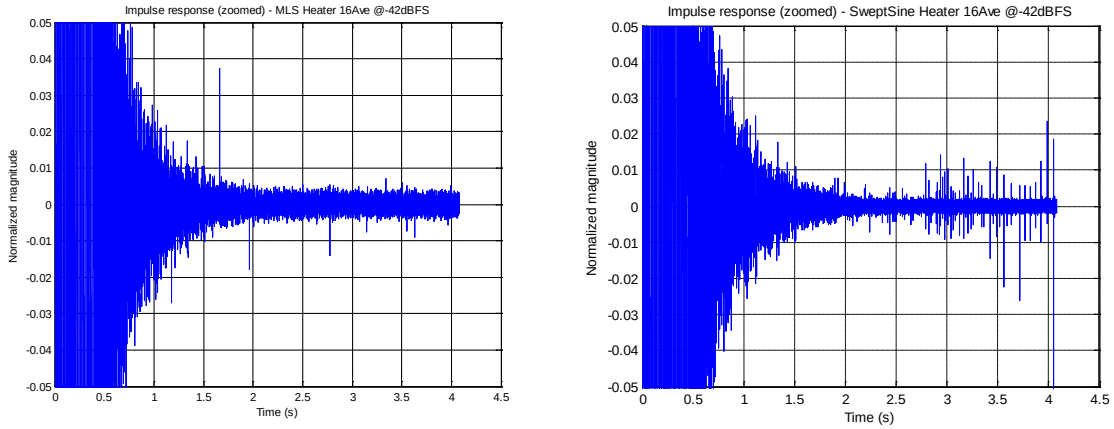


Figure 4.17 – Glitches produced in the IR due to temperature gradients.

As shown, although the IR estimated from the swept sine technique exhibits a high density of glitches, most are located at the very end, thus being able to be eliminated by windowing procedures. However, with the MLS, the sparse glitches fall inside the RI, probably due to slew limiting. Anyway, the glitches density is too low to induce significant distortion on the decay

curves. Nevertheless, this issue should be considered when the impulse responses are to be used for auralization purposes.

4.7.3 Simulation results

Simulations of the time variance effect over the test signals were computed by setting up a model for stretching the waveform by a number of ways, using several degrees of freedom, in an attempt to reproduce results of real measurements, under controlled conditions. The model is based on the fractional delay lines approach with the fractional delay parameter set in two different ways to account for the intra-periodic and inter-periodic effects. The simulated system response is given by $res(n) = x(\alpha n)*h(n)+n(n)$, where α stands for the stretching parameter depending on the effect desired, $x(n)$ is the test signal, $h(n)$ is the IR and $n(n)$ stands for the background noise estimate.

The intra-periodic effect is simulated by using a fractional delay parameter. This parameter is given by a normal distribution random variable $a_i \sim \mathrm{N}(0, \sigma_i^2)$ with sample deviation given by the variance σ_i^2 (maximum values are bounded to two samples) or by a sinusoidal variation, a sort of frequency modulation, $\alpha_i = A(n)\sin(2\pi N_i f_a / f_s n)$, where $A(n)$ stands for the sample deviation from the sampling time instant value, $N_i f_a$ represents the frequency of oscillation applied to the frame i, where N_i is the frequency scale factor and f_s is the sampling frequency, 44.1 kHz for our purposes. $A(n)$ and N_i can be set to a constant value or to a normally distributed random number. This effect is associated to real acoustic situations due to thermal noise in the electronic devices or due to overheating of the transducers, producing phase deviations, often called *jitter*, and rapid fluctuations of the temperature or air mass movement actuating for at least one frame.

The model usually used to simulate the inter-periodic effects is based on stretching the waveform by resampling the test signals with a resampling factor slightly different from the nominal value in order to account for temperature deviations mainly. Therefore, the fractional delay parameter α_r (which is in this case a constant value for each frame) should lead to a resampling factor very near unity. This type of effect is mostly due to the slow variation of the ambient medium temperature, up to few degrees Celsius, or synchronism deviation between the playback and the recording devices (when different equipment is used) during the acoustic

measurements. The resampling factor associated with the temperature increments, *TempIncr*, is defined as *R*= 0.6*(*TempIncr*)/343.4, as mentioned in Chapter 3.

The synthetic IR, *h*(*n*), used in the simulations was sketched assuming diffuse field conditions given by $h(n) = \beta_n \exp\{-C / f_s n\}$, $\beta_n \sim \mathrm{N}(\mu, \sigma_i^2)$ represents independent samples of a normal distribution with μ=0 and σ_i^2 =1, and C = 3ln10/*RT* is the acoustic damping factor. This simple IR model assures linear energy decay curves (shape of the squared impulse response) and equal decay energy rate for all the frequency bands.

Figure 4.18 to Figure 4.22 show some representative results of a number of analyses over the MLS and swept sine test signals related to the models adopted to produce time variance phenomena. The reference frame is considered in two different cases: 1- in quasi-absence of time variance effect, in ideal conditions: in simulation or in well controllable environment scenarios, and 2- in real conditions, *in situ* acoustic measurements, taken from the set of all the frames: the frame of the middle of the ensemble is usually considered. The analysis was performed for the frequency octave bands centered at 500 Hz to 8 kHz plus full band and full band A-weighted.

Intra-periodic model

Normally distributed stretching parameter - Each sample is "shaken" from its nominal position, using a variable following a normal distribution. The reference frame, called Ref, is not "shook". Different averages up to 64 frames are used in the study.

The first experience consisted of simulating the *jitter* effect caused by the time variance with maximum deviation of one sample. The *C* Curvature parameter, background noise and the Valid Decay results are shown and commented.

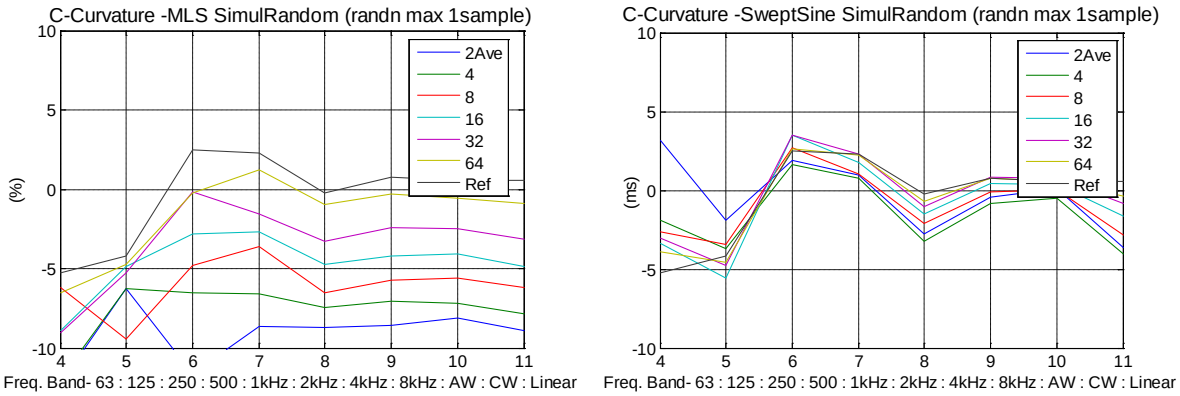


Figure 4.18 – C-Curvature parameter simulation results vs. octave bands.

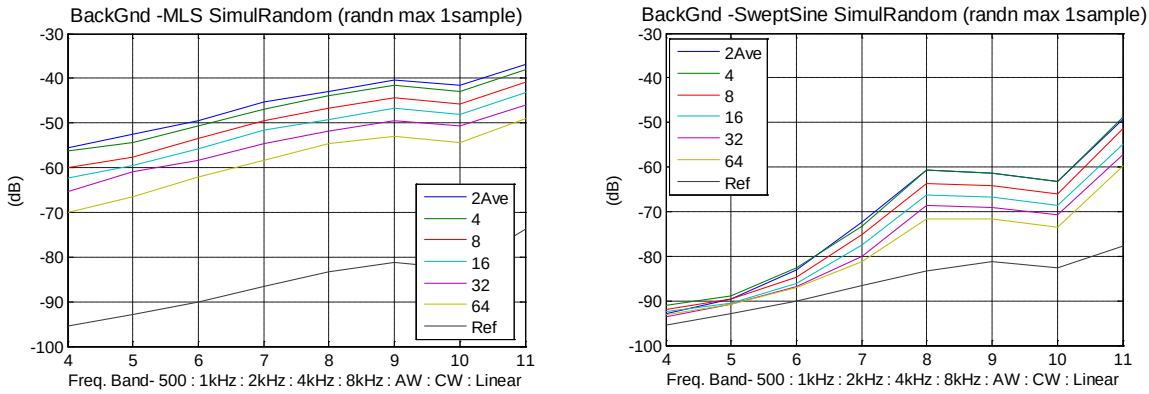


Figure 4.19 – Background noise (or PNR) simulation results vs. octave bands; notice the smashing behavior of the swept sine curves for different number of averages as the frequency decreases; this fact is related with the 16 bit resolution of the IR used.

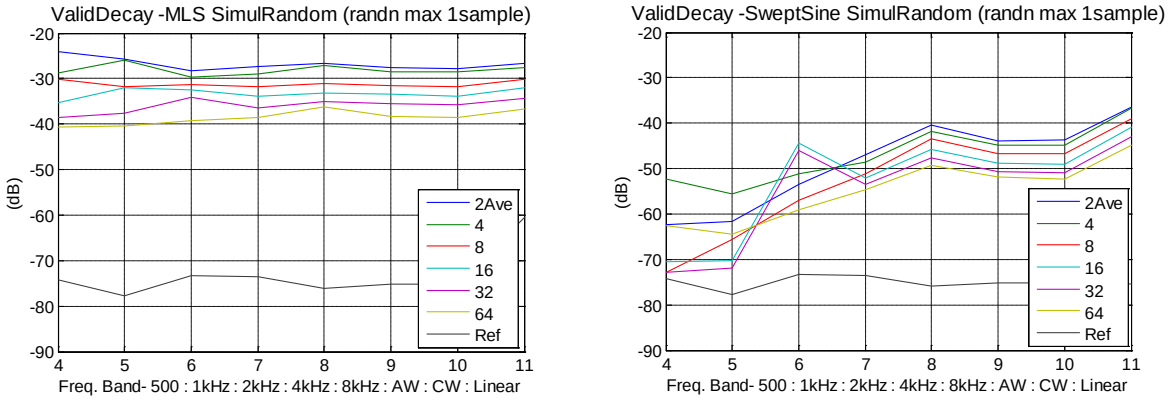


Figure 4.20 – Valid Decay simulation results vs. octave bands.

As a first comment, the SNR seems to be always improved by applying the synchronous averaging procedure. In fact, increase of about 3dB by doubling the number of frames averaged is observed in Figure 4.19.

Since the C-Curvature parameter gives a figure of the deviation of *T20* from *T30*, accurate results for the MLS are only achieved for 64 frame averages. The negative values correspond to T20 higher than T30, which means low SNR. Nevertheless, the results obtained by the swept sine technique show absolute values lower than 5%, proving the results are not degraded severely. These conclusions are confirmed by the background noise and the Valid Decay results. Indeed, Valid Decay results from 28 to 38dB for the MLS and from 40 to 50dB to swept sine are observed for the higher octave bands. The background noise difference between swept sine and MLS is about 17dB for the 8 kHz octave band with a prominent increase in the lower octave bands. These simulated results show a remarkable superiority of the swept sine over the MLS.

These results are not compared with the real acoustical measurements since that issue is out of the scope of this research.

Sinusoidal stretching parameter - Sinusoidal stretch (a sort of frequency modulation) with linear increments or uniform randomly in the range of 0 to 1 (0.99) sample is applied to each frame (the number of sinusoid periods are kept constant and equal to 10). The reference frame, Ref, is not stretched. Different averages up to 64 frames are used.

The first experiment uses a constant maximum sample deviation of one sample and a uniform randomly distributed variable to set the number of sinusoids used to stretch each frame. 4 to 40 sinusoids per frame are used in this case (each frame was split in 4 segments). Figure 4.21 and Figure 4.22 show the results of the PNR and Valid decay.

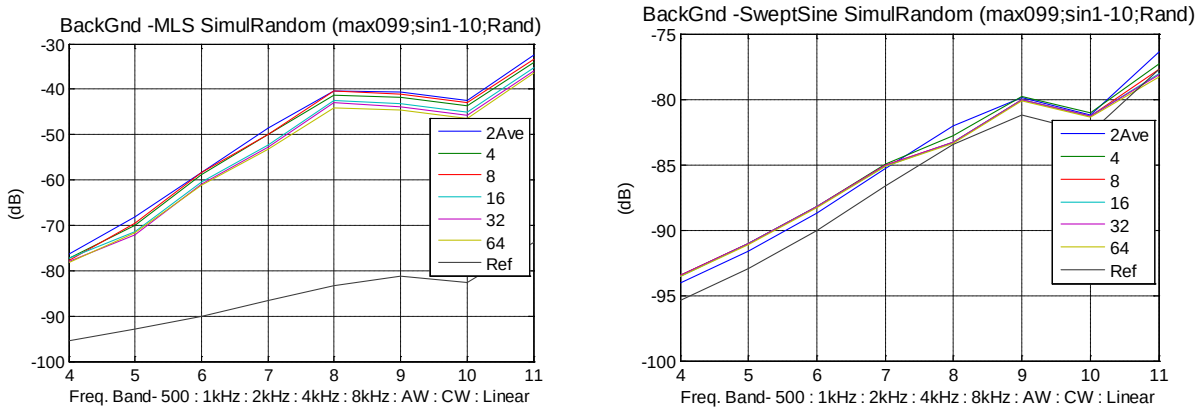


Figure 4.21 – Background noise (or PNR) simulation results vs. frequency octave bands corresponding to the MLS and swept sine test signals; linear increments are applied to each frame; different scales for MLS and for swept sine test signals.

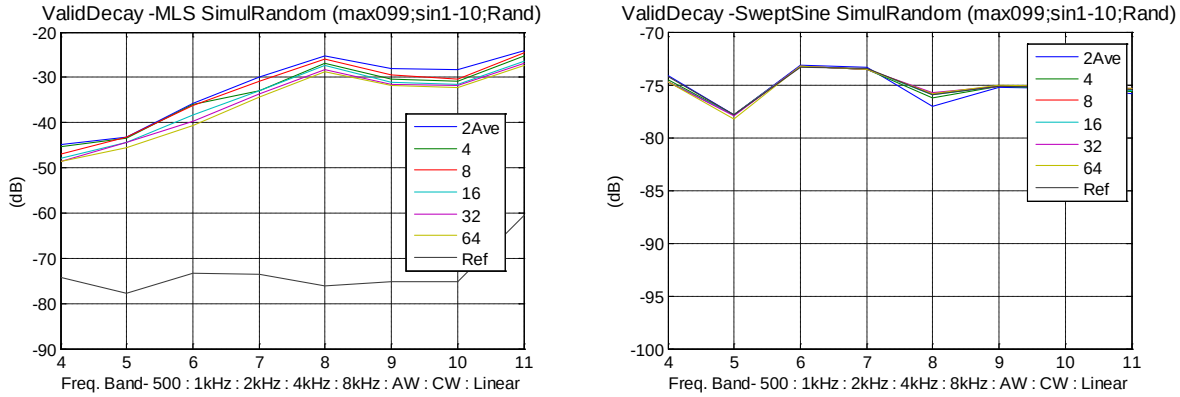


Figure 4.22 – Valid Decay simulation results vs. octave bands corresponding to the MLS and swept sine test signals; linear increments are applied to each frame; different scales for MLS and for swept sine test signals.

The results show no significant improvement by applying the averaging technique. A figure of less than 1dB by doubling the number of frames averaged is achieved. The results obtained by using MLS technique are much degraded in this experiment showing Valid Decay values less than 30dB, which is very low for estimating the acoustical parameters accurately. Instead, the results from swept sine technique are almost unaffected. Actually, the minimum PNR is above 80dB for all octave bands (except for full band), and the Valid Decay is almost about 75dB. Notice the fact that the PNR for the averaged curves are bounded by the Ref curve, having an offset less than 2dB.

This experiment is closely related to fast air mass movements in real situations. The equivalent situation is shown in Figure 4.15 to the case of MLS technique showing similar results.

An additional experiment consisting of linear increments of the deviation parameter in the range of 0 to 1 sample applied to each frame, keeping 40 sinusoids in the tests, was used as a stress test to compare MLS and swept sine signals, as shown in Figure 4.23 and Figure 4.24.

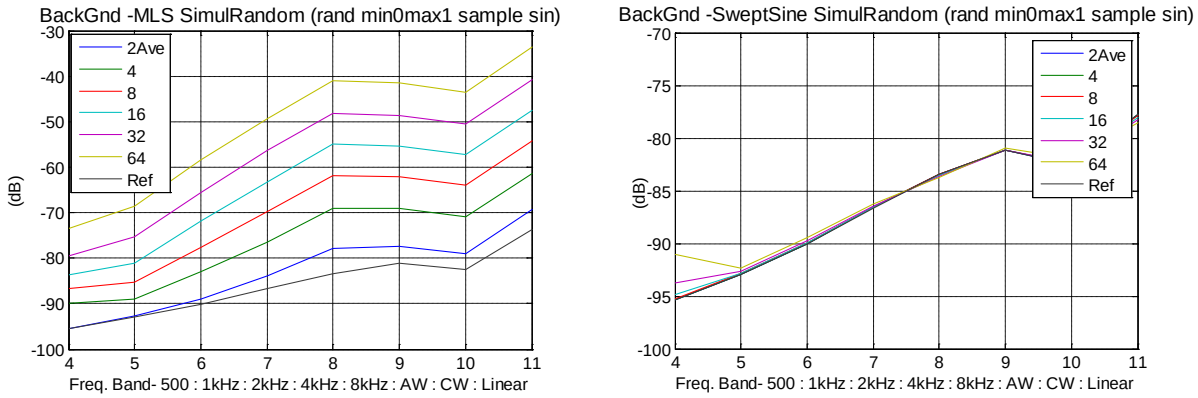


Figure 4.23 – Background noise simulation results vs. octave bands corresponding to the MLS and swept sine test signals; different scales for MLS and for swept sine test.

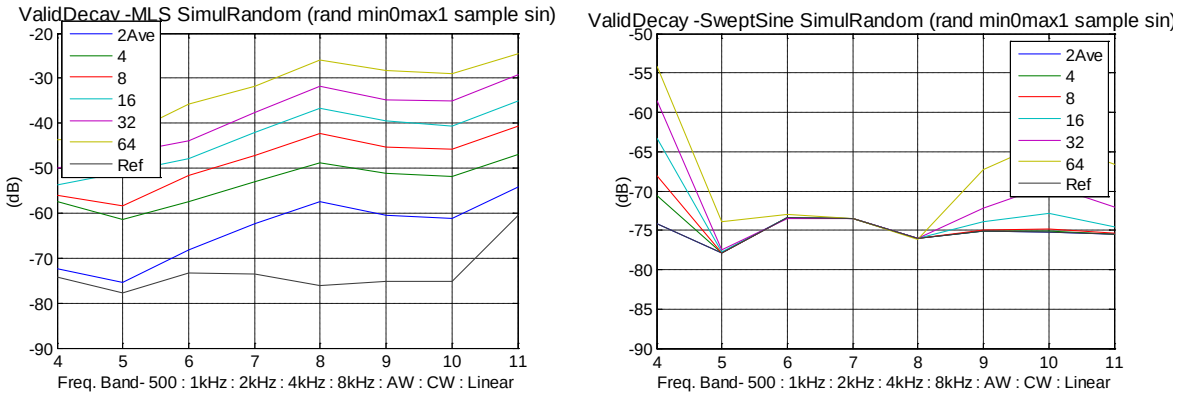


Figure 4.24 – Valid Decay simulation results vs. octave bands corresponding to the MLS and swept sine test signals; linear increments are applied to each frame; different scales used to the MLS and swept sine test.

As a preliminary conclusion, the accuracy of the MLS technique is decreased by applying the average method. This is an expected result since the deviation parameter is increased from frame to frame, thus, increasing the averaged number of frames tends to degrade the results. In fact, a decrease of about 6dB for the PNR results is observed. Apparently, the swept sine results

are not affected, maintaining the same value independently of the number of averages. However, a remarkable distortion is observed for the high frequency part of the IR.

Inter-periodic model

The MLS and swept sine test signals are time stretched by using a resampling procedure to simulate linear increments of the temperature during the acoustic measurements in the range of 0 to 0.5ºC applied to each frame. The reference frame, Ref, is not resampled. The results are depicted in Figure 4.25 and Figure 4.26 for different averages up to 64 frames.

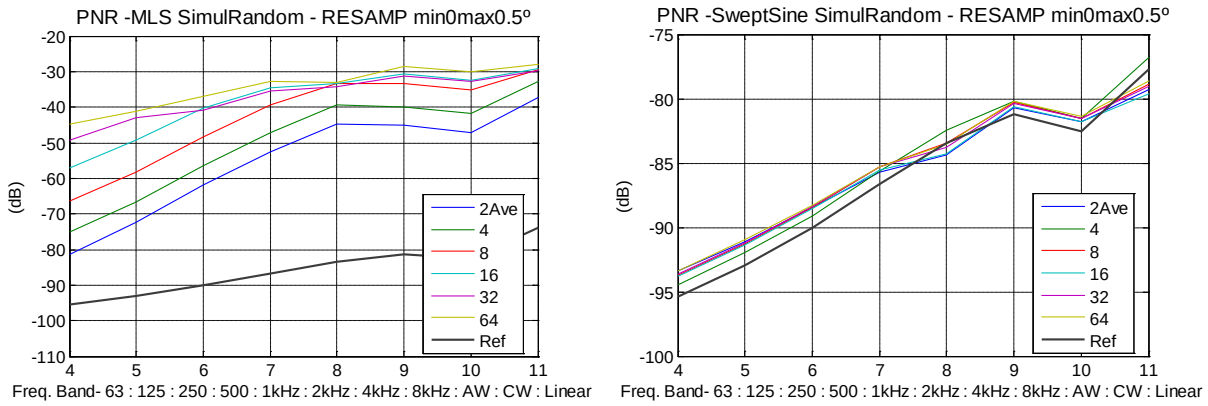


Figure 4.25 – Peak-to-noise ratio simulation results vs. octave bands; different scales for MLS and for swept sine tests.

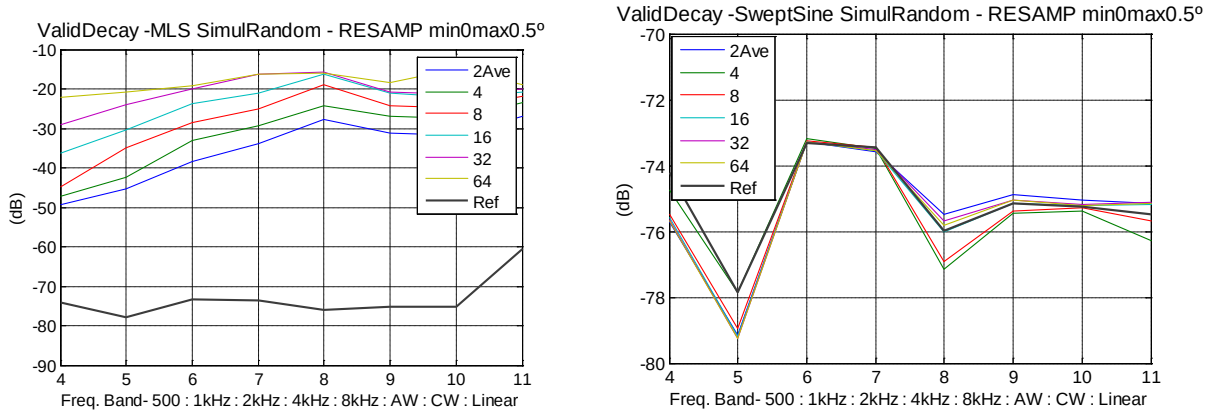


Figure 4.26 – Valid decay simulation results vs. octave bands for the MLS and swept sine test signals; different scales for MLS and for swept sine tests.

These results show a decrease of accuracy for the MLS technique when the averaging procedure is applied. In fact, the overall PNR is degraded by increasing the number of frames averaged. Moreover, the decrease of the PNR is more prominent as the frequency increases.

The swept sine is almost not affected proving once again its superiority against the MLS technique.

A similar experiment was carried out by assigning to the resampling parameter a uniformly distributed random value in order to study the PNR in a different approach. The results are depicted in Figure 4.27 to Figure 4.30.

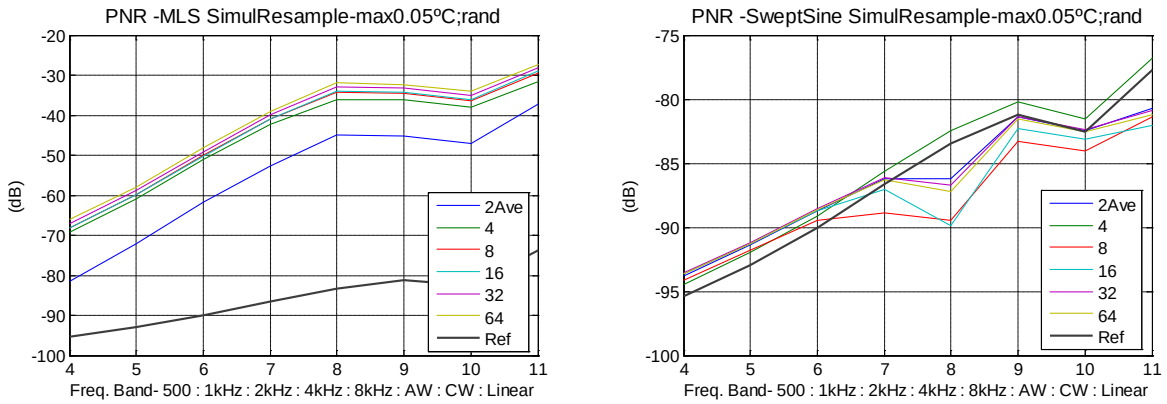


Figure 4.27 – Peak-to-noise ratio simulation results vs. octave bands. Random resampling parameter value is used. Notice the different scales used to the MLS and swept sine test signals results.

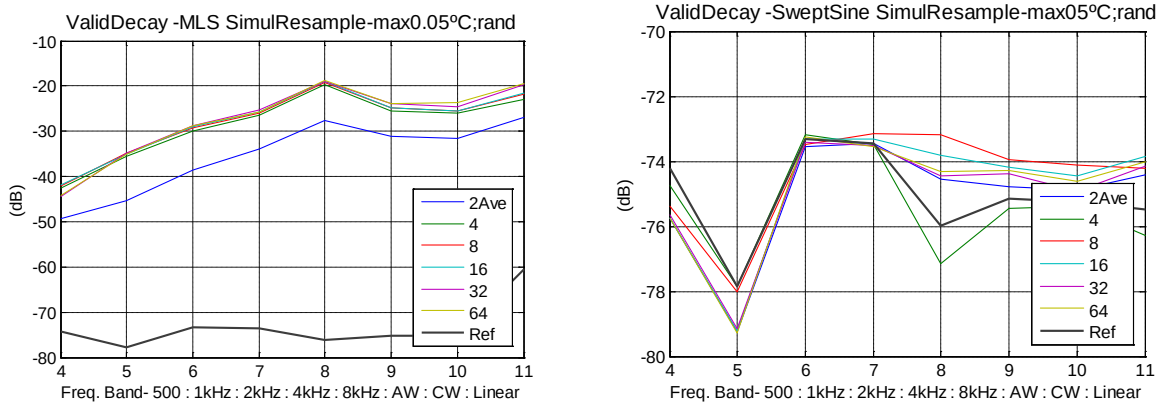


Figure 4.28 – Valid decay simulation results vs. octave bands for the MLS and swept sine test signals Notice the different scales used to the MLS and swept sine test signals results.

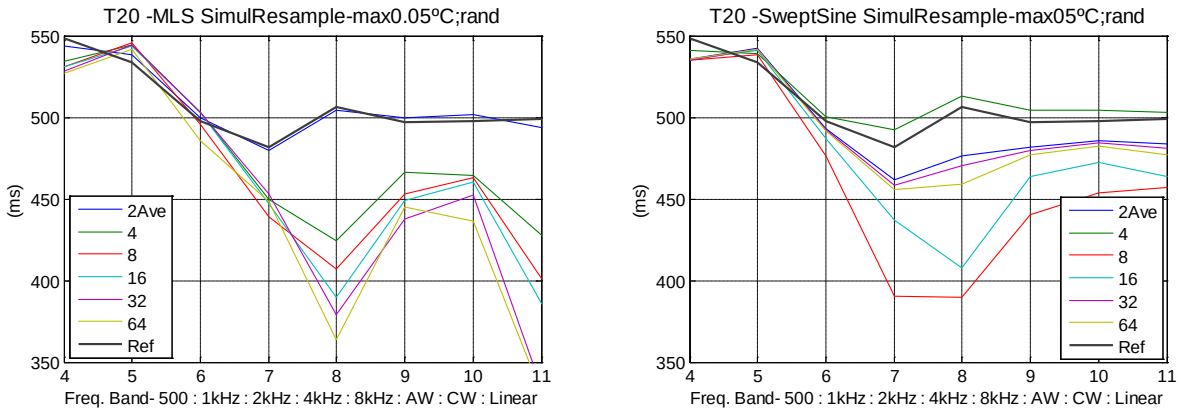


Figure 4.29 – Reverberation time, T20, simulation results vs. octave bands for the MLS and swept sine test signals.

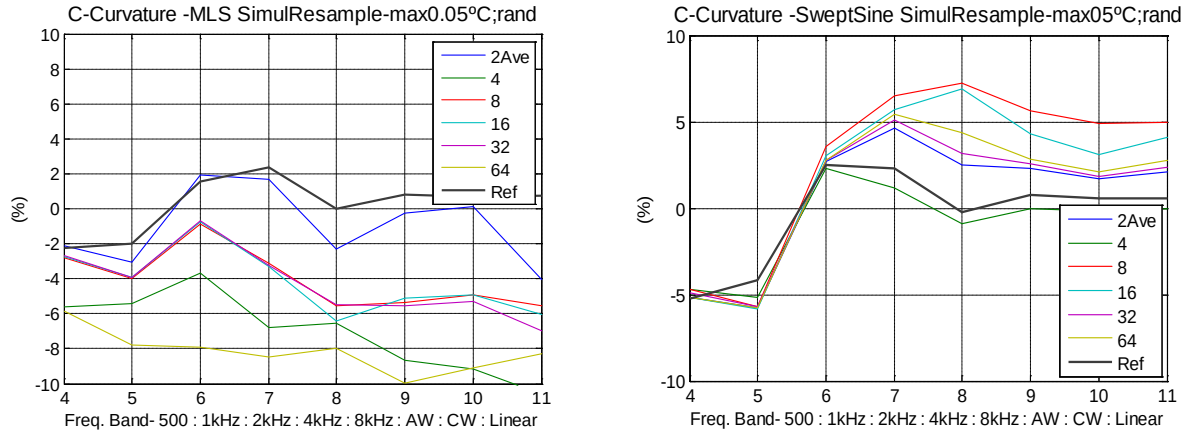


Figure 4.30 – Curvature parameter simulation results vs. octave bands.

Since the frames averaged in each set (2, 4, 8, 16, 32 and 64) are randomly chosen, the PNR for all sets of averages is almost equal (curves in Figure 4.27) for each frequency band. For the MLS method, the *C*-Curvature results show a considerable degradation when the number of averaged frames is increased (note that negative values of the *C*-Curvature correspond to T30 lower than T20 due to the truncation procedure). Consequently, the Valid Decay is too short to accurately estimate the T20. The swept sine technique proves to be the best choice for acoustic measurements for this type of time variance.

The idealized simulated experiments related to the stretch of the test signal samples are analyzed in order to draw guidelines for the time variance effects. Some are related directly to real situations and others are not. Nevertheless, this approach is useful for providing

information on the robustness of the swept sine and MLS measurement techniques to different time variance effects.

4.7.4 Use of ultrasound transducer arrays

Spectral analysis of common audio signals during live event performances, other than clapping hands sounds, show insignificant energy above 40 kHz. This fact suggests the possibility of using an ultrasonic test signal probe with the purpose of correlating the alterations imposed to both the test signals, audio and ultrasonic bands, due to time variance effects. Notice that due to the wavelength difference the ultrasonic sound test is much more sensitive to such effects.

4.7.4.1 Context

A new acoustic measurement technique was devised for low SNR situations where a probe tone signal in the ultrasonic band is sent to the room simultaneously with the test signal frames by using an array of ultrasonic sensors. This device monitors only the direct sound wave path due to its high directivity pattern and because it is pointed directly to the measurement microphone location. This procedure has the advantage of avoiding the perception of the probe ultrasonic testing signals by the audience, keeping a reasonable SNR.

However, care must be taken in the use of the ultrasound loudspeaker array in order to avoid nonlinearities of the sound field [Pompei02, Barbagallo08]. This is achieved by keeping the resulting sound levels below a certain value. This upper-bound sound level is established by performing a calibration test assuring that the air medium does not produce self-demodulation when two tones are applied to the room simultaneously.

In real situations, the SNR is lower at the low frequency audio bands and the time variance effects on the output test signal of the system are more relevant in the high frequency bands. Therefore, the analysis is performed in frequency octave bands.

As a first attempt to investigate the anomalies produced on the signals, a set of tests was carried out in an anechoic chamber in order to evaluate the effect of time variance in the absence of the influence of the room. This could represent an outdoor sound propagation scenario also.

The result of air movements and temperature variations produce some sort of noise around the tonal component in the ultrasonic spectrum. These effects are illustrated in Figure 4.31.

As a first conclusion, the bandwidth of the noise produced on the sides of the tonal component due to the temperature variation is larger than for the air movements. Secondly, a kind of frequency modulation is generated for the *waving pad* case, see the A mark in the Figure 4.31. In fact, this side band noise moves around the tone depending of the direction of the air movement.

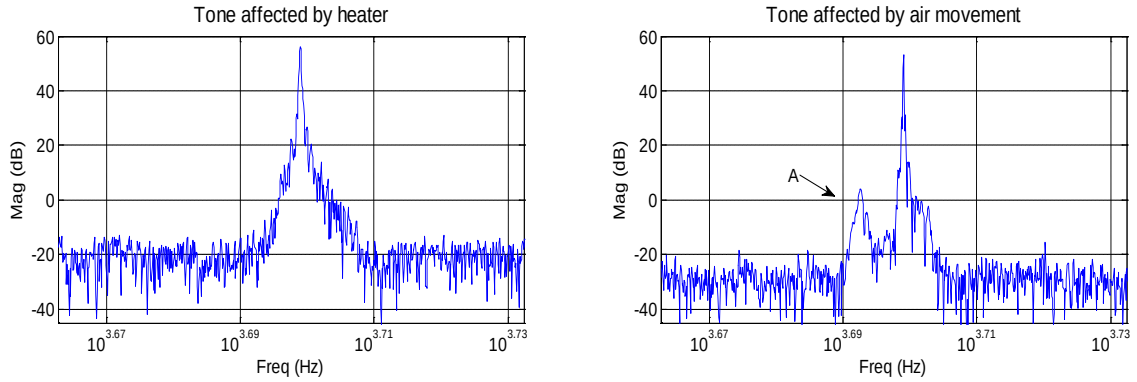


Figure 4.31 – Modified spectra of ultrasonic tonal components due to air heating (left side) and air movement (right side); the tone pilot original frequency was 40 kHz (these scales are moved to 5kHz centre frequency to improve the signal processing techniques performance).

Figure 4.32 shows an example of magnitude variation of the test signal probe by waving a pad. Notice the sound pressure levels amplitude variation is more than 15dB.

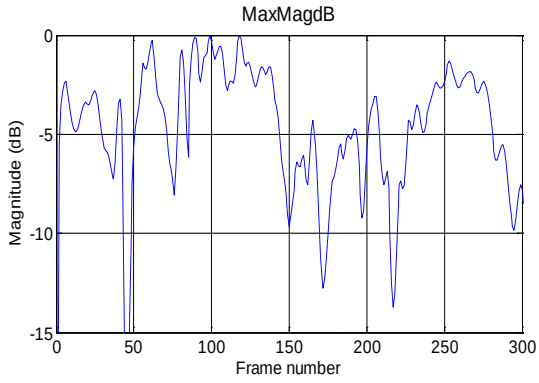


Figure 4.32 – Effect of air movement on the magnitude of an ultrasonic tonal component.

The same analysis was performed for the audio band by using multi-tonal test components. Similar effects have been identified. As expected, due to the different wavelengths involved, the magnitude variation of the tonal components for the audio band is smaller.

The information about these dissimilarities, which is a measure of time-variance, could be taken by the averaging method.

Throughout the second half of the last century, technological advances in understanding how the turbulence influences sound wave propagation have facilitated the continued development of improved ultrasonic (US) devices for flow measurement. The scattering of sound waves in turbulent flows was first considered by Obukov in 1941 [Obukhov41]. Subsequently, other authors have devoted time to the same issue [Andreeva04 and Jakevičius08].

4.7.4.2 Measurement arrangement

The basic list of apparatus used in these experiments consisted of a sound power source, a microphone, a multichannel sound board attached to a personal computer and an ultrasound loudspeaker array [Paulo10d]. The measurement setup is shown in Figure 4.33.

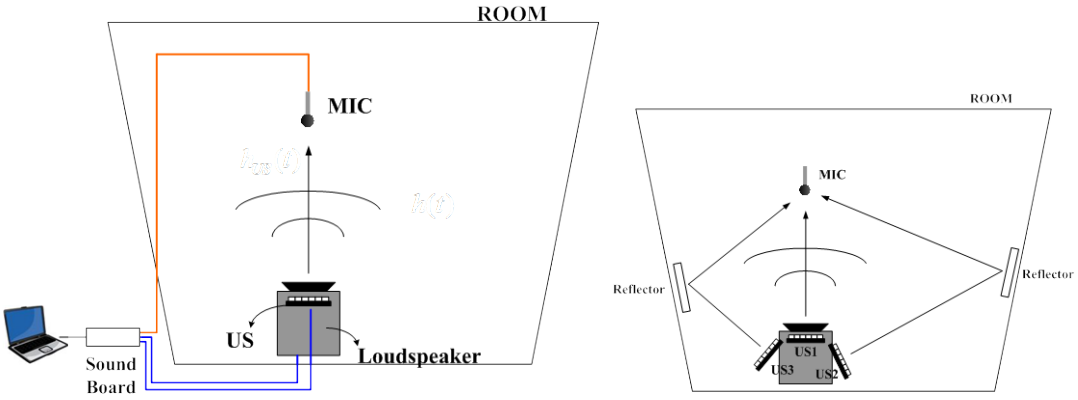


Figure 4.33 – Measurement setup using test signals and ultrasonic test probe simultaneously; on the right, a more realistic arrangement for multipath propagation using several loudspeaker ultrasonic array devices as an improvement to the original system.

The same microphone is used to capture the test signal frames in both the audio band and the ultrasonic range since its bandwidth goes up to 60 kHz.

Only one loudspeaker ultrasonic array was used. Therefore, the correlation between the baseband and ultrasonic results decreases for the case of measurements in a mid-size auditorium. To prevent such situations, a different loudspeaker ultrasonic array arrangement, using multiple-ultrasonic devices, could be applied in order to cover a wider audience area and room volume by establishing multi-path sound propagation, as illustrated in Figure 4.33, right side.

4.7.5 Features analysis and thresholds levels

A set of features extracted from the acquired sound signal was defined from the analysis performed in two distinct cases. For situations of low background noise levels, especially in the high frequency bands, the features are extracted from the room impulse response. Otherwise, for the cases of low SNR, the features are taken from the analysis of the probe tone signal.

Two different threshold levels associated with each feature are defined. In fact, the conformity of the frames is verified by using the lower and the upper threshold levels. Feature values below the lower level indicate conformity of the frame and above the upper level is a measure of inaccuracy of the results and, thus, the frame is removed. Otherwise, the frames are kept for the analysis, though with restrictions and a weight factor is applied as a measure of significance.

4.7.5.1 Cross-correlation based detection

In situations of considerably high SNR, the detection of time variance effects from the study of the cross-correlation function applied to the IR proved to be a good technique. Therefore, a number of features based on the cross-correlation function are examined for use in the subsequent analysis.

As discussed in Chapter 3, the time variance phenomenon affects the high frequency bands more significantly. Therefore, the detection procedures are usually performed for the frequency bands centered at 4 or 8 kHz.

In order to compare the feature analysis correctly, a normalization procedure is applied to the filtered impulse response. Therefore, the bias from the features used in the estimation of the statistical distances between the classes is reduced.

The IR is truncated to a defined fraction of the estimated reverberation time in order to remove the noise floor of the tail since it carries little information due to its stationary random behavior. In this study, ¾ of the IR, IRtrunc, was used and split into U segments of K samples to estimate the features. The features considered are named as follows:

Time-lag function, τ_u (in *s*) – time delay between the peak values of the cross-correlation function, for each segment *u*, performed between two IR where one of them is chosen for reference;

$$\tau_u = index\left\{\max\left(XCorrRef_u(k)\right)\right\} - index\left\{\max\left(CorrSeg_u(k)\right)\right\} \tag{4.29}$$

Correlation Energy error, CorrEnergyErr$_u$ (in dB) – energy of the error of the cross-correlation function, or total quadratic error, related with the reference and the test segment.

$$CorrEnergyErr_u = 10log_{10}\sum_{k=1}^{K}\left(XCorrRef_u(k) - CorrSeg_u(k)\right)^2 \tag{4.30}$$

Correlation maximum ratio, CorrMaxRatio$_u$ (in dB) – Since the maximum amplitude of the cross-correlation function depends of the likelihood between the reference and current segment, this feature is a measure of the deviation from the expected value.

$$CorrMaxRatio_u = 20log_{10}\frac{\max\left(XCorrRef_u(k)\right)}{\max\left(CorrSeg_u(k)\right)} \tag{4.31}$$

Relative Error, CorrRelErr$_u$ (%) – Percentage of the error of the cross-correlation of the reference segment and the current segment.

$$CorrRelErr_u = \frac{XCorrRef_u(k) - CorrSeg_u(k)}{XCorrRef_u(k)}\text{x}100 \tag{4.32}$$

where $XCorrRef_u$ stands for the auto-correlation of the reference segment *u* and $CorrSeg_u$ is the cross-correlation of the reference segment and the current segment *u* in each IRtrunc.

Results of the features Time-lag function, the Energy of the quadratic error and the Correlation maximum ratio are depicted in Figure 4.34 to Figure 4.36. The first frame of a set of eight frames is used as the reference.

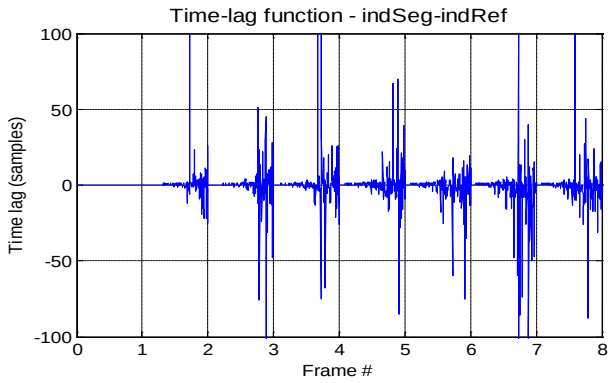


Figure 4.34 – Time-lag function τ – delay time between the max peak values of the cross-correlation function.

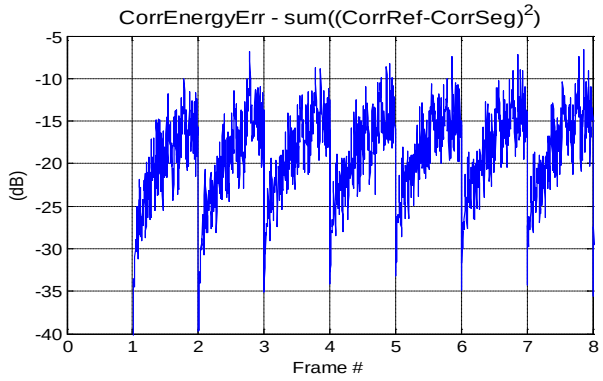


Figure 4.35 – Energy of the quadratic error, $CorrEnergyErr_u$ feature.

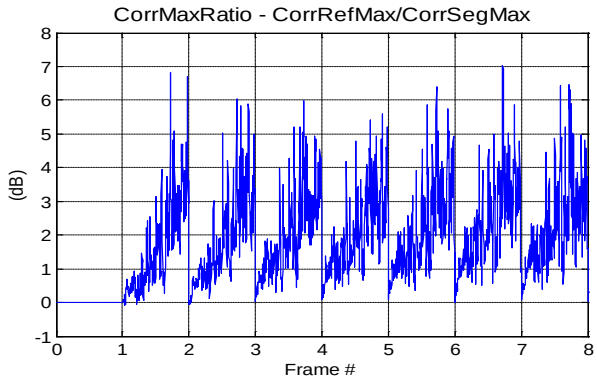


Figure 4.36 – Correlation maximum ratio, $CorrMaxRatio_u$ feature.

Thresholds levels

The threshold levels coupled with each feature, used for the verification of the conformity of each IRtrunc, are found empirically by verifying the error of the energy decay curve. Therefore, the threshold levels are based on the minimization of the IR error.

The threshold levels are found by applying a sliding window along to the feature curves.

4.7.5.2 Ultrasound

As a first attempt to relate the effects of the time variance on the test signal (audio band) and on the tone probe (ultrasound band) a multitone signal in the audio band is used as the test signal. Since both signals are of the same type, sinusoids, the effects are better understood.

The multitone test signal consisted of a set of tonal components centered in each octave frequency band centered at 1, 2, 4 and 8 kHz, generated by the sound source, plus a probe tone of 40 kHz, radiated by the loudspeaker ultrasonic array.

A number of features were extracted by signal analysis in the audio and in the ultrasound bands.

The features considered here are related to the magnitude of the tonal components, the residual part of the ultrasound spectrum over a frequency band centered at the ultrasound tone probe and by cross-correlating the tone probe by time segments.

The magnitude of the tonal components is extracted by using peak continuation and parabolic interpolation techniques, based on FFT analysis, in order to increase the frequency resolution. The features related with the residual part were estimated by using the sinusoidal plus residual noise model approach, as described in Chapter 3. The tonal components are removed more accurately with this approach than by using other filtering techniques such as Constant *Q* filters or adaptive filters due to filter specifications, essentially. In fact, the NLMS algorithm proved to be inappropriate for this purpose. Neither the tonal component is tracked for the needed frequency resolution nor is the bandwidth sufficiently narrow to remove the tonal component only, not degrading the remaining spectrum of the signal.

In this case, the frequency resolution is more important than the time resolution since the signals are quasi-stationary tonal components over time. Therefore, a considerable large sliding Hanning window of 16384 samples with 75% overlapping factor was used.

In order to accommodate the ultrasound tone probe, a sampling rate of 192 kHz was used. However, the different analyses applied to the captured signals were performed at 44.1 kHz by splitting and down-sampling the audio and the ultrasonic frequency bands conveniently. Before down-sampling, a defined frequency band centered in the tone probe is bandpass filtered to select only the ultrasound (US) part which is afterwards demodulated to remain centered at 5 kHz.

Figure 4.37 shows analyses of this signal in an anechoic chamber using the fan device for the low speed and in oscillating mode. The period of oscillation is about 10s.

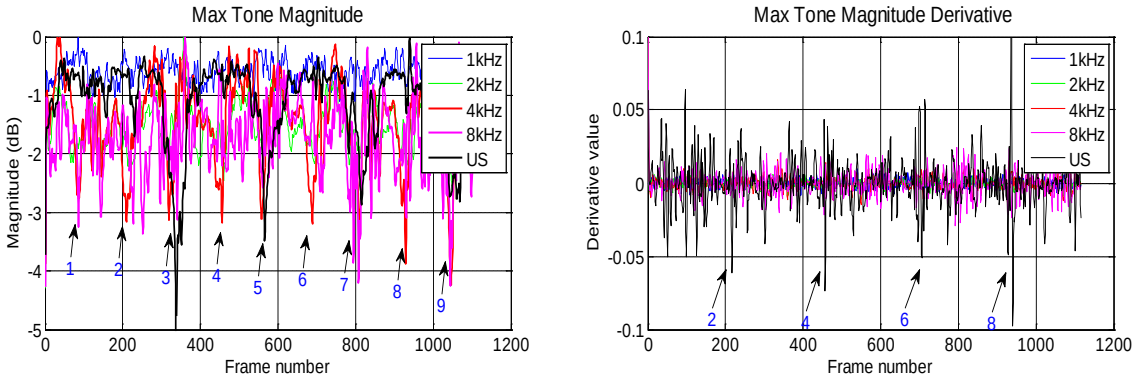


Figure 4.37 - Magnitude and corresponding derivative of the tonal components; the US curve in the Max Magnitude feature is scaled by a factor of 1/10, for better visualization.

The peaks and dips shown in the figures follow a periodical fashion regarding the air movement due to the oscillation movements of the fan device.

The dips in the curves in Figure 4.37, corresponding to the cycles of the oscillating fan, are identified by numbers.

The results show a good relationship between the audio and the ultrasounds (US). Although the Max Tone Magnitude curves for the US do not show significant alteration in the locations marked by the even numbers a pronounced peak appears at these locations in the Max Tone Magnitude Derivative feature. This interesting finding shows that these parameters are complementary and should be used together.

The results in Figure 4.38 show energy offset due to the background noise of the room and to the sound radiated by the fan. Moreover, although the curves for 1 and 2 kHz provide information about the fan movements this energy is mostly due to acoustic noise. Therefore, the

results for the frequency bands up to 4 kHz hold little significance in detecting time variance and, thus this statement can be generalized for all other analyses.

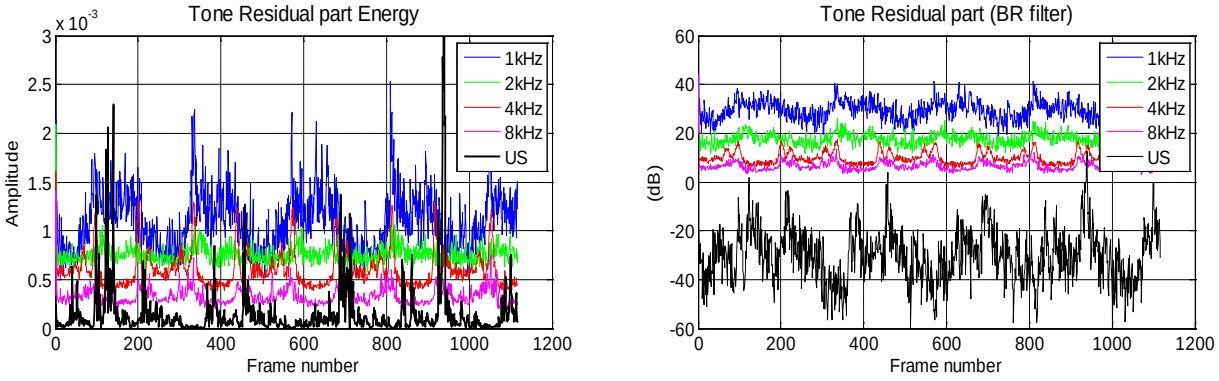


Figure 4.38 – Energy of the residual part of the tonal components; left: energy for the full spectrum with the tonal components removed; right: residual energy corresponding to band-reject a frequency band centered in the tonal components.

Depicted in Figure 4.38 are the results of the Frequency Deviation for different frequency tones.

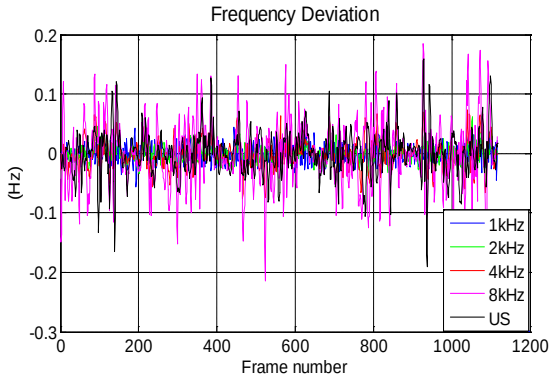


Figure 4.39 – Tone frequency deviation; the US curve is scaled with a factor of 1/10, for better visualization.

The results show fair similarities for the analyses of the residual energy and frequency tone deviation and for magnitude between the multi-tonal and US tests. The different wavelengths involved are responsible for the discrepancy in the amplitude differences (the amplitude variation of the US tone probe is much larger than for the multitonal signal nested in the audio band). Rapid variation in the tone amplitude of the US probe is generally related to strong perturbations in the air and consequently has a significant implication for the time variance effects.

The results show little deviation for the frequency pitch. In fact, the frequency tone deviation barely achieves a deviation of 3 Hz for the worst cases. Like the tone magnitude variation, the higher pitch variation is related to strong air perturbations and can thus be used as a flag.

The features extracted from the ultrasound tone probe considered for the study of the time variance in situations of low SNR are:

Frequency Deviation – frequency shift from the nominal value. This feature is useful to detect strong air mass movements;

Max Tone Magnitude – amplitude variation of the ultrasound tonal component. This feature is very sensitive to alterations of the medium parameters;

Tone Magnitude Derivative – first derivative of the tone magnitude. This feature is useful for detecting air movement directions not identified by the Max Tone Magnitude feature;

Tonal Residual part Energy – energy estimated from the captured US probe signal with the tone removed (bandwidth of 500 Hz centered at the US tone previously demodulated to 5 kHz). The procedure consists of separating the tonal and residual parts by applying peak continuation techniques in the frequency domain;

Tonal Residual part **(BR filter)** – the energy around the US tone frequency is filtered out (50 Hz bandwidth centered at the US tone previously demodulated to 5 kHz). This analysis attempts to identify strong air mass movements, keeping the upper and lower noise bands generated mostly by air movement, corresponding to the example shown on the right side of Figure 4.31.

In addition, the features defined in the standard approach (features extracted directly from the test signals, audio band) are used with the ultrasound tone probe.

The definition of the threshold levels evaluated from the ultrasonic features is performed following the same procedure used for the audio band (baseband).

4.7.6 Analysis methodology

This study aimed at estimating a number of thresholds from parameters related to the time variance, extracted from impulse response or from the probe ultrasonic signal test, to control the averaging procedure in order to increase the overall SNR.

At the reception, each test signal frame is labeled as *valid* or *not valid* by checking the thresholds obtained from the estimated parameters. In the case of a valid frame a weighting factor is applied in order to give more importance to the frames less affected by time variance phenomena. The same procedure can also be applied to a fraction of the frame. In that case, each captured frame is split into *N* segments to provide a finer inspection of the intra time-variance effects.

A procedure was developed with the purpose of assessing and attenuating the time variance effects based on the results of the acoustic tests. The basic concept is illustrated in the flowchart depicted in Figure 4.40.

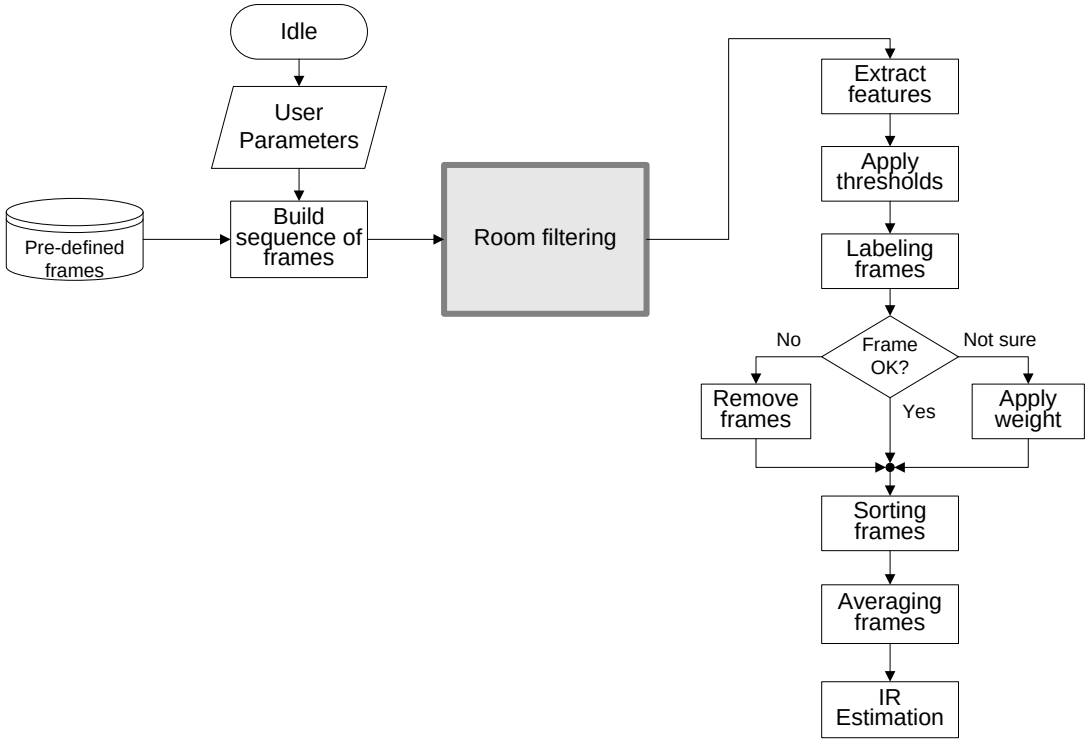


Figure 4.40 – Flowchart of the procedure for handling time variance issues in acoustic measurements.

The relevant procedures of each block are described for the standard and for the ultrasound approaches separately when necessary as follows:

- User parameters – defines the measurement approach: standard or ultrasonic. The standard approach can be evaluated by simulation. Additionally, the type of test signal (swept sine or MLS), the number of frames and the type of sorting the frames are selected;

- Building sequences of frames – composes the sequence of frames to be used in the tests.
- *The standard approach*: by simulation, a database with pre-defined time variance effects applied to the frames, such as stretching the samples randomly to simulate *jitter* (effect occurring frequently due to the overheat of the transducers devices or due to the air movements or rapid temperature variations) or stretching uniformly the frames by altering slightly the sampling rate to simulate slowly temperature variations during the tests is used.
- *The ultrasonic method*: a tone probe signal in the ultrasonic frequency band is added to the frame sequence test signal;
- Extract features – after filtering the sequence of frames by the room, the captured sound is analyzed and the relevant features are extracted. The sampling rate of 192 kHz is used on the tests given the required ultrasonic band. However, the analyses of the acquired signals are performed at 44.1 kHz by splitting and down-sampling each frequency band conveniently.
- *The ultrasonic method*: the frequency of 5 kHz is used for the demodulated ultrasound tone probe. The 500 Hz bandwidth spectrum centered at the 40 kHz tone probe is bandpass filtered to select only the ultrasound signal;
- Apply thresholds – the upper and lower threshold levels are compared with the corresponding feature to verify the conformity of the frame;
- Labeling frames – the frames which pass the conformity test (those below the lower threshold level) are labeled as *valid*. Otherwise, (those above the upper threshold level) are labeled as *invalid*. The remaining frames, those which fall in between these threshold levels, are kept for later consideration with a *not sure* label assigned;
- Frame OK? – the destiny of each frame depends of the attributed label. Obviously, *valid* frames are kept and *not valid* frames are discarded. For the frames assigned with the *not sure* label a weighting factor, following a grade impairment scale, is applied depending of the degree of likelihood from the conformity state (lower threshold level) among all the features;
- Sorting frames – this module has the purpose of sorting the frames in an ascending or descending order according to a pre-defined criterion of significance based on the maximization of the SNR after the averaging procedure, based on the frequency response degradation or by using the mean square error of the features considered.

In the ultrasonic approach, the procedure begins by first inspecting the Frequency Deviation following the Max Tone Magnitude and the Tone Magnitude Derivative features due to their higher sensitivity. If one of these values exceeds the respective threshold, the remaining parameters will be analyzed to find the expected degree of damage inflicted upon the impulse responses.

In order to evaluate the performance of this procedure, some representative simulated and experimental results were obtained and are shown in Figure 4.41.

Simulated results

The results shown in the figure correspond to the simulation of resampling the test signals as described in section 4.7.3 (a uniformly distributed random value assigned to the resampling parameter). The frames were sorted by using the criterion of minimization of the mean square error of the IR (mean of the *Correlation Energy error* feature) among all the IRs. Since minor differences are observed in between the results for the measurement techniques considered in this study the results reported cover the MLS method only.

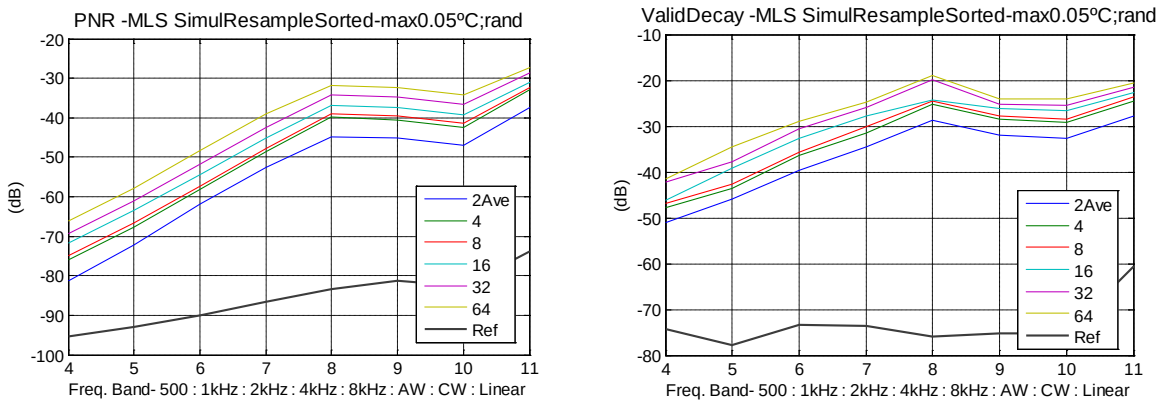


Figure 4.41 – Peak-to-noise ratio and Valid decay simulation results vs. octave bands. Random resampling parameter value is used.

When these results are compared with those shown in section 4.7.3, one concludes that the *PNR* increases by about 6dB for the cases of 4 and 8 averages. The *PNR* and the Valid Decay are degraded by the increase of the number of frames averaged. No advantage is observed by increasing the number of averages. For this case, except for the 2Ave (where the frames are very little affected by the time variance effects, intentionally), the 4 or 8Ave is the best choice, not consuming too much time.

Experimental results

Results of the standard approach and the ultrasonic method using some features are shown in Figure 4.42 to Figure 4.44. Nevertheless, only a subset of the features defined above is used as an attempt to highlight the performance of this new technique.

The standard approach

This experiment uses a test signal composed of 32 frames extracted randomly from all the measurements made with the different devices available to produce time variance in the room (a heater, a mid-size electromechanical fan, and a large pad to create strong air movements), as mentioned in section 4.7.2.

In this experiment, the *Time Lag function* τ and the *Correlation Energy error* features are used simultaneously. The procedure first calculates the maximum index corresponding to the threshold of 45 samples of deviation given by the *Time Lag function* τ. The energy of the *Correlation Energy error* is then estimated up to this index. The frames are sorted by the frame with less energy.

The first two frames in each composed test signal refer to the reference frames (not affected by time variance). Therefore, the T20 results are almost the same for the reference (1Ave) and for 2Ave for which an increment of about 3dB is observed for the PNR and the Valid Decay.

The reference and 2Ave analysis provide the right value of T20 and the 32Ave are not changed by sorting the frames, due to the fact that all the frames are included in this set.

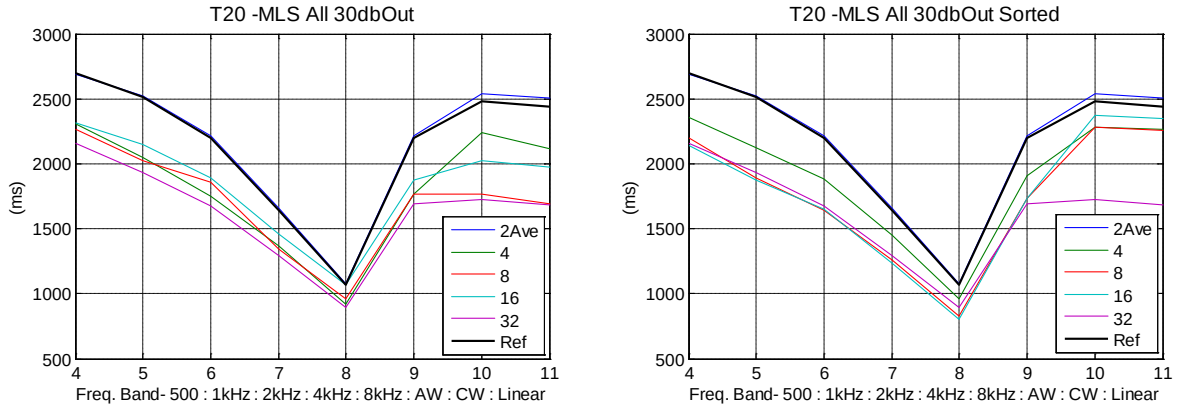


Figure 4.42 – Reverberation time, T20, in frequency octave bands for the MLS test signals with frames sorted.

As mentioned under section 4.3.2, the reverberation times decrease by increasing the number of averages. If the frames are sorted, by increasing the number of averages a decrease of the T20 in the same fashion is expected. Otherwise, the decrease of the T20 is not predictable (depends

of the order of occurrence of the frames). Therefore, the procedure of sorting the frames can be useful in real situations in order to define a stop criterion for the acoustic measurements.

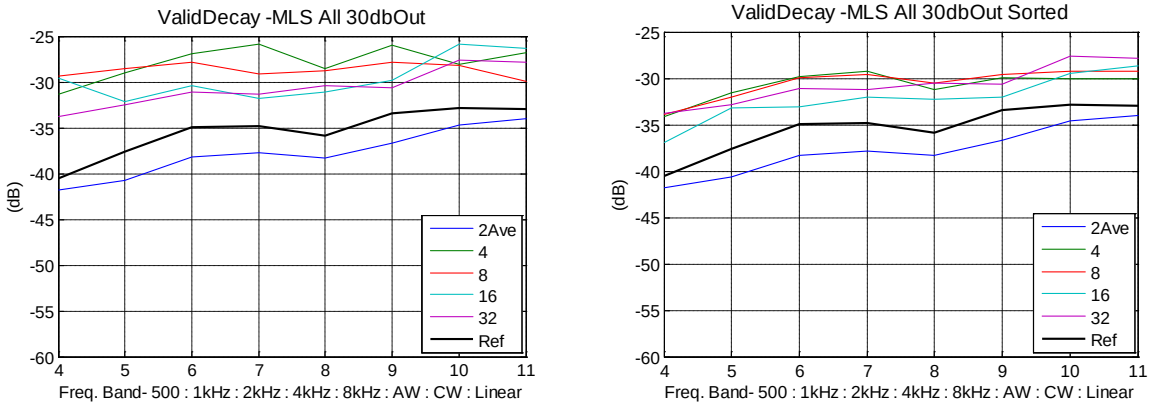


Figure 4.43 – Valid decay experimental results in frequency octave bands for the MLS test signals.

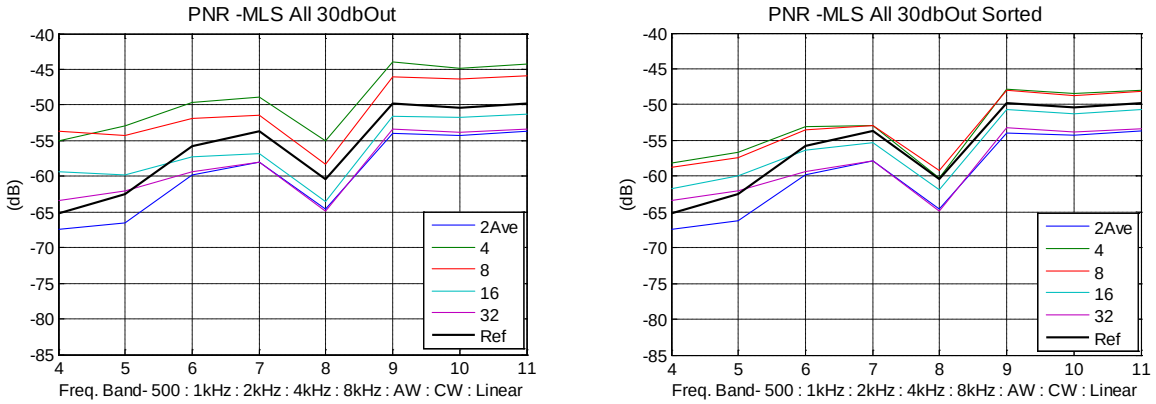


Figure 4.44 – Peak-to-noise ratio experimental results in frequency octave bands corresponding to the MLS test signals.

As a first conclusion, excluding the reference frames, the PNR increases with the number of averages, if no sorting procedure is applied. However, the increase is not 3dB by doubling the number of frames, at least for the transition between 16 and 32Ave. This fact is highlighted on the Valid Decay analysis by showing very similar values.

The Valid Decay and the PNR are improved by about 3 to 5dB for 4 and 8Ave when the frames are sorted. Although a marginal decrease is observed for the PNR in the sorted approach the Valid Decay is increased by about 2dB. This result is important since the Valid Decay is a better measure to guarantee more accurate results for the reverberation time and for the

frequency response. In fact, some deviations are found in IRs between the reference and 16 and 32Ave showing significant audible distortion. Therefore, the procedure adopted to deal with time variance effects should be a balance between the PNR, energy decay and the accuracy of the IR.

Moreover, one can state that better results are achieved for 4Ave if the frames are sorted, at least for this experiment. Nevertheless, the results could be improved with the use of more features simultaneously.

The ultrasonic method

This method consists of analyzing the set of features extracted from the ultrasound tone probe and using the results to label and sort the captured test signals in the audio band.

A simple experiment to validate this method consisted of computing the correlation of the ultrasound tone probe between the corresponding time slots of the reference frame and the remaining frames, extracting a feature similar to the Correlation Energy error used in standard approach, in conjunction with the Frequency Deviation to label and sort the captured test signal frames. A threshold of 1 Hz in the Frequency Deviation feature was used to label the frames as Valid or Not Valid.

The results with the standard approach the PNR and the Valid Decay analysis are shown in Figure 4.45 for comparison purposes.

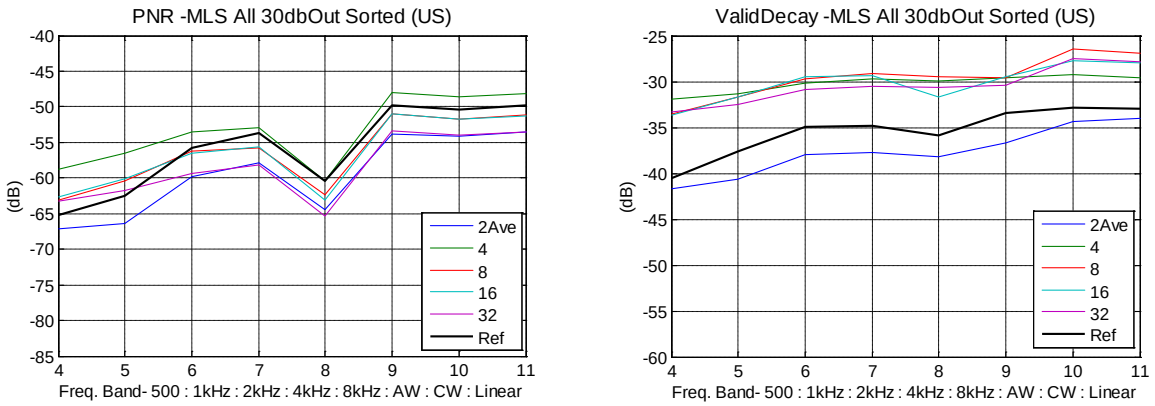


Figure 4.45 – Peak-to-noise ratio, PNR, and Valid Decay experimental results vs. octave bands for the MLS test signals.

The results, from Figure 4.42 to 4.44 and 4.45, confirm the advantage of applying the sorting procedure to the captured frames. In fact, the peak-to-noise ratio and the Valid Decay analysis show a significant increase of about 5 and 3dB, respectively, to up to 8NAve.

Discussion

This last section was all devoted to the inaccuracy of results due to problems of time variance occurring during acoustic measurements and attempts to detect and attenuate its effects. There is no doubt about the importance of this issue regarding the conditions of applicability of the Fourier formalism and, eventually, the correct identification of systems.

Nevertheless, during live performances, especially those occurring in large concert halls where the audience stands up mostly, some degree of time variance due to temperature gradients and air movement is present during the entire event. In such situations, models describing these time variance effects in a statistical approach should be designed to offer an additional level of abstraction in terms of human hearing perception of this effect. In other words, the subjective acoustical parameters are estimated from the IR in ideal conditions even if the acoustic measurements made with an audience do not account for the long-term time variance. This suggests the question: What about the effect on the long-term human perception of sound comfort and perception during a performance?

4.8 Final discussion and conclusion

Two different concepts for improvement of the SNR were presented in this chapter.

The first approach consisting of a modified swept sine and MLS techniques aiming at increasing the SNR was developed and described, together with applications and comparisons.

For this purpose, the average methods such as the standard averaging (Ave), the Frame Weighted Average (WAve), the Segment Weighted Average (SWAve) and the Spectral Segment Weighted Average (SSWAve) were used. Additionally, the Energy Decay Ratio (EDR), Energy Decay Ratio Distribution (EDRD), Power Spectral Density Distribution (PSDRD), Peak-to-noise ratio and Valid Decay analyses were examined.

The EDR and EDRD analysis showed increments of up to 7dB for the Ave and 9dB for the SSWAve method. A remarkable increase of about 13dB between Ave and SSWAve was observed for the same SNR and number of averages (NAve). The PSDRD measure revealed

that the MLS technique always shows an excess of spectral noise higher than for the swept sine and the MLS technique shows much more low frequency noise than the swept sine, especially for low SNR and/or low NAve. The Backgnd noise analysis showed, in most cases, a remarkable improvement of about 10dB of the SSWAve against the Ave method. The Valid Decay for the full band leads to a surprising conclusion that for SNR up to -18dB the threshold of 10dB is only reached when the SSWAve method is used, with the MLS. However, in the case of the swept sine technique, this threshold is achieved for SNR of -30dB with SSWAve or for -18dB for the standard average method. Therefore, the swept sine technique can lead to the same results with much lower SNR or with less number of averages. Acoustic measurements in the presence of an audience where minimization of nuisance is an issue are improved by masking the test signal with music (lowering the SNR). Further, due to time-variance effects the number of the test frames should be reduced [Paulo09b]. Therefore, the swept sine technique is more suitable than the MLS.

Finally, for environments with very low SNR, the swept sine technique always proved to estimate the acoustical parameters better than the MLS technique.

The second approach, consisting of a framework using a new method for identification and attenuation of the time-variance phenomena effects on the acoustic measurements, for situations of low SNR, using an ultrasonic loudspeaker array was developed and implemented. A number of parameters, sound features extracted from an ultrasound tone signal, were used to classify the test signal frames, in the audio band, and to define thresholds in order to improve the SNR and/or increase the accuracy of the estimated IR and consequently the frequency response. The method consists essentially on sorting and weighting each frame before averaging.

Additionally, a similar approach using an extended set of features extracted from the cross-correlation function applied to different IR is used in the case of high SNR. This approach permits to inspect the effect over the audio band and ultrasound band test signals due to time variance and comparing the performance of the swept sine and MLS.

A number of experimental and simulated analysis covering different common acoustical situations in presence of time variance phenomena were carried out. The simulated experiences

consisted of using a model of stretching the waveform to create intra-periodic and inter-periodic distortion over the test signal frames. Due to feasibility considerations these experiments were not applied in the ultrasonic approach.

In order to produce time variance phenomena artificially a set of devices comprising a heater, a mid-size electromechanical fan and a large pad, in an attempt to force significant air masses movements, were used.

The set of features selected were: (1) Time-lag function, (2) Correlation Energy error, (3) Correlation maximum ratio, (4) Relative Error, (5) Frequency Deviation, (6) Max Tone Magnitude, (7) Tone Magnitude Derivative, (8) Tonal Residual part Energy, (9) Tonal Residual part (BR filter). The analysis performed directly to the IR (audio band) uses the first 4 features only.

In a first attempt to verify the performance of the swept sine and the MLS techniques, a set of preliminary analysis to the IR, such as the PNR, Valid decay, T20, and C-Curvature parameter, were carried out. In this situation, the framework was not used, i.e. the frames were not sorted.

The results for the simulated analysis using the intra-periodic model (using a sinusoidal stretching factor) show a superiority of the swept sine against the MLS technique. In fact, the swept sine is almost unaffected by this type of distortion. Instead, a decrease of about 6dB for the PNR results is observed for the MLS. Nevertheless, a significant distortion is observed for the high end of the frequency response of the IR in both techniques. Virtually the same results are observed for the case of the Inter-periodic model proving the robustness of the swept sine against the MLS technique.

The results for the parameters analyzed on the ultrasound tone probe show significant drifts from the static values. The amplitude and phase shift variations amount to 15dB and 130 degrees of variation, respectively, mostly due to temperature gradients. The air mass movement with the fan on was less significant. However, a kind of modulation noise was noticed in the upper and lower side bands of the ultrasonic probe test signal spectrum.

The experimental results using the thresholds and sorting procedure were split in two situations: the standard approach and the ultrasonic method.

Although a number of features were considered in this study for evaluation of the methods, not all features were used. In fact, the standard approach used the Time Lag function and the Correlation Energy error features and ultrasonic method considered the Correlation Energy error used in standard approach, together with the Frequency Deviation. The remaining features are devoted to future phases of this research.

Similar results were achieved for the standard approach and for the ultrasonic method confirming the advantage of the sorting procedure applied to the captured frames. In fact, up to 8NAve the PNR and the Valid Decay analysis show a significant increase of about 5 and 3dB respectively. Moreover, the distortion over the IR and consequently of the frequency response of the system is reduced by applying this procedure.

In mid to large-size volume auditoriums different loudspeaker ultrasonic array arrangement, using multiple-ultrasonic devices, should be applied in order to cover a wider audience area and room volume by establishing multi-path direction sound propagation. Notice that a single loudspeaker ultrasonic array shows a very directive radiation pattern.

As expected, the swept sine technique is much less affected by time variance than MLS. Furthermore, if the frame length is larger than the estimated reverberation time, then the frame segmentation technique can be used to label some frequency bands less affected by a higher weight factor.

The performance of this procedure was assessed by considering the improvement on the SNR and the degradation of the T20.

The use of an ultrasound loudspeaker array has several advantages yielding implications on the sound wave propagation due to time variance. The results shown here are substantially better than the standard averaging techniques achieving 4dB maximum in some octave bands.

Chapter 5

TECHNIQUES FOR ANNOYANCE MINIMIZATION

5.1 Introduction

Acoustic measurements in large enclosures are usually carried out without the presence of people, given the high sound pressure levels of the test signals necessary to excite the room. In concert halls, listening rooms, sport events, railway and bus stations and airport halls, the correct estimation of the acoustical parameters should be performed in a real situation, with the audience inside the enclosure [Rife89].

It is well known that, under specific circumstances, some types of sounds are added to the environment to alter the soundscape quality in order to increase the acoustic comfort of people due to the presence of some particular disturbing noise. In this sense, the use of pleasant sounds, such as music material, may decrease the resulting annoyance during the measurements, while allowing a correct characterization of the acoustic parameters.

From the audience point of view, this procedure has obvious advantages, since the test signal is not actually perceived as a real disturbance of the listening program. In order to increase the signal-to-noise ratio (SNR), either the test signal length can be extended or an averaging technique can be applied, or both. In the latter situation, time-variance issues could lead to a degradation of the SNR due to the usually significant increase of the measurement duration

[Vorlander97b, Juha05]. Therefore, a compromise between annoyance, test duration and SNR was investigated. A modified swept sine measurement technique was developed, aiming at the minimization of nuisance during the acoustic tests.

5.1.1 Windowing

Windows such as the Blackman-Harris and Hanning functions are the most widely used in audio signal processing. The right choice of the analysis window, $w(n)$, is crucial in situations of low SNR. In fact, a compromise between peak factor and bandwidth is necessary to improve the overall SNR after signal processing. Therefore, the Hanning window and the Tukey window, often called the cosine-tapered window, were considered. Although the Hanning window is very often used in audio processing due to its shape, the full power spectrum density of the chirp waveform is not efficiently exploited for swept sine based signals. Actually, the signal energy shows only a maximum at the center frequency bands, and it is a minimum at higher and lower frequencies, thus reducing the effective signal energy level. This is more prominent for higher frequencies bands. As a result, a reduction of effective time-bandwidth product and consequently degradation in SNR is observed [Rao94]. Figure 5.1 shows the shape of the windows used in this study.

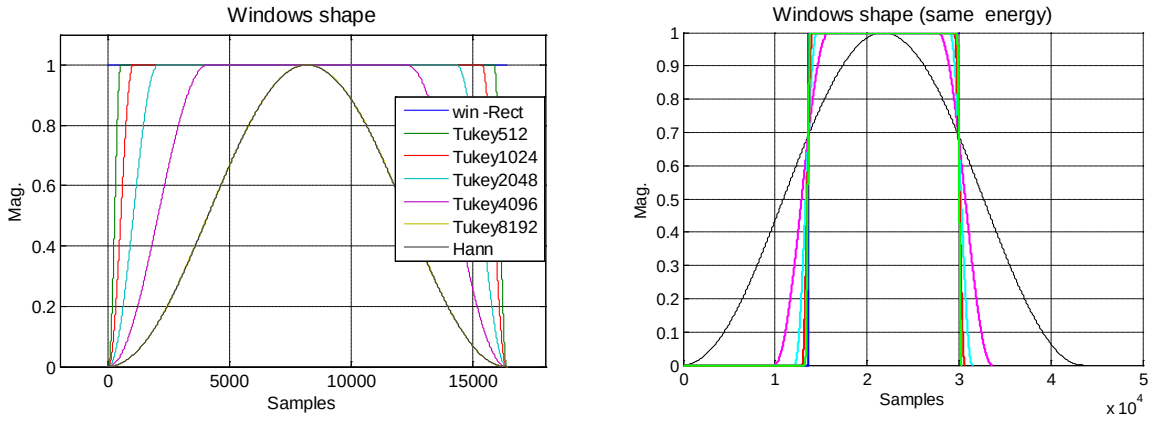


Figure 5.1 – Different shape of windows starting from the rectangular to Hanning window for the cases of: same duration (left side) and same energy (right side).

The swept-frequency signal combined with a rectangular modulation envelope allows both the maximum use of the full system bandwidth and an improvement in signal power. In addition, the required input voltage to the source can be reduced as a result of maximizing the acoustic power in the waveform. However, this approach shows significant amplitude side lobes which

degrade the SNR after the overlap-add procedure by producing errors in the reconstruction of the swept sine frame due to the non-cancelation of the cross-terms of the spectrum following the truncation of the tail of the acquired windowed signals. As a compromise, the Tukey window has the spectrum closest to the rectangular window due to its construction. The Tukey window, $w(n)$, is built from three sections: a tapering increase in amplitude, a central constant amplitude equal to 1, and final tapering decrease in amplitude where each non-constant part. The first and ending parts follow a Hanning shape. The ratio of the tapering length to the total length of each window can be denoted by α, where α=0 and α=1 correspond to the Hanning and rectangular windows, respectively, as shown in Equation (5.1).

$$w(n)\begin{cases} 1 & 0 \le |n| < \alpha\dfrac{M}{2} \\ 1/2\left[1+\cos\left(\pi\dfrac{n-\alpha\dfrac{M}{2}}{2(1-\alpha)\dfrac{M}{2}}\right)\right] & \alpha\dfrac{M}{2} \le |n| \le \dfrac{M}{2} \end{cases} \tag{5.1}$$

where M is the length of the window.

The window represents an attempt to smoothly set the data to zero at the boundaries while not significantly reducing the processing gain of the windowed transform.

Because the spectrum of the Tukey window is close to rectangular, it provides the optimum time-bandwidth product and thus highest energy and maximum SNR. The peak of the pulse compression output using a Tukey windowed chirp signal as input also would become sharper due to the high time-bandwidth product. Figure 5.2 shows the spectra of a windowed swept sine signal corresponding to the octave bands centered at 125 Hz and 4 kHz for rectangular, Tukey and Hanning windows.

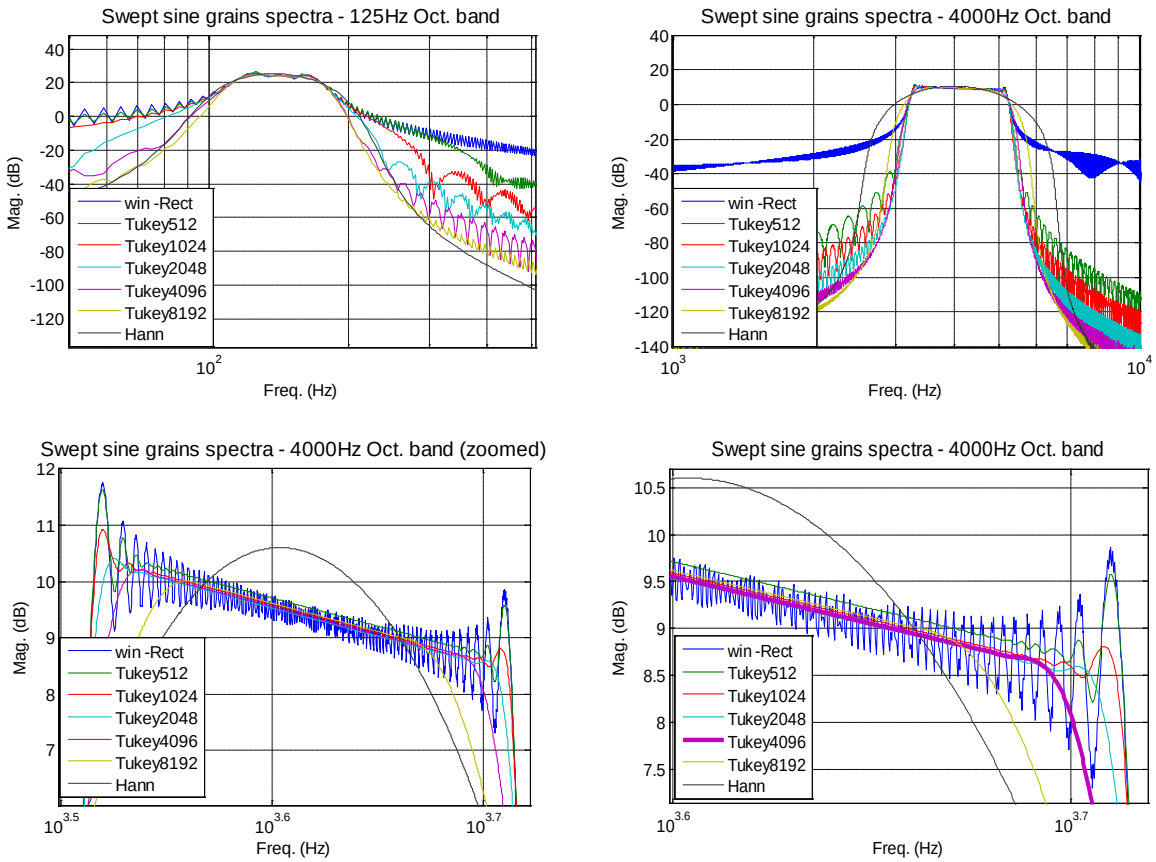
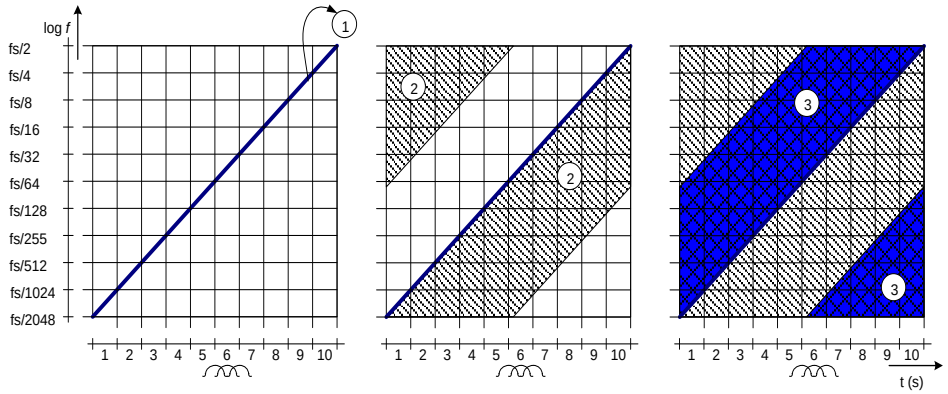


Figure 5.2 – Examples of swept sine spectra for 125 Hz and 4 kHz octave bands (top most); the down most figures depicts a zoom of the 4 kHz case to highlight the different shapes produced by the different windows.

The non-constant energy for these octave band is observed clearly for the Hanning window. On the other hand, the rectangular window and the Tukey window with α close to zero produce a considerable ripple in the pass-band. This leads to an increase in time resolution of the output signal. However, the use of Tukey also causes a ripple in the pass-band frequency spectrum due to non-smooth transition in the central frequency region as a kind of the well known *Gibbs* phenomenon. The overshoot in the pass-band is controlled by α, which for the purposes of this study has the value of 0.25 (the response is outlined in the right down most Figure 5.2). However, although this family of windows exhibit high-sidelobe levels, which could be disadvantageous for a number of audio applications, the noise produced by these sidelobes compared to the Hanning window is not relevant when the swept sine segments are used in situations of relatively low SNR. The gain of the Tukey window, with α=0.7, against the Hanning window achieves approximately 8dB at the mid-high frequency bands.

5.1.2 Band frequency noise

The time/frequency mapping of a swept sine measurement, with the frequency in a logarithmic scale, corresponds to a linear relation, as shown in Figure 5.3a). Therefore, with a convenient use of the windowing technique, according to the window size, each time segment can be assigned to a corresponding octave band or to a fraction of octave band. The chosen grid corresponds to splitting the full audio band in octaves. The effect of the reverberation of the room in the received signal, $s(n)$, can be viewed in Figure 5.3b) as the shaded area assigned by 2, for the case of a frame length longer than the RT (twice of the RT, in this case). The zone with only-noise during the measurements is shown in Figure 5.3c) as the shaded area assigned by 3.



a) b) c)

Figure 5.3 – Time/frequency map of a swept sine measurement. 1- swept sine signal, 2- area of the received signal plus noise, and 3- area of the received noise – only-noise area.

The time/frequency mapping shows several frequency bands that do not belong to the room response. Moreover, the estimation of the RIR does not take these frequency bands into account. They are decorrelated from the swept sine signal. This statement is expressed mathematically by the following considerations, using again additive noise:

$$y(n) = \left[s(n) + d(n)\right] * h(n) = s(n) * h(n) + d(n) * h(n) \quad . \tag{5.2}$$

By windowing the signal $y(n)$ in N segments

$$y(n) = s_1(n) * h(n) + s_2(n) * h(n) + \ldots + s_N(n) * h(n) + \sum_i^N d_i(n) * h(n) \quad . \tag{5.3}$$

Or, in the frequency domain

$$Y(\Omega) = S_1(\Omega)H(\Omega) + S_2(\Omega)H(\Omega) + \ldots + S_N(\Omega)H(\Omega) + \sum_i^N D_i(\Omega)H(\Omega) \tag{5.4}$$

Splitting each windowed segment in B sub-bands and omitting the Ω term for simplicity,

$$\begin{aligned} Y &= \sum_k^B S_{1k}H_k + \sum_k^B S_{2k}H_k + \ldots + \sum_k^B S_{Nk}H_k + \sum_k^B \sum_i^N D_{ik}H_k \\ &= \sum_k^B S_{1k}H_k + \sum_k^B D_{1k}H_k + \ldots + \sum_k^B S_{Nk}H_k + \sum_k^B D_{Nk}H_k \end{aligned} \tag{5.5}$$

The estimation of the RIR is obtained by correlating Y with the inverse spectrum of the swept sine signal, S_{inv}, then

$$\begin{aligned} Y &= S_{inv}\sum_k^B S_{1k}H_k + S_{inv}\sum_k^B D_{1k}H_k + \ldots + S_{inv}\sum_k^B S_{Nk}H_k + S_{inv}\sum_k^B D_{Nk}H_k \\ &= S_{inv}S_{11}H_1 + S_{inv}\sum_k^B D_{1k}H_k + \ldots + S_{inv}S_{NB}H_B + S_{inv}\sum_k^B D_{Nk}H_k \\ &= S_{inv}\left(S_{11}H_1 + \ldots + S_{NB}H_B\right) + S_{inv}\left(\sum_k^B D_{1k}H_k + \ldots + \sum_k^B D_{Nk}H_k\right) \end{aligned} \tag{5.6}$$

Finally

$$Y = S_{inv}SH + S_{inv}\left(\sum_k^B D_{1k}H_k + \ldots + \sum_k^B D_{Nk}H_k\right). \tag{5.7}$$

In Equation (5.7), for each segment, the noise bands belonging to the only-noise areas are discarded, since they are not involved in the estimation of the RIR. To emphasize this result, a test consisting of placing the swept sine stimuli in a filtered noiseless track (noise removed from the area 2) was conducted, as shown in Figure 5.4.

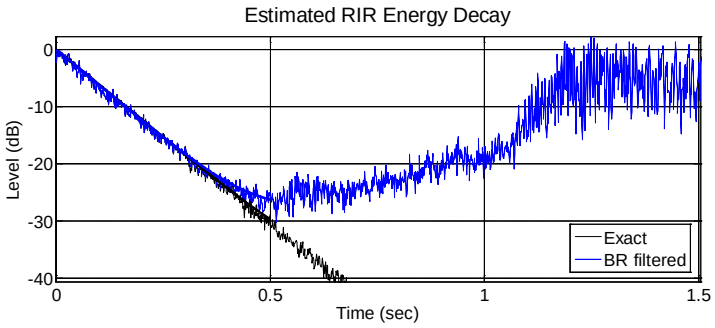


Figure 5.4 – Estimated RIR Energy Decay for the case of noise Band Reject filtering belonging to the area 2;the SNR is -25dB.

This result shows that the estimation of the RIR is achieved by applying a convenient synthesis window to the estimated energy decay curve in order to remove the noise on the tail of the impulse response.

These results are totally correct for the RIR estimation after the application of the time averaging methods, at the end of the Segmented Swept Sine technique. However, the MS value of each segment could be degraded if large amounts of energy concentrated outside the acoustic room information area occur, see area 3 in Figure 5.3c). Therefore, a filtering procedure was implemented by a filter bank [Paulo07a] with pass-band of each filter proportional to the expected RT, according to Figure 5.5a).

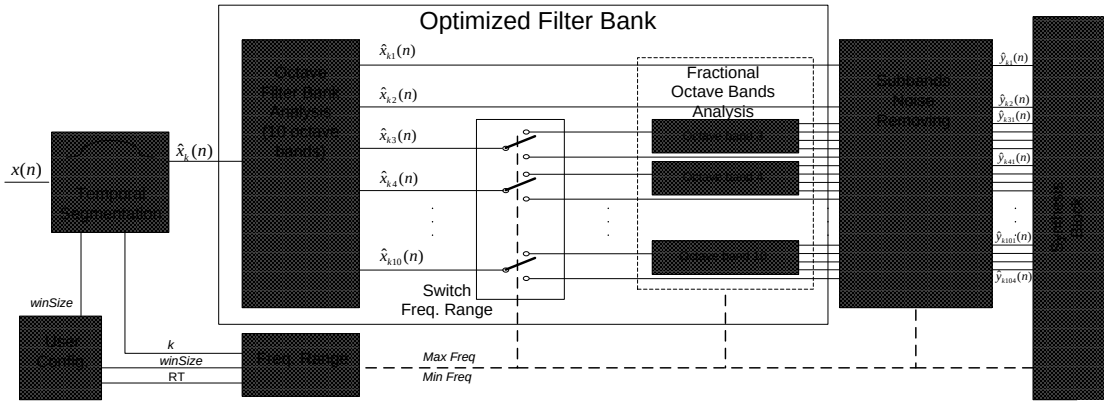
Two possible techniques can be followed to analyze the signal frequency contents: (i) a static filter bank, which is a parallel bank of band-pass filters covering the entire spectrum of interest [Vaidyanathan90], or (ii) a tracking band-pass filter. The right choice is a compromise between narrow band-pass filters for the static filter bank and a continuous filter coefficients updating process.

Static filter bank – This method consists of the following steps:

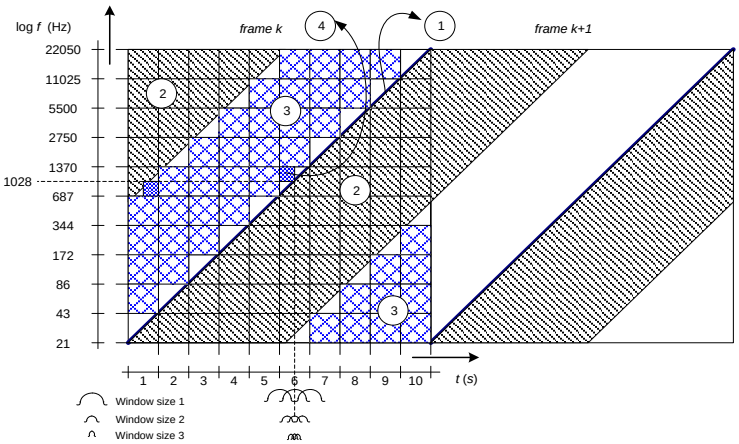
1. The user defines the duration of the swept sine in seconds, the size of the window and the number of averages. This duration must be at least twice the reverberation time expected to guarantee a chance of removing some frequency bands;
2. define the areas where acoustical room information takes place, prohibited area; the remaining area is useless for the impulse response, non-prohibited area;
3. the algorithm adjusts the window length in order to keep a correspondence between each window edge (time domain) and the octave scale (frequency domain);
4. divide the input signal captured, *y*, in segments using the designed window;

5. divide each windowed segments in octave bands and remove the octaves bands of the non-prohibited area; this is the coarse removing process (outer loop);
6. check the band limits that each window corresponds to; usually, each window corresponds to a fraction of an octave;
7. divide the octave frequency band, where this segment belongs, into up to eight subbands (depending of the window size) and remove the fraction octaves bands of the non-prohibited area; this is the finer removing process (inner loop) the inner loop is executed only for the octave bands above the second;
8. apply the average technique, if selected;
9. apply convolution operation between the resulting signal, y_F, and the inverse filter, in order to recover an estimate of the RIR.

The block diagram and the time/frequency mapping are depicted in Figure 5.5b).



a)



b)

Figure 5.5 – Procedure used in the Optimized Filter Bank applied to swept sine technique. a) Block diagram; b) Exemplification of the method. 1- swept sine signal ; 2- prohibited area ; 3- non-prohibited area for octave bands; 4- non-prohibited area for fraction of octave bands; this scheme is evaluated for a sampling frequency of 44100 Hz and a window size 2.

Tracking band-pass filter – The cut-off frequencies of the tracking band-pass filter, fc_1 and fc_2, for each segment, are calculated by

$$f_{ck} = f_i \exp\left\{\frac{Ns_k}{N_T}\ln\left(\frac{f_f}{f_i}\right)\right\} \qquad k = 1,2 \tag{5.8}$$

with
$$\begin{vmatrix} Ns_1 = 1 & Ns_1 \le 0 \\ Ns_1 = (Ns_i - RTe)\, f_S & Ns_1 > 0 \\ Ns_2 = Ns_f\, f_S & Ns_2 \le N_T \\ Ns_2 = N_T & Ns_2 > N_T \end{vmatrix}$$

where f_i is the initial sine sweep frequency, f_f is the final sine sweep frequency, N_{S1}, N_{S2} are the initial and final samples corresponding to cutoff frequencies, $N_T = T_x f_s$ is the frame duration in samples, T is the frame duration, f_s *is* sampling frequency, NS_i, NS_f are the initial and final sample position of the analysis window in the frame (t_i and t_f referenced in Figure 5.6), and *RTe* is the estimated reverberation time.

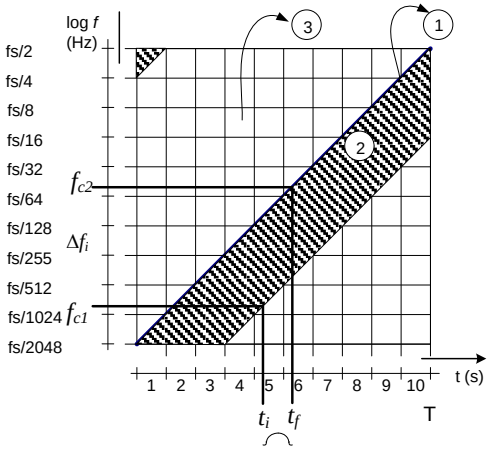
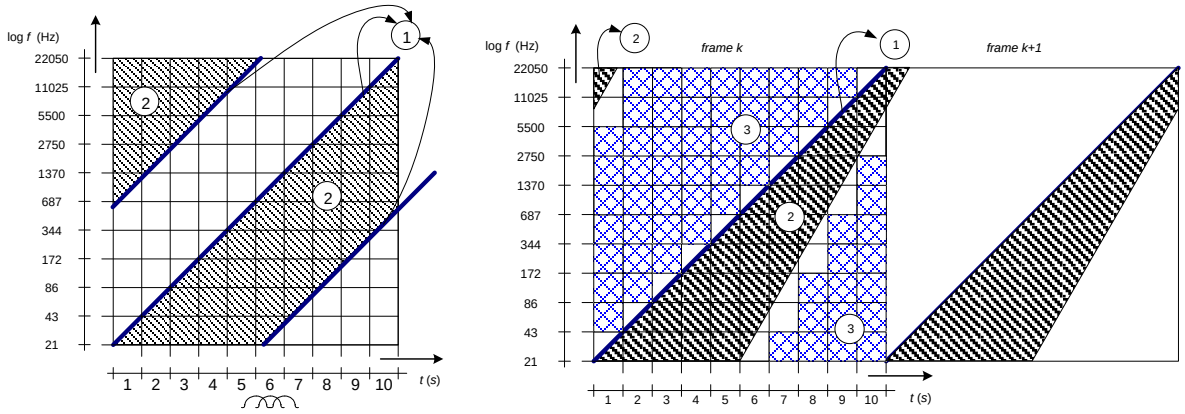


Figure 5.6 – Relationship between the pass-band, Δf, of the BP and the window duration. The marked areas are the same as in Figure 5.3.

This scheme was tested with a sampling frequency of 44100 Hz. The application of these methods has the challenge of prediction the impulse response duration to make the truncation safely not to degrade significantly the low frequency response, tail of the RIR, bearing in mind that the SNR is low. The increase of the test sweep duration permits a narrower bandwidth to nest the test signal. In fact, for the same room, same energy decay curve, decreasing the

frequency rate allows the energy of the past band frequencies to vanish in a closer vicinity of the instantaneous frequency of the sweep.

Furthermore, using different arrangements for the frame ensemble the SNR could be increased after applying the average technique. For instance, using a frame length longer than the expected RT, allows the next frame be sent to the room after a time delay from the previous frame corresponding to the RT. This guarantees that the energy has almost vanished in the room. In practice, a guard margin interval should be considered due to artifacts generated by the filtering process. In this example, for each frame time slot, two complete swept sine frames are sent, implying a gain of 3 dB for the SNR. The signal test frames are conveniently overlapped to avoid time aliasing. This scheme is shown in Figure 5.7 a). Furthermore, in real conditions, the energy decay rate, and consequently the RT60, analyzed in sub-bands, usually decreases with the increasing frequency due to the sound absorption of the enclosure boundaries. Therefore, a larger area without acoustical room information can be achieved as shown in Figure 5.7 b). The areas marked with numbers 2 and 3 correspond to the same as in Figure 5.3.



a) b)

Figure 5.7 – Alternative scheme used to increase the SNR; a) RT values equal to all frequency bands; b) higher energy decay rate for high frequency bands; the areas assigned by numbers are the same as in Figure 5.3.

As a result, the drawback regarding the acoustic test duration inherent to the frame length enlargement can be reduced. In the case of a decrease of the RT with the raising frequency, the

non-prohibited area has been enlarged significantly allowing a better masking effect at the higher frequency bands and larger noise energy removal.

5.2 Test frame based on swept sine grains

A number of authors [Farina00, Muller01] showed that the swept sine technique provides accurate estimates of room impulse responses, $h(n)$, even in situations of low SNR, non-linearity and slight time-variance of the system. Moreover, the logarithmic sine sweep signals belong to a special type of tonal signals in such a manner that the equal length segments acquired by splitting the total swept sine frame in N parts, named grains, when using a convenient window, can be assigned to a fractional octave band [Paulo05b, Paulo07b].

The advantages of the swept sine over the MLS technique are many, such as (i) lower Crest factor for narrowband signals, keeping the SNR with reasonable values, (ii) not needing filtering to produce the frequency bands (notice the elegance of the swept sine signals by using a *PR* filter as the analysis window), (iii) less sensitive to the phase errors occurring in the synthesis stage due to filtering process of the grains. These advantages were sufficient to choose swept sine signals only throughout this work.

This research was aimed at drawing guidelines for the use of swept sine grains in acoustic measurements by creating tempered musical scales by assigning the perceived pitch of each grain to a musical note for technical-artistic compositions, e.g. via midi files. Another application is to use a perceptual human approach, by taking the simultaneous masking effect into account.

5.2.1 Basic concept

Considering the distributive property of the convolution operation, the grains can be sent separately to the room, and at the reception the full response is assembled from the grains by applying a convenient amplitude compensation followed by the overlap-add based procedure given by

$$s(n)=\sum_{k=0}^{N-1} seg_k(n-kSL) \tag{5.9}$$

$$res'(n) = s'(n)*h(n) = h(n)*\sum_{k=0}^{N-1} seg_k(n-kSL-D)$$

$$= \sum_{k=0}^{N-1} h(n) * seg_k(n - kSL - D) \tag{5.10}$$

where $s'(n) = \sum_{k=0}^{N-1} seg_k(n - kSL - D)$ stands for the swept sine signal split into N grains, seg_k, with a delay D in between and SL is the length of the grains.

$$res(n) = \sum_{k=0}^{N-1} h(n) * seg_k(n - kSL) \tag{5.11}$$

This procedure overcomes the task of assigning a specific grain to an octave or fraction of octave band. Therefore, measurements covering the full audio band of interest, as in the standard swept sine technique, are possible by using grains of arbitrary length.

5.2.2 Musical signals

The swept sine is a sort of sinusoidal signal, therefore, the grains built from long test frames (showing slow instantaneous frequency rate) can be assigned to the notes of the musical scale to be used in a creative perspective. Therefore, these musical compositions can be viewed as test excitation signals allowing the estimation of the acoustic parameters with the presence of an audience.

With a properly composed test signal and filtering techniques, using polyphony musical compositions as test signals is possible. In fact, there is a constraint due to the proximity of the frequency band occupied by each grain or/and due to its time duration. Therefore, the grains are discarded, not used to build a swept sine frame, if they have overlapping spectra or if a minimal band guard between them is not verified or if the length of the grain does not guarantee sufficient acoustical energy delivered to the system.

Obviously, the acoustic tests are more time consuming, but after all, the room acoustic parameters are estimated by using a pleasant sound making the bridge between science and music and the performing arts field.

5.2.3 Limitations and enhancements (grain truncation)

Since grains are used to reconstruct the full swept sine frame, by overlapping each grain partially, the SNR shows a theoretical reduction of about 3dB when compared to the standard swept sine method (full frame), considering a noiseless environment (ideal situation). Here, $SNR \approx 10\log_{10}(RT * fact/SL)$ where RT is the estimated reverberation time in seconds and *fact* =

$[20\log_{10}(2^{Nbits})]/60$ is a correction factor accounting for the excess amount of time needed for the room impulse response, RIR, to reach the quantization noise floor, starting from the RT value, with *Nbits* being the number of bits used by the ADC device. This arises since the information about the room carried by the tail of each grain must be kept in the overlap-add procedure. As a first approach, a reduction of the SNR of about 3dB due to the background noise by halving the length of the grains is achieved as

$$R_{gr} = 10\log_{10}\left(1+\frac{RT}{SL}\right). \tag{5.12}$$

Figure 5.8 depicts the synthesis window scheme used at the reception to account for the energy decay of each grain corresponding to a Tukey window of the length of the grain plus Δwin+RT. The Δwin parameter reduces the distortion produced by the edges of the captured grains controlling the degradation inserted by multiplying windows and by enlarging the grain.

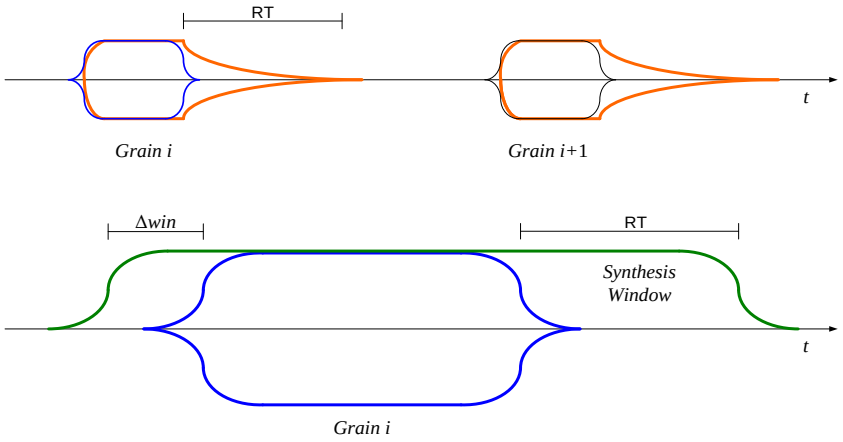


Figure 5.8 – Scheme of the synthesis window used to account for the energy of the tail of the grain *i* used to overlap with the grain *i*+1; Δwin is ¾ of the window length.

Figure 5.9 depicts the Peak-to-noise ratio vs. octave frequency bands for different segment lengths with a SNR of 30dB and white noise as background noise. As expected, about 3dB of gain is achieved by doubling the grain duration. This finding suggests that the averaging method be used in order to keep the SNR acceptable.

The synthetic *IR*, $h_{syn}(n)$, used in the simulations was generated by assuming diffuse field conditions, given by $h_{syn}(n)=\beta_n \exp\{-Cn/fs\}$, where $\beta_n \sim N(0, \sigma_i^2)$ represents independent samples of a normal distribution with $\mu=0$ and $\sigma_i^2 = 1$, and $C = 3\ln 10/RT$ is the acoustical damping

factor. This simple *IR* model assures linear energy decay curves (shape of the squared impulse response) and equal decay energy rate in all frequency bands.

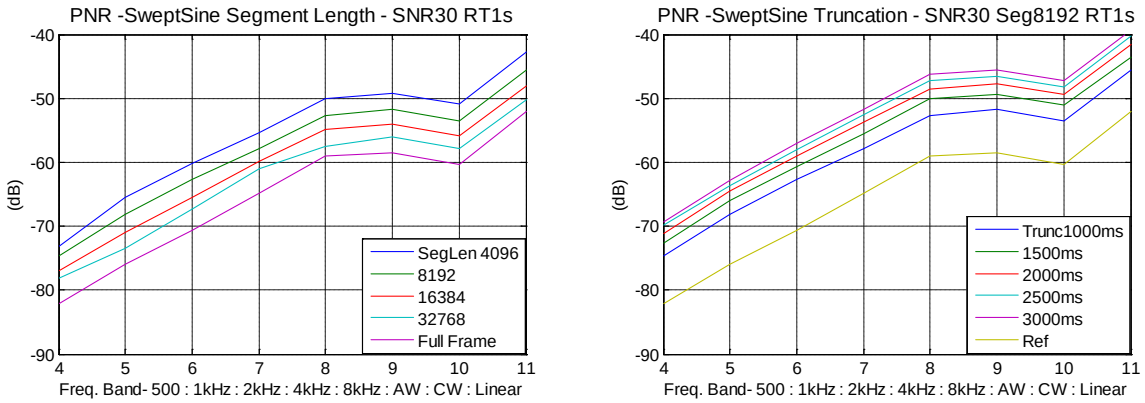


Figure 5.9 – Peak-to-noise ratio, *PNR*, vs. octave bands for different grain lengths (left) and for different amount of grain truncation (right), in samples.

In real situations, a truncation procedure is applied to the tail of each acquired grain in order not to degrade the SNR significantly. Therefore, the amount of the truncation and the overlap of the captured grains are controlled to yield the degree of degradation on the final results. The effect of truncation is sketched in Figure 5.9 (right side) for an input SNR of 30dB.

This results show the importance of using an iterative procedure to decide the amount of truncation that should be applied to achieve the specified overall SNR.

In the next section, the use of swept sine grains with perceptual models is dealt with some detail. Anticipating this approach, the grains are implanted throughout the audio material being properly masked by the audio contents. For this purpose, a band-reject filter with a carefully chosen bandwidth is applied to the zones where the grains are about to be placed following the auditory perceptual issues. The bandwidth is related to the reverberation time and the sweep rate is bounded by the human perception of the artifacts. The maximum allowed bandwidth depends of the type of the masking effect that has to be dealt with.

Two different approaches are alternatively used in the filtering process. Nevertheless, both approaches use a fixed band-reject filter (with constant coefficients) centered in the chosen frequency.

The first approach simply consists of applying the filter to the specified spectral zone. The grain is implanted with a delay corresponding to the time needed to give a chance for the energy of the music material to vanish before the grain starts.

The second approach consists of three steps. In the first step, a fixed filter (of constant coefficients) centered at the frequency is applied to attain the considerations stated above. In the second step, a band-reject tracking filter with adapting coefficients is applied to assure a varying frequency response corresponding to the sweep sine rate. The third step follows the same principle of the first step, i.e. a fixed band-reject filter is applied after the ending point of the grain to allow the grain energy decay do not fall over the noise. Figure 5.10 depicts a basic scheme of this approach. Notice the fact that the reverberation time of the room can be estimated by the slope of the grain spectrum.

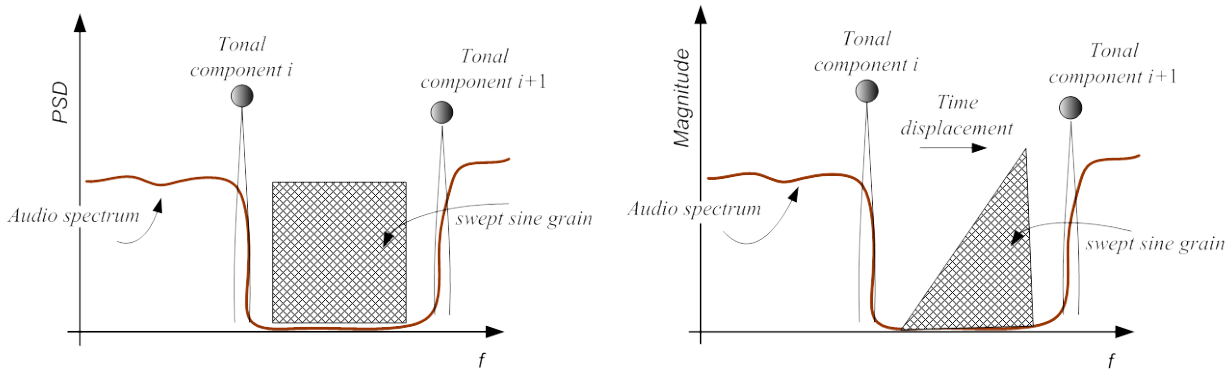


Figure 5.10 – Scheme of filtering process used to nest the swept sine grains, PSD (left side) and spectrum (right side).

A number of tests using white noise $\beta_n \sim N(0, \sigma_i^2)$ with $\mu=0$ and $\sigma_i^2 = 1$ and a synthetic impulse response signal were carried out in the laboratory in order to account for the limitations of the effect of the truncation in controlled conditions. The results correspond of using a RT60 of about 0.5 s and an overall SNR of –35dB.

A first test was carried out with the synthetic *IR* to evaluate the performance of this technique for different bandwidths of the band-reject filter used to nest the grains and to assure equal decay energy rate in all frequency bands. For example, the value 120% RT gives the bandwidth used related to the estimated RT and the sweep rate, which in this case an excess of 20% of the

RT is used to control the SNR. A considerable amount of degradation for the RIR energy error is observed for the Hanning window, as shown in Figure 5.11 and 5.12.

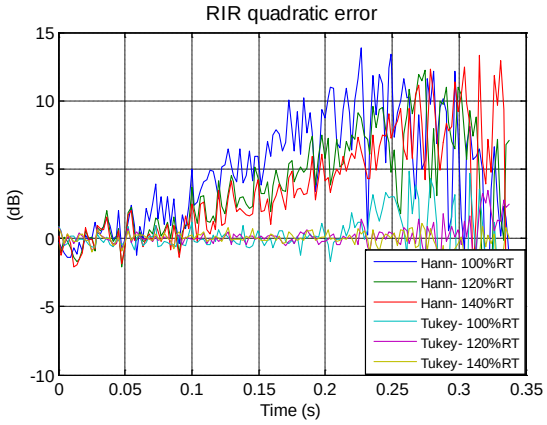


Figure 5.11 –RIR energy error (simulated results) for Hanning and Tukey windows.

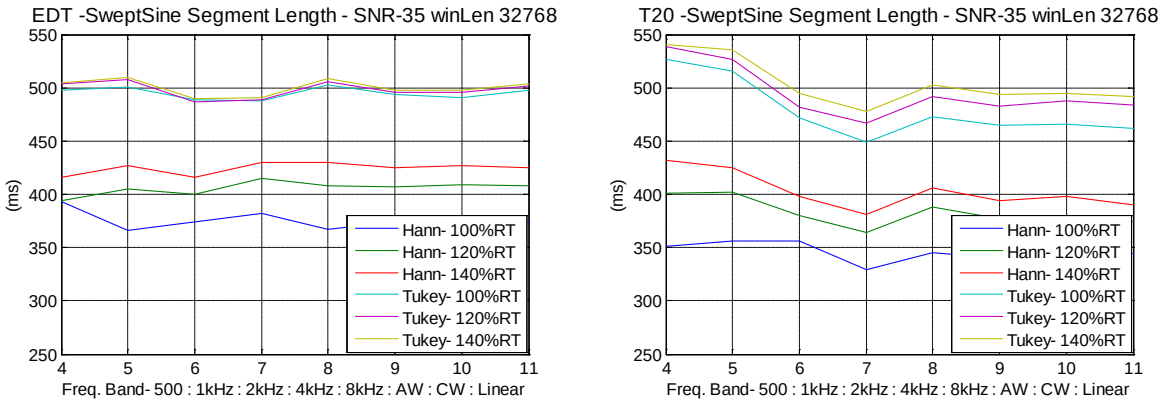


Figure 5.12 – Early Decay Time and Reverberation Time vs. octave bands for Hanning and Tukey windows; the reference EDT and T20 are 500 ms for this study.

The figures show that the resulting error is low when the Tukey window is used (approximately 6% of error for the worst case). Moreover, the use of Hanning window leads to a PNR of about 20dB, implying an insufficient dynamic range for accurate results.

5.3 Perceptual models based techniques

A new approach using the music program to mask the test signal by using an auditory perceptual model (PM), with the aim of reducing audience nuisance, was considered. From the perspective of the acoustic measurements, the music is the disturbing signal (background noise). Therefore, the music tracks are referred to in this study invariability as noise or music and a swept sine signal, based on a logarithmic sinusoidal sweep, is used as the test signal.

The swept sine technique was chosen since (i) it can lead to high signal-to-noise ratios, (ii) it is robust against non-linearities and time-variance, and (iii) being a special kind of a tonal signal, it can be masked more easily by the music according to the properties of the human auditory system.

5.3.1 General

The challenge of this approach consists of choosing a variety of music tracks, $d(n)$, that can be used together with the swept sine method, $s(n)$, with the trade-off between minimizing nuisance and keeping the SNR high enough to estimate the room impulse response accurately. Moreover, the acoustic measurements can be made with the presence of people assuming the test signals are not perceived by the audience.

A general model diagram is depicted in Figure 5.13. The spectral modification stage is responsible for analyzing/processing the spectral contents of the music track to nest the test signal conveniently. At the reception, on the test signal assembling stage, the filtered swept sine frame is identified from the captured composed signal, $y(n)$, and the resulting frame, $res(n)$, is assembled with the help of the *Side Information*. The *Side Information* consists basically of a set of parameters such as the location, magnitude shape and length of the signal test.

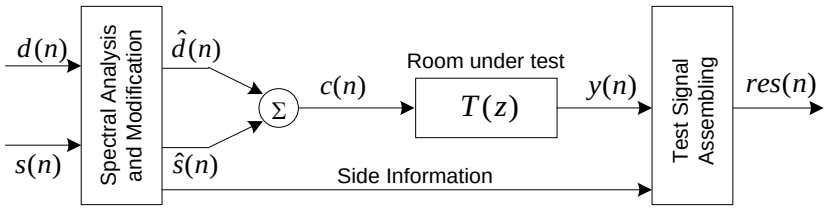


Figure 5.13 – General model diagram to nest the grains with noise added to the test signal.

This model does not consider the ambient noise that is here made up of the noise generated by people or by ancillary electro/mechanical systems. This is a reasonable assumption for situations where the SNR between the test signal and the ambient noise is large.

The perceptual model used in this approach considers only the simultaneous masking effects, does not use the time masking effects and is applied mostly in rooms with reverberation time above 200 ms. Therefore, when the masking sound stops, its energy still remains for a period of time, depending on the reverberation time, making the time masking effect inappropriate [Painter00, Ferreira98].

The spreading functions in simultaneous masking are level dependent of the tonal masker and do not follow a linear behavior. In fact, the shape of the maskers is broader for higher sound levels and asymmetries around the tonal masker arise (for narrow band signals, the masking effect covers a larger zone for the frequencies above the masker component), thus allowing a higher flexibility (larger bandwidth and more acoustic power delivered to the room) to insert the test signal. These facts are more prominent in the higher frequency bands (above 500 Hz).

The perceptual model used in this procedure is basically a modified version of that described in the ISO 11172-3 (MPEG-1) Psychoacoustic Model 1 standard [ISO/IEC11172-3].

The noise maskers are estimated as residual power spectral densities within each critical band not associated with the tonal maskers. All energy associated with the spectral components in each critical band that have not contributed to a tonal masker is then concentrated into a single noise masker. Therefore, this frequency band cannot be removed to insert the test signal. However, the identification of the noise maskers by the algorithm is important in order to decide adequately where the test signal can be placed. Basically, these locations correspond to the frequency spectrum ranges where the tonal masking threshold exceeds the noise masking threshold.

An error function is then defined as the difference between the tonal masking threshold, *TMT*, and the global masking threshold, *GMT*, referred to as the *Tonal to Global Masking Ratio* function, *TGMR*. This function is a measure of the noise masking threshold contribution to the global masking threshold. The algorithm then seeks for the frequency bands with the lower *TGMR*.

These aspects are shown in Figure 5.14. In this example (upper image, corresponding to the spectrum analysis of a period of the STFT) the *TGMR* indicates that for the low and high frequency bands (approximately for the frequency up to 1000 Hz and for the band of 10 to 15 kHz, bins up to 12 and between 125 to 180) the algorithm is not able to mask the test signal conveniently. The *GMT* is mostly influenced by the *Noise Masking Threshold*. However, for low frequency bands, the noise shows a strong harmonic behavior. The results shown in the lower part of **Error! Reference source not found.** correspond to the analysis of a piece of audio of

about 7.5 s long in 3D. The left side image shows the Tonal Masker of the audio (with the noise part removed) and the right side image shows the *TGMR*.

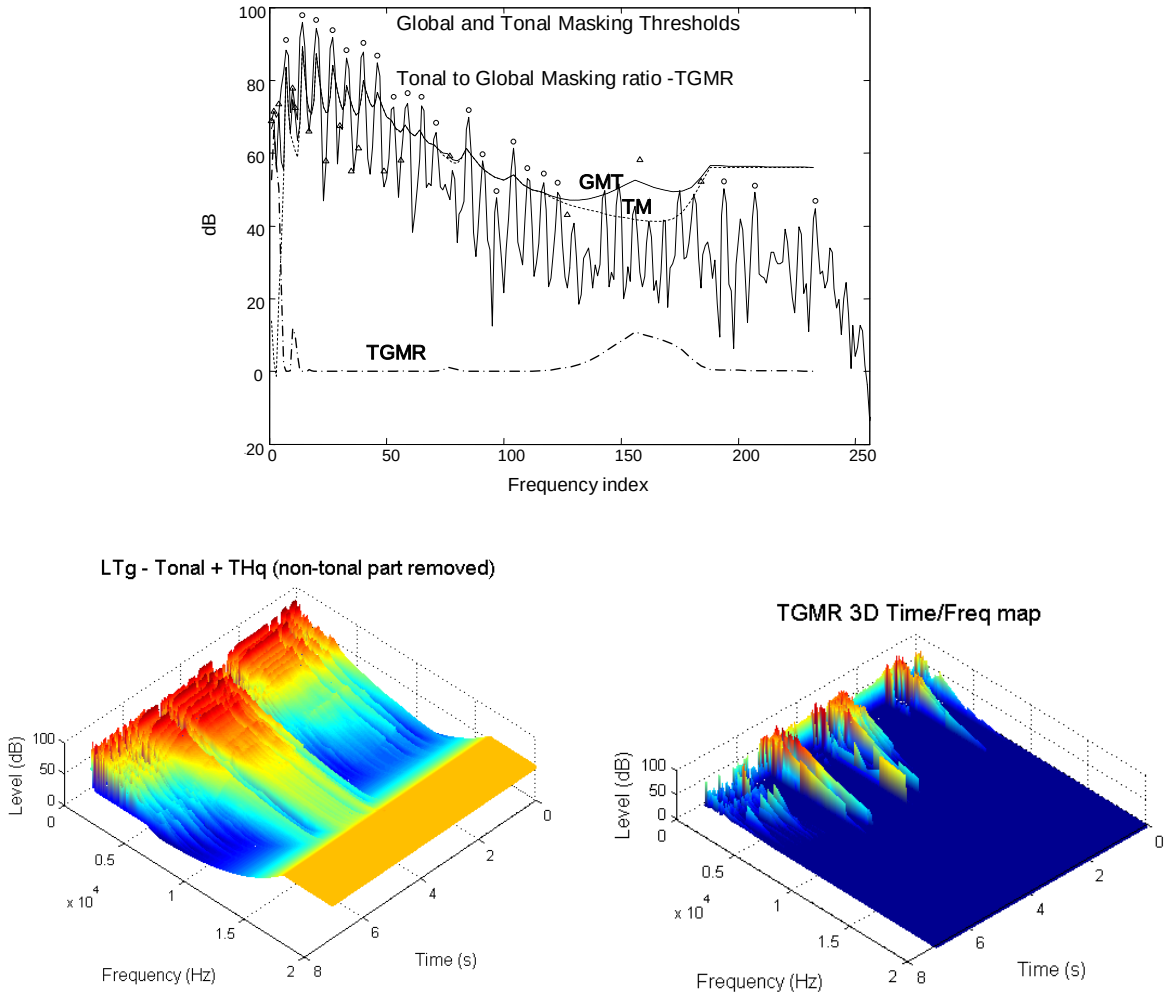


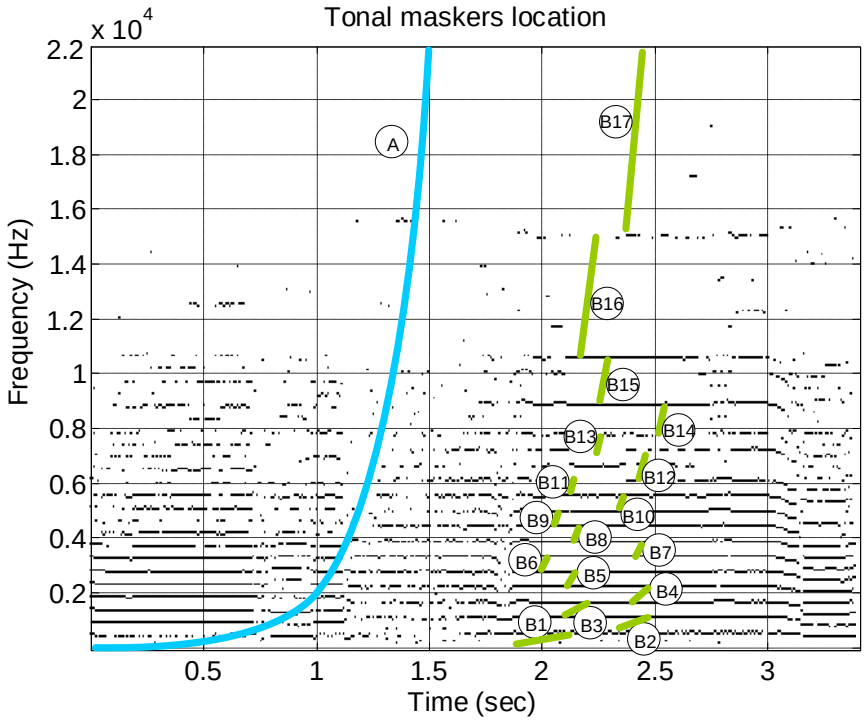
Figure 5.14 – Tonal to Global Masking Ratio function, and the Global and Tonal Masking Thresholds; the circle and the triangle refer to the tonal maskers and to the noise maskers, respectively. 512 points are used in the *fft* and 44100 Hz for the f_s.

For isolated tonal maskers, when the masking shapes of two adjacent tonal maskers do not interfere with each other (the tonal maskers are located in different critical bands), the algorithm tries to explore just the spectrum above the tonal masker frequency. However, if the algorithm is not able to cover the full audio spectrum both sides are used.

Because playback levels are unknown during the psychoacoustic signal analysis, in perceptual audio coding a normalization procedure is required in order to estimate the sound pressure

levels conservatively from the input signal. The calibration of the maximum levels is based on the choice of a range of pre-stored values according to the maximum expected playback sound pressure level applied to the room or by means of an *in situ* acoustic measurement. The perceptual model can thus be optimized by considering each particular situation.

Two different approaches for nuisance minimization are considered, according to the scheme depicted in Figure 5.15. This figure shows the spectral location of the tone maskers of a piece of audio, estimated by the use of pitch continuation techniques, for the two possible arrangements, A and B, regarding the insertion of the swept sine test signal in the music material. Scheme A uses a full swept sine frame. In scheme B, the swept sine frame is split into several segments, the grains (the contribution of all grains covers the whole audio band of interest). Anticipating our description of the techniques, these approaches are named *Test signal plus noise with modification*, TSNM, and *Segmented test signal plus noise with modification*, STSNM, respectively, as will be described in detail in the following sections.



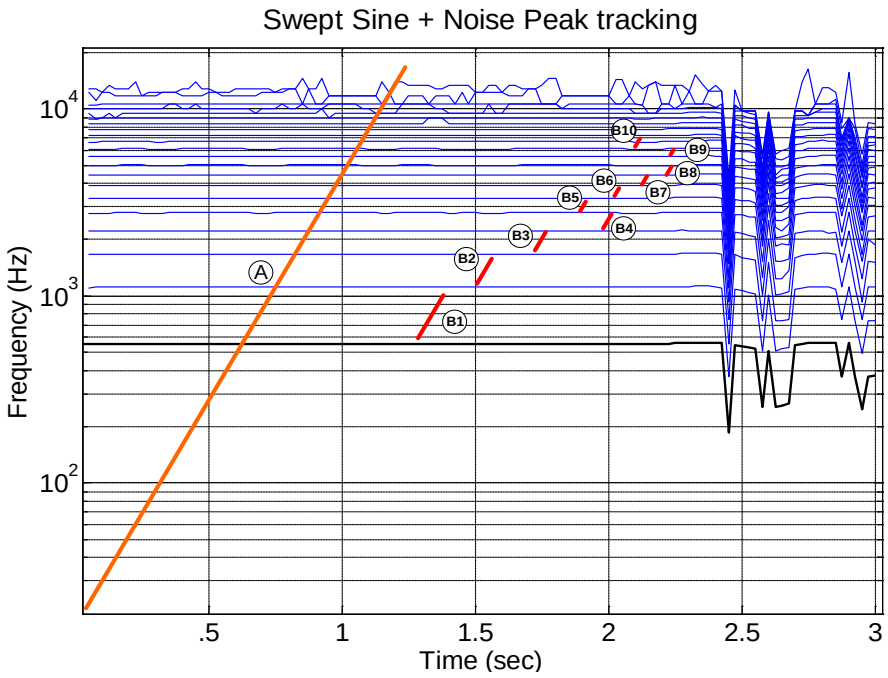


Figure 5.15 – Tonal masker tracks location of a piece of audio (with fundamental frequency of about 550 Hz) vs. time and the two schemes of insertion of the test signal for frequency in linear scale (top most image) and logarithmic scale (bottom most image).

As observed in Figure 5.15, the full swept sine signal forces the test signal to be placed over the tonal components of the music material, TSNM approach. This procedure can degrade the overall SNR and distort the room response significantly and there is a good chance that the test signal will not be conveniently masked for a number of frequency bands. As an attempt to overcome this weakness, the swept sine signal is split into several segments, grains, which are smeared throughout the time/frequency map of a music track, STSNM approach. However, as the frequency is increased, the time-frequency rate also increases due to the logarithmic law of the test signal, and consequently some grains have not the chance to be placed in between the tonal components, especially for the high frequency bands, keeping a reasonable SNR in those frequency bands. This constraint can be overcome by increasing the length of the swept sine frame.

5.3.2 Sinusoidal modelling of tonal music events

Most of the techniques presented in this chapter are based on the simultaneous masking effect using tonal components according with the human auditory system, as mentioned above. Therefore, the algorithms developed search for the tonal zones of the music material to place

the test signals conveniently. However, as pointed in the previous section, some grains have to cross some tonal components inevitably in order to have reasonable time duration, providing sufficient sound power delivered to the system under test for those frequency bands. Thus, an additional research with the aim of investigating the degradation of the test signal due to the tonal components that fall over the grains was carried out, keeping in mind that the masked signals, the grains, show intensity levels below the tonal components of the music of more than 20dB in order not to be perceived by the people.

Moreover, the tonal components of the musical instruments are not stationary in the sense that they are not pure tones.

Figure 5.16 shows a typical spectrogram of a saxophone music piece emphasizing the steady-state tonal components parts. Typically, associated with this type of signals there is some kind of amplitude and frequency modulation due to the musical expressivity applied by the performer, also called *tremolo* and *vibrato* audio effects. In the zone marked by B this result is visually clear. In fact, a frequency deviation of tens of hertz and an amplitude depth of 50% are often expected for the sounds produced by a range of musical instruments. In general, all musical instruments, with the exception of the classic keyboards, are able to produce frequency modulation and associated with this type of modulation the amplitude of the sound is somehow modulated too. However, the control of the amplitude modulation by the musician is achieved for the wind and brass instruments only and, of course, for the human natural sound producing instrument, the voice, by controlling the airflow using the diaphragm muscle. Zone A is characterized essentially by the absence of modulation, at least produced intentionally. Nevertheless, some degree of variation of the pitch and intensity of the tonal components is observed providing an enrichment of the global sonority of the musical instruments.

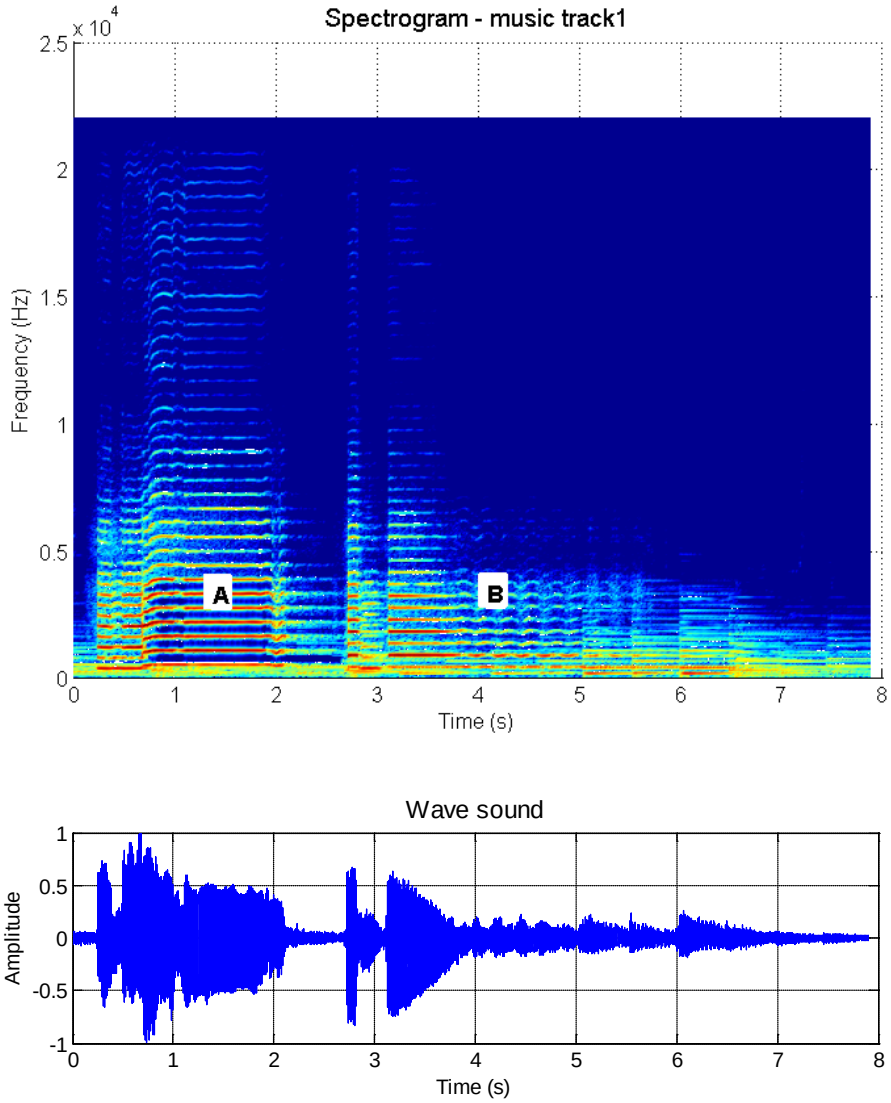


Figure 5.16 – Spectrogram and the sound waveform extracted from a saxophone music piece emphasizing the stationary tonal components parts. Shown in the B zone is a strong modulation both in amplitude and frequency applied by the musician.

The spectral richness of tonal components for some musical instruments allows the usage of the full audio band by sending several grains to the room simultaneously, at different frequency bands.

To accomplish the evaluation of the damage produced over the test signal due to the tonal components, a number of arrangements using a composite signal, $c(n)$ composed by multi-tonal components masking partially the swept sine signal, $ss(n)$, with *AWGN – additive white Gaussian noise*, $n(n)$ were carried out. The level of degradation was evaluated by the estimated

error of the amplitude frequency response using the composed signal (a set of sinusoids plus the test signal) filtered by a synthetic IR. In this laboratory approach, a synthetic IR, $h_{syn}(n)$, showing an all-pass frequency response, as described in the previous section, was used. Obviously, considering a *LTI* system, the system output to the composed test signal shows the same spectral content with each sinusoid, $s_i(n)$, altered in the amplitude and phase, $sFilt_i(n)=A_i |H_{syn}(\Omega_i)| \cos[\Omega_i n + arg H_{syn}(\Omega_i) + \theta_i]$, where i stands for the index of each sinusoid, A_i is the amplitude of each input sinusoid, θ_i is the phase at the time origin, and Ω_i is the angular frequency of each sinusoid.

The synthesized set of sinusoids has different amplitudes and phases located at the frequencies of 200, 400, 1000, 2000, 5000, and 8000 Hz. The tests were performed for four different cases: (i) pure sinusoids, (ii) the ensemble was modulated in amplitude, $s_{AMi}(n) = [1+\mu\ c_r(n)]\ s_i(n)$, (iii) the ensemble was modulated in frequency, $s_{FMi}(n) = A_i \cos[\Omega_i n + \Delta F/F_c \sin(\Omega_{ci} n)]$, and (iv) both the amplitude and frequency modulation was applied, where μ stands for modulation index, $c_r(n) = \cos(\Omega_c n)$ is the carrier signal, the ΔF is the frequency deviation, the F_c is the carrier frequency and $\sin(\Omega_{ci} n) \propto \int s_i(n)$.

In order to increase the SNR, a filtering technique capable of eliminating, or at least attenuating, the tonal components is applied. Two types of filtering approaches were tested. The first approach, NLMS_QConst, follows two steps: (1) the NLMS algorithm is applied to the captured composed signal, $c\_filt(n)$, to estimate the frequency of each partial, and (2) a notch filter with a high quality factor Q is used implemented by second-order IIR filters, which are very fast and provide very precise resolution, even at lower frequencies. However, increasing the quality factor also increases the duration of the transient process in the filter after the excitation. The second filtering approach, *SSTC*, is based on the spectral subtraction of tone components. Basically, a peak detection algorithm extracts the sinusoid parameters (amplitude, frequency and phase) from each partial and the peak continuation algorithm is used to build a tonal component in opposite phase used to destroy the target tone belonging to the noise (music). Figure 5.17 shows the basic block diagrams of these two filtering approaches.

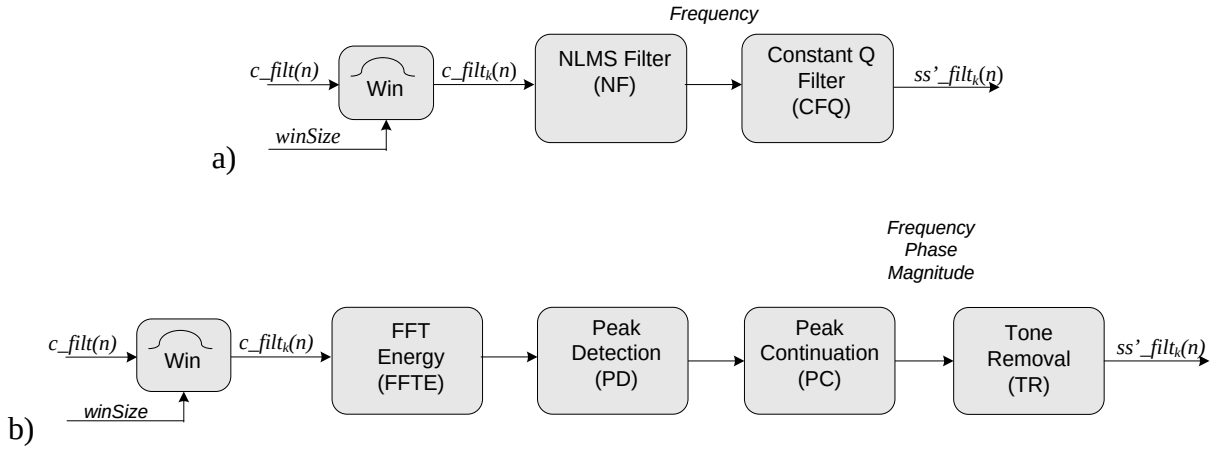
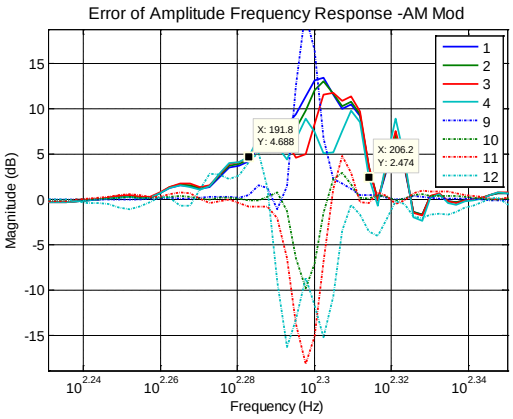


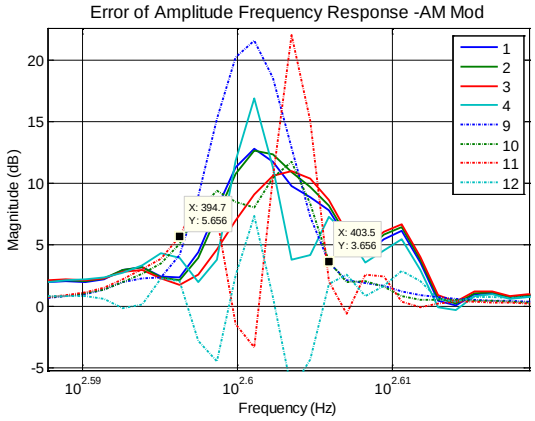
Figure 5.17 – Block diagrams of the two filtering approaches used to assess the degradation of the test signal due to the presence of tonal components with high sound pressure levels; a) refers to the adaptive NLMS filter followed by a notch filter; b) uses the spectral subtraction of tones, SSTC.

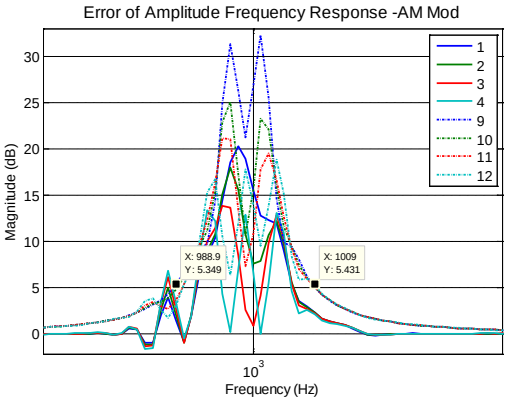
Figure 5.18 shows the error of the amplitude frequency response for a number of different tests. The dotted and the filled curves refer to the NLMS_QConst and to the SSTC algorithms, respectively. The numbers in the legend of the figure refer to the different analysis evaluated, according to Table 5.1.

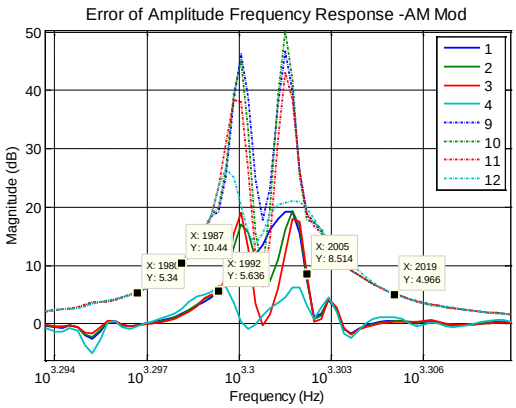
Table 5.1 – Parameters used to modulate the multi-tone components.

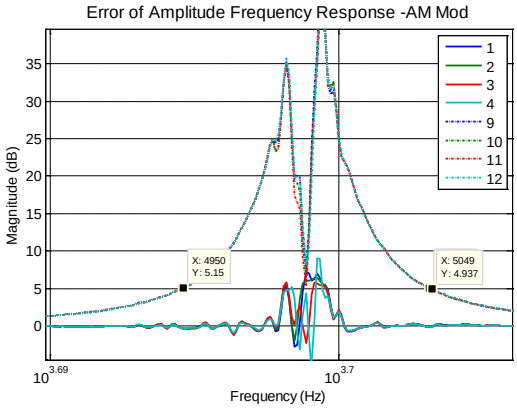
Legend #	AM		FM	
	Mod. Index, μ (%)	Freq. Mod. (Hz)	Freq. Dev, ΔF (Hz)	Freq. Mod. (Hz)
1, 9	-	-	-	-
2, 10	20	1	-	-
3, 11	50	1	-	-
4, 12	20	2	-	-
5, 13	-	-	1	1
6, 14	-	-	2	1
7, 15	-	-	5	1
8, 16	20	2	1	1

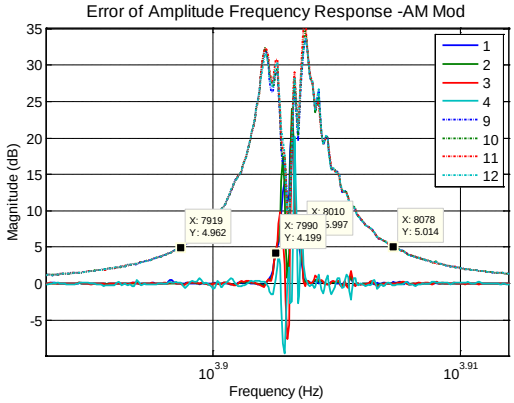

Error of Amplitude Frequency Response -AM Mod
Magnitude (dB)
Frequency (Hz)
X: 191.8
Y: 4.688
X: 206.2
Y: 2.474
1
2
3
4
9
10
11
12


Error of Amplitude Frequency Response -AM Mod
Magnitude (dB)
Frequency (Hz)
X: 394.7
Y: 5.656
X: 403.5
Y: 3.656
1
2
3
4
9
10
11
12


Error of Amplitude Frequency Response -AM Mod
Magnitude (dB)
Frequency (Hz)
X: 988.9
Y: 5.349
X: 1009
Y: 5.431
1
2
3
4
9
10
11
12


Error of Amplitude Frequency Response -AM Mod
Magnitude (dB)
Frequency (Hz)
X: 1987
Y: 10.44
X: 1992
Y: 5.636
X: 2005
Y: 8.514
X: 2019
Y: 4.966
1
2
3
4
9
10
11
12


Error of Amplitude Frequency Response -AM Mod
Magnitude (dB)
Frequency (Hz)
X: 4950
Y: 5.15
X: 5049
Y: 4.937
1
2
3
4
9
10
11
12


Error of Amplitude Frequency Response -AM Mod
Magnitude (dB)
Frequency (Hz)
X: 7919
Y: 4.962
X: 7990
Y: 4.199
X: 8078
Y: 5.014
1
2
3
4
9
10
11
12

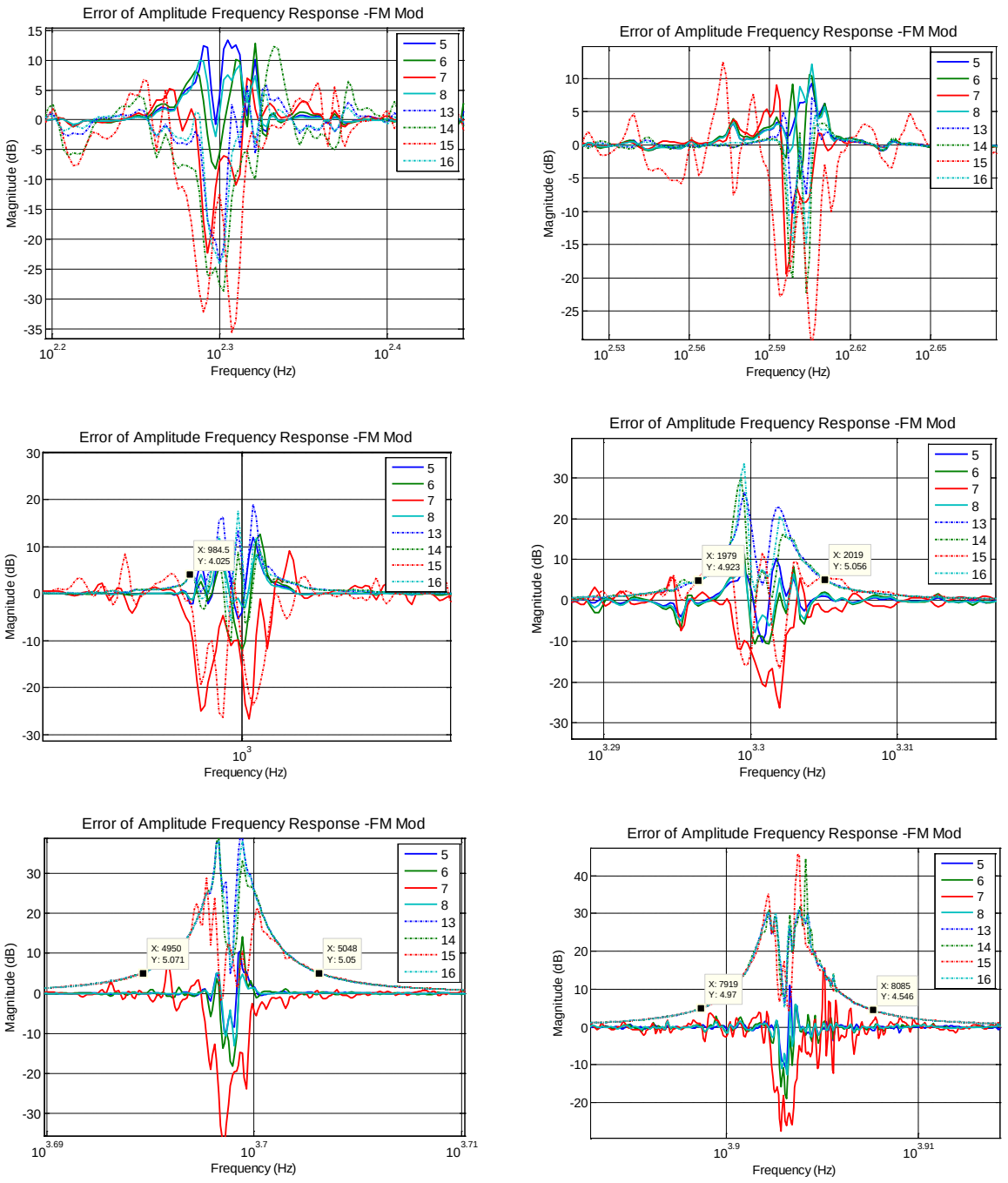


Figure 5.18 –RIR amplitude frequency response error due to the presence of tonal components during the acoustical tests; the dotted and the filled curves refer to the NLMS_QConst and to the SSTC algorithms, respectively.

As expected, the room response is more degraded when the multi-tones are frequency modulated, the magnitude of the error is higher and the range of frequencies affected is larger (this result is shown for the tests #7 and #15). Moreover, the amplitude of the room response is

more affected when the NLMS_QConst is used, in all analysis. In particular for the high frequency bands, an amplitude response error of more than 30dB is expected. The SSTC approach barely achieves 20dB for the worst case scenario. Since a *Q* factor of 40 is used for the NLMS_QConst, the bandwidth is fixed for each central band frequency. For the SSTC, the bandwidth affected is fixed at about 15 Hz (depending of the width of the window used).

The SSTC approach has additional advantages such as not introducing transient delays in the test signal and corresponds to the natural way of synthesizing the tonal parts of the music signal. Therefore, the SSTC method was adopted in the subsequent study.

These findings suggest synthesizing some tonal parts of the music material, forcing maximum values of the frequency modulation, modulation index and frequency deviation parameters of each partial, in order to facilitate the extraction process of the tonal components. It should be kept in mind that by pre-processing the spectral contents of some parts of the music to synthesize artificially some tonal components it is not a lossless process in the sense that the music will be damaged somehow. However, if the spectral zones to be processed and the modulation parameters are chosen with care, minimal artifacts will be detected by an audience. In this sense, a set of music samples built with the replacement of some tones by the synthesized counterparts with controllable modulation parameters was carried out. These music samples were used in psychoacoustic tests made available for a number of people. The tests consisted of synthesizing some tonal components (in this first quiz only one tone was synthesized simultaneously) using different combinations of amplitude and frequency modulation with/without smoothing techniques applied. The smoothing approach applied to the amplitude and the frequency of the tonal components improves the process of extraction of the tonal components. The results were obtained by the following steps:

1. Eliminate both the amplitude and the frequency modulation, using pure sinusoids. The results show that the naturalness is degraded significantly (the pure tone is detectable after the onset time of about 200ms). However, the sensation of discomfort seems to be more noticeable for lower sound pressure levels.

2. Eliminate the frequency modulation:
   a. keep the amplitude modulation – the synthetic tones of the music excerpts used were not detectable for most of the people enquired;

b. reduce the amplitude modulation (smoothing the amplitude of the tones) – it was found that keeping a minimal amount of amplitude modulation the spectral modification introduced is not noticeable in almost all situations.

3. Eliminate the amplitude modulation,

   a. keep the frequency modulation – same artifacts are evident by eliminating the information of the amplitude of the tones;

   b. smooth the frequency modulation – although the filtering process of the tones is more simplified, the artifacts are still present.

As a conclusion, the synthesis of the tonal parts of the music that crosses the test signal is made frequency modulation free with the smoothing of the amplitude of the tones (case 2.b. above) and the tones of the music that stand outside the test signal (on the borders) are synthesized as a replication of the original.

5.3.3 Tone or narrow-band noise probes plus noise with modification, TPNM

The sampled RT for a number of discrete frequencies provides an a priori estimate of the RT for each frequency band. In fact, the smoothed decay energy rate of a narrow-band signal, such as tonal component or filtered white noise, is a simple approach for the estimation of the RT in situations where, for instance, the test signals are supposed not to be perceived by the audience. Therefore, this method is used essentially in conjunction with the perceptual mechanisms of the human hearing system. Nevertheless, different approaches based only on the transient parts of signals such as speech and musical sounds are also proposed in the literature [Ratnam03, Vieira04].

The method consists of implanting a number of narrow-band signal probes throughout the audio spectrum in order to cover each octave band, or fraction of octave band, and following the probe decaying energy after switching-off the sound source. The averaging method is then applied to several decay curves in order to make a more accurate estimation of the RT for a particular frequency band. Figure 5.19 shows the simplified block diagram of the method that was implemented.

Figure 5.19 – Block diagram of the tone probes plus noise method to estimate the RT.

Although narrow-band signals can be used in the method, at least theoretically, they were not used in this study due to the fact that a bandwidth less than 1/12 octave band is usually needed for reasonable perceptual masking results.

Obviously, the results of this method are only a first approach of the RT for different frequency bands, but the IR cannot be estimated. Nevertheless, these results will be used to configure the time extension needed for the *prohibited zone* as shown in section 5.1 (Figure 5.20) which are used in the TSNM and STSNM methods as described in the following sections.

Simulated tests with synthetic and measured IR were used to validate the method. Figure 5.20 shows the spectrogram of the IR measured in an auditorium emphasizing the increase of the energy decay rate (decrease of the RT) for the raising frequencies. This room shows an average RT of about 2.4s for the mid-band frequencies of 0.5, 1 and 2 kHz as in the black curve in the figure.

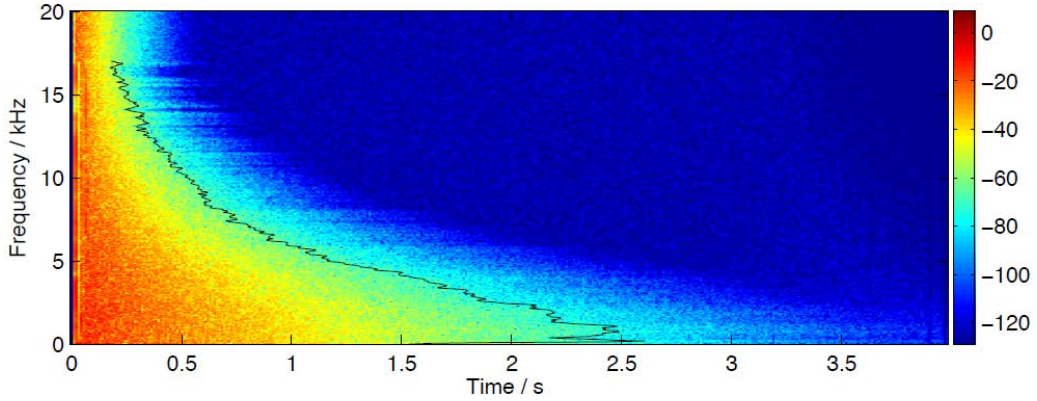


Figure 5.20 – Spectrogram of the RIR estimated from the measurements made in the Promenadikeskus concert hall in Pori, Finland [Juha05].

Table 5.2 summarizes the results for four different cases: (i) pure tone signals located at the central frequency of each 1/3 octave band, (ii) Amplitude Modulated tones – the same signal referred in (i) is amplitude modulated with modulation index of 50%, (iii) the signal described in (ii) is frequency modulated with a modulation frequency of 1 Hz and frequency deviation of

5 Hz, and (iv) pure tones masked by noise, used accordingly to the perceptual model in order to be masked by the stationary tonal parts of the music contents. The amplitude and frequency modulation applied to the signals have two different purposes: (a) using more realistic test signals such as signals produced due to the time variance phenomena and (b) it is an attempt to search for the tone-based probe arrangement which produces best results, more smooth decay curves. The *T30* results are shown in the table for a synthetic *IR*, $h_{syn}(n)$, of RT value of 1 s with equal decay energy rate for all the frequency bands, as mentioned before, and for the measured IR shown in Figure 5.20.

Table 5.2– Room acoustical parameters for octave bands averaged over all responses with receiver positions in the audience area.

IR synthetized RT (s)	1/3 Octave (Hz)	pure tones	Octave	AM_FM AM =on IndMod=0.5 FM=off	Octave	AM_FM AM =on IndMod=0.5 FM=on freqDev=5Hz freqMod=1	Octave	pure tones embeded
Measured values		SNR =inf	Ave 1/3 Oct	SNR =inf	Ave 1/3 Oct	SNR =inf	Ave 1/3 Oct	SNR =60dB
	62.5	1.16	1.12	1.08	1.09	1.05	1.04	
	78.7	1.07		1.09		1.03		
	99.2	1.03		1.12		1.02		
1	125	1.08	1.04	1.1	1.03	1.03	1.01	1.17
	157	1.02		0.88		0.99		
	198	0.96		0.92		0.93		
1	250	0.87	0.98	0.97	0.97	0.97	0.95	0.95
	315	1.12		1.02		0.95		
	397	0.96		0.96		0.96		
1	500	1.01	0.99	1.02	0.98	1.06	1.00	1.02
	630	0.99		0.97		0.97		
	794	1.12		1.1		0.97		
1	1000	0.99	1.08	0.99	1.07	0.99	0.99	0.96
	1260	1.13		1.12		1.02		
	1587	1.07		1.05		1.01		
1	2000	0.95	1.03	0.95	1.01	0.97	0.96	1.07
	2520	1.08		1.04		0.9		
	3175	1.03		1.02		0.94		
1	4000	0.98	1.00	0.98	0.99	0.96	0.96	1.02
	5040	0.99		0.97		0.97		
	6350	0.99		0.99		0.99		
1	8000	1	0.96	0.95	0.97	0.93	0.96	0.91
	10079	0.88		0.98		0.96		
	12699	0.97		0.85		0.94		
	16000	1.07	1.02	1.06	0.96	1.04	0.99	

IR measured in an auditorium

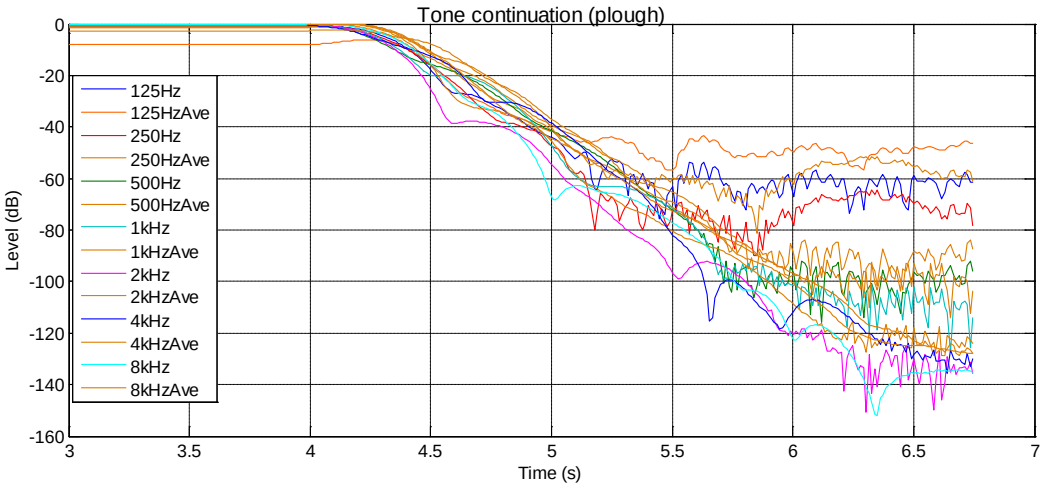
Measured

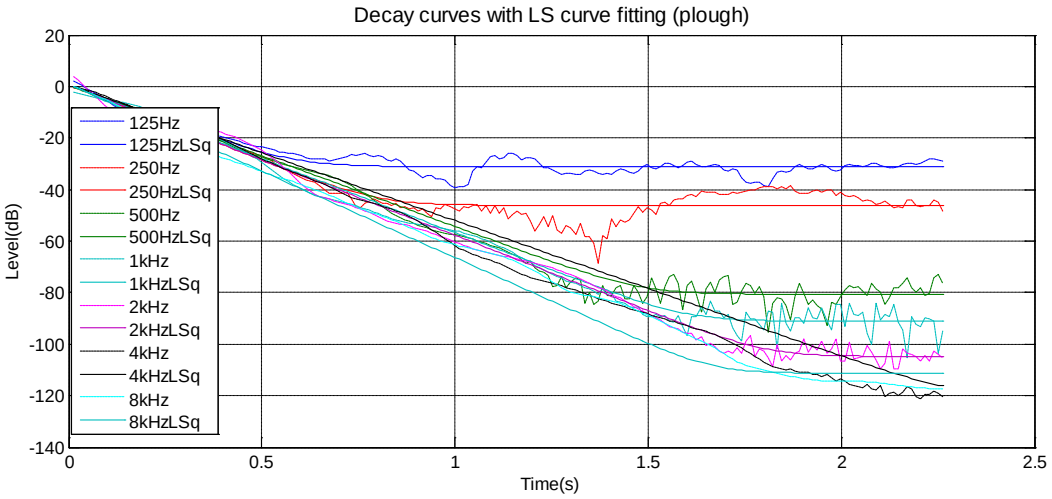
values		SNR =inf	Ave 1/3 Oct	SNR =inf	Ave 1/3 Oct	SNR =inf	Ave 1/3 Oct	SNR =60dB
	62.5	2.45	2.42	2.48	2.45	2.25	2.35	
	78.7	2.38		2.41		2.44		
	99.2	2.39		2.39		2.52		
2.6	125	2.6	2.57	2.52	2.55	2.47	2.52	2.8
	157	2.73		2.73		2.58		
	198	2.43		2.45		2.79		
2.4	250	2.28	2.44	2.28	2.46	2.47	2.55	2.35
	315	2.61		2.64		2.4		
	397	2.5		2.46		2.25		
2.4	500	2.54	2.63	2.51	2.57	2.33	2.31	2.67
	630	2.84		2.75		2.36		
	794	2.41		2.41		2.35		
2.3	1000	2.25	2.26	2.26	2.26	2.38	2.29	2.26
	1260	2.11		2.11		2.14		
	1587	2.35		2.31		2.13		
2.1	2000	1.95	2.17	1.96	2.14	2.12	2.08	1.96
	2520	2.21		2.16		1.99		
	3175	1.98		1.98		1.83		
1.7	4000	1.94	1.76	1.89	1.76	1.69	1.63	2.08
	5040	1.36		1.42		1.36		
	6350	1.2		1.19		1.14		
1.2	8000	0.87	0.89	0.89	0.90	0.9	0.89	1.01
	10079	0.61		0.62		0.64		
	12699	0.51		0.52		0.59		
	16000	0.31	0.41	0.32	0.42	0.33	0.46	

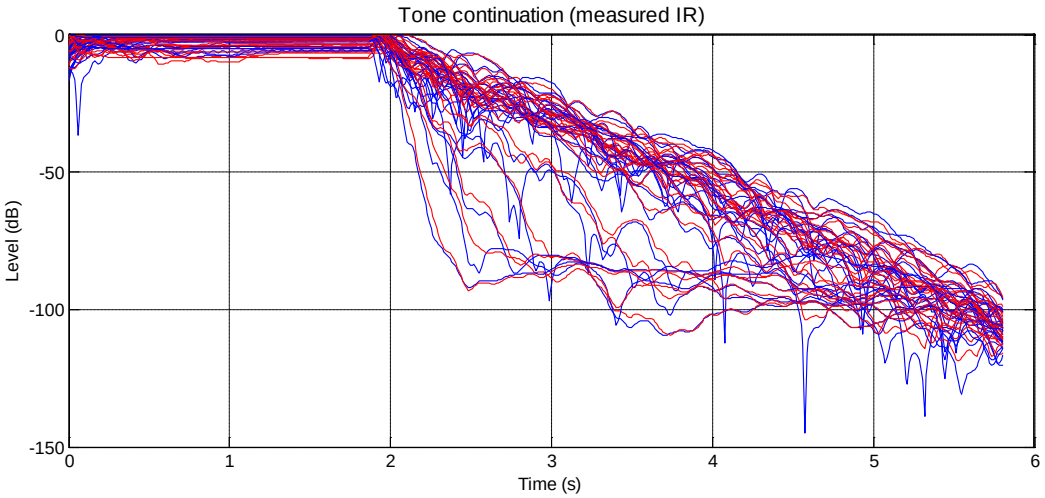
Since the data for the measured RT is in octave bands, the results for the analysis are averaged for comparison purposes.

The results in Table 5.2 show that this method is able to provide good estimates of the RT, with an error of less than 5% and 10% for the mid-band frequencies for all the cases tested of synthetic and measured *IR*, respectively. Moreover, the results are not significantly changed when the modulated-based probe signals are used.

The energy decay and the smoothed energy decay curves for the case of synthetic IR are depicted in Figure 5.21 The smoothed curves used to estimate the RT are produced by using the *lsqcurvefit* function (Least Squares sense non linear approach [Karjalainen02]). This parametric model approach proved to be superior to most standard methods, including the Schroeder integration method, to estimate the RT in situations of low SNR. The curves are versions of the tone continuation with the peak interpolated with the three nearest neighbors to increase the accuracy.


Tone continuation (plough)
Level (dB)
Time (s)
125Hz
125HzAve
250Hz
250HzAve
500Hz
500HzAve
1kHz
1kHzAve
2kHz
2kHzAve
4kHz
4kHzAve
8kHz
8kHzAve


Decay curves with LS curve fitting (plough)
Level(dB)
Time(s)
125Hz
125HzLSq
250Hz
250HzLSq
500Hz
500HzLSq
1kHz
1kHzLSq
2kHz
2kHzLSq
4kHz
4kHzLSq
8kHz
8kHzLSq


Tone continuation (measured IR)
Level (dB)
Time (s)

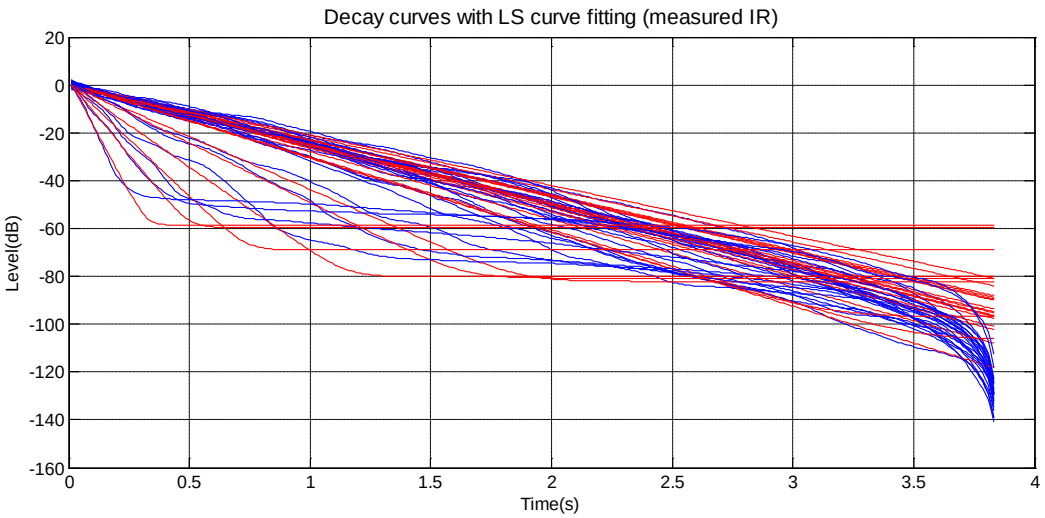


Figure 5.21 – Decay curves of the concert hall for synthesized (two top most) and measured IR (two down most); only octave bands are shown for simplicity; the filled curves in down most figure are the line fitted of the dashed curves using the *lsqcurvefit* function.

This technique uses the perceptual mechanisms of the human hearing system to mask the test signals conveniently according to the simultaneous masking effect. Therefore, since the tonal part of the music material is needed to mask the test probe tone, it is suggested that the spectral contents of the music be used itself as the probe tone. As shown in the previous section, the truncation of harmonics is not perceived provided that they are higher harmonics belonging to a certain music note. Hence, this additional approach could have advantages, especially at the higher frequency bands, where more constraints are applied to the filter specifications used to implant the tone test probes.

5.3.4 Noise modified perceptually-based technique

To keep the room response free from the influence of the noise, i.e. from the music energy, considering the chart in Figure 5.13, the signal $c(n)$ can be assembled simply by removing the noise from the corresponding *prohibited area* (see Figure 5.22). This procedure, obviously, increases the SNR since the music energy is not superimposed to the room response. However, it could alter the spectral contents of the music, thus implying a higher nuisance to the audience. A better approach consists of using this concept together with a perceptual model. This PM evaluates the *Signal-to-Mask ratio* (*SMR*) for each windowed segment. The path which best fits nuisance constraints, referred to as the *Preferred Path* (the curve A in Figure

5.15), is then considered. This path is achieved by minimizing the perception of the test signal by the audience using the simultaneous masking effect of the human auditory system.

The procedure attempts to keep the spectrum of the music with as minimal changes as possible in order to reduce the risk of altering the harmonic contents of the sound (changing its timbre and tonality) and, thus, to be less perceptible by the audience. Therefore, the tonal sounds will appear in the same frequency bin of the test signal. In this situation, as the correlation between the frequency components of the noise and of the test signal is high, an accurate estimation of the RIR for this frequency band was not possible. Another important related issue is the onset and attack time of a particular sound. This issue is used by the human auditory system in agreement with the harmonic contents of a sound, in a subjective way, to characterize and distinguish between different sounds and musical instruments. Therefore, no transformation should be applied when an onset of a sound is detected. For these reasons, the procedure should be applied together with the weighted averaging technique in order to increase the SNR. Figure 5.22 shows a possible scenario for the implementation of this procedure.

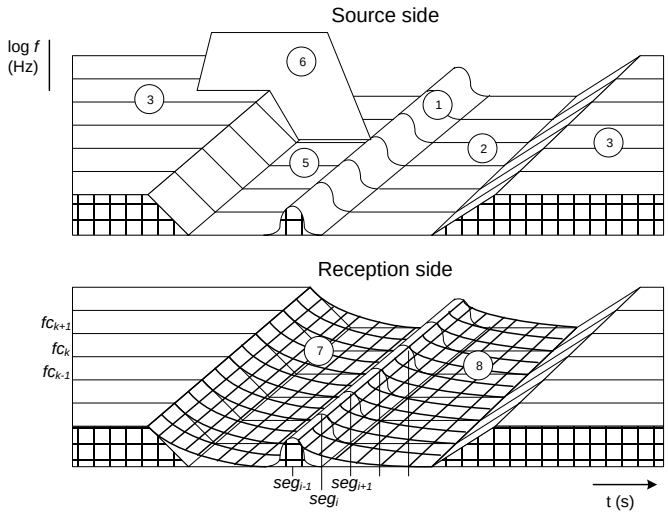


Figure 5.22 – 3D scheme of the insertion of the swept sine in the noise; 1- swept sine; 2- prohibited area; 3- noise; 5- dead zone; 6- tonal component that cannot be removed 7- noise energy decay; 8- swept sine energy decay.

In order to increase the SNR, a dead zone (with no signal) is considered to reduce the energy decay of the noise in the prohibited area. The results of the psychoacoustic model are shown

just for the Source side. The swept sine envelope, keeping a pre-defined SNR, is not represented for reasons of simplicity.

Naturally, the performance of such procedure increases for the long term tonal quasi-stationary music tracks in order to guarantee an accurate masking of the test signal energy decay. This method can thus be said to be acoustically environment dependent.

5.3.5 Test signal plus noise with modification, TSNM

Based on the *SMR* results, this method estimates a number of parameters such as (i) the best starting time delay from the beginning of the music track, (ii) the best swept sine frame envelope, (iii) the tonal components which cross the swept signal that cannot be removed, and (iv) the onset locations. These issues identify the *Preferred Path*.

The method consists of the implementation of the steps described next before sending the test signal to the system (at the source side), and after capturing the filtered test signal (at the reception side).

Source side

The algorithm generates a composite test signal, $c(n)$, made up of a music track, as noise, with a modified spectrum, and at least one swept sine frame, as follows:

1. The user defines the duration of the sweep in seconds, the size of the window and the number of averages. This duration should be several times the expected reverberation time to guarantee narrower band removal and, consequently, to better mask the test signal (UC block);
2. Define the bandwidth for noise removal purposes based on the duration of the sweep;
3. Apply the PM to the music track identifying the paths that correspond to the better masking effect and fixing the tonal components (frequency, magnitude and phase) that cannot be removed, and the onsets. All these parameters are stored (SCPP block);
4. For the preferred path, a band rejection tracking filter is applied for a convenient accommodation of the swept sine. The depth of this trail is defined for a chosen SNR (TBRF block);
5. Replace the tonal components that could not be removed (TR block) ;
6. Filter the test signal, $s(n)$, according to the inverse absolute threshold of hearing and noise shape (THIFS block). This pre-emphasis filtering process is important to keep the same SNR for all the frequency bands, thereby, reducing the annoyance;
7. Add the filtered music signal to the weighted swept sine;
8. Play the composed signal, $c(n)$, in the room under test.

Reception side

9. Filter the windowed signal by a bandpass filter, corresponding to the inverse filter of the band reject filter applied to the source side, taking account the Side Information (TBPF block);
10. Apply the Threshold of Hearing filter and unshape, de-emphasis filtering process, according to the Side Information (THFU block);
11. Apply the SSTC filtering approach to attenuate the tonal components of the music present in the system response;
12. Apply the averaging technique to the resulting signal, *res*(*n*), if selected, and/or an additional method such as the *Time Spectral Segmented Swept Sine* technique in order to further increase the SNR [Paulo05a, Paulo05b];
13. Operate the convolution between the resulting signal, *res*(*n*), and the inverse filter, in order to estimate the RIR accordingly.

Figure 5.23 depicts the block diagram for the source and reception stages.

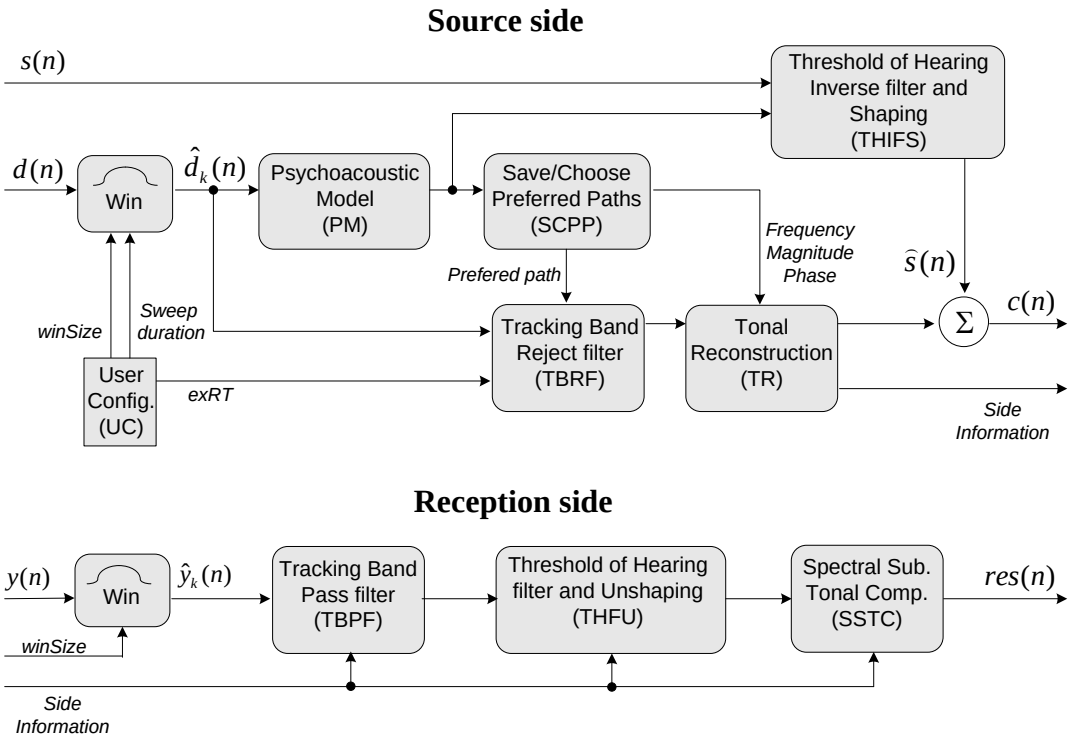


Figure 5.23 – Composite test signal assembler/de-assembler block diagram for the TSNM method.

This method presents an important issue by predicting the impulse response duration to make its truncation safe so as not to degrade significantly the low frequency response, bearing in mind that the SNR is low. Moreover, the duration of the sinusoidal sweep is related to the

removed noise bandwidth. The increase of the sweep duration allows for a narrower bandwidth. This is an iterative procedure that starts with the maximum expected energy decay duration.

5.3.6 Segmented test signal plus noise with modification, STSNM

With the TSNM procedure described previously, the masking process is only partially achieved. To minimize the nuisance, the music levels should be increased. However, the overall SNR will then decrease. Therefore, a more sophisticated arrangement was investigated [Paulo07a].

For a satisfactory use of the perceptual model, the test signal should be band limited for a better simultaneous masking. As shown above, with the windowing procedure, each segment corresponds to a fraction of the total swept sine spectrum. Therefore, a set of windowed segments, referred to as perceptual grains (the assigned B segments referred in Figure 5.15), covering a specific frequency band where tonal masking effect takes place, is produced. The amplitude levels of each perceptual grain are to be adjusted carefully according to the masking shape of each tonal sound. The 20dB difference level between the tonal masker and the test signal is a reasonable value to mask the perceptual grains adequately.

Each perceptual grain is used as a band-pass test signal and is implanted in the appropriated spectral locations, assuring that they are masked by the noise. The algorithm must ensure that the concatenation of all perceptual grains encompasses a full swept sine frame. At the reception, the complete filtered swept sine is assembled by using the overlap/add method or instead, each grain is processed individually to estimate the *IR* for the respective frequency band.

Figure 5.24 depicts a basic scheme for the procedure used in this technique for one particular perceptual grain. The noise shape is represented by a linear decay rule, for simplicity. In fact, the energy content of music tracks usually decreases with the raise in frequency. The parameters Δf_i and ΔSPL_i define the maximum frequency bandwidth and the maximum energy allowed for each perceptual grain before being perceptible to the *i*th tonal masker. The Δf_i is estimated by the PM algorithm. TheΔ SPL_i depends of the *Tonal to Global Masking ratio* function, TGMR, and the *Maximum Allowed Perceived Noise*, MAPN. The MAPN is needed

for reasons of algorithm convergence, as a compromise between perceived noise and SNR. This parameter is selected by the user.

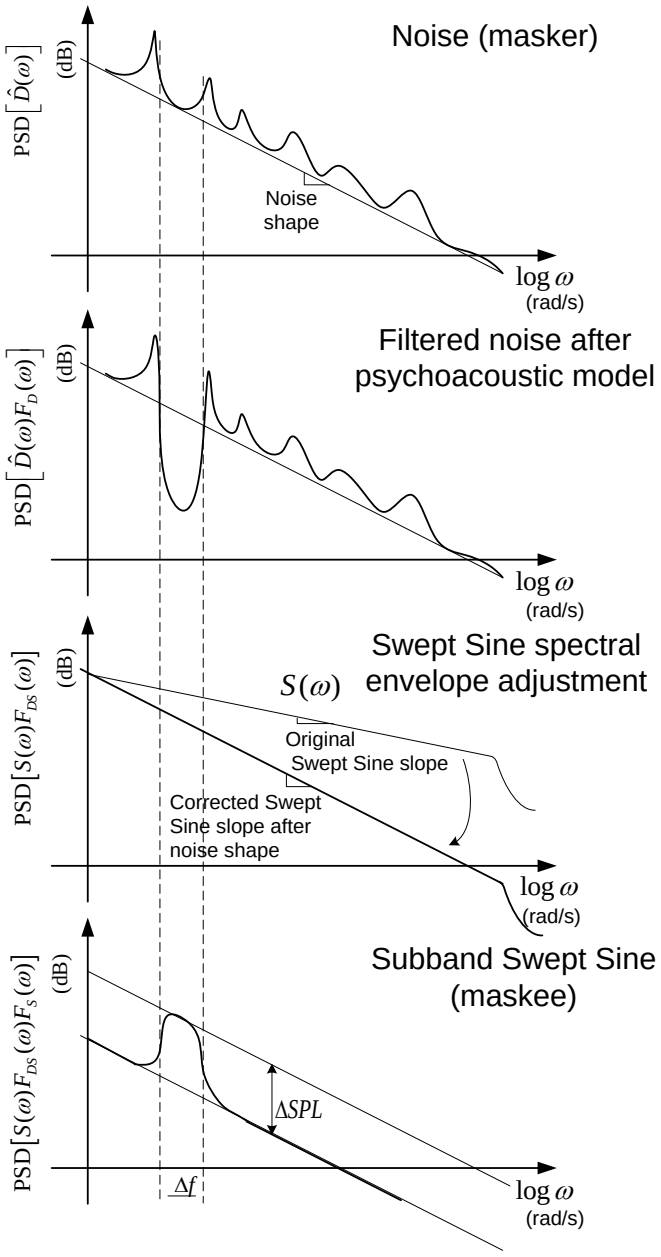


Figure 5.24 – Steps to build the narrow band test signal, perceptual grain, following the psychoacoustic model constraints.

Instead of using the perceptual audio coding standard algorithms, which consider the PCM material as dry audio, this technique strongly depends of the acoustical parameters. Therefore,

an additional issue should be considered and added to the PM algorithm: the spectral characteristics of music should remain quasi-stationary for a period of time of no less than the RT, implying that all the parameters evaluated should remain almost constant. This ensures that the acoustic energy almost vanishes avoiding superimposing the grains over the noise. In a realistic scenario, this implies that the total sweep duration must be much longer than needed. However, one should keep in mind that this study focuses on the minimization of annoyance, not on imposing restrictions to the measurement time duration (only restricted by the variations of the ambient parameters, such as humidity and temperature). The TSNM and STSNM algorithms (see section 5.1.2) incorporate a dead zone (a guard zone, with absence of signal) to prevent that the noise energy decay almost vanishes before the test signal starts. For the STSNM algorithm, the energy in this zone is controlled by a predefined minimum SNR. This procedure controls the perceived distortion produced in the music.

Ideally, the noise should be kept stationary for a minimum time duration of $T_d = 2\text{RT}$ (1x RT for the dead zone plus 1x RT for the segment energy decay), which is difficult to achieve in real conditions. The non accomplishment of these recommendations implies a decrease in the SNR. However, the dead zone can be controlled by taking the results of the statistical analysis of the noise into account.

As a consequence, the algorithm is most often applied to music material whose spectral contents show long term stationary tonal components.

The scheme in Figure 5.25 illustrates this constraint emphasizing the relation between the RT and the total sweep duration and the Δf_i range. This scheme shows a complete swept sine frame, for simplicity. However, the test signal is split into several perceptual grains, as described above and observed by the conceptual drawing in Figure 5.26.

The total test signal duration, T, is calculated from

$$T = \Delta t \frac{\ln\left(\omega_2 / \omega_1\right)}{\ln\left(\omega_f / \omega_i\right)} \tag{5.13}$$

where $\omega_f - \omega_i = 2\pi\Delta f_i = 2\pi(f_f - f_i)$ and $\Delta t = \alpha\text{RT}$, $\alpha \geq 1$.

As an example, consider f_i =1000 Hz; f_f =1100 Hz; f_1 =20 Hz; f_2 =20 kHz, RT=1s; α=1; then T ≈ 72.5 s.

This expression is also used to estimate the amount of time spent for the test signal to evolve between two tonal components. In fact, considering the masking tonal components are belonging to a musical note structure and a minimum grain length is required to ensure sufficient acoustical power delivered to the room, the total test signal duration, T_{ap}, should be higher than a certain value, i.e. is lower bounded. Therefore, T corresponds to the maximum value of the expected RT and the value of T_{ap}. As an example, consider tonal maskers extracted from the spectrum of a musical instrument note with a 1000 Hz fundamental frequency. If a minimum grain length of 1/3 second (about 16000 samples with sampling frequency of 44100 Hz) is used, a T_{ap} of about 21 s is needed to cover the frequency band from f_i =8000 Hz to f_f =9000 Hz. This value increases to a prohibitive value of about 100 s for a fundamental frequency of 200 Hz. As a result, the minimum value of T_{ap} is dictated by the higher musical notes, producing reasonable acoustic power, appearing in the music material due to the logarithmic rule of the swept sine signal. Fundamental frequencies of 1 to 2 kHz for acoustic musical instruments are considered, hence, the minimum value of T_{ap} is approximately 10 s.

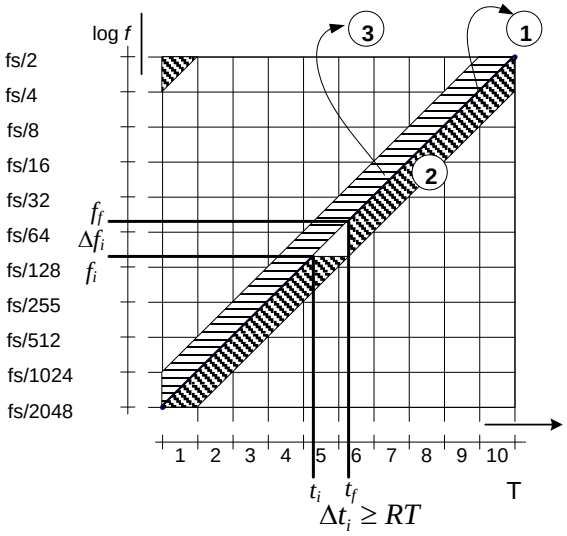


Figure 5.25 – Relationship between Δf, RT and total swept sine frame duration, T. 1- swept sine, 2- swept sine energy decay and 3- controlled energy zone, CEZ.

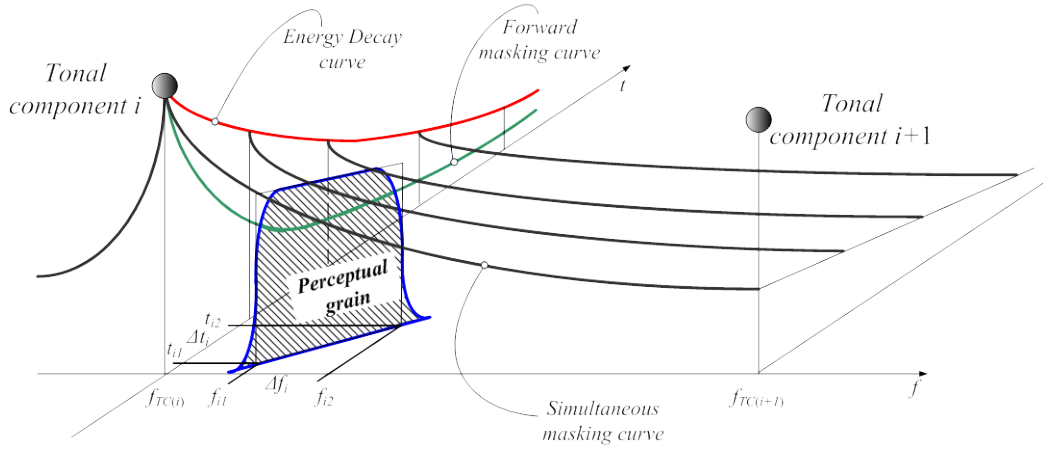


Figure 5.26 – Relationship between Δf, Δt and the masking threshold of a grain implanted in the music program.

As a description of the STSNM method to produce the composite test signal (grains ensemble implanted over the music signal) consider the music track whose spectrogram is depicted in Figure 5.27.

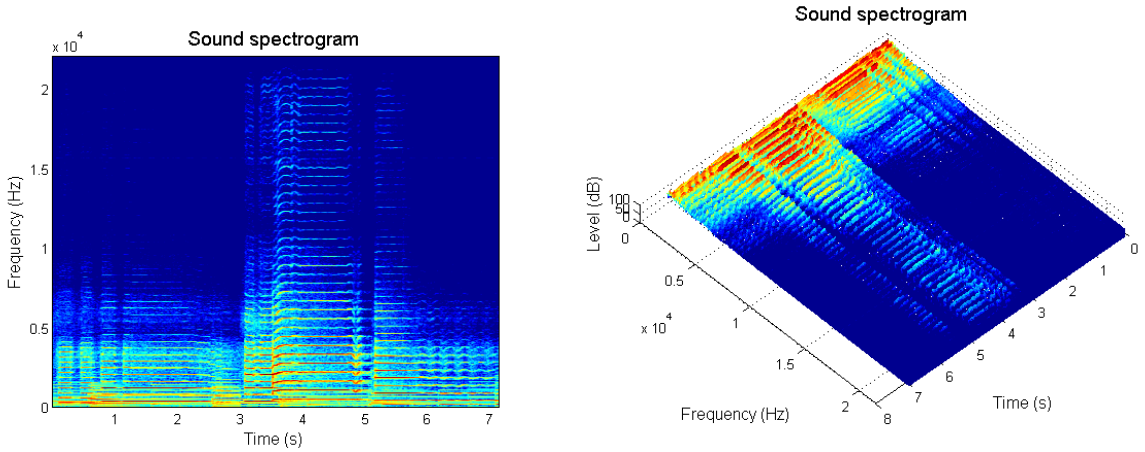


Figure 5.27 – Spectrogram of a musical track of a piano and a saxophone duet; 2D and 3D views.

Notice the relative high acoustic power of the harmonics ensemble in the sound track segment of length of about 1 s centered at 4 second, approximately. The appearance of such spectrum eases the covering of the full frequency spectrum of interest.

The TGMR analysis shown in a time/frequency map provides detailed information about the parts of the spectrum that are masked / not masked by the tonal components of the sound, as depicted in Figure 5.28. The dark blue area indicates that the spectrum is properly masked by

the LTtATH, tonal part, LTt, plus the *Absolute Threshold of Hearing*, ATH. Although almost all the spectrum is masked by the sound track LTtATH this is not sufficient to ensure reliable acoustical measurement results. In fact, the information concerning to the SNR throughout the spectrogram is not included in the TGMR analysis. The perceptual tonal tracks responsible for the masking process (tonal tracks that are not masked by other spectral components of the music) are also shown in the figure.

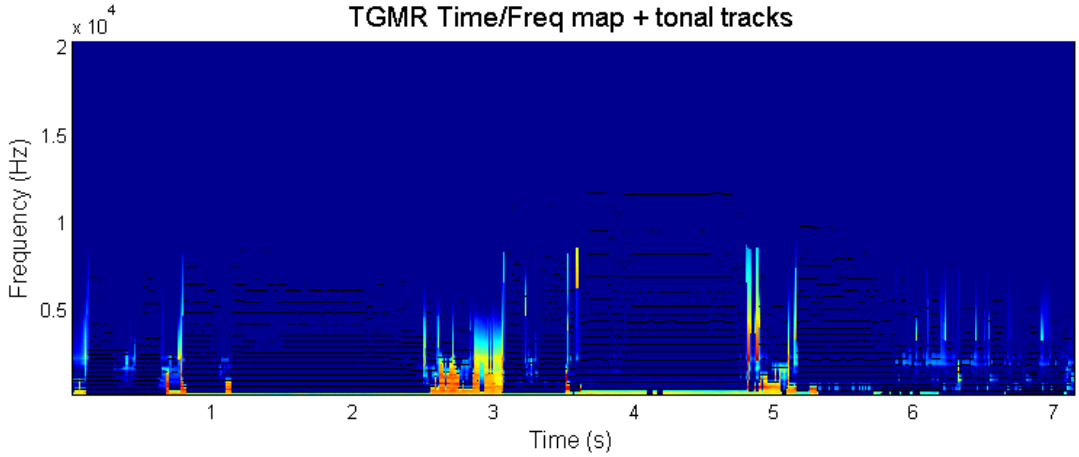


Figure 5.28 – TGMR analysis results showing the perceptual tonal tracks.

An important finding is that much less tonal tracks than those in the original spectrum are needed, especially at the higher frequency bands due to the simultaneous masking effect. In fact, since some harmonics are masked by another, the effective bandwidth between two adjacent tone components increases, allowing grains with higher length. This fact reduces the amount of grains needed to cover the entire audio spectrum improving the performance of the method.

Nevertheless, due to the non-linear behavior of the hearing system the mask shape is sound pressure level dependent. Therefore, the number of tonal tracks for situations of low sound levels is expected to increase. This fact does not affect significantly the masking areas in the TGMR analysis because the tonal and non-tonal parts of the sound are affected almost in the same way.

Figure 5.29 shows the 3D TGMR analysis for the standard approach (TGMR = GMR – (LTt + ATH)) and for the cleaned version. The cleaned 3D TGMR analysis is calculated by removing

the spectral zones, islands, corresponding to consecutive time and frequency indexes of less than *minTimeExtSec*= 20ms and *minFreqExtend*= 2 frequency lines, respectively. The values of these parameters are chosen as the thresholds providing the artifacts inserted by this procedure are not perceived by the people. This procedure also removes the spectral zones produced by the processing errors of the TGMR analysis.

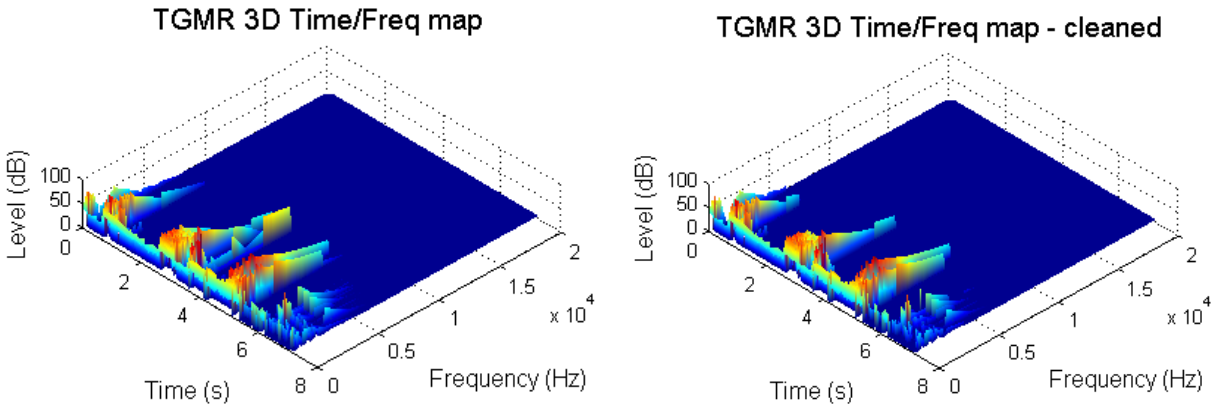


Figure 5.29 – 3D TGMR analysis results; left side view: same as in Figure 5.28; right side view: cleaned version.

The Tonal-to-Mask ratio plus the Absolute Threshold of Hearing analysis, LTtATH = LTt + ATH, is shown in Figure 5.30. Notice the fact that in this perceptual model the frequency bands above about 16k Hz are inaudible.

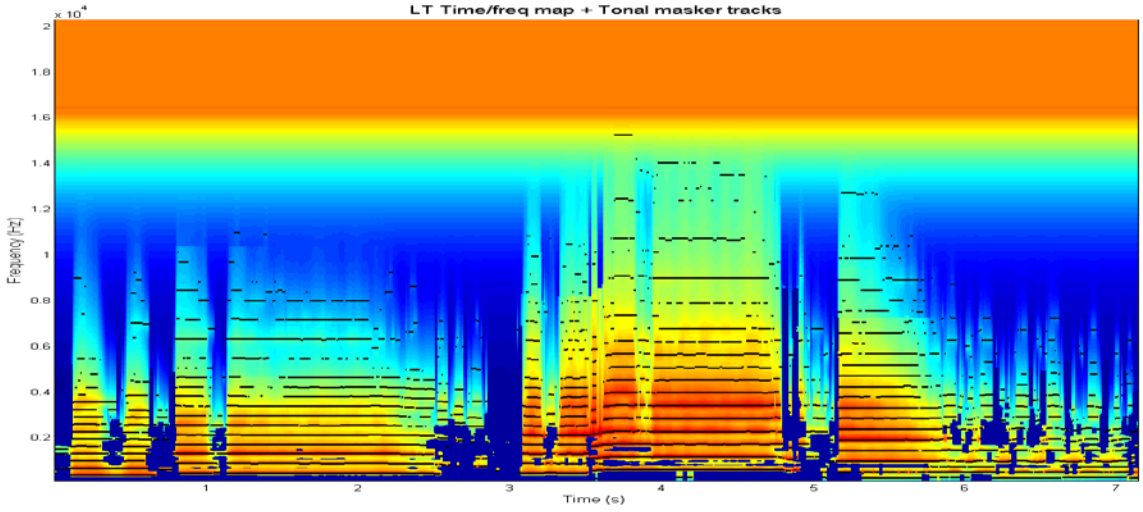


Figure 5.30 – LTtATH analysis results showing the tonal tracks used to mask the spectrum; the darker blue area, lower frequency bands part of the spectrum, indicates the non-tonal zones.

The information provided by the LTtATH analysis is more useful than the TGMR since the sound levels are shown on the mask. Actually, the STSNM method uses this information to define the masked zones by attributing a minimum threshold to the masking sound levels in order to keep a reasonable SNR for the bands of lower masking levels. Figure 5.31 shows a 3D view of the LTtATH analysis and a 3D view of the LTtATH analysis with the levels limited to a pre-defined value. In this example, a threshold of 30dB was applied as shown by the light blue color on the right side. This threshold is important to limit the minimum SNR allowed. Therefore, visually, it is easier to identify the usable area of the spectrogram which covers the entire spectrum and the respective allowed levels.

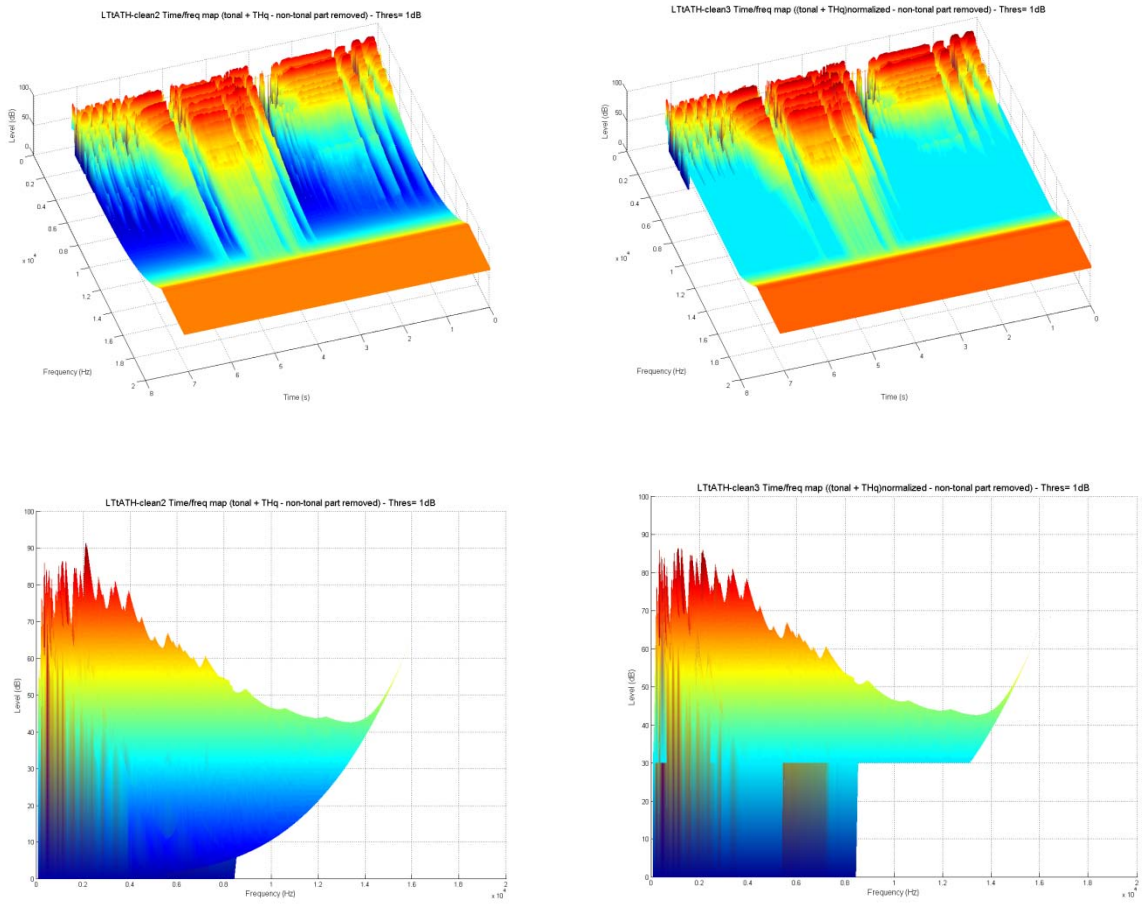


Figure 5.31 – *Tonal-to-Mask ratio* plus the *Abslotute Threshold of Hearing* analysis, LTtATH =LTt + ATH, shown as 3D spectrogram (top most) and the frequency vs. sound levels (down most); the right views refer to the LTtATH analysis limited to a lower admissible sound level of 30dB (the non-tonal part is left unchanged to 0dB level).

The next step consists of identifying the candidate zones with higher sound pressure levels and larger areas following a pre-defined criterion. Figure 5.32 shows the LTtATH spectrogram with three zones identified.

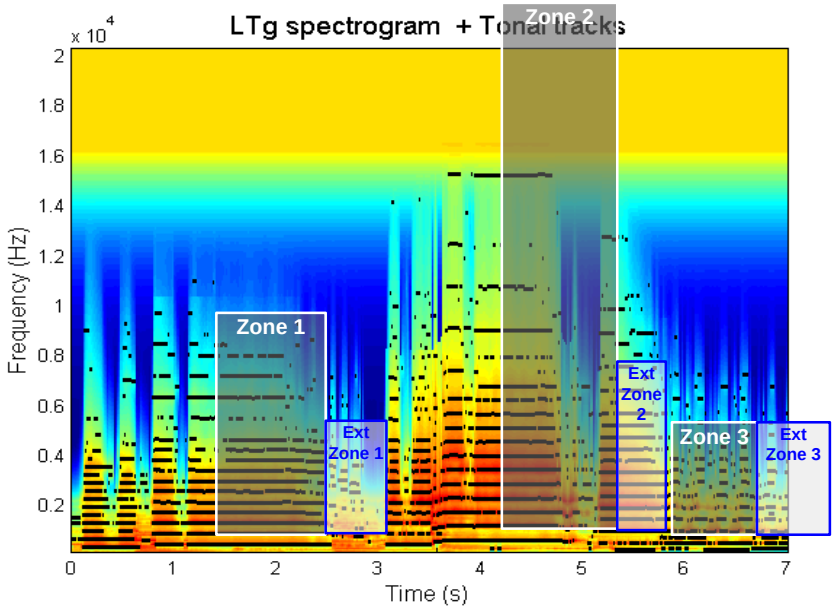


Figure 5.32 – *Tonal-to-Mask ratio* plus the *Absolute Threshold of Hearing* analysis, *LTtATH*; zones 1 to 3 are candidates to nest the perceptual grains.

Notice the areas marked by *Ext. Zone 1 to 3*, extended zones. These zones are used in case that extra time is needed to accommodate the decaying energy part of the grains. Actually, the grains duration cannot exceed the candidate zones in order to assure the test signal is accurately masked by the music. However, the decaying energy part of the grain is masked in Ext. Zones since the energy decay rate of the masker follows the same fashion in this zone. Nevertheless, some music degradation and artifacts are expected due to the fact that the masking spreading functions are level-dependent (the grains could be not masked properly) and if some music contents of the Ext. Zones have to be removed to keep the target SNR.

The next step consists of estimating the allowed Δf_i in between the tonal tracks for each zone. Provided the tonal tracks are stable for the zones (the frequency deviation of each track is under a pre-defined value and each track stays for almost the time duration of the zone), a simple

algorithm based on a histogram of the frequency indexes for the time length of the zone is applied [Paulo11b].

Figure 5.33 shows the results of applying this approach to the tonal tracks of the three zones mentioned in Figure 5.28. The results are shown in percentage for better comparisons of the zones. Therefore, it is easy to verify the behavior of each tonal track by observing the percentage of occurrence of each tone and the spreading around the index mean value. To improve the accuracy of this approach, the tones with occurrences lower than the threshold of 5% are discarded.

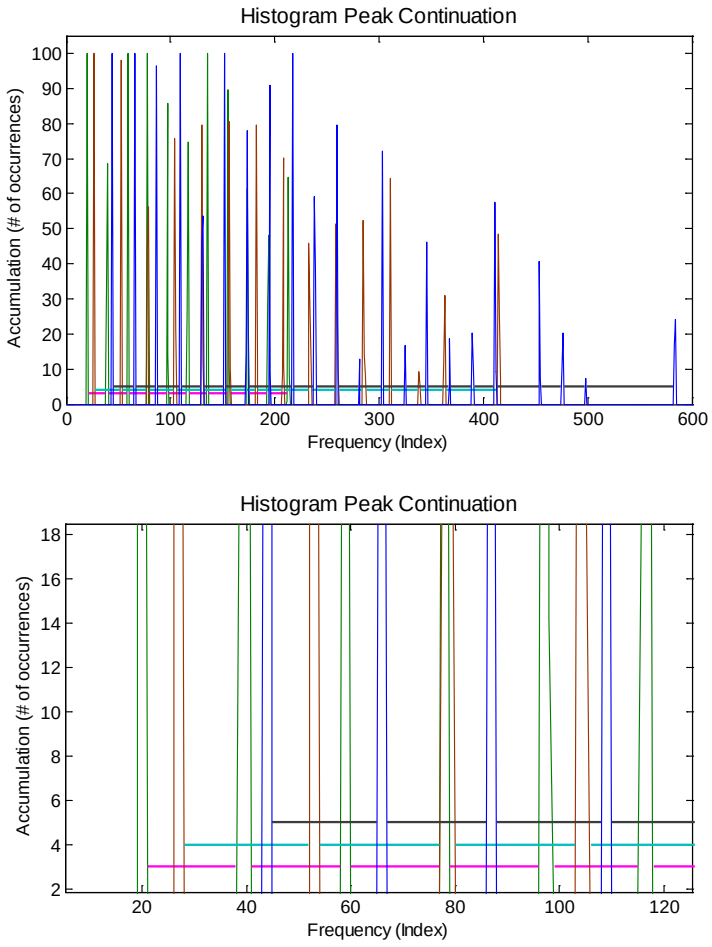
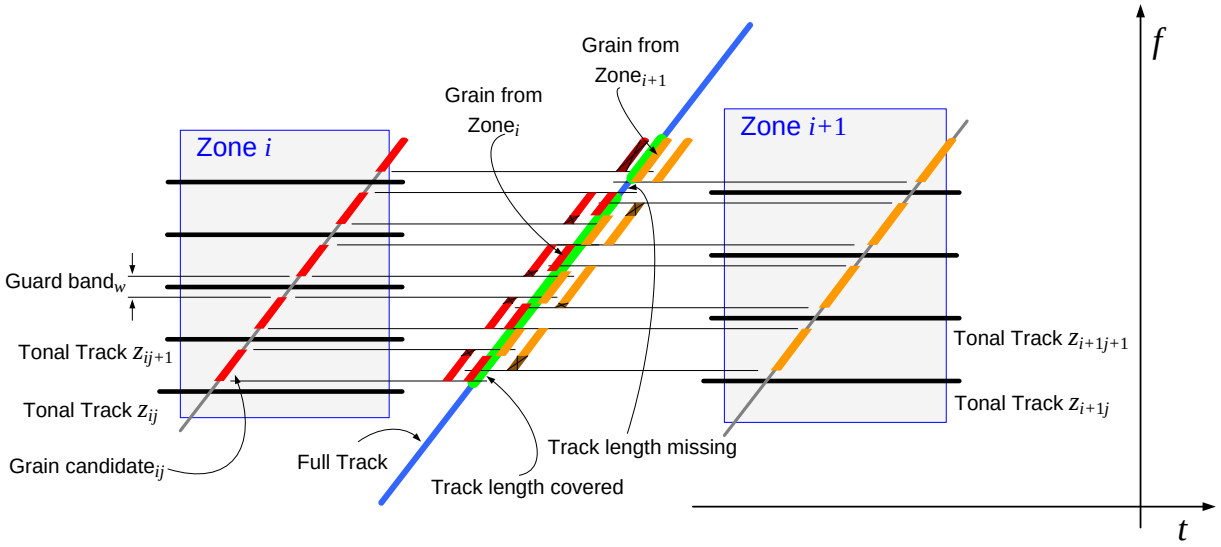


Figure 5.33 – Histogram of the occurrences of tonal components belonging to the tonal tracks for the three candidate zones mentioned in Figure 5.32; lower image: zoom-in of the upper image for a better visualization of the frequency ranges candidates, Δf_i, used to nest the perceptual grains; occurrences in brown, green and blue are assigned to zones 1, 2 and 3, respectively.

After the zone database is populated and the Δf_i identified, an adaptive algorithm is used to pick up and adjust the grains selected for each zone. The candidate grains are segments of the swept sine frame following the constraints of the Δf_i less a pre-defined bandwidth to provide a guard band, $GuardBand_w$, centered in the tonal track, to ease the filtering process. The full track frame is built by all the grain candidates available in the *GrainsC Pool*, reservatory holding the grain candidates that were not used yet.

The full track frame is obtained by the fusion of the grain candidates through the audio frequency band of interest. Therefore, each grain candidate is tested, chosen and adjusted in length and level by an adaptive algorithm to produce a grain if a number of requisites are held. The requisites used in the adaptive algorithm are based on (i) the maximization of the length of the grain and (ii) the maximization of the Noise-to-Mask ratio, *NMR*.

A scheme of this technique is depicted in Figure 5.34 where the construction of the full track frame from two distinct candidate zones is shown.



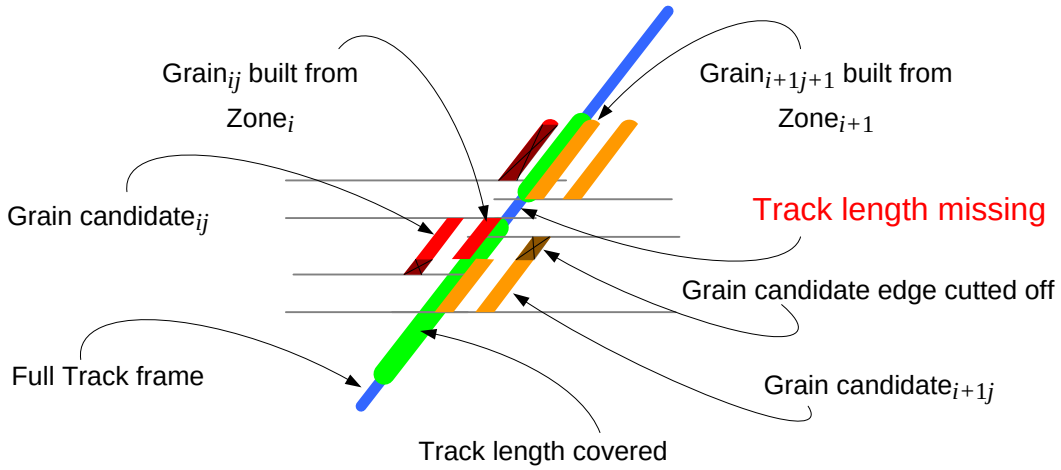


Figure 5.34 – Technique used to build the full track frame.

Figure 5.35 shows the procedure used to concatenate the grains from the perceptual zones (in this case three zones are used). Notice the minimum time length needed to assure sufficient acoustic energy delivered to the room by each grain (intervals assigned by the red circles with the number 1 inside). Moreover, if possible, the consecutive grains belong to different perceptual zones in order to reduce the distortion produced by the filtering procedure.

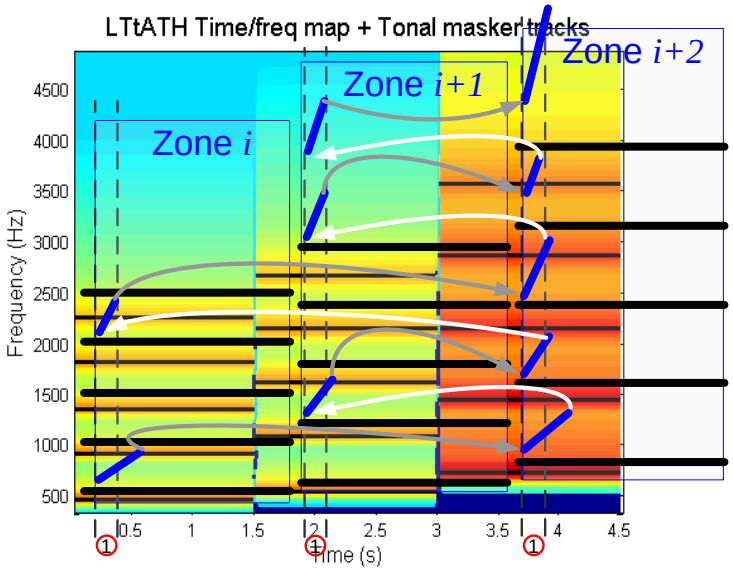


Figure 5.35 – Basic technique used to build up the full track frame based on perceptual grains minimizing a cost function; time intervals assigned by the red circles with the number 1 inside show the minimum grain length allowed.

If after emptying the *GrainsC Pool*, some parts of the spectrum are not covered, as observed in the down most Figure 5.34, the Spectral Subtraction of Tone Components, *SST*C, algorithm is applied to minimize the consequences of the lack of information in the measurement results.

The process of emptying the *GrainsC Pool* could result in the production of several full track frames with some of them not completely finished. Therefore, the average method is used among groups of grains with the same length.

The block diagram describing the proposed algorithm for the composite signal assembler and de-assembler is depicted in Figure 5.36.

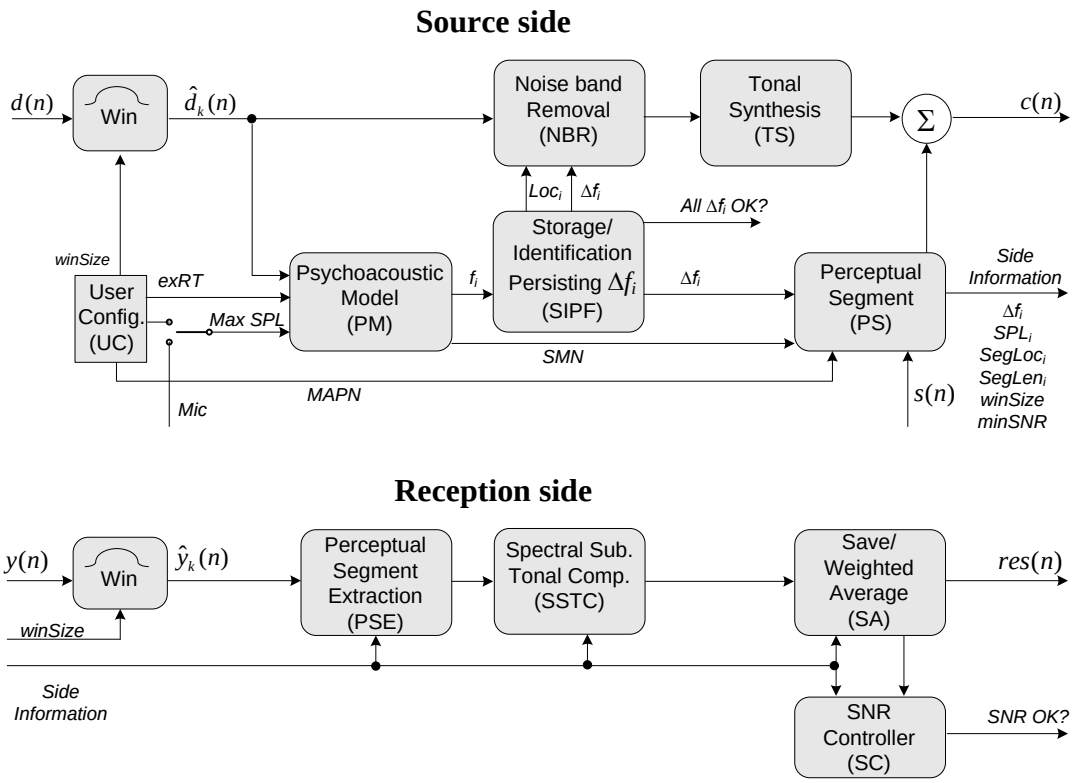


Figure 5.36 – Composite test signal assembler/de-assembler for the STSNM method.

The procedures for this technique are summarized in the following steps:

Source side

The algorithm generates a composite test signal, $c(n)$, based on a selected music track and the swept sine on form of grains as test signal as follows:

1. The user specifies the *Maximum Allowed Perceived Noise*, MAPN, the expected reverberation time, RT, and the segment length, *WinSize* (UC block);
2. Specify the maximum sound pressure level for energy normalization of the PM. This is made by user selection or estimated by sound level from a reference microphone in the room;
3. Segment the noise by using an analysis Hanning window with size specified (Win block);
4. Identify the f_i maskers and estimate the TGMR, see chapter 2 - Perceptual Model (PM block);
5. Store the f_i evaluating the Δf_i for each segment (SIPF block);
6. Identify the persisting Δf_i locations (segments repeated a number of times attaining by *PT*) (SIPF block);
7. Remove the noise from the Δf_i frequency range and f_i location (NBR block);
8. Synthesize the tonal components which crosses the perceptual grain (*TS* block);
9. Build the perceptual grains for the identified persisting f_i locations according to the noise masking shape (PS block) and the TGMR function, and form the Side Information for sending to the Reception side;
10. Test the $All\Delta f_i$ output. This variable provides information on whether the entire audio spectrum was covered by perceptual grains for the expected RT. If not, one of two situations occur: (i) the algorithm ends the first iteration - a second iteration must be computed using the patterns of both sides of tonal masking effect, see Chapter 3, Perceptual Model, or (ii) the algorithm ends the second iteration; even using the entire tonal masking pattern, the full audio spectrum could not be covered. In this latter situation, the algorithm TSNM, described in the previous section, should be applied;
11. Assemble the composite signal $c(n)$;
12. Run the de-assembler algorithm to compute the achieved SNR for each frequency sub-band (SC block). This operation should be made in order to guarantee a reasonable overall SNR. Otherwise, a new call to the assembler algorithm should be made with an increased MAPN value. However, the possible perception of the test signal will cause more annoyance to the audience.

Receiving side

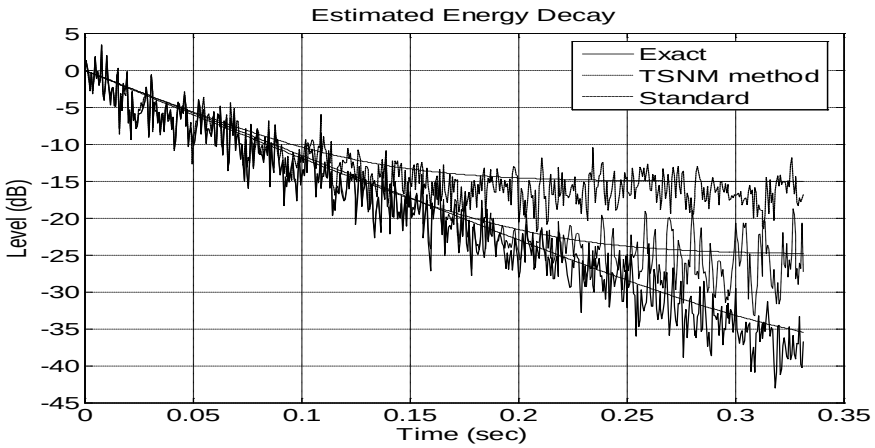
1. Apply the windowing procedure to the input signal according to the size window and type specified (Win block);
2. Extract the sub-bands having test signal information, for several sub-bands, regarding the Side Information (PSE block);
3. Remove the tonal components left in the $c(n)$ signal (SSTC block);
4. Compose the filtered swept sine, $res(n)$, weighting and averaging the perceptual grains, based on the Side Information (SA block);
5. The SNR Controller block explanation was already covered at the Source side;
6. Apply the convolution operation between the resulting signal, $res(n)$, and the inverse filter, in order to estimate the RIR accordingly.

In real conditions, the energy decay rate, and consequently the RT, analyzed in sub-bands, usually decreases with the increasing frequency due to the sound absorption of the materials used at the boundaries surfaces of the room enclosure. Therefore, the condition of long term quasi-stationary noise becomes dependent on the frequency, and the methods presented here can take advantage of that.

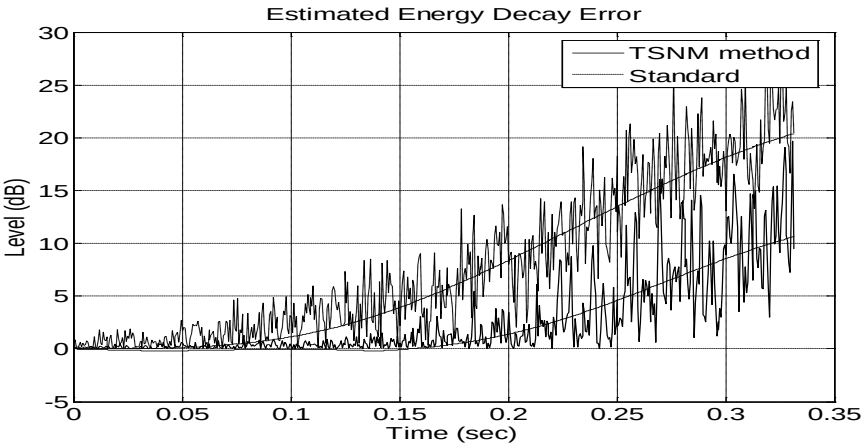
Another important issue consists of assessing the performance, capabilities and limitations of this approach in real time situations. In fact, in the future phases of the development of these methods, the sound signal used to mask the grains test signal will be the live musical program of the venue. The most serious problem is the constraints related to the procedure to build the full track frame from the perceptual grains. Actually, this procedure uses the tonal tracks of the selected zones which are defined for the full sound signal. Therefore, the problem of predicting the tonal tracks arises.

5.3.7 Results and discussion

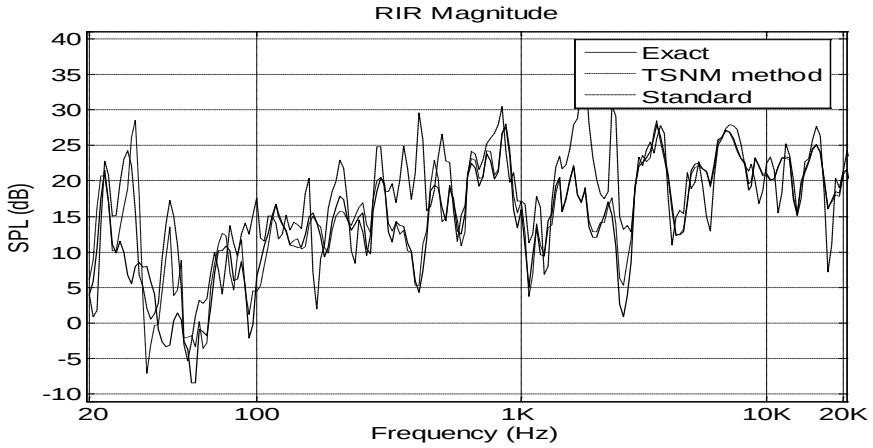
A number of tests were conducted for the full audio frequency range. A recorded impulse response signal (noiseless environment) with a RT60 of about 0.5 s (which corresponds approximately to the reverberation time of an ordinary living room) was used to simulate the characteristics of the system. Different types of signals with different SNR values were investigated. The techniques presented in this work were compared with the standard swept sine method for evaluation purposes. Figure 5.37 shows the calculated results for the RIR *Energy Decay*, RIR *Energy Decay Error* and RIR *Amplitude Frequency* response for the proposed TSNM method.



a)



b)



c)

Figure 5.37 – Calculated responses based on the RIR: a) Energy Decay, b) Energy Decay Error and, c) Amplitude Frequency Response, for an original music track with spectrum modification.

The figure shows an improvement of SNR of about 10 dB due to the noise band removing scheme, reported by the RIR Energy Decay error response. Notice that the RT cannot be estimated accurately by the standard swept sine technique (the regression curves for that technique and for the exact RIR have different slopes at the origin).

Identical results were obtained for the STSNM method, thus the corresponding curves are not presented here. However, the increase of the SNR is lower than for the TSNM method due to the overlap-add of the grains in STSNM method.

The accuracy in the RIR Magnitude Frequency Response obtained for TSNM and STSNM methods is much better than with the standard swept sine technique, as shown in Figure 5.37. However, for the low frequency bands, up to about 60 Hz, all the techniques fail to provide adequate results.

This experiment was carried out with a jazz music track (sample of *I'm Okay, Stan Getz, Café Montmartre)* as noise. An overall SNR of -30dB and a sampling frequency of 44.1 kHz were used. The time duration of the swept sine was about 65 seconds corresponding to a sweep time rate of 2 seconds per 1/3 octave band, approximately. This allows applying the test signal to a narrow bandwidth of the music in order to reduce the degradation of the spectral music contents due to the phase shift introduced, the well-known *phaser* audio effect, and, thus, be less perceptible by the audience. However, the bandwidth decrease carries also a decrease in the SNR due to phase problems of the bandpass filters used in the swept sine removal process. The results were obtained for a bandwidth of about 1/6 octave band.

A survey on the nuisance to people in the audience was carried out. No significant annoyance due to the measurements test signals was reported. The test signal is hardly noticeable with the TSNM method and is imperceptible with the STSNM method. However, the measurement duration, for the same final SNR, is much larger for the STSNM. This shows that the right choice of technique must be a compromise between annoyance and measurement time duration. Nevertheless, in some cases, the perceptual grains can appear superimposed in time, preventing the application of the necessary guard band from being applied.

This method uses the noise spectral contents (music tracks) in order to take a better advantage of the masking phenomena of the human auditory system using a dedicated scheme to remove noise before the estimation of the RIR. In this sense, special signals based on swept sine and music were developed.

In most cases, the TSNM procedure uses a complete swept sine frame plugged into the musical material. In the STSNM procedure, the swept sine frame is split into several segments and is implanted only in the frequency bands masked by the audio contents. This more restrictive arrangement minimizes nuisance but degrades the SNR due to the bandpass filtering process and the overlap-add procedure.

The application of this method assumes that the duration of each sweep frame is much longer than the RT60 value. This allows the algorithm to correctly find some non-prohibited area (area of the time/frequency mapping that hold no relevant acoustical room information) for spectral removal and for keeping the bandwidth of the removed noise as narrow as possible, thus minimizing the loss of the music spectrum contents, especially for the case of TSNM.

These two procedures are in a first approach considered as separate techniques. However, in some circumstances it may be advantageous to use them together. In fact, for music tracks where long term stationary tonal sounds cover the entire audio spectrum of interest, applying the STSNM method has benefits over the TSNM. If this constraint is not met, some parts of the spectrum could be covered by using the TSNM method in conjunction with the STSNM.

The survey conducted over a group of people revealed that by adopting this procedure the music effectively masks the test signal, making acoustical tests with the presence of people quite possible.

Chapter 6

PRACTICAL APPLICATIONS

6.1 Introduction

In this chapter, case studies of practical application of the methods and techniques developed are presented with experimental results and discussion. In the cases where the relevant results were presented in the previous chapters with the purpose of a proper explanation of the experimental technique, only a brief description summarizing the experimental results will be shown.

6.2 Ultrasound transducer arrays to account for time-variance

The use of ultrasonic loudspeaker arrays to account for time variance effects in the acoustical measurements was described in Chapter 4, Section 4.7.

Although simple methods are often used to detect time variance phenomena based on the cross-correlation function, the results are not accurate, showing a considerable deviation in situations of low SNR. Therefore, a non-intrusive innovative approach, using probe signals out of the audio band, was developed in order not to conflict with the test signals and not to be perceived by an audience.

6.2.1 Introduction

In real room acoustical measurements, assuming time-invariance condition is usually not correct due to the non-ideal electroacoustic measurement chain and the variation of the air medium characteristics. A measurement technique was then set-up with the purpose of monitoring the acoustical media, searching for time variance phenomena, for low SNR situations.

A test signal in the ultrasonic band, more precisely at 40 kHz, is applied to the room by using an ultrasound loudspeaker array, with high polar pattern directivity, simultaneous with the test signal frames. The relevant parameters (features) for establishing time-variance and associated thresholds are then estimated from the acquired ultrasonic sound signal. The valid test signal frames, which pass the thresholds test, are labeled with a weighting factor depending on its significance. Otherwise, the frames are rejected, and are not considered in the averaging process.

The measurement of selected acoustical parameters was carried out, as mentioned in Section 4.7, in a mid-size reverberant chamber with a volume of 150 m^3 (5x5x6m), approximately, and reverberation time for the mid-frequencies in the range of 2-2.5 s. Since this enclosure is assumed to be a more controlled environment than a regular auditorium, the time variance occurrence can be better observed.

The medium in the enclosure was disturbed by a number of devices such as a heater (H) and a medium size electromechanical fan with three rotating speeds (F). Additionally, the air was disturbed by manually waving with a pad of area 0.6m2 (WP), in an attempt to force significant air masses movements.

6.2.2 Ultrasonic loudspeaker array set-up

The main characteristics of the developed ultrasound array are: (i) polar pattern with a main lobe of about 8 degrees @ -3dB, (ii) more than 15dB attenuation from on axis to secondary lobe of radiation, (iii) centre frequency of 40 kHz, (iv) maximum sound pressure level of 145 dB @ 30cm, and (v) bandwidth of 2.5 kHz (improved to 10 kHz by filtering compensation). The array set-up is shown in Figure 6.1.

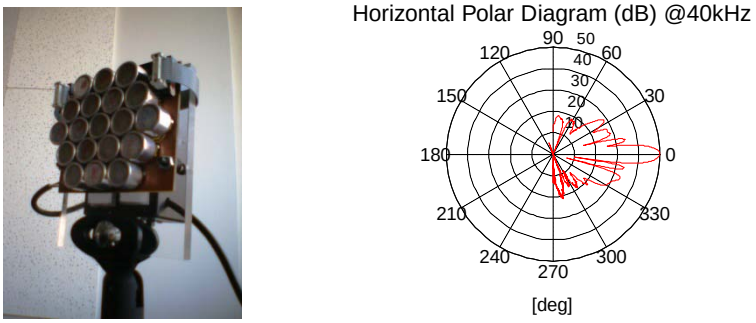


Figure 6.1 - Image and pattern diagram of the ultrasound loudspeaker array.

6.2.3 Experimental results

Figure 6.2 shows results for wavingPad 20 times during the tests with the heater, wavingPad 7 times, heater, fan low and for the reference.

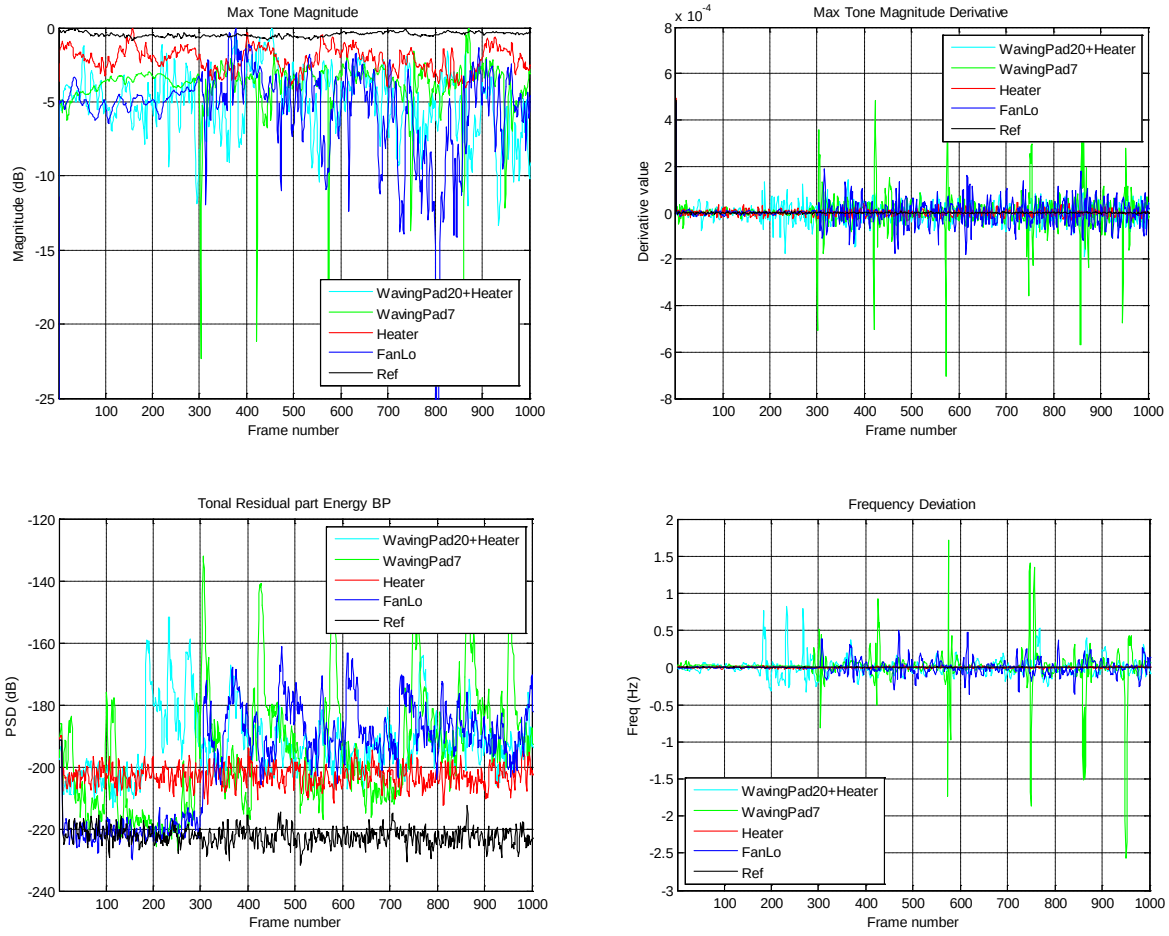


Figure 6.2 – Features Max Tone Magnitude and its derivative, Tonal Residual part Energy BP and Frequency Deviation for the case of WavingPad20 plus Heater, WavingPad7, Heater, FanLo and for the Reference.

In all the tests, except for those using the heater, the medium was not disturbed for the first two frames, the Reference frames, in order to calibrate the system and perform the relative analysis.

Based on the results, a number of thresholds were established in order (i) to detect the occurrence of time variance problems in the enclosure under test, and (ii) as a tool of help for the decision to be used on the averaging stage.

Comparison was made to evaluate the improvement in the SNR. The residual energy analysis was synchronized with the 64 test signal frames. The labeling procedure was used for the cases of full frame or for the *N* frame segments, implying validation and application of the weights before the averaging process.

Table 6.1 shows results for the SNR (in dB), by using the swept sine technique, for the situations of the waving pad, standard averaging and the procedure implemented, respectively. Only the residual energy analysis was used. To highlight the gain by sorting the frames, the columns for 4 and 8 NAve are shaded.

Table 6.1 – SNR for the situation of Pad Waving, standard averaging (top most) and the procedure implemented with US (down most).

Band/#Ave	Ref	2	4	8	16	32
0.5	65.1	67.4	55.0	53.7	59.4	63.4
1	62.5	66.5	53.0	54.2	59.8	62.1
2	55.8	59.9	49.6	51.8	57.3	59.3
4	53.6	57.9	48.9	51.4	56.8	57.9
8	60.4	64.5	55.1	58.3	63.5	64.8
AWeight	49.8	54.0	44.0	46.1	51.5	53.3
Full band	49.8	53.7	44.3	45.9	51.3	53.4

Band/#Ave	Ref	2	4	8	16	32
0.5	65.1	67.1	58.8	63.1	62.7	63.3
1	62.5	66.4	56.5	60.5	60.1	61.8
2	55.8	59.7	53.5	56.1	56.5	59.3
4	53.6	57.9	53.0	55.8	55.7	58.2
8	60.4	64.4	60.4	62.3	63.1	65.3
AWeight	49.8	53.9	48.0	50.9	51.0	53.4
Full band	49.8	53.6	48.2	51.2	51.3	53.5

The measurement results for the MLS are very poor when compared to the swept sine in situations of time variance. This finding was already mentioned by previous researchers [Farina00, Muller01].

6.3 Statistical classification of road pavements

Road noise abatement policies frequently make use of acoustical barriers. However, the financial cost of these devices can be high and their applicability and performance in urban areas is usually limited. Moreover, their visual impact tends to decrease their acceptance by local communities. Sound insulation reinforcement of the most exposed facades is another frequent measure, though its efficiency strongly depends of the door and window sealing conditions; this may cause some discomfort, even when an adequate air conditioning system is installed. This measure has no effect on outdoor noise, such as in gardens, balconies or outdoor public spaces. With noise reduction at the source being the most effective strategy, and since noise regulations tend to favor increasingly lower noise limits, research and development initiatives are pursuing innovative road noise reduction solutions, especially those associated with road surfaces to reduce the tire/road noise. Given the stringent limitations on vehicle noise emissions established in Europe, USA and Japan in the last decades, the tire/road interface is recognized as the major source contributing to road traffic noise [Bento-Coelho07].

Low-noise road surfaces have been increasingly viewed by road planners as a cost-effective solution. This is particularly relevant for road sections where speed typically exceeds 50 km/h [Sandberg02], since, for light cars, tire/road noise is prevalent at this speed range. Low-noise road surfaces are basically built by changing the texture and/or the porosity of the mixtures of which they are made. In some conditions, a noise reduction of up to 10dB can be achieved [Freitas08a]. Therefore, an increasingly wide variety of road surfaces is becoming available.

Different methods for road quality certification and assessment have become standard and are currently being used. The most common methods are based on the measurement of the surface texture and roughness, often by profiling the transversal and longitudinal road surface [Sayers98]. However, these techniques are not established for noise purposes. Therefore, road engineers and planners, as well as contractors, often have insufficient data on the correlation between the type of road surface and the resulting noise emission profile, throughout the road's operational life. Furthermore, the problem of not knowing exactly the correlation between the texture of surfaces and the noise emission profiles is an unsolved issue. It became clear that there was a need to summarize and incorporate the information regarding noise and texture into a single resource document. This recognition led to the present research, which aims to provide

additional information to complement data, such as texture, roughness, and related parameters, acquired by existing standard procedures.

A further issue, relevant for road decision-makers, regards the noise characteristics of different segments in the same road section. Road noise levels can be quite different due to different pavement surfaces, maintenance conditions, and surface defects.

This work describes a new acoustic measurement system that was developed to analyze and classify the sound signal generated by the tire/surface interaction of a moving vehicle. A system capable of identifying/classifying the type of road surface was built by using statistical learning tools [Paulo08] and [Paulo09P].

6.3.1 Methodology

The vehicle rolling acoustic signal is acquired by following the Close Proximity procedure (CPX), as described in the ISO CD 11819-2 standard [ISOCD11819-2]. This standard regards the measurement of the influence of road surfaces on traffic noise under near field conditions, where tire/road noise predominates (i.e., where engine noise, intake noise, and exhaust noise can be discarded), which is known to happen for speeds above 50km/h, for light vehicles. For the purpose of the present research, the sound signal was captured by two microphones mounted on the vehicle.

The macrotexture was acquired with a High Speed Profilometer (HSP) following the ISO 13473-1 standard [ISO13473-1]. Therefore the indicator used for comparing macrotexture and noise was based on the Mean Profile Depth (MPD).

Two types of analysis were carried out: (i) long-distance road analysis, to survey a road network, for surface inconsistencies between two locations (the road under test could include different types of surfaces due to maintenance interventions and/or pavement distress anomalies – only noise), and (ii) analysis of short road segments, usually of a few hundred meters, composed of a single type of pavement (noise and texture). For the first type of analysis, the classification results are presented with the noise parameters and the location of the surface types on a map, using geographical information provided by a GPS device. The equipment also measures the vehicle's speed for each geographical coordinate. A block diagram of the noise system is shown in Figure 6.3.

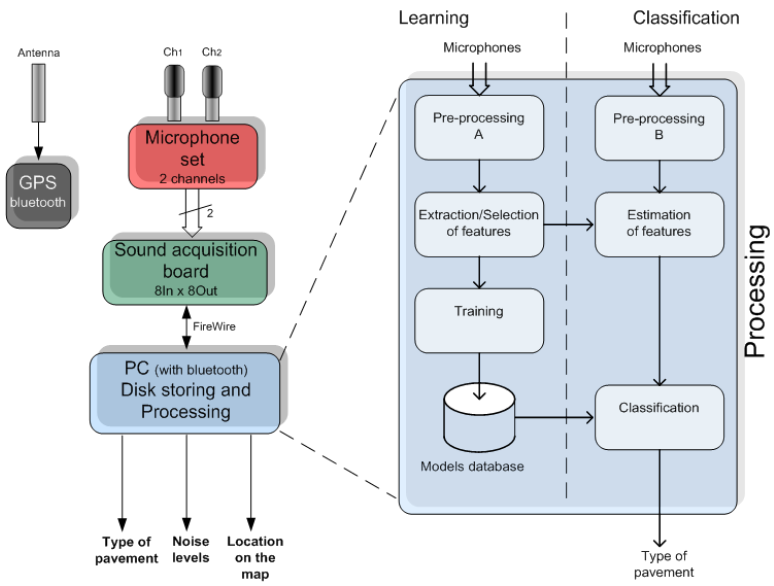


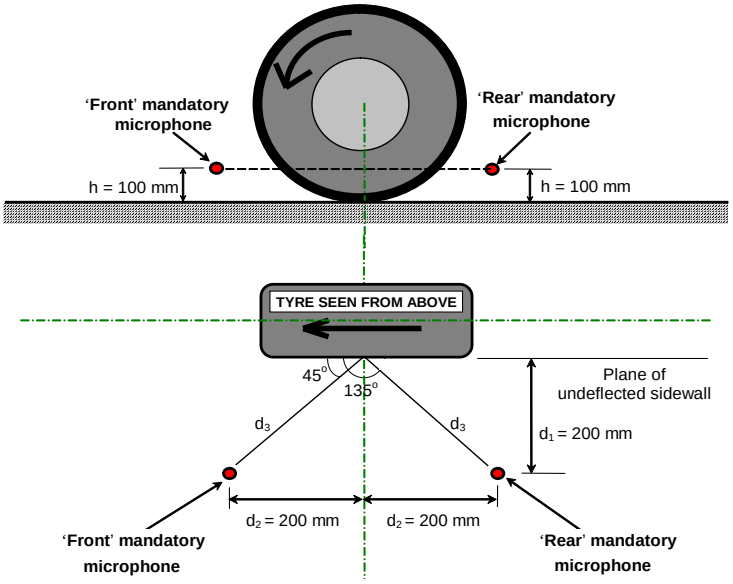
Figure 6.3 – Block diagram for the road pavement classification system.

Although the reference speeds recommended in the ISO CD 11819-2 standard are 50, 80 and 110 km/h, in some experiments only the average speed of 80km/h was used. This speed was chosen for the purpose of model validation. To compensate for speed deviations, a speed correction factor was applied to the sound signal captured by the microphones. This factor is estimated by using $\hat{L}\max = 32*\log(V) + b$, based on the linear regression of $L\max$ measurements with the logarithm of the speed V, where b is determined by using the mean value of the sound samples acquired for the reference speed of 80km/h, with a maximum deviation of 1 Km/h [Freitas08b]. The corrected maximum levels are determined by $L\max = \hat{L}\max_{vref} + \varepsilon$ where $\varepsilon = L\max_{vcur} - \hat{L}\max_{vcur}$ is an estimate of the error resulting from speed deviation, and the *vcur* and *vref* subscripts indicate current speed, and reference speed of 80 Km/h, respectively [Paulo10a] and [Paulo10b]. The error resulting from this approximation is about 1dB for a speed variation of 5 Km/h, which is lower than the variance of the results. The spectral envelope characteristics of the sound are almost unchanged for small changes in vehicle speed.

6.3.1.1 Signal acquisition

A support structure was designed and built to steadily hold the microphones in a suitable position in a standard light vehicle, for CPX measurements. The hardware system includes the following components: (a) an aluminium structure attached to the vehicle's chassis, on which the

microphones were mounted following the standard recommendations, (b) a set of two ½ inch omnidirectional condenser microphones, (c) an external sound board with 2 in/out lines, (d) a speed metering device coupled to one wheel, (e) a GPS device, (f) a personal computer, and (g) a DC-AC power inverter (12Vdc to 230Vac) to feed the devices [ISOCD11819-2]. This experimental arrangement is shown in Figure 6.4.



a)

b)

Figure 6.4 – Microphone measuring positions scheme (after [Sandberg02]) (top most) and Close Proximity system developed for the CPX tests (down most).

The captured sound signals were stored and processed in a specific audio module based on the Matlab toolbox platform. The overall sound emission signals were post-processed by applying a linear average approach to the signal for several road sections. The length of the pavement

surfaces tested extended for more than 100 m avoiding the need of averaging different runs on the same road. The sound pressure levels measured at the 'front' and at the 'rear' microphones were arithmetically averaged.

The macrotexture information was provided by a Road Surface Profilometer® system mounted on a vehicle, as shown in Figure 6.5.

Figure 6.5 – Picture of the High Speed Profilometer (HSP) system.

6.3.1.2 Pavement types

Two road pavement databases were considered within the scope of this work.

The first pavement database, db1, was built including eight general surface types, as depicted in Figure 6.6. These surface types cover those most commonly used in Portuguese roads. Some have been traditionally used: porous asphalt1 (PA), cubic paving stones (PCS), and cement concrete (CC), while others have been introduced more recently, namely in silent pavement studies: gap graded asphalt rubber (GGAR), rough thin layers (RTL) and asphalt recycled rubber (ARR). The age of the tested pavements is up to three years, except for a type of dense asphalt (denoted herein as 'old') which is a five-year old sample surface. The mean traffic density is about 40,000 vehicles per year. Figure 6.6 shows the visual aspect of these types of pavements.

Figure 6.6 – Pavement types tested in this study (database db1). 1- porous asphalt1 (PA1) (new), 2- gap graded asphalt rubber (GGAR), 3- rough thin layers (RTL), 4- dense asphalt (DA) (old), 5- cubic paving stones (PCS), 6- cement concrete (CC), 7- asphalt recycled rubber (ARR-BMB), 8 - porous asphalt 2 (PA2).

The second pavement database, db2, consists of a set of five road surfaces also commonly used in Portugal: one of Dense Asphalt (DA), one of Slurry Surfacing (SS) and three surfaces of Open Graded Asphalt Rubber (OGAR), which at the present time are used extensively.

Table 6.2 presents the maximum aggregate size and the age of each surface. The maximum aggregate size together with the texture and the porosity and the age of the surface were reported in literature as the main parameters affecting tire/surface noise [SILVIA06]. Figure 6.7 shows the visual aspect of these pavement surfaces.

Table 6.2 – Properties of the asphalt mixtures in the second database of pavements.

Type of mixture	OGAR1	OGAR2	OGAR3	SS	DA (0/16)
Maximum aggregate size (mm)	10	12	10	10	16
Age (years)	1	4	4	2	1

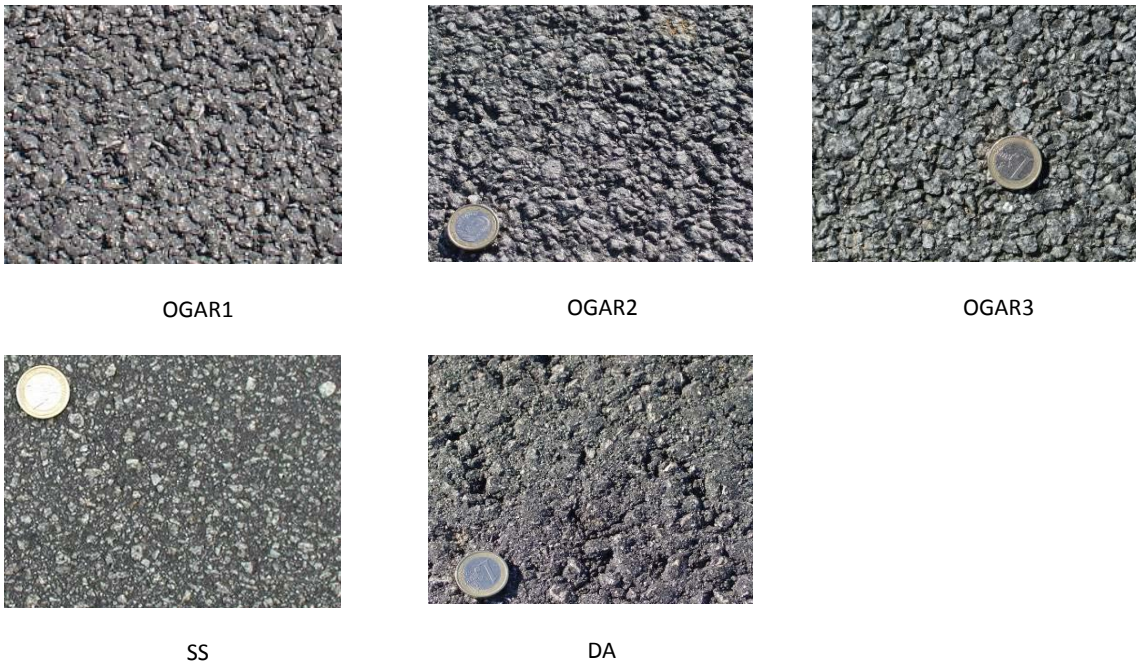


Figure 6.7 – Visual aspect of the surfaces used in database db2.

6.3.1.3 Signal pre-processing

This module applies the A-weighting curve to the acquired sound signals. This procedure has the purpose of relating the noise emission with the perceived sound by people. In the same time, the low frequency energy generated mostly by the moving vehicle's air flow, i.e. a spurious signal contribution, is also attenuated.

The tire/road noise generation phenomenon is basically caused by two different types of mechanisms: (i) excitation of the tire structure by vibrations, leading to sound radiation, and (ii) aerodynamic effects. Sound is generated mostly by the radial vibration of the tire structure, as well as from the vibration of the sidewalls, with the latter being less important and occurring mostly at low frequencies, up to about 300 Hz. The changes in contact forces over long time periods can result from inconsistencies (e.g. tire non-uniformity), defects in the tire structure tread pattern geometry, and road texture, and mechanisms due to the tire/surface contact, such as stick slip and stick snap. Firstly, the tangential and radial forces lead to friction mechanisms, which in turn lead to stick-slip, when the tread blocks are in contact with the road. Secondly, when the tread blocks are about to lose contact with the road, certain adhesive forces will tend to keep them in contact with the road, causing stick-snap.

In general, tire vibrations are expected to be the main source of noise, for frequencies below 1000 Hz. For higher frequencies, the noise generation is interpreted as air-pumping. As the tread of the tire enters the leading edge of the road contact area, air is squeezed out as the tread is compressed and penetrates into the road surface. At the trailing edge, the tread is decompressed and lifts up from the road surface at which point air rushes back in to fill the voids [Sandberg02, Kropp07, Descornet06, Paje09, Lui04, Berengier97, Attenborough80].

The scores of generation mechanisms, which in turn lead to a plethora of tire/road noise scenarios, make it rather difficult to decide which modeling strategy to use. The issue is even more problematic because of the complexity of the radiation conditions for both vibration and aerodynamic sources. Therefore, the noise spectrum should be split into several complementary frequency bands in order to monitor the tire/pavement noise generation phenomena so that the performance of the classifier is enhanced. In this study, three frequency bands were used: a low band (200 - 1400 Hz), a middle band (1400 - 3000 Hz), and a high band (3000 - 8000 Hz). Additionally, a 1/3 octave band analysis was carried out in some cases, for improved frequency discrimination. A full band frequency analysis was also used to account for the overall signal energy. In order to improve the results in the classification stage, and for comparison purposes, a smoothing technique was also applied to the features.

All these procedures were carried out using time segmentation (windowing) of the signal acquired from the microphones and from the HSP device. The window length was adjusted to identify short duration events such as vehicles passing-by and surface deficiencies, and these disturbances were removed to keep only the stationary part of the signals.

6.3.2 Feature extraction

Although a number of systematic methods can be considered for selecting the best features, this choice depends heavily on the knowledge about the nature of the subject [Li01]. The features vector adopted in this work corresponds to a set of characteristics that best suits our goals with regard to noise and texture, while additional characteristics can be used to improve classification accuracy. The features which characterize the pavement were extracted from measurements of texture based on the Mean Profile Depth, MPD.

The road pavement classification process starts with the extraction of a set of features from the sound signal. Features were obtained for frames of 10 and 20 seconds, depending on the extension of the road section, using Hanning sliding windows of length equal to 2048 samples. The sliding factor was 25% of the window length (hop-size of 512 samples), with sampling frequency of 32 kHz. The use of this sliding factor (lower than 50%) allows a better identification of the variability of the pavement texture and of the location of any surface inhomogeneities. The feature vectors were grouped into classes according to the type of pavement.

The feature selection procedure has several goals: to remove any redundant characteristics, to decrease the dimensionality of the feature vector, and to decrease the computational cost of the classifier. However, the vector of selected features has to include all the fundamental characteristics of the road surface that are relevant to the classification task.

The set of features used (and their acronyms) are presented next.

Noise features

The sound pressure RMS value is the feature that, in a first approach, best describes the type of surface, given its close relationship to the noise energy perceived by the human auditory system. For this reason, the A-weighting curve was applied to the acquired signals, thereby removing the low frequency energy contributed by the moving vehicle's air flow, as mentioned above. Therefore, the RMS is the primary noise feature to be used.

1. Overall sound pressure levels – *LAeqFull*. This is a measure of the total energy of the noise effectively perceived by the human, in the range of 200 to 8000 Hz.
2. Narrow band sound pressure levels: low band – *LAeqLo* (200 - 1400 Hz), middle band – *LAeqMid* (1400 - 3000 Hz) and high band – *LAeqHi* (3000 - 8000 Hz).

   The above two types of features, following A-weighting and filtering, are calculated by using the short time average of energy according to

$$LAeqX_i = \frac{1}{W}\sum_{i=1}^{W} s_A{}^2(i)w(n-i) \ , \qquad (6.1)$$

where *X* denotes *Full, Lo, Mid* or *Hi* suffixes, *w* is the Hanning window of length *W*, and $s_A(i)$ is the noise signal (after A-weighting), in the *i*-th frame.

3. Spectral energy centroid – *Ct*. The computation of this feature is based on the squared magnitude of spectrum, using the short-time Fourier transform (STFT), performed frame-by-frame along the time axis [Li01]. The spectral centroid of the *i*-th frame *u* is defined as

$$Ct_i = \frac{\sum_{u=0}^{M} u\left|f_i(u)\right|^2}{\sum_{u=0}^{M}\left|f_i(u)\right|^2}, \qquad (6.2)$$

where *M* is the index of the highest frequency band and $f_i(u)$ represents the STFT of the *i*th frame.

The use of the squared magnitude of the spectrum, a measure of the noise energy, is aimed at a better correlation with the stimulus perceived by the human auditory system.

4. Band spectral energy centroid: low band - *CtLo*, middle band - *CtMid* and high band – *CtHi*. These features are computed in the same way as *Ct* with the sound signal band filtered.

5. Spectral Energy Maximum – *FreqMax1LAeq*. This feature corresponds to the main peak of the power spectral density estimate. The tire/road pavement noise spectrum profile is characterized often by a maximum sound intensity centered in the range of about 800 to 1800 Hz.

6. Secondary Spectral Energy Maximum - *FreqMax2LAeq*. This feature corresponds to the secondary peak of the smoothed power spectral density of the noise. A well-defined secondary peak appears in the noise spectrum for some types of surfaces. This feature has a significant meaning for thin layer pavement surfaces.

These two last features are evaluated by averaging *N* consecutive frames (moving average method, with *N*=2) in the frequency domain in

$$FreqMaxXLAeq_i = \max\left\{\frac{1}{2N+1}\sum_{k=i-N}^{i+N}\left|f_k(u)\right|^2\right\}, \quad (6.3)$$

with $X = 2$, FreqMax2LAeq, for the Secondary Spectral Energy Maximum, and where i is the current frame.

7. Maximum spectral energy modulation index – *FreqModIndex*. It was observed that, for some types of pavements, the *secondary spectral energy maximum* feature shows a kind of frequency modulation. This feature is applied to the *FreqMax2LAeq* feature, giving information of the frequency drift of the secondary spectral energy maximum. This feature, estimated by the derivative of the *FreqMaxLAeq2*, is relevant for thin layer pavement surfaces, as an indicator of the pavement quality.
8. 1/3 octave band sound pressure levels - *Oct13B*.
9. Delta spectrum magnitude – *DiffSpecMag*. For each feature analyzed in the frequency domain, the energy difference in consecutive frequency bands in the same frame is computed.
10. Spectral Energy Maximum – *FreqMaxLAeq*. This feature corresponds to the main peak of the power spectral density estimate. The tire/road pavement noise spectrum profile is characterized often by a maximum sound intensity centered in the range of about 700 to 1200 Hz. This feature is evaluated by averaging N consecutive frames (moving average method) in the frequency domain.

The sound pressure RMS value is highly dependent on several factors, such as the speed of the vehicle, the type of tires, and the age of the pavement. Therefore, a combination of features is required. Examples of the time evolution of some of the features (*LAeqFull*, *LAeqLo*, *LAeqMid*, *LAeqHi*, *CtLo*, *CtMid* and *CtHi*) for some of the pavements are depicted in Figure 6.8. This figure also shows the smoothed versions of these curves. The curves show that RTL pavements lead to noise levels of about 4dB lower than ARR pavements, on average, though the variance is much higher. For instance, the standard deviation of the *CtMid* feature for ARR pavements is twice that for the RTL pavements (31 and 59 Hz, respectively).

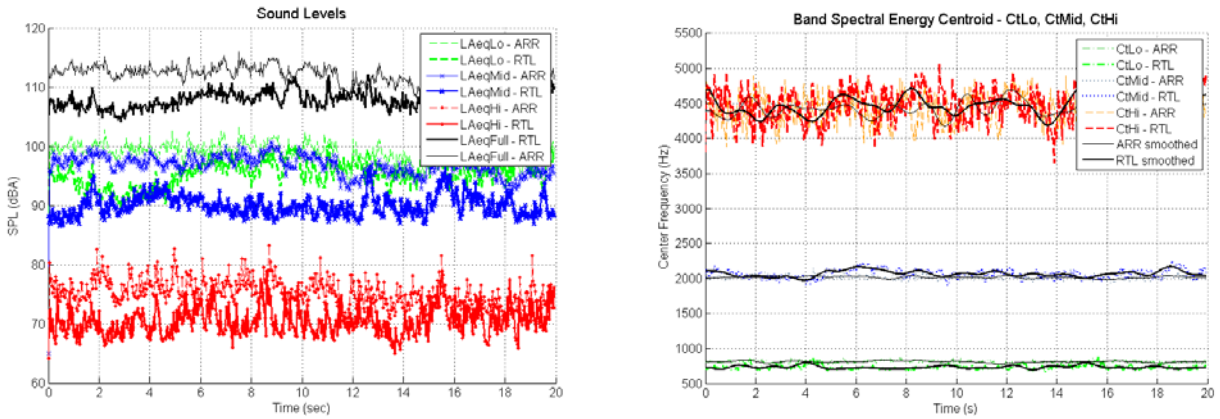


Figure 6.8 – Pavement types features; left: LAeqFull, LAeqLo, LAeqMid, LAeqHi (an offset of 10dB was applied to LAeqFull curves for better visual purposes); right:CtLo, CtMid and CtHi for GGAR (thin curve) and RTL (thick curve), for a car speed of 80 Km/h and an observation interval of 20s (~440 m).

Texture features

The MPD is used to characterize the macrotexture of pavements, giving a measure of the average of the profile depth only. Therefore, the intent of using this feature to relate noise and texture is limited. The following development provides a step forward on this issue. The estimation of MPD is given by MPD = [peak level (1^{st}) + peak level (2^{nd})]/2 - Average level, after the ISO 13473-1 standard. The diagram depicted in Figure 6.9 illustrates the concept.

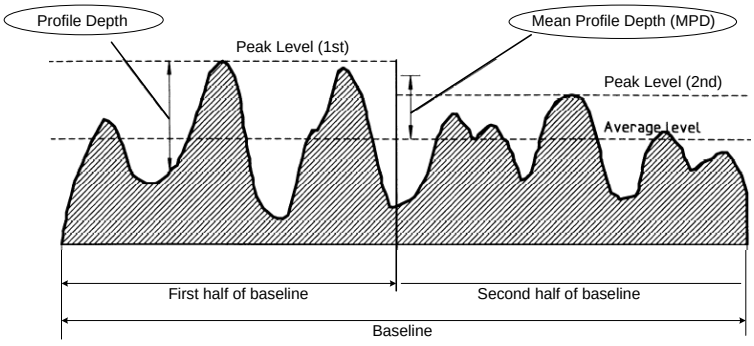


Figure 6.9 – Scheme for the calculation of MPD.

The MPD method does not address some surface texture features such as the pavement aggregate particle shape, size, and distribution. The method is not meant to provide a complete assessment of pavement surface texture characteristics. In particular, care should be exercised in interpreting the result if the method is applied to porous surfaces or to grooved surfaces [ISO13473-1].

Despite these considerations, several characteristics, defining different pavements, can be extracted.

Figure 6.10 shows the MPD variation for all the tracks (only smoothed curves are drawn for clarity).

Regardless of the type of mixtures, the MPD is considerably high, except for the SS surface.

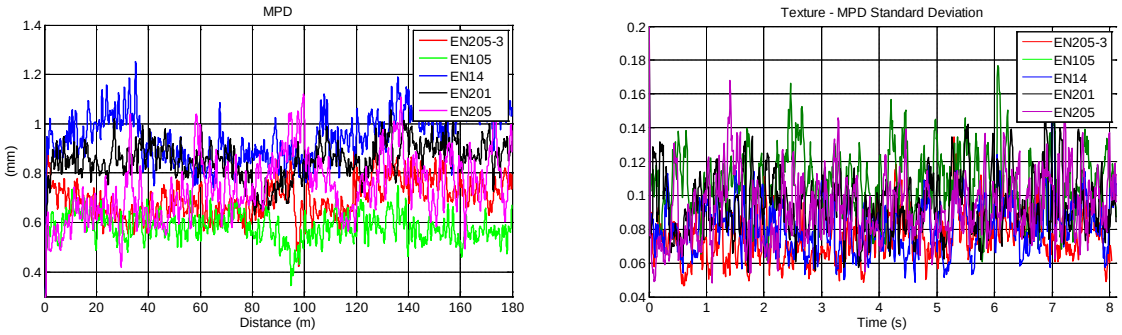


Figure 6.10 – Mean Profile Depth, MPD, and Standard deviation of MPD of the road tracks.

The criteria adopted for choosing the set of texture features was based on the physical meaning of the data given by the MPD. Since the MPD is itself a mean measure, all the parameters extracted are mean values too. The parameters analyzed are:

1. Mean value of MPD, meanMPD (in mm) – this feature represents the MPD mean value using the moving average method;
2. Standard Deviation of MPD, stdMPD (in mm) – gives information about the regularity of the surface;
3. Maximum of MPD, maxMPD (in mm) – this measure is the maximum deviation around the mean MPD. The absolute maxima is a biased version of this feature;
4. Crest Factor of MPD, CrestFMPD (in dB) – consisting of the maxMPD to rms ratio in dB. Like the stdMPD, it gives information about the regularity of the surface;
5. Peak-to-peak of MPD, pk2pkFMPD (in mm) – consisting of the difference between the maximum value of MPD and the minimum value of MPD;
6. Mean Correlation MPD, meanCorrMPD - Mean Correlation between two halves of a MPD frame. Each frame is split into two parts with the same length and cross correlated;
7. Centroid MPD, CtMPD – this feature gives information about the centre of gravity of the spectral energy.

6.3.3 Classification

The feature vectors from each pavement class are divided in two sets: a training set and a test set. Therefore, the module used for classification was the result of a supervised learning approach. A database populated by entries describing different types of surfaces was built beforehand

[Duda73]. A randomly chosen subset of 30% of the vectors is assigned to the training set and the remaining 70% correspond to the test set. The Bayesian and Neural Network classifiers were tested. An *N*-dimensional Gaussian density model (where *N* is the length of the feature vector) is then adjusted to each class, based on the training set, assuming that each road surface class comprises a single centroid with its own mean and co-variance.

Although different techniques to implement pattern classifiers are available, a set of discriminant functions were adopted, each one corresponding to a pavement type, which are simply given by the posterior probability of each class, given the observed feature vector *x*, that is, $\delta_k(x) = P(k/x)$, for $k \in \{1,\ldots,K\}$ [Li01]. These posterior probabilities are calculated by using Bayes law, using the normal distribution assumed for the classes, with mean and covariance estimated from the training dataset. The standard maximum a-posteriori (MAP) classifier corresponds to choosing the class which maximizes the discriminant functions for all possible classes *k* [Duda73] and [Hastie01], that is,

$$\hat{k} = \arg\max_k \ \delta_k(x) \tag{6.4}$$

where $\hat{k}$ denotes the class estimate for feature vector x.

Because of the adoption of separate estimates of the covariance matrix for each class, this classifier corresponds to a quadratic discriminant analysis (QDA). An alternative procedure would be to adopt a common covariance for all the classes (linear discriminant analysis – LDA), which can usually be estimated more robustly from limited data. However, it was found that QDA achieves better results, without an insignificant increase of computational load, and it was thus adopted in this study.

Given the assumption of Gaussian densities for the classes, each with its own mean μ_k and covariance matrix C_k,, for $k = 1,\ldots, K$, the quadratic discriminant functions are simply given by

$$\delta_k(x) = -\frac{1}{2}\log\left|C_k\right| - \frac{1}{2}(x-\mu_k)^T C_k^{-1}(x-\mu_k) + \log P(k), \tag{6.5}$$

where $P(k)$ is the a priori probability of class *k* and *T* denotes matrix transpose.

After obtaining each covariance matrix, C_k, estimating the mean, μ_k, and calculating the a priori probability of each class, $P(k) = N_k/N$ (where N_k is the number of training samples of class *k* and

N is the total number of training samples), the classifier is ready to classify new samples by computing the values of $\delta_k(x)$ for each $k = 1,\ldots,K$, and applying the decision rule in Eq. (6.4).

The Neural Network classifier is a feed-forward network with tan-sigmoid transfer functions for both the hidden layer, which uses 9 neurons, and output layer.

In this study, the a priori probability of occurrence of each pavement class, $P(k)$, was simply considered equal to $(1/K)$. However, in real world applications, in the presence of a reasonable acknowledge of the type of surfaces of a road network, the results can be improved by considering different a priori probabilities for each class of pavement. For instance, if one is interested in listing the types of pavements during a road trip predominantly using highways, the PCS pavement type (cubic stones) is very unlikely to appear. This can be formalized by letting the corresponding a priori probability be very low.

In order to test different combinations of statistical classifiers with the purpose of classification/identification of road pavements, a few case studies were carried on, as described next.

6.4 Case study I – Noise-based Bayesian approach

This study used the first database.

The assumption of normal distribution for the selected features is supported by the observation of the distribution function of each feature, as depicted in Figure 6.11.

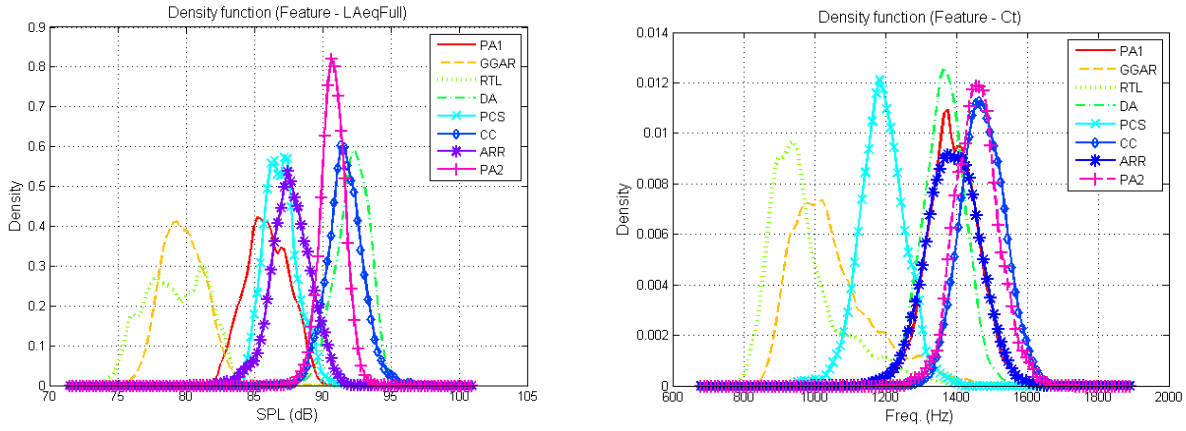


Figure 6.11 – Distribution functions for the features LAeqFull (A-weighted broadband equivalent sound pressure levels) and Ct (Centroid).

6.4.1 Experimental Results

In order to have some insight into the classifier results, the corresponding decision boundaries were plotted. Since nine features were used, and it is impossible to graphically represent the 9-dimensional decision boundaries, 2D projections of these boundaries for three choices of pairs of features, *LAeqFull - CtLo*, *LAeqFull - CtMid*, and *LAeqFull – CtHi*, are depicted in Figure 6.12.

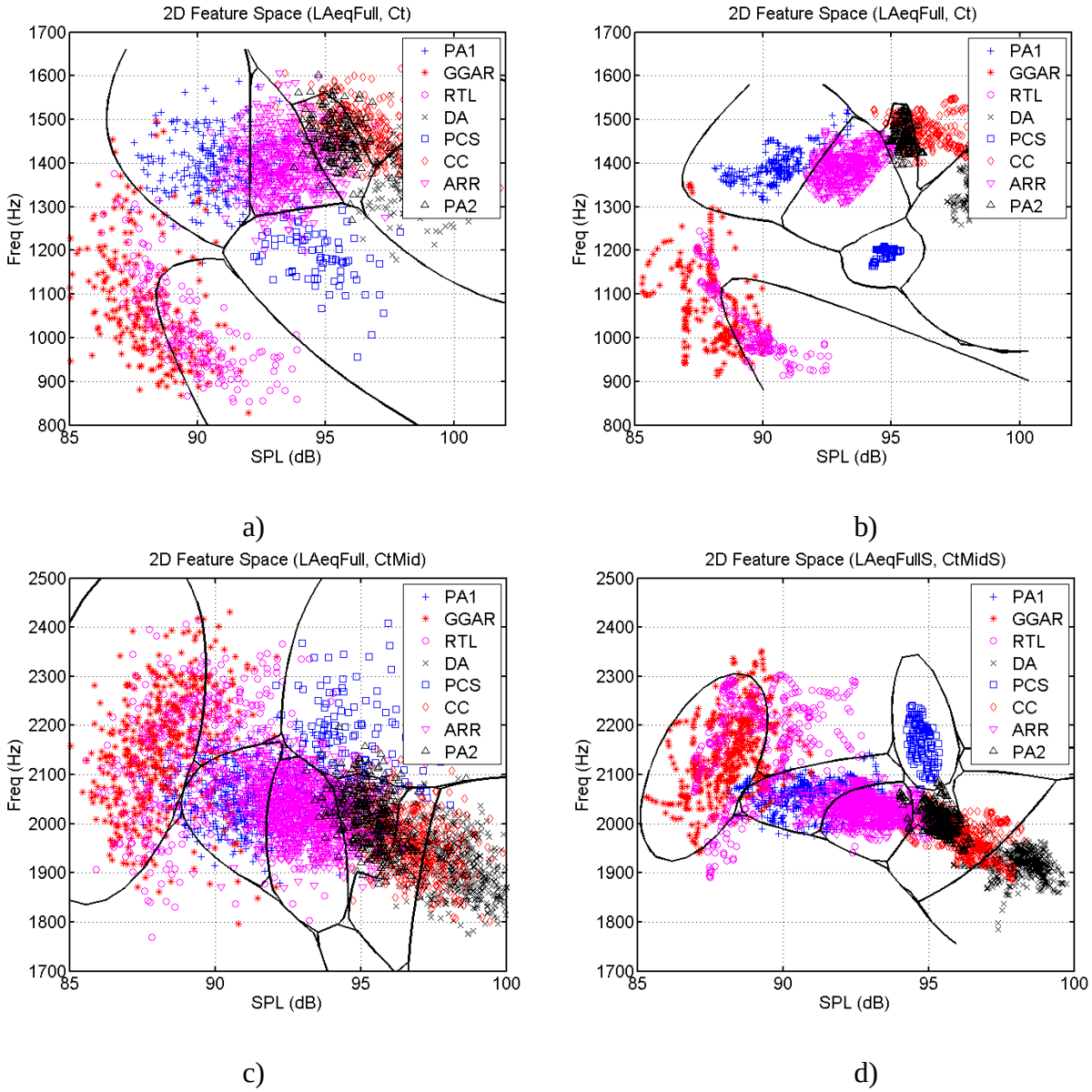


Figure 6.12 – Feature space in 2D of a) LAeqFull vs. Ct and c) LAeqFull vs. CtMid with the respective smoothed versions b) and d).

The figure shows that the GGAR and RTL pavement types behave similarly for the low and middle frequency bands. However, in the high frequency band (feature space *LAeqFull – CtHi*) these two classes can be easily separated.

The feature *Oct13B*, averaged over the different frames, yields relevant information on the stationary behavior of each pavement type, with respect to frequency as depicted in Figure 6.13.

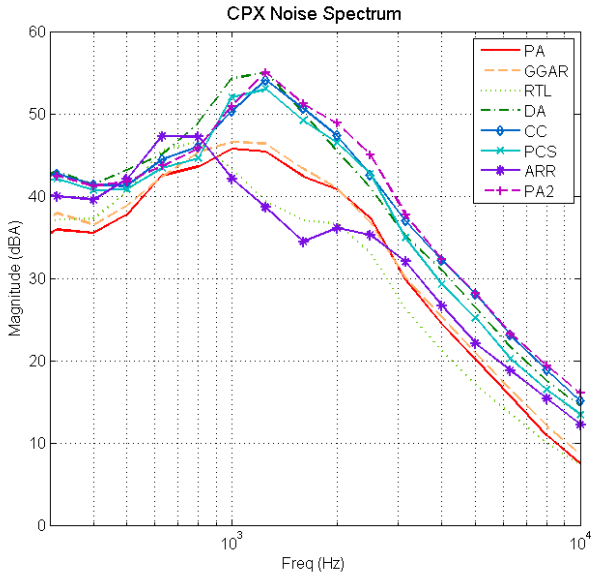
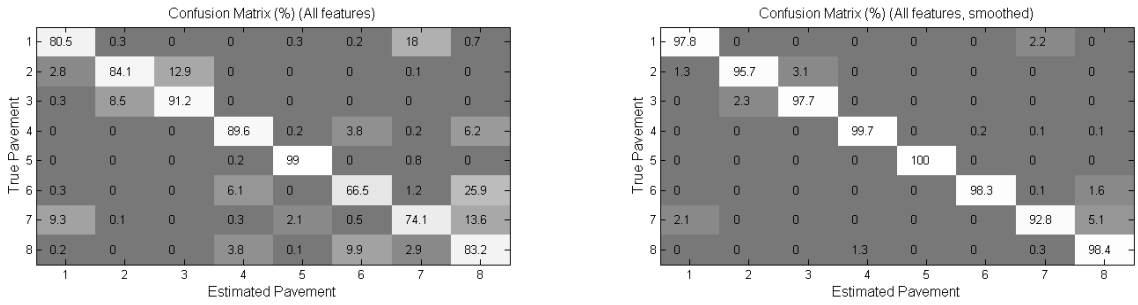


Figure 6.13 – Pavement spectra for the different surface types.

The performance of the classifier is assessed by means of the corresponding confusion matrix, obtained with the test dataset, see

Table **6.3**.where the results are presented in percentage of samples with true pavement classification. This is a square matrix, with dimension equal to the number of classes, in which each entry, say (k_1, k_2), is an estimate of the probability that an observation (from true) pavement class k_1 is classified as belonging to class k_2. The main diagonal of the confusion matrix corresponds, of course, to the correct classifications. Therefore, for an ideal error free classifier, in which each pavement sample is classified according to its true class, only the elements on the main diagonal of the confusion matrix would be non-zero.

Table 6.3 – Confusion Matrix using the Bayesian classifier. 1 - original features (left) and 2 - features with smoothing (right).

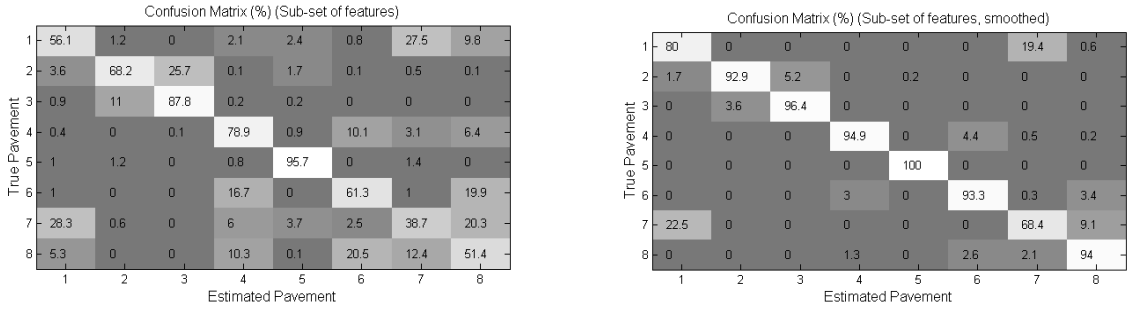


The classification results for the tested road pavements show a clear agreement between the true pavement class and the classifier output. The results also show a lower error when the smoothing technique is applied to the features. However, for a complete analysis, the original data should also include all the information concerning the spreading behavior of each feature, thus providing the classifier with all the information. This increases the robustness of the method to other different types of pavements.

Since the sound levels generated are directly related to the speed of the vehicle, an accurate surface classification should be based on measurements carried out at different constant speeds. However, such a procedure involves long term measurements and is not always a practical option. In an attempt to overcome this issue, an additional approach that used a feature space subset, formed by selecting some features by inspection (a manual feature selection method), was tested. The subset used contains features that are independent of the absolute values of the sound levels, normalized features, such as the *Ct*, the *CtLo*, the *CtMid*, the *CtHi*, the *FreqMaxLAeq*, the *DiffSpecMag*, the *FreqModIndex* and the *SpecEnergyR*.

The classification results using this subset of features show minor errors on identifying the correct pavement, as shown in Table **6.4**. This suggests that a reduced number of features can be used in the classifier, without a significant decrease of performance.

Table 6.4 – Confusion Matrix using the Bayesian classifier with all the features excluding LAeqFull, LAeqLo, LAeqMid, LAeqHi. 1 - original features (left) and 2 - features with smoothing (right).



Preliminary results show that the accuracy of the road pavement classifier is not notably affected by the vehicle's speed, at least for the tested pavement types.

The classified road section pavements can be plotted on a geographical map, showing the results of a road network survey. Figure 6.14 depicts the sound pressure levels together with the pavement types identified on a 30km journey (from Lisbon to Montijo).

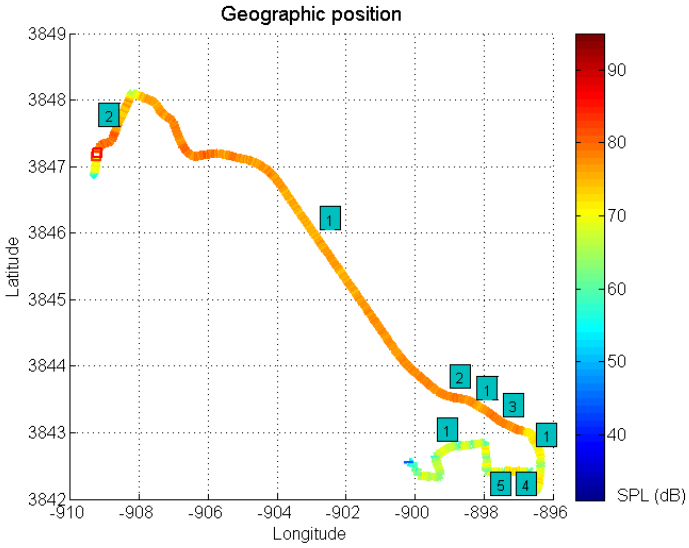


Figure 6.14 – Road pavement classification/localization survey using a GPS device (the sound emission levels are shown by using a color scale).

This approach can contribute to increase the accuracy of road noise prediction models, since they take into account the actual type of pavement of each road section, by assigning corresponding corrections according to the noise emission characteristics. For the implementation of this

procedure, surveys on a number of road sections were carried out by using a GPS device synchronously with the CPX system. The types of road surface sections identified are marked by the numbered squares, corresponding to the pavement types.

6.5 Case study II – Noise texture-based Bayesian approach

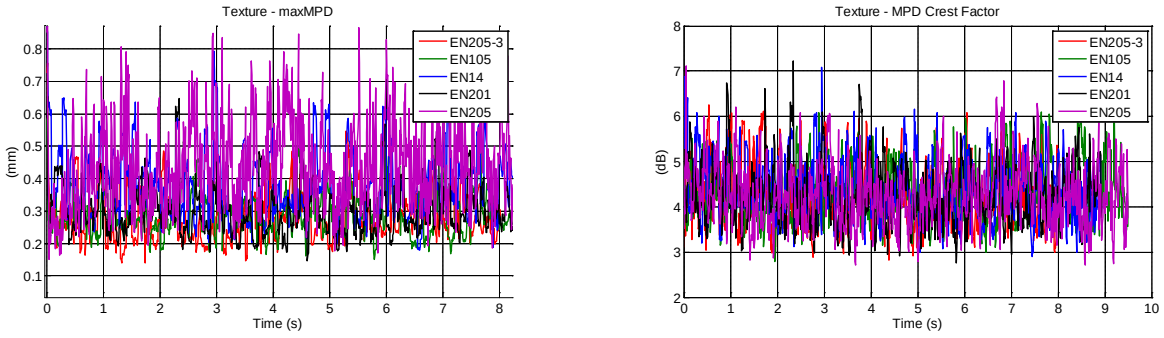
This case study uses the second database of road pavements with the Bayesian classifier.

The feature vector of adopted in this work includes a set of characteristics which are the most appropriated to relate noise and texture.

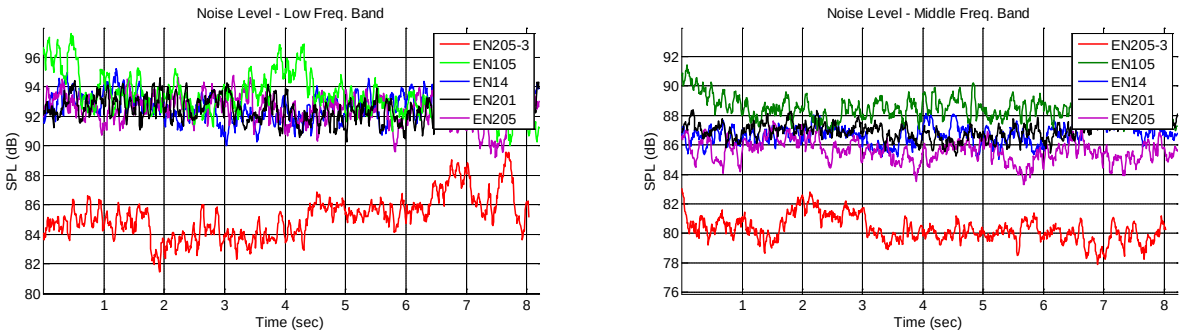
The features were obtained for frames of 2048 samples using rectangular or Hanning sliding windows, depending on the analysis. The sliding factor was 12.5% of the window length (hop-size of 256 samples), with a 32 kHz sampling frequency. The use of this sliding factor (lower than 50%) allows a better identification of the variability of the pavement texture and location of any surface irregularities. In order to make the rate of samples for texture and noise data measurements equal, the resampling technique was applied to the texture data (the data provided by the MPD measurements is 40000 samples/km (one sample for each 2.5 cm) for a vehicle speed of 80 km/h (resampling factor of R= 36).

6.5.1 Selection of the relevant features

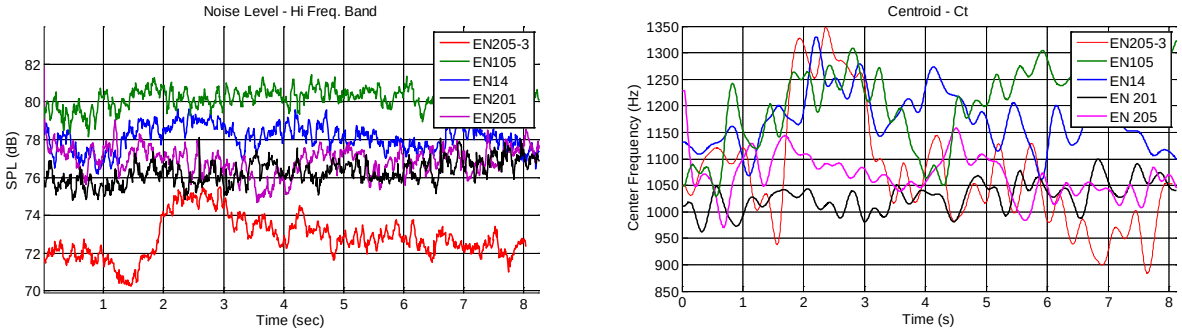
Examples of the time evolution of some of the features such as, maxMPD, stdMPD, LAeqLo, LAeqMid, LAeqHi, and Ct are depicted in Figure 6.15.



a)



b)



c)

Figure 6.15 – Time evolution of some features of the pavements; a) maxMPD (topmost left) and stdMPD (topmost right), b) LAeqLo (middle left) and LAeqMid (middle right; c) LAeqHi (downmost left) and Ct (downmost right).

Table **6.5** collects the relevant global values of the parameters observed by the features used on this study.

Table 6.5 – Relevant global values of the road parameters.

Feature	Statistics	OGAR1	SS	OGAR2	OGAR3	DA
LAeq (dBA)	ave	88	96	95	94	95
	std	0.9	1.1	0.8	0.9	0.7
LAeqLo (dBA)	ave	85	93	93	92	92
LAeqMid (dBA)	ave	80	88	87	87	85
LAeqHi (dBA)	ave	73	80	78	76	77
Ct (Hz)	ave	1000	1200	1150	1000	1050
	std	110	79	55	27	38
maxFreqBand (Hz)	ave	660	920	930	920	850
meanMPD (mm)	-	0.69	0.58	0.95	0.85	0.75
StdMPD (mm)	-	0.13	0.12	0.17	0.13	0.18
maxMPD (mm) (@1% max)	max	1.4	1.1	1.7	1.4	1.5
CtMPD	ave	51	72	49	54	57
CrestF_MPD (dB)	ave	4.3	4.4	4.4	4.3	4.3
	std	0.71	0.61	0.68	0.66	0.67
Pk2pkMPD (mm) (@1% max)	max	0.7	0.75	0.7	0.85	0.93

Figure 6.16 shows the contours of the statistic centroids, corresponding to the center of mass of each pavement type, to appraise the correlation between the different characteristics of the pavements.

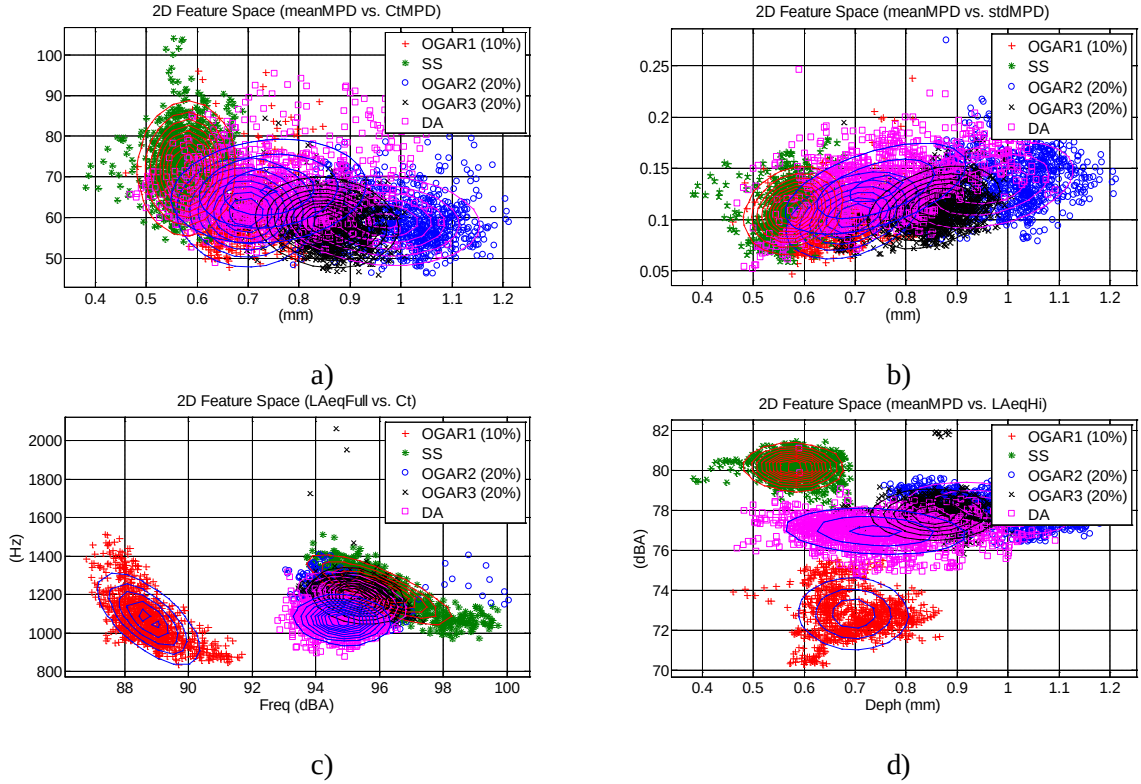


Figure 6.16 – 2D plots of pairs of different features; a) meanMPD vs. CtMPD, b) meanMPD vs. StdMPD, c) LAeq vs. Ct and d) meanMPD vs. LAeqHi.

Some conclusions can be stated from the figures and tables above:

1- The OGAR1 pavement exhibits the lowest global noise emission (a difference of about 7dBA compared to the other pavements) and the lowest frequency for the maximum LAeq (feature maxFreqBand). However, the LAeq and Ct samples are considerably spread and with some level of dependency. This means that the spectral energy has a non-stationary behavior (it changes very often along the track) and they depend inversely with the Centroid feature (for an increase of the noise levels a decrease of the centroid is observed, and vice versa). This characteristic was already found for different pavements [Freitas08b] and [Paulo08];
2- The OGAR2, OGAR3, SS and DA pavements generate almost the same global noise. However, a number of differences are observed: (i) although some differences are observed on the meanMPD (specially for the OGAR2, OGAR3 and DA) little difference is noticed on the global noise levels, (ii) the OGAR2 and DA pavements show the highest deviation of the meanMPD, though little correlation with the noise levels is observed;
3- Some features, such as the StdMPD, maxMPD, CrestFMPD and maxCorrMPD, have little relevance to the correlation because of their similarity. This finding is related to the

fact that the MPD is itself a mean parameter, estimated by averaging methods, thus giving less significance to the features estimated from the maximum values.

6.5.2 Analysis of the results

The vector of features extracted was applied to a statistical classifier as a first approach to correlate texture and noise data. The density functions of each feature were analyzed in order to verify in advance the level of success on the separation of the pavement surfaces. With this feature selection procedure, the features maxMPD, CrestFMPD, pk2pkMPD and maxCorrMPD were excluded because they were not completely independent from each other. Moreover, little difference was observed on the density functions for the different road pavements. This permits reducing the dimension of the feature vector, increasing the classification process without reducing the accuracy of the results.

The classifier system is trained by using the MPD and the sound signal for each type of pavement.

Again, by observation of the distribution function of the majority of features, a normal distribution of the selected features was assumed.

6.5.3 Experimental results

In order to provide some insight into the classifier obtained, the corresponding decision boundaries have been plotted. Ten features were used, after the feature selection procedure. Since it is impossible to graphically represent the *N*-dimensional decision boundaries, 2D projections of these boundaries for the pair of features LAeqFull – meanMPD, which are the most important relating noise and texture, are depicted in Figure 6.17.

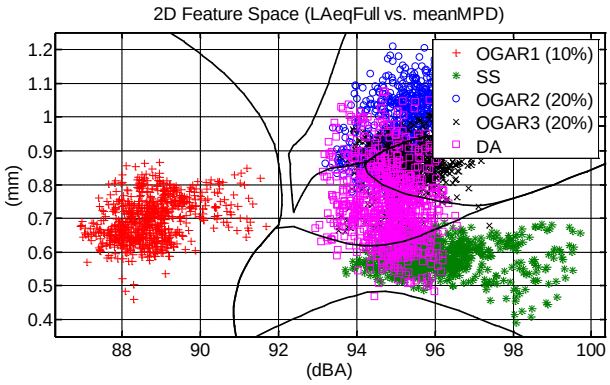


Figure 6.17 – Feature space in 2D showing for LAeq and meanMPD with indication of decision boundaries of the classifier.

The figure reveals that for these features some classes of pavement types are similar. However, if the remaining features are used the classes can be practically separated. Each sub-feature space provides information on the best separation between the classes using the statistical classifier. Figure 6.18 depicts the spectra for the different surface types.

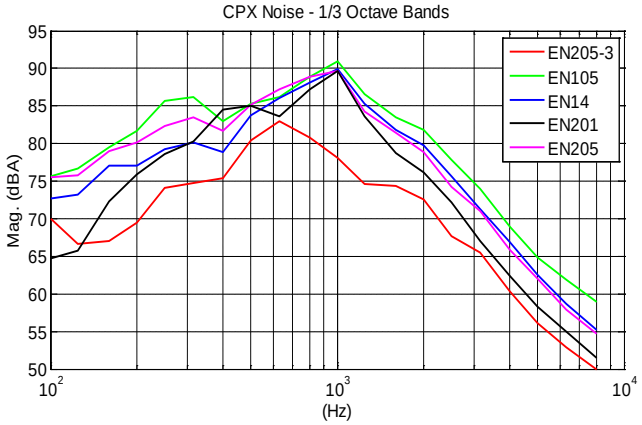
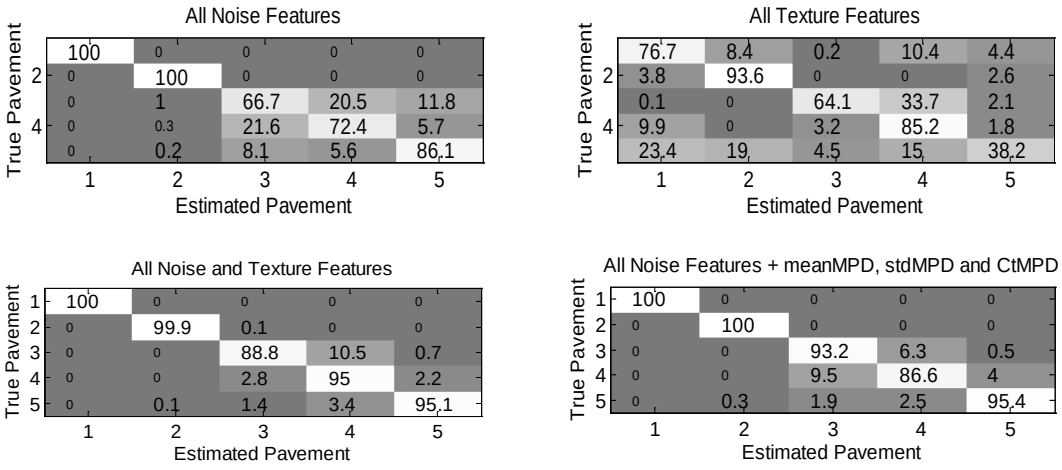


Figure 6.18 – Pavement spectra for the different surface types.

The result of the classification is assessed by means of the related confusion matrix, as shown in Table 6.6. This matrix yields the error of the classifier.

Table 6.6 – Confusion matrices using the Bayesian classifier. Topmost left – noise features only; Topmost right – texture features only; downmost left – all the features suggested and downmost right – all noise plus the selected texture features.



As a trail for evaluating the importance of the noise and textures related features on the classification, several tests were carried out. The first attempt consisted of the classification with only noise features. The result showed that, for only to the OGAR1 and SS pavements, the

classification error is not significant, the remaining pavements are identified with a degree of confidence ranging from about 60% to 80%. The test with texture features only proves to be unsatisfactory for the identification of the correct road surface, however, for the OGAR3 (pavement 4 in Table 6.6) the results are better than using noise features only. In general, the noise features are more robust, showing more statistical independence, than texture features. The main reason for this finding is because only features based on macrotexture are used, which are insufficient to characterize the road texture completely.

The use of all the features leads to a good accuracy with the true pavement, except for the OGAR2 pavement. A remarkable finding is the fact that the same results are achieved with only three textures features (statement already mentioned before), with the advantage of reducing the computational load. Obviously, the increase of the amount of different road pavements types could alter this statement.

6.6 Case study III – Noise texture -based Bayesian–Neural Network approach

This study used the second database of road pavements (db2) with the Bayesian and the Neural Network classifiers for comparison purposes [Paulo10c].

The sound signal was acquired along the road segments for the vehicle speeds in the range of 30 to 110 km/h (in most tracks), in steps of 10km/h, using the Close Proximity procedure (CPX).

Figure 6.19 depicts the ranking of the features (by backward searching method and based on the inter-intra criterion) originally introduced in the texture and noise correlation study [Paulo08] and [Paulo10a], namely, 1-*LAeqFull* (overall sound pressure levels), 2-*LAeqLo* (low freq. band), 3-*LAeqMid* (middle freq. band), 4-*LAeqHi* (high freq. band), 5-*Ct* (centroid), 6-*CtLo* (centroid low freq. band), 7-*CtMid* (centroid middle freq. band), 8-*CtHi* (centroid high freq. band), 9-*FreqMaxSPL* (frequency band of maximum SPL), 10-*maxLAeqBand* (max. SPL band), 11-*max2FreqBand* (secondary maximum SPL band), 12-*meanMPD* (Mean of MPD), 13-*stdMPD* (Standard Deviation of MPD), 14-*CtMPD* (centroid MPD), 15-*CtLoMPD* (centroid MPD low freq. band), 16-*CtMidMPD* (centroid MPD middle frequency band), 17-*CtHiMPD* (centroid MPD high freq. band), 18-*maxAmpMPD4* (max. amplitude above mean of MPD), 19-*rmsmaxAmpMPD* (maxima rms of MPD), 20-*maxMPD* (maxima of MPD), 21-*CrestF_MPD*

(Crest Factor of MPD), 22-*pk2pkMPD* (Peak-to-peak MPD), 23-*ZCrossMPD* (zero crossings of MPD), 24-*meanCorrMPD* (Mean Correlation MPD).

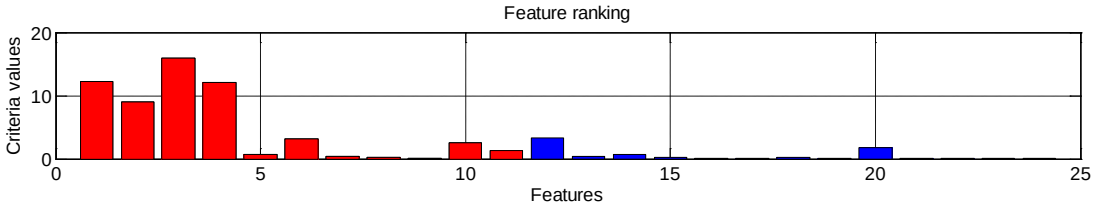


Figure 6.19 – Rank of the noise (red bars) and texture (blue bars) features for classification proposals.

The features are sorted by descending order of importance as {3 1 4 2 12 6 10 20 11 14 5 13 7 15 8 18 24 22 9 19 23 17 21 16}.

Figure 6.19 shows that the features extracted from noise have more significance, are more decorrelated, than the features from texture. Effectively, the MPD method does not address some surface texture characteristics such as the pavement aggregate particle shape, size, and distribution. The method is not meant to provide a complete assessment of pavement surface texture characteristics [ISO13473-1, Sayers98]. Despite these considerations, several characteristics, defining different pavements, can be selected.

In order to evaluate the characteristics of the MPD in a number of simulations of different textures, a simple model was developed to generate the longitudinal pavement profile. The model is able to generate two types of profiles: the first type is assigned to periodic profiles based on the selected road wavelength and the second type consists of generating the profile randomly using a table with the pavement grading curves which gives information about the grade percentage of the aggregates. Some examples of the type of aggregates considered are round shape, triangular shape, raised cosine shape and square shape. Figure 6.20 shows an example of a surface profile segment built from round shape aggregates, the most appropriate for our purposes.

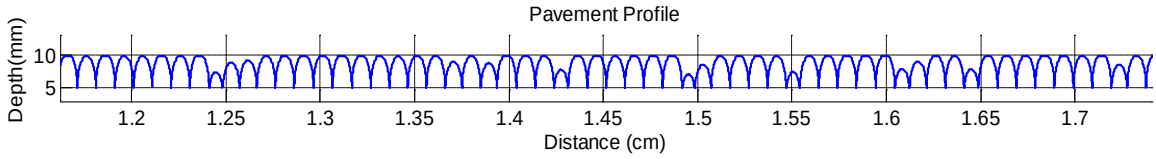


Figure 6.20 – Surface profile segment of round shape aggregates with maximum grade of 10mm.

Figure 6.21 depicts the MPD and mean MPD curves for different values of sliding window (sw) and baseline length (bl) parameters for a 4 m long road segment. Figure 6.22 shows a zoom for a better visualization.

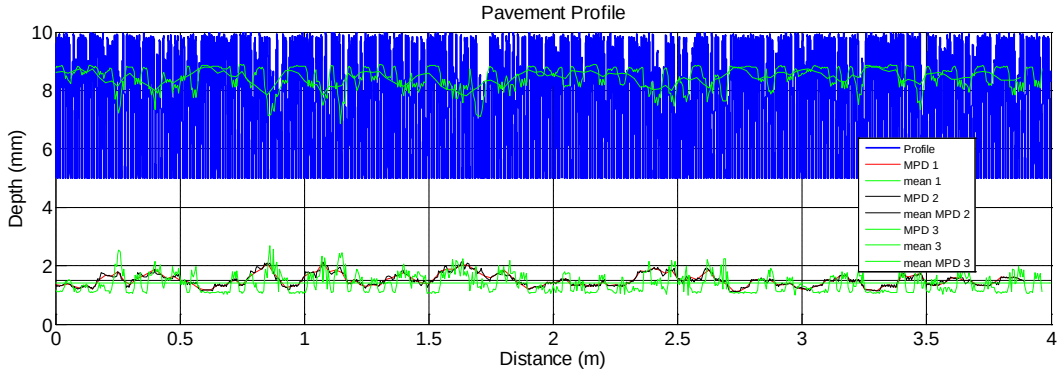


Figure 6.21 – Surface profile segment; MPD1 @ sw=2.5 cm and bl=10 cm (red line); MPD2 @ sw=1 cm and bl=10 cm (black line); MPD3 @ sw=1 cm and bl=4 cm (red line); mean MPD – global mean of MPD; the global mean of the profile is also shown.

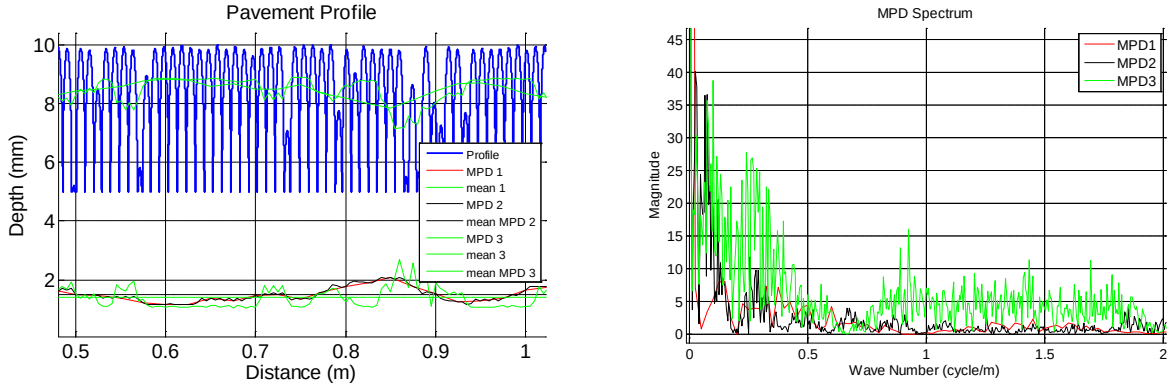


Figure 6.22 – Surface profile segment (zoom of Figure 6.21) and spectrum of MPD curves.

6.6.1 Selection of the relevant features

As expected, the MPD according to the standard, which refers to a baseline of 10 cm length (curve MPD1), is not able to reproduce the rapid variations of the profile (the MPD is intended for macrotexture only). A sliding window of 2.5 cm is an internal characteristic of the profilometer device used for the acquisition/estimation of the MPD. By decreasing the sliding window parameter, an insignificant improvement is obtained (MPD2 curve). This is due to the fact that half of the baseline is much longer than the maximum grade of the aggregates. Using a baseline length of 4 cm and a sliding window of 1 cm significantly the accuracy of the profile depth (MPD3 curve) can be improved. Notice the MPD evolution for the distance about 0.9 cm

(left part of Figure 6.22); the curve MPD3 detects much better the variation on the profile than MPD1. The MPD spectrum reinforces this statement. Notice the increase of resolution for the MPD3 spectrum curve, for a low wave number.

Since all parameters extracted are mean values, the best noise and texture features (selected from a subset of the first 12 features of the rank, given by Figure 6.19) to be analyzed in this study are then: Overall sound pressure levels – LAeqFull (in dBA), Narrow band sound pressure levels, Spectral energy centroid – Ct, Spectral band energy centroid – CtLo, CtMid and CtHi, Band spectral energy centroidThird octave band sound pressure levels - Oct13B, Spectral Energy Maximum – FreqMaxLAeq, Mean of MPD, meanMPD (in mm) Standard Deviation of MPD, stdMPD (in mm), Centroid MPD, CtMPD, Maxima of MPD, maxMPD (in mm).

6.6.2 Evolution of features with vehicle velocity

The complete analysis of the features should comprise the dependence of each feature with the speed of the rolling vehicle. Therefore, the characterization/identification of a pavement could be carried out for a specific velocity only. Although for a complete analysis three velocities should be used, as recommended in ISO CD 11819-2, in some circumstances, such procedure is difficult to implement, due to the logistics associated to the measurements in situ, for instance. Figure 6.23 and Figure 6.24 show a set of features related to the speed dependence.

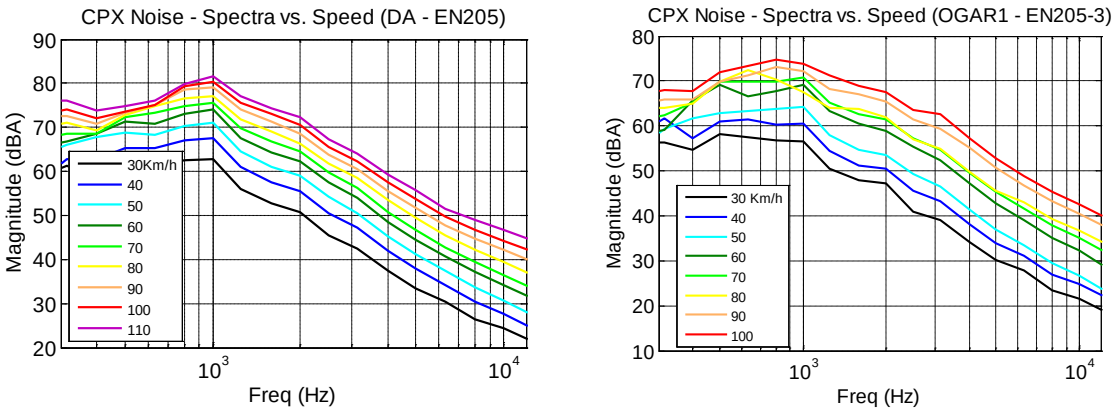


Figure 6.23 – Global noise spectra for different vehicle speed.

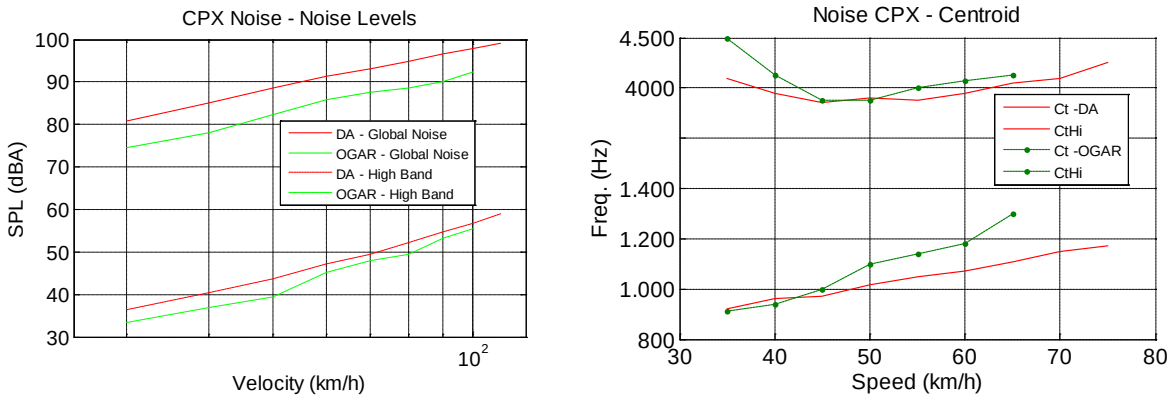


Figure 6.24 – Global and high frequency band noise for different vehicle speed.

As shown in Figure 6.23 and Figure 6.24, although the global noise levels (strongly frequency dependent for high noise levels) follows a linear fashion with the decimal logarithm of the speed, this is not true for high frequencies, which is better approximated by a linear law with speed. Furthermore, as observed, an increase of the speed implies to an increase on *Ct* feature, following a non-linear law, which is dependent of the pavement type.

6.6.3 Feature distribution

The density function associated to each feature provides guidelines to choose the most convenient type of classifier. As suspected by observation of the contours depicted in Figure 6.25 and confirmed by Figure 6.26, the density functions of the features for OGAR1 pavement do not follow a normal distribution for some types of pavements.

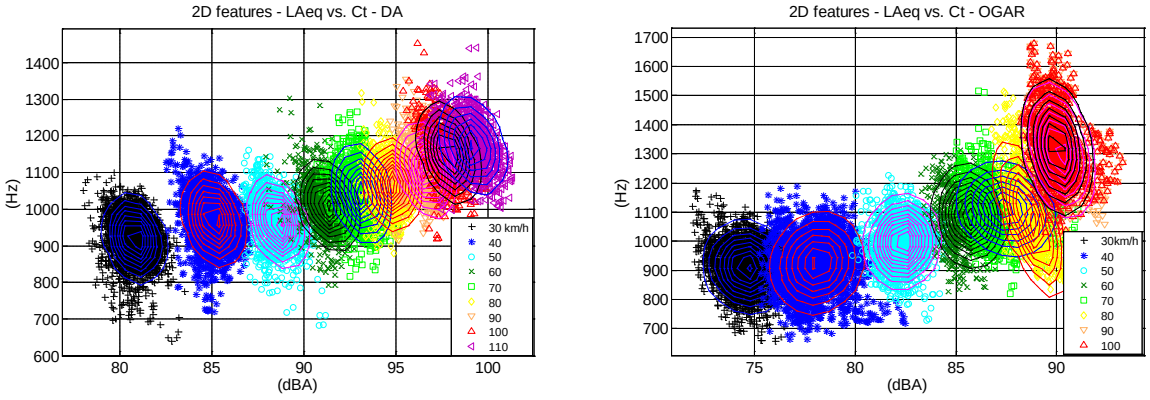


Figure 6.25 – 2D feature space of LAeq vs. Ct for different vehicle speed and for the pavements dense asphalt (DA), left, and Open Graded Asphalt Rubber (OGAR1), right. The contours give a measure of the density of the observations on the feature space.

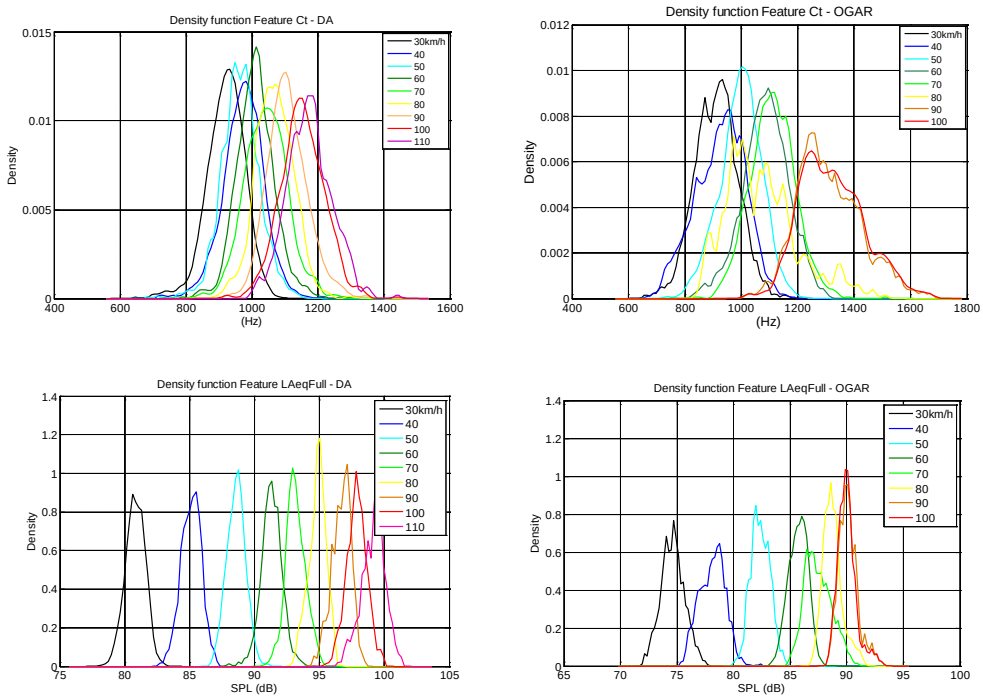


Figure 6.26 – Density functions of feature Ct for different vehicle speed.

Therefore, Bayesian and neural networks classifiers were tested for comparison. One first advantage of using Neural Networks against Bayesian classifiers is that the assumption of normal distribution for the selected features is not needed.

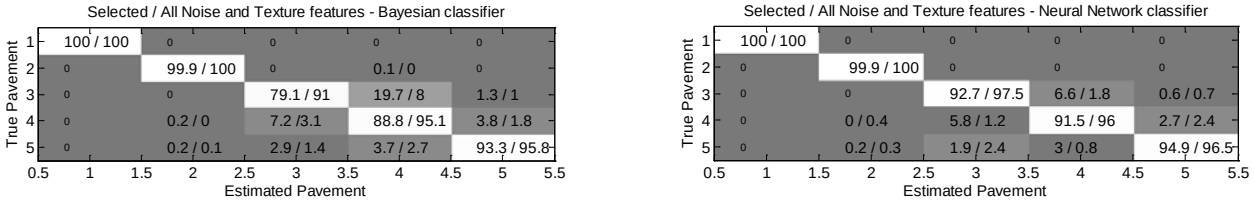
6.6.4 Experimental results

The same procedure as for the Case study II was applied to the features.

The result of the classification is assessed by means of the related confusion matrix, as sketched in Table **6.7**. This matrix yields the error made by the classifier. Each non zero cell holds the result for the selected features (first 12 features given by the ranking method) and for all the set of features (24 features). The values in each cell give the classification result in percentage.

The performance of the Bayes classifier was analysed against the Neural Network classifier, by using two different sets of features; (i) selected features and (ii) the whole set of features. The results showed that for both classifiers the OGAR1 and SS pavements are properly identified.

Table 6.7 – Confusion matrices using the Bayesian (left) and Neural Network (right) classifiers.



With the Bayes classifier, the OGAR2 pavement is barely identified with the set of 12 selected features, about 79% classified on the true pavement, increasing to 91% with all features. With all the features (and excepting the OGAR2) the true pavements are identified with a degree of confidence above 95%. With the Neural Network, a degree of confidence above 90% is observed for the 12 selected features, for all pavements, and above 96% using all the features. The OGAR2 pavement (pavement 3) is well identified with this classifier.

6.7 Discussion

A new method for the classification and identification of different types of road pavements, using statistical classification techniques, comprising an array of procedures to acquire and pre-process road sound sinals near a test tire and macrotexture information provided by a profilometer device, was developed and implemented.

Since the assumption of Gaussian distribution of the features extracted from the texture and acoustic measurements is not always fulfilled, Bayesian and Neural Network classifiers were considered.

The results of road surface classification showed in general a good agreement with the true road pavements for the two classifiers assessed, above 90% for the Bayesian and above 95% for the Neural Network. However, with the feature selection method, reducing the number of features used on the classification, the accuracy of the Neural Network classifier is significantly better than the Bayesian classifier. This is due to the fact that the assumption of features with normal distribution is not correct for some types of road pavements. Nevertheless, the time for the training step is much larger and a great care must be taken in the pre-processing step of the data used to define the features in order to achieve the correct neural network parameters. Moreover, with the use of smoothing techniques applied to the features an improvement of the results are expected.

The method can be used for assessing road pavements in terms of age and degree of erosion. The procedure offers a good and accurate characterization of the road surface noise emission and provides valuable data both for low noise pavement designers and for road engineers, builders and contractors. This method can also be used for conformity tests on new road pavements, by comparing it with accurate data from a reference road surface. A further application of this road pavement classification system is the improvement of the accuracy of road pavement data for road traffic noise modeling and mapping purposes.

Due to the discrepancy of measurement data of some types of pavements, care must be taken with the pre-processing of the data, excluding, as much as possible, discrepancies related with problems of the construction/degrading process of the roads (unless the goal of the study is conformity road assessment).

Since the speed of the vehicle is related to the resulting noise levels, the correct surface classification should be carried out with measurements at several speeds and at constant velocity.

In order to evaluate the dependence of the vehicle speed on the feature set, line fitting methods were applied to some 2D feature space. It was found that the increase of noise levels for high frequencies do not follow a linear behavior with the decimal logarithm of speed, which occurs for the global noise levels. The same result is observed for the Ct feature. This type of conclusion is useful for the estimation of features for a specific speed other than the reference speed.

The features based on the MPD are insufficient to describe the texture accurately as shown from the model for simulating the road profile developed in this study.

As a final remark, the correct correlation between noise and texture characteristics should take into account not only the level of macrotexture but also the micro and megatexture in order to provide more information to correlate texture and noise conveniently, thus, achieving better classification results.

Chapter 7

CONCLUSIONS

This Chapter presents the main conclusions of the work developed within the scope of this thesis. A brief review of the main points discussed in Chapters IV to VI, the contribution of the thesis to the state-of-art, is presented and the practical applications implemented for evaluation of the techniques developed for system impulse response measurement are summarized.

Finally, some weak points of the current approaches are identified and presented, and proposals for further improvement of the methods and applications are suggested.

7.1 Contributions and general conclusions

Acoustic measurement techniques especially designed to improve the accuracy of the measured system impulse response were investigated and tested. The methods were developed in order to reduce the well identified weaknesses of the standard measurement techniques used in specific but common applications.

The first class of methods, the Segmented Weighted Average technique, SWAve, and the Spectral Segmented Weighted Average technique, SSWAve, are aimed at increasing the overall SNR, considering the non-stationary characteristics of the background noise present in the measurements. The second class of methods, the Ultrasound probe test method, attempts to

reduce the influence of time variance effects on the estimation of the system impulse response by using a number of features and thresholds estimated from the analysis of the ultrasound probe signals. Finally, the third class of methods, the Test Signal Plus Noise with Modification, TSNM, and the Segmented Test Signal Plus Noise with Modification, STSNM, use the operation mechanisms of the human auditory system, emitting music to perceptually mask the test signals, thus making the acoustic measurements possible with the presence of an audience causing a minimal nuisance to the people.

Segmented Weighted Average technique, SWAve, and Spectral Segmented Weighted Average technique, SSWAve

These techniques take the non-stationary characteristics of the background noise into account in order to increase the overall SNR. The music signals are used as background noise.

The standard average method (Ave) was used in tests for comparison purposes in order to evaluate the resulting relative gains. The Energy Decay Ratio (EDR), Energy Decay Ratio Distribution (EDRD), Power Spectral Density Distribution (PSDRD), Peak-to-noise ratio (PNR) and Valid Decay analyses were examined to compare the performance of the swept sine against the MLS methods. The EDR and EDRD analysis showed increments of up to 7dB for the Ave and of 9dB for the SSWAve method. A remarkable gain of about 13dB between Ave and SSWAve was observed for the same SNR of the captured signal and for the same number of averages (NAve). The PSDRD measure revealed that the MLS technique always shows an excess of spectral noise higher than the swept sine. Also, much lower frequency noise is observed, especially for low SNR and/or low NAve. The Backgnd noise analysis showed, in most cases, a significant improvement of about 10dB for the SSWAve against the standard averaging method. The Valid Decay for full band leads to an interesting conclusion that for SNR up to -18dB the threshold of 10dB is only reached when the SSWAve method is used together with the MLS. Instead, in the case of the swept sine technique, this threshold is reached for SNR of -30dB, with SSWAve or for -18dB using the standard averaging method. Therefore, the swept sine technique can lead to the same results with considerably lower SNR or, in another perspective, a lower number of averages is needed to obtain the same SNR.

The nuisance caused by the acoustic measurements performed in the presence of an audience can be minimized by masking the test signal with music (thus lowering the SNR) and due to

time-variance effects the number of the test frames should be reduced Therefore, the swept sine technique prove to be more suitable for critical conditions than the MLS.

As a final conclusion, for environments with very low SNR, the swept sine technique proved always to estimate the acoustical parameters better than the MLS technique.

Ultrasound transducer arrays to account for time-variance

The performance of this method was assessed by the improvement of the SNR and T20.

The results for the simulated analysis using the intra-periodic model (by applying a sinusoidal stretching factor) show the superiority of the swept sine against the MLS technique. In fact, the swept sine is almost unaffected by this type of distortion. Instead, a decrease of about 6dB for the PNR results is observed for the MLS. Nevertheless, a significant distortion is observed at the high end of the frequency curve of the IR for both techniques. Almost the same results are observed for the case of the inter-periodic model proving again the robustness of the swept sine over the MLS technique.

The results of the ultrasound tone probe approach show significant drifts from the steady-state values (room air medium not disturbed). The amplitude and phase shift variation parameters achieve 15dB and 130 degrees, respectively, mostly due to temperature gradients. The air mass movement with the fan switched-on produced less deviation from the steady-state conditions. However, a kind of modulation noise was noticed in the upper and lower side bands of the ultrasonic probe test signal spectrum.

The experimental results using the thresholds and sorting procedure were tested for the standard and the ultrasonic approaches.

The similarity of the results obtained for the standard approach and for the ultrasonic method confirms the advantage of the sorting procedure applied to the captured frames. In fact, using up to 8Nave, a considerable increase of about 3 and 5dB in the Valid decay and the PNR analysis, respectively, is observed. Moreover, the distortion on the IR and consequently on the frequency response of the system is reduced by applying this procedure for the sets of averages up to 8 frames, 4Ave and 8Ave.

As expected, the swept sine technique is much less affected by time variance than the MLS procedure. Furthermore, if the frame length is larger than the estimated reverberation time, then

the SSWAve can be used to label also the frequency bands less affected by applying a higher weighting factor.

Test signal plus noise with modification, TSNM, and Segmented test signal plus noise with modification, STSNM

Relevant responses based on the RIR, namely the Energy Decay, the Energy Decay Error and the Amplitude Frequency response, were investigated. When music is the disturbing component, the method showed a gain on the PNR of more than 10 dB for the standard swept sine technique.

Although these two procedures can be considered as independent techniques, in some circumstances it may be advantageous to use them together. For music tracks where long term stationary tonal sounds cover the entire audio spectrum of interest, applying the STSNM method has benefits over the TSNM. However, if this constrain is not met, some parts of the spectrum could be covered by using the TSNM method in conjunction with the STSNM. The trade-off is a decrease on the masking perception sensation.

The results of a survey on a group of people revealed that with this procedure the music effectively masks the test signal, making acoustic tests with the presence of people quite possible.

Statistical classification of road pavements

The results of road surface classification provided by the method, using Bayesian classifiers, showed a good agreement with the actual road pavements. The error corresponding to a mismatch classification, i.e. classifying an unknown road pavement in a wrong pavement type, is found to be in the range of 1 to 33% (using the original features) and falls to less than 8% with the use of smoothing techniques applied to the original features, which can be considered quite good given the similarity of some road surface characteristics.

To complement the previous road classification method, a new set of features related with pavement texture was added and the Bayesian analysis was complemented by a neural network classifier in order to improve the classification results. The results showed in general a good agreement with the true road pavements for the two classifiers assessed, above 90% for the Bayesian and above 95% for the Neural Network. However, with the feature selection method the accuracy of the Neural Network classifier is significantly better than the Bayesian classifier.

This is due to the fact that the assumption of normal distribution features is inaccurate for some types of road pavements. Nevertheless, the training step time is much larger and care must be taken in the data pre-processing stage to define the features in order to achieve the correct neural network values of the nodes and the bias.

7.2 Discussion (applicability and limitations) of the new concepts

A number of comments emphasizing the practical applications and weaknesses of each method developed in this thesis are presented next.

Segmented Weighted Average technique, SWAve, and Spectral Segmented Weighted Average technique, SSWAve

These techniques proved to be better than the standard averaging method for all situations considered. However, their best performance is achieved when the background noise shows a strong non-stationary behavior and high sound pressure levels.

Theoretically, there is no processing gain for the cases of stationary noise. However, with the STSNM approach, this statement is only achieved if perfect reconstruction properties are considered on the implementation of the filter bank. Nevertheless, this constraint compromises the selectivity of the amplitude frequency response of the filters.

Otherwise, the filter bank implementation used in the STSNM has some restrictive issues to attend to due to problems of phase distortion introduced in the cutoff frequencies of each band-pass filter.

This approach can be applied to the TSNM and STSNM methods in order to increase the SNR, assuming each homologue perceptual grain holds different energy.

A dedicated framework with a user-friendly interface (Matlab's GUI) was developed to test and demonstrate these techniques as well as the TSNM and STSNM. However, considerable amounts of calculation time inherent to the Matlab programming language are required. This, in fact, drastically limits the use of this measurement platform in the STSNM for total test duration longer than 60s, test signal duration x Nave.

Ultrasound transducer arrays to account for time-variance

The method developed in this thesis represents a step forward in the field. In fact, the use of ultrasonic sensor arrays for monitoring the environment parameters variation and consequently

the time variance effects in situations of low SNR in acoustic measurements is a novel approach.

The use of the ultrasound loudspeaker array permits an insight into implications on the sound wave propagation due to the time variance leading to substantially better results than the standard averaging techniques. In fact, for a regular measurement up to 4 dB of the SNR gain is expected in some octave bands. With a larger number of features and with the application of statistical learning methods these results might be improved.

Test signal plus noise with modification, TSNM, and Segmented test signal plus noise with modification, STSNM

This approach is highly advantageous over standard procedures since it allows acoustic measurements to be made in the presence of people assuming the test signals are not perceived by the audience. Additionally, measurements for checking the compliance with building acoustics regulations or other legislation could follow this approach in order to minimize any possible annoyance, especially if high sound pressure levels are employed.

Another important application of this technique lies on the evaluation of the system response of broadcast systems. The method allows the testing of the broadcast chain, partially or globally, using the radio program itself over a regular broadcast emission.

The procedure associated with the production of the music plus test composite signal to be sent to the system is performed offline. Therefore, the applicability of this approach, at this stage of research, is restricted to situations of music reproduction only, such as public address or broadcast systems and in the breaks of live performances.

Statistical classification of road pavements

This method can be used for assessing road pavements in terms of age and degree of erosion. The classified road section pavements can be plotted on a geographical map, showing the whole road network properties.

The procedure offers a complete and accurate characterization of the road surface noise emission and provides valuable data for silent pavement designers, road engineers, builders, and contractors. This method can also be used for conformity tests on new road pavements, by comparing their characteristics with accurate data from reference road surfaces. An important application of this road pavement classification system is the improvement of the accuracy of

road pavement data for road traffic noise modeling and mapping purposes. This approach can contribute to the increase of the accuracy of road noise prediction models, since they consider the actual type of pavement of each road section, by assigning corrections according to their noise emission characteristics.

The methods developed in this thesis already led to a national patent [Paulo09P].

7.3 Future work

Particular suggestions and comments on the different new techniques are outlined next.

Ultrasound transducer arrays to account for time-variance

Although a number of features were considered in the present work for the evaluation of the method, not all of them were used. The use of the remaining features will be dealt with in future phases of this research.

In the case of mid to large-size volume auditoriums different loudspeaker ultrasonic array arrangement, using multiple-ultrasonic devices, should be applied in order to cover a wider audience area and room volume by establishing a multi-path direction sound propagation. Notice a single loudspeaker ultrasonic array shows a very directive radiation pattern, though only 16dB of attenuation is measured between the main and the secondary lobe. To overcome this limitation, four new loudspeaker ultrasonic arrays with improved radiation pattern are currently being developed and tested in order to cover different wave propagation paths.

Also, statistical classification procedures will be applied to the frame segments of the test signal on the labeling process.

Test signal plus noise with modification, TSNM, and Segmented test signal plus noise with modification, STSNM

The eigenvalues of the room under test (the spatial room resonances) could play an important role in the perception of the masked test signals by the people in a venue depending of the acoustical and geometrical characteristics of the enclosure. Actually, a specific perceptual grain masked by a particular tonal component in laboratory conditions could be perceived in a venue in different seats if, for instance, a room resonance is located exactly in the frequency band covered by a particular perceptual grain (thus amplifying this part of the spectrum) leaving the tonal masker out. Nevertheless, one believes that the critical situations occur for the cases of

enclosures with very reflective boundary surfaces and with regular geometries, which are not the most usual situations.

Even so, this issue should be considered in the next stage of this research by testing in detail a larger number of auditoriums and enclosures.

Real time possessing avoids the limitation of handing the music material (background noise as far as the measurement is concerned) offline, giving much more freedom for the application of the techniques in live performances

Statistical classification of road pavements

The correct correlation between noise and texture characteristics should take into account not only the macrotexture but also the micro and the megatexture in order to provide more information, thus achieving better classification results. Also, an additional set of features based on the sonic sensation produced by each type of road pavement, using perceptual models, should be considered.

Due to the discrepancies of measurement data of some types of pavements, care must be taken in the pre-processing of the data, excluding, as much as possible, discrepancies related with the road construction (unless the goal of the study is conformity road assessment).

As an extension of this research, each pavement should be tested for different stages of its age and traffic density. Therefore, a more complete characterization of the pavement could be available to provide guidelines for the pavements builders.

A comprehensive framework to study and compare different approaches and algorithms is currently being developed for each module considered in this thesis featuring real time characteristics, or at least reducing drastically the processing time, and the possibility of working as standalone devices. In fact, limitations concerning the processing time will be overcome with the migration of the Matlab code to a faster programming language, such as C++ or Java.

REFERENCES

General bibliography

[Agerkvist93] F. T. Agerkvist, F. Jacobsen, "Sound power determination in reverberation rooms at low frequencies", *Journal of Sound and Vibration*, vol. 166, pp. 179-190, 1993.

[Argenti98] F. Argenti, B. Brogelli, E. D. Re, "Design of pseudo-QMF banks with rational sampling factors using several prototype filters", *IEEE Transactions on Signal Processing*, vol. 46, no. 6, pp. 1709-1715, 1998.

[Andreeva04] T. A. Andreeva, W. W. Durgin, "Experimental investigation of ultrasound propagation in turbulent, diffractive media", *Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1532-1536, 2004.

[ATSC95] ATSC, United States Advanced Television Systems Committee Digital Audio Compression (AC-3) Standard, Doc. A/52/10, 1995.

[Attenborough80] K. Attenborough, S. I. Hayek, J. M. Lawther, "Propagation of sound above a porous half-space", *Journal of the Acoustical Society of America*, vol. 68, no. 5, pp. 1493-1501, 1980.

[Barbagallo08] M. Barbagallo, M. Kleiner, "Modulation and Demodulation of Steerable Ultrasound Beams for Audio Transmission and Rendering", *11th International Conference on Digital Audio Effects - DAFx-08*, pp. 227-232, Espoo, Finland, September, 2008.

[Barron73] M. Barron, "Growth and decay of sound intensity in rooms according to some formulae of geometric acoustics theory", *Journal of Sound and Vibration*, vol. 27, pp. 183-196, 1973.

[Barron88] M. Barron, L. J. Lee, "Energy Relations in Concert Auditorium", *Journal of the Acoustical Society of America*, vol. 84, pp. 618-628, 1988.

[Baumgarte95] F. Baumgarte, C. Ferekidis, H. Fuchs, "A Nonlinear Psychoacoustic Model Applied to ISO/MPEG Layer 3 Coder", *99th International Convention of the Audio Engineering Society*, Toronto, Canada, 1995.

[Beaton96] R. J. Beaton, J. G. Beerends, M. Keyhl, W. Treurniet, "Objective Perceptual Measurement of Audio Quality", in Collected Papers on Digital Audio Bit-Rate Reduction (N. Gilchrist and C. Grewin, eds.), *Audio Engineering Society*, pp. 126–152, 1996.

[Beauchamp69] J. W. Beauchamp, "A Computer System for Time-Variant Harmonic Analysis and Synthesis of Musical Tones", *Music by Computers*, H. F. von Forester and J. W. Beauchamp, eds., J. Wiley, New York, pp. 19–62, 1969.

[Békésy60] G. von Békésy, "Experiments in Hearing", *McGraw-Hill*, New York, 1960.

[Bello04a] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. E. Davies, M. B. Sandler, "A Tutorial on Onset Detection in Music Signals", *IEEE Transactions On Speech And Audio Processing*, vol. 13, no. 5, pp. 1035 - 1047, 2005.

[Bello04b] J.P. Bello, C. Duxbury, M.E. Davies, M.B. Sandler, "On the use of Phase and Energy for Musical Onset Detection in the Complex Domain", *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 553-556, 2004.

[Coelho07] J. Luis Bento Coelho, "Community Noise Ordinances", *Handbook of Noise and Vibration Control*, Ed. Malcolm J. Crocker, John Wiley & Sons Inc., New Jersey, Chapter 130, pp. 1525-1532, 2007.

[Beranek88] L. Beranek, "Acoustical Measurements" *Acoustical Society of America*, 1988.

[Berengier97] M. C. Berengier, M. R. Stinson, G. A. Daigle, J. F. Hamet, "Porous road pavements: Acoustical characterization and propagation effects", *Journal of the Acoustical Society of America,* vol. 101, no. 1, pp. 155-162, 1997.

[Blesser01] B. Blesser, "An Interdisciplinary Synthesis of Reverberation Viewpoints", *Journal of the Audio Engineering Society*, vol. 49, no. 10, 2001.

[Bodiund77] K. Bodiund, "A normal mode analysis of the sound power injection in reverberation chambers at low frequencies and the effects of some response averaging methods", *Journal of Sound and Vibration*, vol. 55, pp. 563-590, 1977.

[Borish83] J. Borish, J. B. Angell, "An Efficient Algorithm for Measuring the Impulse Response Using Pseudorandom Noise", *Journal of the Audio Engineering Society*, vol. 31, pp. 478-488, 1983.

[Borish85] J. Borish, "An Efficient Algorithm for Generating Colored Noise Using Pseudorandom Sequence", *Journal of the Audio Engineering Society*, vol. 33 no. 3, pp. 141-144, 1985.

[Bradley86] J. S. Bradley, "Speech Intelligibility Studies in Classrooms", *Journal of the Acoustical Society of America*, vol. 80, pp. 846-854, 1986.

[Bradley96] J. S. Bradley, "Optimizing the Decay Range in Room Acoustics Measurements Using Maximum Length Sequence Techniques", *Journal of the Audio Engineering Society*, vol. 44, pp. 266-273, 1996.

[Brown92] J. C. Brown, "Musical fundamental frequency tracking using a pattern recognition method", *Journal of the Acoustical Society of America*, vol. 92 no. 2, pp.1394-1402, 1993.

[Brown94] G. J. Brown, M. Cooke, "Computational auditory scene analysis", Computer Speech and Language, vol. 8, pp. 297-336, 1994.

[Brown99] Brown, J.C. "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features", *Journal of the Acoustical Society of America*, vol. 105, pp. 1933-1941, 1999.

[Brown01] J. C. Brown, O. Houix, S. McAdams, "Feature dependance in the automatic identification of musical woodwind instruments", *Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1064–1072, 2001.

[Bryan10a] N. J. Bryan, J. S. Abel. "Methods for Extending Room Impulse Responses Beyond Their Noise Floor", *129th Audio Engineering Society Convention*, San Francisco, CA, November, 2010.

[Bryan10b] N. J. Bryan, J. S. Abel. “Impulse Response Measurements in the Presence of Clock Drift”, *129th Audio Engineering Society Convention*, San Francisco, CA, November, 2010.

[Chu78] W. T. Chu, “Comparison of Reverberation Measurements Using Schroeder's Impulse Method and Decay-Curve Averaging Method”, *Journal of the Acoustical Society of America*, vol. 63, pp. 1444-1450, 1978.

[Chu95] W. T. Chu, “Time-variance effect on the application of the m-sequence correlation method for room acoustical measurements”, *15th International Congress on Acoustics*, Trondheim, Norway, vol. 4, pp. 25-28, 1995.

[Ciric02] D. Ciric, Milosevic, M., “Inaccuracies in Sound Pressure Level Determination from Room Impulse Response”, *Journal of the Acoustical Society of America*, vol. 111, no. 1, Pt. 1, pp. 210-216, 2002.

[Cox86] R. V. Cox, “The design of uniformly and nonuniformly spaced pseudo quadrature mirror filters”, *IEEE Transactions on Acoustics, Speech, And Signal Processing*, vol. ASSP-24, pp. 1090-1096, 1986.

[Cox01] T. J. Cox, F. Li, P. Darlington, “Extracting room reverberation time from speech using artificial neural networks”, *Journal of the Audio Engineering Society*, vol. 49, no. 4, pp. 219–230, 2001.

[Davies66] W. D. T. Davies, “Generation and Properties of Maximum-Length Sequences”, *Control*, 1966.

[Davy03] M. Davy, S. J. Godsill, “Bayesian harmonic models for musical signal analysis”, *Bayesian Statistics*, vol. 7, 2003.

[Depalle97] P. Depalle, T. Hélie, “Extraction of Spectral Peak Parameters Using a Short-Time Fourier Transform Modeling and No Sidelobe Windows”, *Workshop on Applications of Signal Processing to Audio and Acoustics - IEEE ASSP*, New Paltz, NY , USA, October, 1997.

[Depalle93a] P. Depalle, G. Garcia, X. Rodet. Tracking of partials using hidden Markov models. *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP93*, Minneapolis, Minesota, April, 1993.

[Depalle93b] P. Depalle, G. Garcia, X. Rodet, “Tracking of partials for additive sound synthesis using Hidden Markov Models”, *IEEE International Conference on Acoustics, Speech, and Signal Processing* -93, Minneapolis, Minesota, April, 1993.

[Descornet06] G. Descornet, L. Goubert, “Noise Classification of Road Pavements, Task 1: technical background information 1, Draft Report 05”, *Directorate-General Environment, European Commission*, 2006.

[Dolson86] M. Dolson, “The Phase Vocoder: A Tutorial”, *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[Dubois07] C. Dubois, M. Davy, “Joint detection and tracking of time-varying harmonic components: a flexible Bayesian approach”, *IEEE Transactions on Audio and Speech Language*, vol.15 no. 4, pp.1283-1295, 2007.

[Duda73] R. O. Duda, P. E. Hart, "Pattern classification and scene analysis", *Wiley*, New York, 1973.

[Dudley39] H. Dudley, "Remaking speech", *Journal of the Acoustical Society of America*, vol. 11, pp. 169–177, 1939.

[Dunn93] C. Dunn, M. O. Hawksford, "Distortion immunity of mls-derived impulse response measurements", vol. 41, no. 5, pp. 314–335, 1993.

[Ellis96] D. W. Ellis, "Prediction-driven computational auditory scene analysis", PhD thesis, MIT, 1996.

[Lundeby95] A. Lundeby, T. E. Vigran, H. Bietz, M. Vorlander, "Uncertainties of Measurements in Room Acoustics", Acustica, vol. 81, pp. 344–355, 1995.

[Faiget98] L. Faiget, C. Legros, "Optimization of the Impulse Response Length: Application to Noisy and Highly Reverberant Rooms", *Journal of the Audio Engineering Society*, vol. 46, no. 9, pp. 741-749, September, 1998.

[Farina00] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-sine technique" *108th AES Convention, Preprint 5093*, Paris, France, February, 2000.

[Farina93] Farina, A., "An Example of Adding Spatial Impression to Recorded Music: Signal Convolution with Binaural Impulse Responses", *Conference of Acoustics and Recovery of Spaces for Music*, Ferrara, Italy, October, 1993.

[Fausti98] P. Fausti, A. Farina, R. Pompoli, "Measurements in opera houses: comparison between different techniques and equipment", *International Conference on Acoustics - ICA98*, pp. 26-30, Seattle, USA, June, 1998.

[Ferreira98] A. J. Ferreira, "Spectral Coding and Post-Processing of High Quality Audio", *Phd thesis, Faculdade de Engenharia da Universidade do Porto*, Portugal, 1998,

[Fletcher33] H. Fletcher, W. A. Munson, "Loudness, its definition, measurement and calculation", *Journal of the Acoustical Society of America*, vol. 5, pp. 82–108, 1933.

[Fletcher40] H. Fletcher, "Auditory Patterns", Revue of Modern Physics vol. 12, pp. 47–65, 1940.

[Foote97] J. Foote, Content-based Retrieval of Music and Audio. Multimedia Storage and Archiving Systems II, *Proceedings of SPIE*, 1997.

[Freedman67] M. D. Freedman, "Analysis of Musical Instrument Tones", *Journal of the Acoustical Society of America*, vol. 41, no. 4, pp. 793–806, 1967.

[Freedman68] M. D. Freedman, "A Method for Analyzing Musical Tones", *Journal of the Audio Engineering Society*, vol. 16, no. 4, pp. 419–425, 1968.

[Freitas08a] E. Freitas, J. Preto Paulo, J. L. Bento Coelho, P. Pereira, "Towards Noise Classification of Road Pavements", *3rd European Conference on Pavement and Asset Management*, Coimbra, Portugal, 2008.

[Freitas08b] E. Freitas, J. Preto Paulo, J. L. Bento Coelho, "Silent Surfaces: An Experience in Portugal", *6th Symposium on Pavement Surface Characteristics – SURF08*, in CD-ROM, Portoroz, Slovenia, 2008.

[Fulton08] Fulton, J., Hearing: A 21st, Century Paradigm. Victoria, BC. CA: Trafford, 2008.

[Gade94] A.C. Gade, *Sabine Centennial Symposium*, Cambridge, Mass., *Acoustical Society of America*, pp. 191, Woodbury, New York, 1994.

[Gold99] B. Gold, N. Morgan, "Speech and Audio Signal Processing: Processing and Perception of Speech and Music", John Wiley & Sons, 1999.

[Goodwin96] M. Goodwin, "Residual modeling in music analysis-synthesis", IEEE *International Conference on Acoustics, Speech, and Signal Processing – ICASSP96*, May, 1996.

[Grant01] D. Grant, G. Davidson, L. Fielder, "Subjective Evaluation of an Audio Distribution Coding System", *111th AES Convention*, pp. 5443, New York, November, 2001.

[Grey77] J. M. Grey, J. A. Moorer, "Perceptual evaluations of synthesized musical instrument tones", *Journal of the Acoustical Society of America*, vol. 62, no. 2, pp. 454–462, 1977.

[Grey78] J. M. Grey, J. W. Gordon, "Perceptual effects of spectral modifications on musical timbres", *Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1493–1500, 1978.

[Griesinger96] D. Griesinger, "Beyond MLS – Occupied Hall Measurements with FFT Techniques", *101st Audio Engineering Society Convention* – AES, vol.44, p. 1174, preprint 4403, December, 1996.

[Hartmann97] W. M. Hartmann, Signals, Sound, and Sensation. Springer Verlag, 1997.

[Hastie01] T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning", Springer-Verlag, New York, 2001.

[Hellman72] R. Hellman, "Asymmetry of Masking Between Noise and Tone", *Perception and Psychophysics*, vol.11, pp. 241-246, 1972.

[Hirata82] Hirata, Y., "A Method of Eliminating Noise in Power Responses", *Journal of Sound and Vibration*, vol. 82, no. 4, pp. 593–595, 1982.

[Hoang89] P. Q. Hoang, P. P. Vaidyanathan, "Non-uniform multirate filter banks: Theory and design", in Proc. Int. Symp. Circuits Syst., pp. 371-374, May 1989.

[Houtgast80] T. Houtgast, H. J. M. Steeneken, "Predicting Speech Intelligibility in Rooms from the Modulation Transfer Function. 1. General Room Acoustics", *Acustica*, vol. 46, pp. 60-72, 1980.

[Howard09] D. M. Howard, Jamie A.S. Angus, "Acoustics and Psychoacoustics", *Focal Press - Elsevier*, 4th edition, 2009.

[ISO/DIS 3382] ISO/DIS 3382. "Measurement of reverberation time of rooms with reference to other room acoustical parameters", Draft, 1996.

[ISO/IEC11172-3] ISO/IEC 11172-3 "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s, Part 3: Audio", 1992.

[ISO266] ISO 266 "Acoustics - Preferred frequencies", 1997.

[ISO13473-1] ISO 13473-1. “Characterization of pavement texture by use of surface profiles – Part 1: Determination of Mean Profile Depth”. International Organisation for Standardisation (ISO), Geneve, Switzerland, 1997.

[ISO3382-2] ISO 3382-2. “Acoustics — Measurement of room acoustic parameters – Part 2: Reverberation time in ordinary rooms”. International Organisation for Standardisation (ISO), Geneve, Switzerland, 2008.

[ISOCD11819-2] ISO CD 11819-2. “Acoustics – Method for Measuring the Influence of Road Surfaces on Traffic Noise – Part 1: The Close Proximity Method”, International Organisation for Standardisation (ISO), Geneve, Switzerland, 2000.

[ITU-R BS.1387] Recommendation ITU-R BS.1387, “Method for objective measurements of perceived audio quality”, 1998.

[Jacob89] K. D. Jacob, “Correlation of Speech Intelligibility Tests in Reverberant Rooms with Three Predictive Algorithms”, *Journal of the Audio Engineering Society*, vol. 37, pp. 1020-1030, December, 1989.

[Jakevičius08] L. Jakevičius, A. Demčenko, “Ultrasound attenuation dependence on air temperature in closed chambers”, ISSN 1392-2114, Ultragarsas - Ultrasound, vol. 63, no. 1, 2008.

[Jayant93] N. Jayant, et al., “Signal Compression Based on Models of Human Perception”, Proc. IEEE, pp. 1385-1422, October, 1993.

[Jayant84] N. Jayant, P. Noll, “Digital Coding of Waveforms Principles and Applications to Speech and Video”, *Englewood Cliffs: Prentice-Hall*, 1984.

[Jensen01] K. Jensen, “The timbre model”, Mosart, Barcelona, Spain, 2001.

[Johnston88] J. D. Johnston, “Transform coding of audio signals using perceptual noise criteria”, IEEE J. Selected Areas in Comm., vol. 6, pp. 314-323, February, 1988.

[Juha05] “Concert Hall Impulse Responses – Reference”, *Laboratory of Acoustics and Audio Signal Processing Helsinki University of Technology*, Pori, Finland, 2005.

[Kaneda95] Y. Kaneda, “A Study of Non-linear Effect on Acoustic Impulse Response Measurement”, *Journal of the Acoustical Society of Japan*, vol. 16, no. 3, pp. 193-195, 1995.

[Karjalainen85] M. Karjalainen, “A New Auditory Model for the Evaluation of Sound Quality of Audio Systems”, *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP85*, Tampa, 1985.

[Karjalainen02] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, V. Välimäki, “Estimation of Modal Decay Parameters from Noisy Response Measurements”, *Journal of the Audio Engineering Society*, vol. 50, no. 11, pp. 867-878, 2002.

[Kovacevic93] J. Kovacevic, M.Vetterli, “Perfect reconstruction filter banks with rational sampling factors”, *IEEE Transations of Signal Processing*, vol. 42, no. 6, pp. 2047-2066, June 1993.

[Krasner79] M. A. Krasner, "Digital Encoding of Speech and Audio Signals Based on the Perceptual Requirements of the Auditory System", *Technical Report 535, MIT*, Lincoln Laboratory, Lexington, 1979.

[Kropp07] W. Kropp, T. Kihlman, J. Forssén, L. Ivarsson, "Reduction Potential of Road Traffic Noise", *Chalmers University of Technology, Applied Acoustics*, 2007.

[kuttruff01] H. Kuttruff, "Room Acoustics". *4th edition, Taylor & Francis e-Library*, 2001.

[Kuttruff07] H. Kuttruff, "Acoustics". *English edition. Taylor & Francis*, London and New York, 2007.

[Lagrange03] M. Lagrange, S. Marchand, M. Raspaud J. B. Rault, "Enhanced partial tracking using linear prediction", *Digital Audio Effects Conference - DAFx'03*, London, 2003.

[Lagrange05] M. Lagrange, S. Marchand, J. B. Rault, "Long Interpolation of Audio Signals Using Linear Prediction in Sinusoidal Modeling", *Journal of the Audio Engineering Society*, vol. 53, no. 10, pp. 891-905, 2005.

[Lee95] J. Lee, B. G. Lee, "A design of nonuniform cosine modulated filter banks", IEEE Trans. Circuits and Systems-II: Analog and Digital Signal Processing, vol. 42, no. 11, pp. 732-737, November, 1995.

[Levine98] S. Levine, J. Smith, "A sines + transients + noise audio representation for data compression and time/pitch scale modifications", *105th Audio Engineering Society Convention*, preprint no. 4781, September, 1998.

[Li01] D. Li, I. K. Sethi, N. Dimitrova, T. McGee, "Classification of general audio data for content-based retrieval", *Pattern Recognition Letters*, vol. 22, pp. 533-544, 2001.

[Lidy05] T. Lidy, A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification", *International Conference Music Information Retrieval – ISMIR05*, London, U.K., pp. 34–41, 2005.

[Liu93] B. Liu, L. T. Bruton, "The design of N-band nonuniform-band maximally decimated filter banks", *Asilomar Conference*, pp. 1281-1285, 1993.

[Li97] J. Li, T. Q. Nguyen, S. Tantaratana, "A simple design method for near-perfect-reconstruction nonuniform filter banks", *IEEE Transactions on Signal Processing*, vol. 45, pp. 2105-2109, Aug. 1997.

[Lochner64] J. P. A. Lochner, J. F. Burger, "The Influence of Reflections on Auditorium Acoustics", *Journal of Sound and Vibration*, vol. 1, pp. 426-454, 1964.

[Luce63] M. Jr. Clark, D. Luce, R. Abrams, H. Schlossberg, J. Rome, "Preliminary experiments on the aural significance of parts of tones of orchestral instruments and on choral tones", *Journal of the Audio Engineering Society*, vol. 11, no. 1, pp. 45–54, 1963.

[Lui04] W. K. Lui, K. M. Li, "A theoretical model of the horn amplification of sound radiated from tires above a porous road pavement", *Journal of the Acoustical Society of America*, vol. 116, pp. 313-320, 2004.

[Lundeby95] A. Lundeby, T. E. Vigran, H. Bietz, M. Vorlander, "Uncertainties of Measurements in Room Acoustics", *Acustica*, vol. 81, pp. 344-355, 1995.

[MacWilliams76] F. J. MacWilliams, N. J. A. Sloane, "Pseudo- Random Sequences and Arrays", *Proceedings of IEEE*, vol. 64, no. 12, pp. 1715-1729, December, 1976.

[Martin93] J. Martin, D. Van Maercke, J.-P. Vian, "Binaural simulation of concert halls: A new approach for the binaural reverberation process", *Journal of the Acoustical Society of America*, vol. 94, no. 6, pp. 3255–3264, 1993.

[McAulay86] R. J. McAulay, T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744-754, 1986.

[Meng02] Z. Meng, K. Sakagami, M. Morimoto, G. Bi, K. Chichung Alex, "Extending the Sound Impulse Response of Room Using Extrapolation", *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, March, 2002.

[Misdariis98] N. R. Misdariis, B. K. Smith, D. Pressnitzer, P. Susini, S. McAdams, "Validation of a multidimensional distance model for perceptual dissimilarities among musical timbres", *16th International Congress of Acoustics – ICA98*, Woodbury, NY, pp. 3005–3006, 1998.

[Moore96] B. C. J. Moore, "Masking in the Human Auditory System", *Collected Papers on Digital Audio Bit-Rate Reduction (N. Gilchrist and C. Grewin, eds.), Audio Engineering Society*, pp. 9–19, 1996.

[Moore97] B. C. J. Moore, An Introduction to the Psychology of Hearing. *Academic Press, fourth edition*, 1997.

[Muller01] S. Muller, P. Massarani, "Transfer-function measurement techniques with sweeps", *Journal of the Audio Engineering Society*, vol. 49, no. 6, pp. 443, 2001.

[Nagai97] T. Nagai, T. Futie, M. Ikehara, "Direct design of nonuniform filter banks", in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing – ICASSP97*, vol. 3, pp. 2429-2432, April, 1997.

[Nayebi93] K. Nayebi, T. P. Barnwell, M. J. T. Smith, "Nonuniform filter banks: A reconstruction and design theory", *IEEE Transactions on Signal Processing*, vol. 41, no. 3, pp. 1114-1127, Mar. 1993.

[Nielsen97] J. L. Nielsen, "Improvement of Signal-To-Noise Ratio in Long-Term MLS Measurements with High Level Nonstationary Disturbances", *Journal of the Audio Engineering Society*, vol. 45, no. 12, pp. 1063-1066, December, 1997.

[Ning05] L. Ning, "Vibrato analysis and synthesis for violin: an approach by using energy density spectrum analysis", Internal report, National University of Singapore, 2005

[Noll93] P. Noll, "Wideband Speech and Audio Coding", *IEEE Communications Magazine*, vol. 31, no. 11, pp.34-44, November, 1993.

[Norcross95] S. Norcross, J. Vanderkooy, "A Survey of the Effects of Nonlinearity on Various Types of Transfer-Function Measurements", *99th International Convention of Audio Engineering Society*, pp. 4137, 1995.

[Obukhov41] A. M. Obukhov, "Scattering of sound in turbulent flow". Dokl. Akad. Nauk SSSR, vol. 30, pp. 611–614, 1965.

[Painter00] T. Painter, A. Spanias, "Perceptual Coding of Digital Audio". *Proceedings of the IEEE*, vol. 88, no. 4, April, 2000.

[Painter05] T. Painter, A. Spanias, "Perceptual Segmentation and Component Selection for Sinusoidal Representations of Audio", *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp.149-162, 2005.

[Paje09] S. E. Paje, M. Bueno, U. Viñuela, F. Terán, "Toward the acoustical characterization of asphalt pavements: Analysis of the tire/road sound from a porous surface", *Journal of the Acoustical Society of America*, vol. 125, no. 1, 5-7, 2009.

[Peeters02] G. Peeters, X. Rodet, "Automatically selecting signal descriptors for Sound Classification", *International Computer Music Conference – ICMC02*, Goteborg, Sweden, 2002.

[Peeters04] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project", *Cuidado project report*, *Institut de Recherche et de Coordination Acoustique Musique - IRCAM*, 2004.

[Portnoff76] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast fourier transform", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 243-248, June, 1976.

[Pompei02] F. J. Pompei, "Phased array element shapes for suppressing grating lobes", *Journal of the Acoustical Society of America*, vol. 111 no. 5, 2002.

[Princen87] J. Princen, A. Johnson, A. Bradley, "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation", *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP87*, pp. 2161-2164, 1987.

[Princen95a] J. Princen, J. D. Johnston, "Audio Coding with Signal Adaptive Filterbanks", *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP95*, pp. 3071- 3074, 1995.

[Princen95b] J. Princen, "The design of nonuniform modulated filter banks", IEEE Transactions on Signal Processing, vol. 43, no. 11, pp. 2550-2560, November, 1995.

[Rabiner93] L. Rabiner, B. Juang. "Fundamentals of speech recognition". New York, Prentice Hall, 1993.

[Rao94] N. A. H. K. Rao, "Investigation of a pulse compression technique for medical ultrasound: A simulation study", *Medical & Biological Engineering & Computing - MBEC*, vol. 32, no. 2, pp. 181–188, 1994.

[Ratnam03] R. Ratnam, D. L. Jones, B. C. Wheeler, J. William D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time", *Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[Ream70] N. Ream, "Nonlinear identification using inverse-repeat m sequences", *Proceedings of IEE (London)*, vol. 117, pp. 213–218, 1970.

[Rife89] D. Rife, J. Vanderkooy, "Transfer-Function Measurement with Maximum-Length Sequences", *Journal of the Audio Engineering Society*, vol. 37, no. 6, pp. 419-444, June, 1989.

[Rife92] D. Rife, "Modulation Transfer Function Measurements with Maximum-Length Sequences", *Journal of the Audio Engineering Society*, vol. 40, no. 10, pp. 779-790, 1992.

[Risset69] J.-C. Risset, M. V. Mathews, "Analysis of Musical-Instrument Tones", *Physics Today*, vol. 22, no. 2, 23–30, 1969.

[Röbel02] A. Röbel, "Estimating partial frequency and frequency slope using reassignment operators", *International Computer Music Conference – ICMC02*, Göteborg. 2002.

[Rumelhart86] D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning Internal Representations by Error Propagation", *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Foundations* D.E. Rumelhart and J. L. McClelland, Eds., (MIT Press Cambridge, MA), vol. 1, pp. 318-362, 1986.

[Sandberg02] U. Sandberg, J. A. Ejsmont, "Tyre/Road Noise Reference Book", INFORMEX, 2002.

[Satoh05] F. Satoh, J. Hirano, S. Sakamoto, H. Tachibana, "Sound propagation measurement using Swept-Sine signal", *Inter-Noise05*, 1691, 2005.

[Satoh07] F. Satoh, J. Hirano, Sano, Masato, Hayashi, Yukiteru, Sakamoto, Shinichi, Tachibana, Hideki, "Acceptable temperature Changes During Synchronous Averaging for Reverberation Time Measuring by Swept-Sine Method", *Inter-Noise07*, 2007.

[Satoh98] F. Satoh, Y. Hidaka, H. Tachibana, "Reverberation Time Directly Obtained from Squared Impulse Response Envelope", *International Congress on Acoustics – ICA98*, vol. 4, pp. 2755–2756, Seattle, WA, June, 1998.

[Satoh99] F. Satoh, Y. Hidaka, H. Tachibana, "Influence of time variance on room impulse response measurement", *Acustica united with Acta Acustica*, vol. 85, pp. 441, 1999.

[Sayers98] M. Sayers, S. Karamihas, "The Little Book of Profiling", *University of Michigan*, 1998.

[Scharf70] B. Scharf, "Critical Bands", *Academic Press*, 1970.

[Schroeder66] M. R. Schroeder, "New Method of Measuring Reverberation Time", *Journal of the Acoustical Society of America*, vol. 38, pp. 329-361, 1965, and vol. 40. pp. 549-551, 1966.

[Schroeder79a] M. R. Schroeder, "Integrated-Impulse Method Measuring Sound Decay without Using Impulses", *Journal of the Acoustical Society of America*, vol. 66, 1979.

[Schroeder79b] M. R. Schroeder, et al., "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear", *Journal of the Acoustical Society of America*, vol, 64, no. S1, pp. 1647-1652, 1979.

[Schroeder81] M. R. Schroeder, "Modulation Transfer Functions: Definition and Measurement", *Acustica*, vol. 49, pp. 179-182, 1981.

[Serra89] X. Serra, A System for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition, *PhD. Thesis, CCRMA, Stanford University*, 1989.

[Serra90] X. Serra, J. Smith, "Spectral Modeling Synthesis A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition". *Computer Music Journal*. vol. 14, pp. 12-24, 1990.

[Serra97] X. Serra, "Musical sound modeling with sinusoids plus noise", *Musical signal processing*, Swets & Zeitlinger Publishers, 1997.

[Shaughnessy00] D. O'Shaughnessy, "Speech Communications: Human and Machine". *IEEE Press, 2nd edition*, 2000.

[SILVIA06] SILVIA. "Guidance manual for the implementation of low-noise road surfaces", FEHRL report 2006/02, *Forum of European National Highway Research Laboratories*, Brussels, Belgium, 2006.

[Sinha93] D. Sinha, A. H. Tewfik, "Low Bit-Rate Transparent Audio Compression Using Adapted Wavelets", *IEEE Transactions Acoustics, Speech, and Signal Processing*, vol. 41, no, 12, pp. 3463 - 3479, 1993.

[Smirnov98] A. Smirnov, "Proto musique concrete: Russian futurism in the 10s and 20s and early ideas of sonic art and art of noises", *Inventionen 98 Festival, Haus des Rundfunks*, Berlin, Germany, September, 1998.

[Smith87] J. O. Smith, X. Serra, "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation", *International Computer Music Conference*, Urbana, IL (Int. Computer Music Assn., San Francisco), pp. 290–297. Also available as Report no. STAN-M-43, Dept. of Music, Stanford Univ., 1987.

[Soulodre98] G. A. Soulodre, "Adaptive Methods for Removing Camera Noise from Film Soundtracks". *PhD thesis, McGill University*, Montreal, Canada, 1998.

[Stan02] G. Stan, J. Embrechts, D. Archambeau, "Comparation of Different Impulse Response Measurement Techniques", *Journal of the Audio Engineering Society*, vol. 50, no. 4, 2002.

[Steeneken80] H. M. Steeneken, T. Houtgast, "A Physical Method for Measuring Speech-Transmission Quality", *Journal of the Acoustical Society of America*, vol. 67, pp. 318-326, 1980.

[Sterian98] A. Sterian, G. H. Wakefield, "A model-based approach to partial tracking for musical transcription", *SPIE Annual Meeting - Advanced Signal Processing Algorithms, Architectures, and Implementations VIII*, San Diego, 1998.

[Streit90] R. L. Streit, R. F. Barrett, "Frequency line tracking using hidden Markov models", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.38, no. 4, pp. 586-598, 1990.

[Strong67a] W. Strong, M. Clark, "Synthesis of Wind-Instrument Tones", *Journal of the Acoustical Society of America*, vol. 41, no. 1, pp. 39–52, 1967.

[Strong67b] W. Strong, M. Clark, "Perturbations of Synthetic Orchestral Wind-Instrument Tones", *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 277–285, 1967.

[Svensson95] P. Svensson, J.L. Nielsen, "Short-term time-variance in electoracoustic channels and its effect on MLSmeasurements", *15th International Conference on Acoustics – ICA95*, vol. 4, pp. 163-166, 1995.

[Svensson99] U. Svensson, J.L. Nielsen, "Errors in MLS measurements caused by time-variance in acoustic systems", *Journal of the Audio Engineering Society*, vol. 47, pp. 907-927, 1999.

[Tatarskii61] V. I. Tatarskii, “Wave Propagation in a Turbulent Medium”. *McGraw Hill*, New York, vol. 134, no. 3475 pp. 324-325, 1961.

[Terhardt82] E. Terhardt, G. Stoll, M. Seewann, \Algorithm for extraction of pitch and pitch salience from complex tonal signals”, *Journal of the Acoustical Society of America*, vol. 71, 1982.

[Thiede00] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidner, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, B. Feiten, “PEAQ - the ITU standard for objective measurement of perceived audio quality”, *Journal of the Audio Engineering Society*, vol. 48, 2000.

[Tolonen99] T. Tolonen, “Methods for separation of harmonic sound sources using sinusoidal modeling”, *106th AES Convention*, Munich, 1999.

[Treurniet00] W. C. Treurniet, G. Soulodre, “Evaluation of the ITU-R objective audio quality measurement method”, *Journal of the Audio Engineering Society*, vol. 48, 2000.

[Tsutsui92] K. Tsutsui, H. Suzuki, O. Shimoyoshi, M. Sonohara, K. Akagiri, R.M. Heddle, “ATRAC: Adaptive Transform Acoustic Coding for MiniDisc”, *preprint 3456, 93rd AES Convention*, 1992.

[Vaidyanathan90] P. P. Vaidyanathan, “Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications”. *IEEE*, vol. 78, no. 1, pp. 56-93, 1990.

[Vanderkooy93] J. Vanderkooy, “Aspects of MLS Measuring Systems”, *Journal of the Audio Engineering Society*, vol. 42, pp. 314-335, 1993.

[Veldhuis92] R. Veldhuis, “Bit Rates in Audio Source Coding”, *IEEE Journal on Selected Areas in Communications*, vol. 10, pp. 86–96, 1992.

[Verma98] T. Verma, T. Meng, “An analysis/synthesis tool for transient signals that allows a flexible sines + transients + noise model for audio”, *IEEE International Conference on Acoustics, Speech, and Signal Processing – ICASSP98*, May, 1998.

[Verma97a] T. Verma et al., “Transient modeling synthesis: A flexible analysis/synthesis tool for transient signals”, *International Computer Music Conference – ICMC97*, 1997.

[Verma97b] T. Verma, S. Levine, and T. Meng, “Transient modeling synthesis: A flexible analysis/synthesis tool for transient signals”, *International Computer Music Conference*, Thessaloniki, Greece, pp. 164–167, 1997.

[Vettefii95] M. Vettefii, J. Kovacevic, “Wavelets and Subband Coding”, Prentice Hall, Englewood Cliffs, 1995.

[Vieira04] J.Vieira. Automatic Estimation of Reverberation Time. 116th AES International Convention, Berlin, Germany, May 2004.

[Vorlander94] M. Vorlander, H. Bietz, “Comparison of Methods for Measuring Reverberation Time”, *Acustica*, vol. 80, pp. 207-217, 1994.

[Vorlander97a] M. Vorlander, M. Kob, “Practical Aspects of MLS Measurements in Building Acoustics”, *Aplications of Acoustics*, vol. 52, pp. 239-258, 1997.

[Vorlander97b] M. Vorlander, E. Mommertz, "Guidelines for the Application of the MLS Technique in Building Acoustics and in Outdoor Measurements", *Inter-Noise97*, Budapest, Hungary, August, 1997.

[Wada95] S. Wada, "Design of nonuniform division multirate FIR filter banks", *IEEE Transactions on Circuits and Systems II*, vol. 42, pp. 115-121, 1995.

[Walmsley99] P. J. Walmsley, S. J. Godsill, P. J. W. Rayner, "Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters", *IEEE Workshop on* applications of Signal Processing to Audio and Acoustics - WASPAA'99, New Paltz, 1999.

[Watanabe98] T. Watanabe, Y. Shibahara, T. Kida, N. Sugino, "Design of nonuniform FIR filter banks with rational sampling factors", in Proc. IEEE APCCAS, pp. 69-72, 1998.

[Wen07] Wen, Xue, "Harmonic Sinusoid Modeling of Tonal Music Events", *PhD thesis, Queen Mary*, University of London, 2007.

[Wen08] J. Y. C. Wen, E. A. P. Habets, P. A. Naylor, "Blind Estimation of Reverberation Time Based on the Distribution of Signal Decay Rates", *IEEE International Conference on Acoustics, Speech, and Signal Processing – ICASSP08*, Las Vegas, USA, March, 2008.

[Wright01] M. Wright, J. Beauchamp, K. Fitz, X. Rodet, A. Robel, X. Serra, G. Wakefield, "Analysis/synthesis comparison", Organized Sound, vol. 5, no. 3, pp. 173–189, 2001.

[Xiang95] N. Xiang, "Evaluation of Reverberation Times Using a Nonlinear Regression Approach", *Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 2112–2121, 1995.

[Xiang06] N. Xiang, T. Jasa, S. U. Zuhre, "Schroeder Decay Decomposition for Sound-Energy Decay Analysis in Acoustic Coupled-Volume Systems". Wespac IX 2006, *9th Western Pacific Acoustics Conference*, Seoul, Korea, 2006.

[Zölzer02] U. Zölzer et al., "DAFX - Digital Audio Effects", ISBN: 0-471-49078-4, *John Wiley & Sons*, 2002.

[Zwicker91] E. Zwicker. T. Zwicker, "Audio Engineering and Psychoacoustics. Matching Signals to the Final Receiver, the Human Auditory System ", *Journal of the Audio Engineering Society*, vol. 39, pp. 115–126, 1991.

[Zwicker99] E. Zwicker. H. Fastl, "Psychoacoustics: Facts and Models". Springer-Verlag, 1999.

PAPERS PUBLISHED ON JOURNALS (or ready to be submitted)

[Paulo09a] J. P. Paulo, J. L. Bento Coelho, C. A. Martins, "Hybrid MLS Technique for Room Impulse Response Estimation", *Applied Acoustics*, vol. 70, no. 4, pp. 556-562, 2009.

[Paulo10a] J. P. Paulo, J. L. Bento Coelho, M. A. T. Figueiredo, "Statistical Classification of Road Pavements from Noise Measurements", *Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1747-1754, 2010.

[Paulo1xx] J. P. Paulo, J. L. Bento Coelho, "Swept Sine Against MLS With Music Signals As Background Noise Using Different Averaging Techniques", *Journal of the Audio Engineering Society* (201x) (Ready to be submitted);

[Paulo1xx] J. P. Paulo, J. L. Bento Coelho, "Use of Ultrasound Transducer Arrays to Account for Time-variance on Room Acoustics Measurements", *Journal of the Audio Engineering Society* (201x) (Ready to be submitted).

[Paulo1xx] J. P. Paulo, J. L. Bento Coelho, "A Room acoustics measurement technique based on a Perceptual model", *Journal of the Audio Engineering Society* (201?) (To be submitted soon).

PAPERS PRESENTED IN CONFERENCES (selection)

[Paulo03] J. P. Paulo, C. R. Martins, J. L. Bento Coelho, "Hybrid M Sequences for Room Impulse Response Estimation", *Pre-prints of the 115th AES Convention*, New York, USA, October, 2003.

[Paulo05a] J. P. Paulo, C. R. Martins, J. L. Bento Coelho, "Time Segmented Swept Sine Technique for Room Impulse Response Estimation", *ICSV12* –12th *International Congress on Sound and Vibration*, Lisbon, Portugal, July, 2005.

[Paulo05b] J. P. Paulo, C. R. Martins, J. L. Bento Coelho, "Time Segmented Swept Sine Technique for SNR Enhancement", *Forum Acusticum-05*, pp. 2229-2233, Budapest, Hungary, August, 2005.

[Paulo07a] J. P. Paulo, C. R. Martins, J. L. Bento Coelho, "A Room acoustics measurement technique based on a Perceptual model for Annoyance Minimization", *AES 30th International Conference on Intelligent Audio Environments*, Saariselkä, Finland, March, 2007;

[Paulo07b] J. P. Paulo, C. R. Martins, J. L. Bento Coelho, "Segmented Swept sine technique for room impulse response estimation", *153rd Meeting of the Acoustical Society of America*, Salt Lake City, USA, June, 2007.

[Paulo08] J. P. Paulo, J. L. Bento Coelho, "Road Pavement Classification Based on Noise Emission Characteristics", *Acoustics'08- Euronoise2008*, Paris, France, June, 2008.

[Paulo09b] J. P. Paulo, J. L. Bento Coelho, "Swept Sine against MLS in Room Acoustics with Music Signals as Background Noise", *16th International Conference on Sound and Vibration - ICSV16*, Krakow, Poland, July, 2009.

[Paulo10b] J. P. Paulo, E. Freitas, J. L. Bento Coelho, "Texture and Noise Features for Road Pavement Identification and Classification", *InterNoise10*, Lisbon, Portugal, June, 2010.

[Paulo10c] J. P. Paulo, J. L. Bento Coelho, "Method For Identification of Road Pavement Types using a Bayesian Analysis and Neural Networks", *16th International Conference on Sound and Vibration - ICSV16*, Cairo, Egypt, July, 2010.

[Paulo10d] J. P. Paulo, J. L. Bento Coelho, "On the use of Ultrasound Transducer Arrays to Account for Time-variance on Room Acoustics Measurements". *129th AES Convention*, San Francisco, USA, November, 2010.

[Paulo11a] J. P. Paulo, J. L. Bento Coelho, "Accounting for Time Variance Effects on swept sine and MLS Techniques", *InterNoise11*, Osaka, Japan, September, 2011.

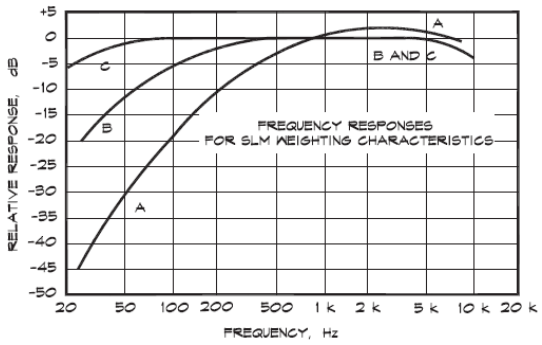
[Paulo11b] J. P. Paulo, J. L. Bento Coelho, “Swept sine Grains Applied to Acoustical Measurements using Perceptual Masking Effects”, *131st AES Convention*, New York, USA, October, 2011.

PATENTS

[Paulo09P] Patent Portugal, Patente de Invenção Nacional PT104200. “Sistema de Análise de Conformidade de Pavimentos Rodoviários com Base no Ruído”, Authors: J. P. Paulo, J. L. Bento Coelho. Institutions: IST and ISEL. Published by INPI – Instituto Nacional de Propriedade Intelectual, in Boletim da Propriedade Industrial nº 78/2009, edited on 2009.04.22. Approved on 2010.02.26. (*in portuguese*)

APPENDIX A1.

Weighting curves - ANSI-S1.4



Frequency Response Characteristics in the American National Standards Specification for Sound Level Meters (ANSI-S1.4 – 1971).

The relative weightings in each third-octave band for the A and C filters are set forth in Table Electrical Weighting Networks

Frequency Hz	A-Weighting Relative Response, dB	C-Weighting Relative Response, dB
12.5	−63.4	−11.2
16	**−56.7**	**−8.5**
20	−50.5	−6.2
25	−44.7	−4.4
31.5	**−39.4**	**−3.0**
40	−34.6	−2.0
50	−30.2	−1.3
63	**−26.2**	**−0.8**
80	−22.5	−0.5
100	−19.1	−0.3
125	**−16.1**	**−0.2**
160	−13.4	−0.1
200	−10.9	0
250	**−8.6**	**0**
315	−6.6	0
400	−4.8	0
500	**−3.2**	**0**
630	−1.9	0
800	−0.8	0
1,000	**0**	**0**
1,250	+0.6	0
1,600	+1.0	−0.1
2,000	**+1.2**	**−0.2**
2,500	+1.3	−0.3
3,150	+1.2	−0.5
4,000	**+1.0**	**−0.8**
5,000	+0.5	−1.3
6,300	−0.1	−2.0
8,000	**−1.1**	**−3.0**
10,000	−2.5	−4.4
12,500	−4.3	−6.2
16,000	**−6.6**	**−8.5**
20,000	−9.3	−11.2

APPENDIX B1.

List of the data of the critical bands

Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz)
1	50	-100	10	1175	1080-1270	19	4800	4400-5300
2	150	100-200	11	1370	1270-1480	20	5800	5300-6400
3	250	200-300	12	1600	1480-1720	21	7000	6400-7700
4	350	300-400	13	1850	1720-2000	22	8500	7700-9500
5	450	400-510	14	2150	2000-2320	23	10,500	9500-12000
6	570	510-630	15	2500	2320-2700	24	13,500	12000-15500
7	700	630-770	16	2900	2700-3150	25	19,500	15500-
8	840	770-920	17	3400	3150-3700			
9	1000	920-1080	18	4000	3700-4400			

APPENDIX C1.

Preferred frequencies of the filter banks used on acoustical measurements

The international standard ISO 266 [ISO266] specifies preferred frequencies for acoustical measurements used for the filter banks. For most acoustical measurements and presentations of data, a frequency spacing based on a constant percentage increment is generally preferred and then the test frequencies form a geometric series.

The preferred frequencies, in Hz, for octave or fractional octave bands are computed from the theoretical values given by:

Filter centre frequency

$$f_{m0} = 2^{FOct} \times 2^{FOct} \times \ldots \times 2^{FOct} \quad \text{and} \quad f_{m0} = \sqrt{f_{Cm1} \times f_{Cm2}}$$

Theoretical filter, lower cutoff frequency

$$f_{Cm1} = (1/2)^{\frac{1}{2} FOct \times f_0}$$

Theoretical filter, upper cutoff frequency

$$f_{Cm2} = 2^{\frac{1}{2} FOct \times f_0}$$

Preferred filter, lower cutoff frequency

$$f_{Cm1} = \frac{f_{m0}}{\sqrt[6]{2}} = 0.8909\, f_{m0}$$

Preferred filter, upper cutoff frequency

$$f_{Cm2} = f_{m0} \times \sqrt[6]{2} = 1.1225\, f_{m0} \quad \text{or} \quad f_{Cm2} = f_{Cm1} \times \sqrt[3]{2} = 1.2599\, f_{Cm1}$$

Frequency range	Attenuation (dB)
$\left[\frac{f_{m0}}{\sqrt[12]{2}} = 0.9439\, f_{m0},\ f_{m0}\sqrt[12]{2} = 1.0595\, f_{m0}\right]$	$-0.5 \le \Delta \le 1$
$\left[\frac{f_{m0}}{\sqrt[6]{2}} = 0.8909\, f_{m0},\ f_{m0}\sqrt[6]{2} = 1.225\, f_{m0}\right]$	$-0.5 \le \Delta \le 6$
for $\frac{f_{m0}}{\sqrt[3]{2}} = 0.7937\, f_{m0}$ and for $f_{m0}\sqrt[3]{2} = 1.2599\, f_{m0}$	≥ 13
below $\frac{f_{m0}}{4}$ and above $4\, f_{m0}$	≥ 50
below $\frac{f_{m0}}{8}$ and above $8\, f_{m0}$	≥ 60

where FOct is the fraction of the octave (FOct=1 is the octave, FOct=1/3 is the third of octave, and so on)

APPENDIX D1.

The Johnston model

The implementation of the psychoacoustic model is as follows:

Step 1 - Critical Band Analysis

The first step calculates the energy present in each critical band from the complex spectrum of each signal segment computed by the FFT, assuming discrete non-overlapping critical bands. The summation is

$$B(i) = \sum_{m=bl_i}^{bh_i} \left| X(m) \right|^2, \quad i = 1, \ldots, i_{\max}$$

where bl_i and bh_i are the lower and upper bounds of the i_{th} critical band. The value of i_{max} depends on the sampling frequency. Johnston notes that a true critical band analysis would calculate the power within one critical band at every frequency m. This would create a continuous (higher resolution) critical band spectrum.

Step 2 - Spread of Masking Function

To calculate the excitation pattern, Johnston uses the spreading function as proposed by Schroeder et al. in [Schroeder79]. The spreading function SF has lower and upper slopes of +25 dB/Bark and -10 dB/Bark respectively. The spreading function as given in [Hartmann97] is used to estimate the effects of masking across critical bands. The spreading function is calculated for $abs\ (j - i) \leq 25$, where i is the bark frequency of the masked signal, and j is the bark frequency of the masking signal, and placed into a matrix $SF(i,j)$.The term bark is often used to indicate a frequency difference of 1 critical band. It is a reasonable approximation (at intermediate speech levels) to the experimental data given by Zwicker [Zwicker99]. This spreading function is then convolved with the bark spectrum to account for inter-band masking, to give

$$C(i) = SF(i,j) * B(i),$$

where $C(i)$ stands for the spread critical band spectrum.

Step 3 - Calculation of the Noise Masking Threshold

Two masking thresholds are used, one for a tone masking noise and another for noise masking a tone. Tone-masking-noise is estimated at 14.5+i dB below $C(i)$. Noise-masking-tone is estimated as being a uniform 5.5 dB below $C(i)$ across the whole critical band spectrum.

To determine if the signal is tone-like or noise-like, the concept of Spectral Flatness Measure (SFM) is used [Jayant84]. The SFM is defined as

$$SFM = \frac{\mu_g}{\mu_a},$$

where μ_g and μ_a represent the geometric and arithmetic means of the power spectrum respectively. *SFM* is bounded by 0 and 1. Values close to zero will occur if the spectrum in a particular band is nearly sinusoidal and values approaching 1 indicate a flat spectrum. From this value, a totality coefficient α is generated, by

$$\alpha = \min\left(\frac{SFM_{dB}}{-60}, 1\right),$$

For an entirely tone-like signal $\left(SFM_{dB} \to -\infty\right)$ the resulting tonality coefficient is α = 1. Conversely, an entirely noise-like signal would have SFM_{dB} = 0 and thus α = 0.

The offset, in decibel, for each band is calculated with the tonality coefficient as

$$O(i) = \alpha(14.5+i) + 5.5(1-\alpha),$$

The set of JND estimates in the frequency power domain are obtained by subtracting the offsets from the spread critical band spectrum by

$$T(i) = 10^{\log_{10} C(i) - \frac{O(i)}{10}},$$

Step 4 - Converting the Spread Threshold back to the Bark Domain and including the Absolute Threshold

This step attempts to make the deconvolution of $B(i)$ with the spreading function. However, this process is very unstable due to the shape of the spreading function, and a correction factor (renormalization) is used instead. In fact, the spreading function, because of its shape, increases the energy estimates in each band due to the effects of spreading. The

renormalization takes this into account, and multiplies each $T(i)$, by the inverse of the energy gain, assuming a uniform energy of 1 in each band. In other words, given a flat $B(i)$, and a condition where all $O(i)$ are equal, it will return a flat renormalized $T(i)$ denoted by $T_j(m)$. Then, each $T(i)$ is checked against the absolute threshold of hearing $T_{ABS}(i)$ and replaced by

$$T_j(m) = \max\left\{T(i), T_{ABS}(i)\right\},$$

Although this alternative is a crude approximation of the committed error it is shown to be insignificant for our purposes.

Since the actual playback level is not known, it is assumed that the playback level is set such that the quantization noise is inaudible. Specifically, it is assumed that a signal of 4 kHz with peak magnitude of ± 1 least significant bit of a 16 bit integer value is at the absolute threshold of hearing (-5 dB SPL at 4 kHz).

APPENDIX E1.

Assessment of the quality of the coded signals

The common idea behind perceptual quality measures is to mimic the situation of a subjective test, where human beings would have to score the quality of sound samples in a listening laboratory environment. The result is called 'mean opinion score', MOS. For a long time, subjective test procedures were the only means of assessing the sound quality impression.

The International Telecommunications Union (ITU), as the agency of the United Nations which regulates information and communication technology issues, launched a number of standards for the assessment of audio coding subjectively and objectively.

Historically related to the assessment of telephone connections, useful methods for assessing the listening quality of telephone band systems were first standardized within the ITU-T. Recommendation P.800 defines for instance the absolute category rating test method (ACR) which has been used for the assessment of speech codecs since 1993. Within the ACR test method, the ITU five grade impairment scale is applied (see Table E.1 below). In the telecommunication environment, testing is done without a comparison to an undistorted reference. This copes with a typical situation of a phone call, where the listener has no access to a comparison with a reference, e.g. the original voice of the other party. However, it

should be noted that the listening test according P.800 could be regarded as a comparison between a test signal and a reference "in the mind" of the listener. The reason for this is that the listener is very familiar with the natural sound of a human voice.

For comparison reasons, and in order to be able to merge the results of different individuals, it is necessary to adjust the listeners' opinions to an absolute scale. For this purpose, predefined examples with well defined noise insertions of fixed modulated noise reference units (MNRU, [ITUT810]) are presented at the beginning of a test. Each sample represents an example distortion corresponding to the ITU-T version of the five grade impairment scale.

Table E.1 - The ITU-T five-grade impairment scale.

Impairment	Grade
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Based on these test conditions a population of typically 20 to 50 test subjects will be presented with an identical series of speech fragments. Every test subject will be asked to score each sample by applying the impairment scale. After statistical processing of the individual results, a Mean Opinion Score (MOS) can be calculated. With thorough setups, such test results can be reproduced quite well, even at different locations. Further tests defined by P.800 are the 'Comparison Category Rating' (CCR) and 'Degradation Category Rating' (DCR) procedures. It goes without saying that the effort needed in terms of subjects and time is tremendous. It is clear that such test methods can not be applied within a practical or field environment in the daily life.

The ITU recommended a test procedure to assess wide band audio codecs on the basis of subjective tests. Subjective assessments of low bit rate audio codecs in the past always targeted at almost transparent quality. For this reason, the test method focuses on the comparison of the coded/decoded signal to the unprocessed original reference. The relevant recommendation is known as BS.1116, titled "Methods for the Subjective Assessment of small Impairments in Audio Systems including Multichannel Sound Systems" [ITUR1116] which was issued by the ITU-R in 1994 and was updated in 1997.

The test method, which is recommended by BS.1116, is referred to as "double-blind triple-stimulus with hidden reference". It is extremely sensitive and allows for the accurate detection of small impairments of the sound material. The grading scale used should be treated as continuous with "anchors" derived from the ITU-R five-grade impairment scale according to ITU-R BS.562 [ITUR562]. It is depicted in Table E.2.

Table E.2- The ITU-R five-grade impairment scale.

Impairment	Grade	SDG
Imperceptible	5.0	0.0
Perceptible, but not annoying	4.0	-1.0
Slightly annoying	3.0	-2.0
Annoying	2.0	-3.0
Very annoying	1.0	-4.0

The analysis of the results from a subjective listening test is generally based on the Subjective Difference Grade (*SDG*) and is defined as:

$$SDG = Grade_{\text{signalUnderTest}} - Grade_{\text{ReferenceSignal}}$$

Differently to the listening test according to ITU-T P.800, an explicit comparison between the test signal and a reference signal is needed in the case of BS.1116, since the listener never knows how the original (music) signal would sound like. This method was applied in a variety of international verification tests in the past. However, keep in mind that because of the scope of ITU-R it can be applied to small impairments only, which means a practical limitation to almost "transparent" studio quality.

Because of the restrictions to small impairments, other methods are needed for the quality assessment of very low bit rates (i.e. of large impairments). The methods according to ITU-T P.800 were adopted for some assessments to overcome the problem of a "gap" for a useful recommendation on testing significantly impaired wide band audio signals. While in principle they seem to be better suited for impaired music signals when compared to the BS.1116 method, its exploitation for very low bit rate audio coding applications remains questionable, as there are no clearly defined example distortions in such a case. Therefore, an advanced listening test procedure has been advised by an EBU expert group, known as "MUSHRA". MUSHRA stands for "Multiple Stimulus with Hidden Reference Anchors".

The new method targets testing significantly impaired audio signals, such as those derived at very low bit rates. MUSHRA was adopted as an international recommendation by the ITU-R working party 6Q in 2001. Although first experience has been gained, also MUSHRA exhibits some major drawbacks which make it very difficult if not impossible to derive an *absolute* quality metric, which one would expect from a measurement. Still, perceptual measurements cope well with MUSHRA results, if properly correlated.

Furthermore, careful experimental design and planning is needed to ensure that uncontrolled factors do not contaminate the listening test so that ambiguities are not caused. Of course, subjective experiments require a huge number of subjects to achieve statistically relevant results, and thus are very costly and time consuming. For quality testing in the laboratory as well as in the field, subjective tests are clearly not an option. Consequently, perceptual models can be applied to generate an objectively derived quality metric ('objective MOS' or OMOS) that can be compared to a mean opinion score scale. In the course of perceptual processing, many other detailed model output values, such as FFT spectra, dynamically measured bandwidth, distortions, modulation and excitation will be generated and reported, too, making this technology universally applicable.

Assessing audio quality was a pending issue during the years of developing compression schemes. Consequently, the idea of substituting the subjective tests by objective computer based methods has been an ongoing focus of research and development. Early work motivated through the development in speech coding was reported in [Karjalainen85]. Since then, several methods were introduced.

The International Telecommunications Union (ITU) developed a psychoacoustic model for use in a standard for objective measurement of perceived audio quality (adopted as ITU-R BS.1387) [Thiede00]. An evaluation of this method performed by Treurniet and Soulodre [Treurniet00] found that this measure produced results that correlate well with subjective ratings obtained in formal listening tests.

The following description covers only the parts that are necessary in the context of noise reduction algorithms. However, since this masking model is implemented for use by the noise reduction methods described in the following sections, the calculation of the excitation pattern and the masking threshold will be covered in detail.

The Perceptual Evaluation of Audio Quality (PEAQ) model, which is full compliance with ITU-R BS.1387, provides accurate and repeatable estimates of audio quality degradation occurring through e.g. coding procedures modeling human perception by a computer model of the ear. It compares the audio signal input to a device under test (DUT) with the corresponding (degraded) audio signal output from that device on a perceptual basis.

More recently, a large-scale listening test was performed to evaluate the effect of cascading multiple generations of coding, and the results of those listening tests were published in [Grant01]. Afterwards, PEAQ was used to evaluate the audio quality degradation for each combination of codec and audio program item. In general, good correlation between the average subjective scores assigned by the listeners and the PEAQ ODG are verified.

This model consists of two versions: one that is intended for applications requiring high processing speed, called the basic version, and another for applications requiring highest achievable accuracy, called the advanced version. The basic version only uses an FFT-based ear model, whereas the advanced version uses an FFT-based model to determine the masking threshold and a filter bank based ear model to compare internal representation.

The model output values are based partly on the masked threshold concept and partly on a comparison of internal representations. In addition, it also yields output values based on a comparison of linear spectra not processed by an ear model. The model outputs the partial loudness of nonlinear distortions, the partial loudness of linear distortions (signal components lost due to an unbalanced frequency response), a noise to mask ratio, measures of alterations of temporal envelopes, a measure of harmonics in the error signal, a probability of error detection, and the proportion of signal frames containing audible distortions. Selected output values are mapped to a single quality indicator by an artificial neural network with one hidden layer [Rumelhart86].

The steps involved in computing the excitation pattern in the basic version are as follows.

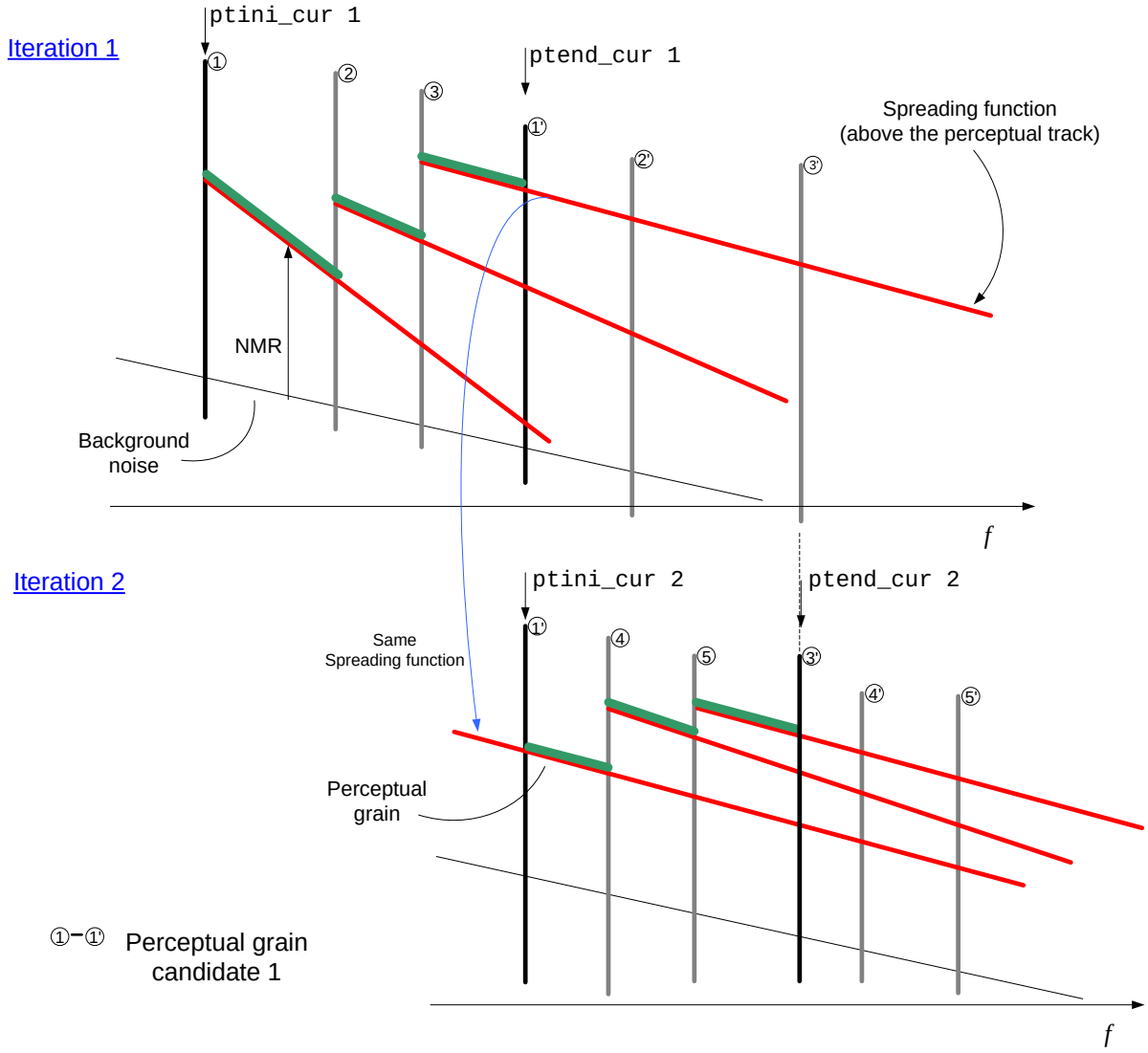
- Time to Frequency conversion
- Frequency-dependent weighing
- Mapping into perceptual (Bark) domain
- Adding internal noise
- Applying the spreading function
- Applying nonlinear superposition
- Temporal Spreading

APPENDIX F1.

The full track frame module construction

The algorithm to implement the Full Track Frame for the STSNM technique is described in this appendix.

BASIC SCHEME



where 1-1' correspond to the beginning of the grain candidate 1 and so on.

PSEUDO-CODE

The inputs for this module are:

- `fsegZones` – structure of the perceptual zones constituted by the perceptual tracks;
  - `fsegZones.fini(i,j)`– initial frequency of the grain candidate (in bins);
  - `fsegZones.fend(i,j)`– end frequency of the grain candidate (in bins);

  Index *i* is the perceptual tracks and the index *j* is the zone.

- `minGRLEN` – minimum allowed grain length (in samples);
- `minNMR` – minimum allowed Noise-to-Mask ratio (in dB);
- `minGUARDB` – minimum allowed guard band from the Perceptual Grain and the perceptual track (in bins);
- `minSNR` – minimum allowed Signal-to-Mask ratio;
- RTe – estimated Reverberation Time

The outputs for this module are:

- `FullTrFrame` – full track frame. This binary array informs the missing track parts of the full frame;
- `FullfsegZones` – structure of the perceptual zones constituted by the perceptual tracks;
  - `FullfsegZones.fini(i,j)`– initial frequency of the grain (in bins);
  - `FullfsegZones.fend(i,j)`– end frequency of the grain (in bins);
  - `FullfsegZones.lvl(i,j)`– level of the grain (in dBSPL);

The intermediate variables for this module are:

- `ptini_cur` – pointer to the initial frequency interval of the current iteration (in bins);
- `ptend_cur` – pointer to the end frequency interval of the current iteration (in bins);
- `Refzone` – reference zone (the first zone) in the interval of the current iteration;
- `GrCandNMR_Acc` – estimate of the NMR accumulation for each grain candidate in a sub-interval;

- Step 1 – Initialization of the variables
  1. Initialize `GrCandpool` with `fsegZones;`
  2. Initialize to zero the `FullTrFrame` array with length NFFT;
  3. Select the grain candidate with lower `GrCandpool.fini` from all zones and store in `GrCandCur.fini, GrCandCur.fend` and `GrCandCur.zone`;
  4. Initialize `Refzone` with `GrCandCur.zone;`
  5. Initialize `ptini_cur` with `GrCandCur.fini;`
  6. Initialize `ptend_cur` with `GrCandCur.fend;`

7. WHILE grains candidates in `GrCandpool` were not used DO

- Step 2 - Select the perceptual grain of the reference zone (first part of the perceptual track)
    a. FOR all zones different of `Refzone` DO
        i. Find the grain candidates with `GrCandpool.fini` lower than `ptend_cur` and `GrCandpool.fend` higher than `ptend_cur`
        ii. and store in `GrCandCur.fini, GrCandCur.fend` and `GrCandCur.zone`;
    END FOR
    b. Update `ptini_sec` with `min(GrCandCur.fini);` % find the beginning of the second grain candidate in the current interval;
    c. Find `Gr_sec = GrCandCur(ptini_cur to ptini_sec)` of the `Refzone`; % Choose the segment of the `GrCandCur` from the beginning of the current interval to the beginning of the second grain candidate;
    d. Store in `FullTrFrame` array;
    e. Store `Gr_sec` in `FullfsegZones` structure;
    f. Remove the grain candidate `Gr_sec` from the `GrCandpool`;
    g. `NumSubInt = length(GrCandCur)-1;` % number of sub-intervals. Excludes the first grain in the interval;

- Step 3 – Choose sub-intervals and select the perceptual grain with higher NMR
    h. FOR `NumSubInt` DO
        i. Find `ptini`; % pointer to the initial frequency of the sub-interval. Corresponds to the `GrCandCur.fini` of the grain candidate *i*;

        ii. IF *i*+1 is equal to `NumSubInt`
            1. Find `ptend`; % pointer to the final frequency of the sub-interval. Corresponds to the `ptend_cur`;
        ELSE
            2. Find `ptend`; % corresponds to the `GrCandCur.fini` of the grain candidate *i*+1;
        END IF

        iii. Find `NumGrCand`; % number of grain candidates belonging to the current sub-interval;

        iv. FOR `NumGrCand` DO
            1. Calculate `GrCandNMR_Acc;` % Noise-to-Mask ratio accumulation of the sub interval;
            2. Mark grain candidate as already used;
        END FOR
        v. Calculate max of `GrCandNMR_Acc`; % `GrCandCur` with higher `GrCandNMR_Acc;`
        vi. IF (`ptend` – `ptini`) is less than `minGRLEN AND`

`GrCandCur.fini == ptini`

a. `ptend = ptini + minGRLEN -1;` % extends the length of the grain

vii. END IF

viii. Find `Gr = GrCandCur(ptini to ptend);` % Choose the

ix. Store in `FullTrFrame` array;

x. Store `Gr` in `FullfsegZones` structure;

xi. Remove the grain candidate `Gr` from the `GrCandpool`; % the grain candidates used are removed;

END FOR

i. Update `ptini_cur` with `ptend`; % update the pointer of lower edge of the next interval;

j. Update `ptend_cur` with `max(GrCandCur.fend);` % update the pointer of lower edge of the next interval with the grain candidate with higher `GrCandCur.fend`;

k. Update `Refzone` with `Gr.zone`; % notice that the current Gr corresponds to the last chosen grain candidate;

END WHILE

- Step 4 % fusion the consecutive grains belonging to the same zone.

l. FOR all grains of `FullfsegZones` DO

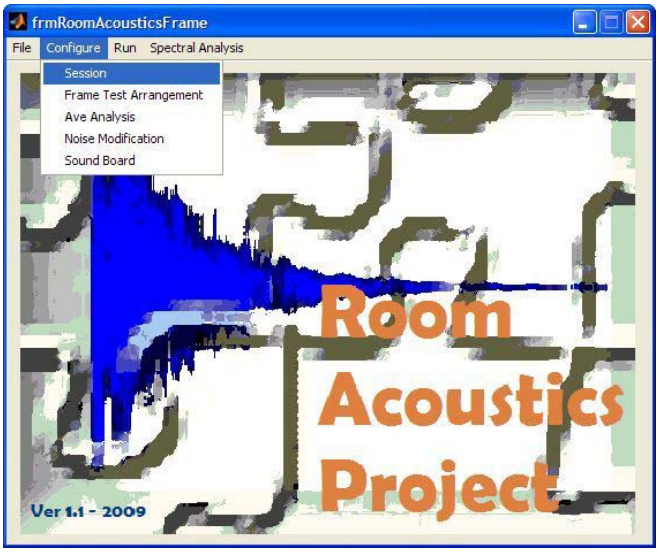
END FOR

- Step 5 % take care of the grains needed to cover the last part of the spectrum. Notice the grains candidates used in Step 2 consist of the frequency band in between two consecutive perceptual tracks in the same zone. The mask effect of the top most perceptual track (higher frequency of each zone) is not used yet.

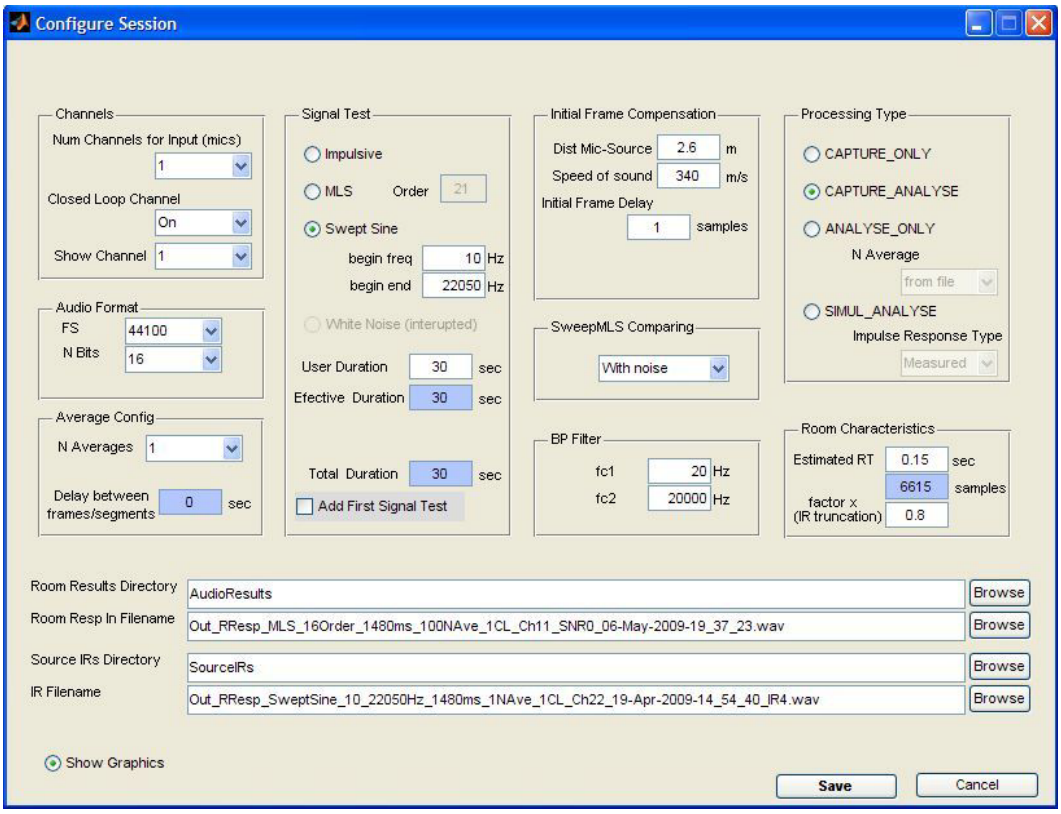
APPENDIX G1.

The educational platform room acoustics project

Main frame. Main desktop from where all the configurations supporting the measurements and analyses are accomplished.

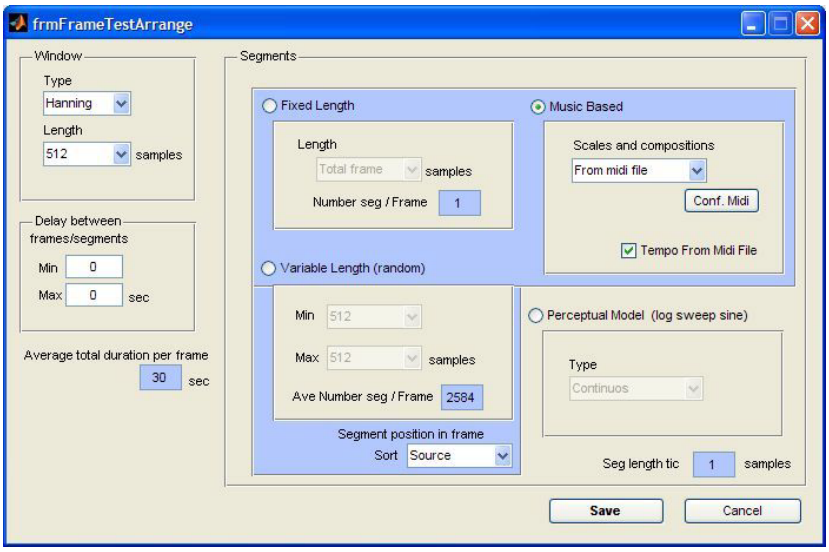


Session configuration module. Configuration of the basic parameters for the measurement session. Makes the configuration of the number of channels for input (microphones), the audio format, number of averages, the type of test signal used for session, basically. Allows selecting the kind of analysis: CAPTURE_ANALYSE makes the full analyses, CAPTURE_ONLY captures the sound only (intended for later processing analyses), ANALYSE_ONLY performs an analyse from a signal previously captured and stored and SIMUL_ANALYSE permits to simulate an acoustical environment from an impulse response stored (synthetic or recorded from a room). The directories for store and load the results are also configured.

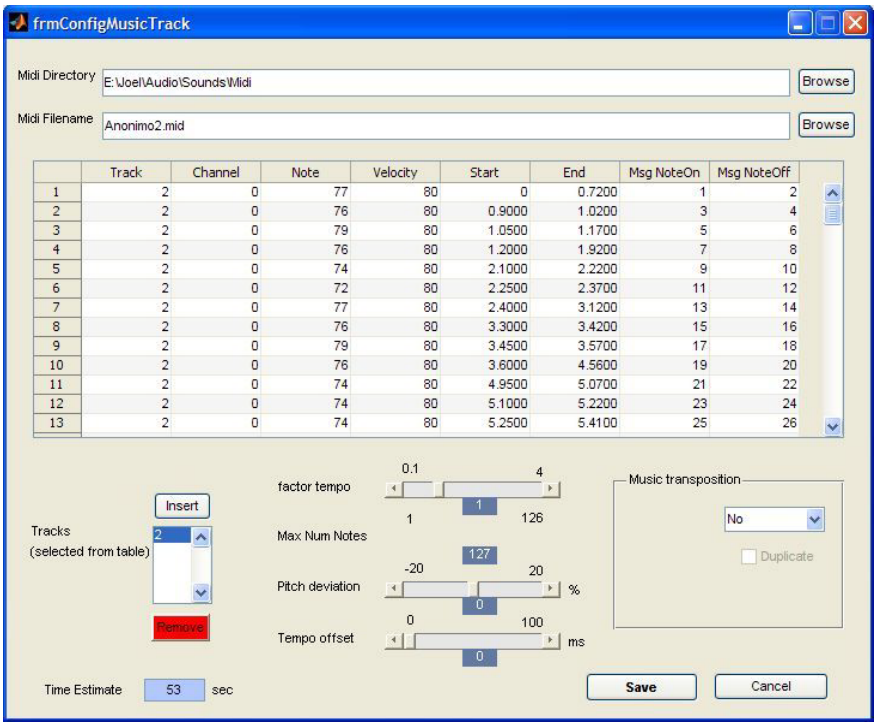


Frame Arrangement Module. Configuration of the frames ensemble patterns allowing the selection of the time delay between frames (or segments), in case of using multiple frames in the acoustical test.

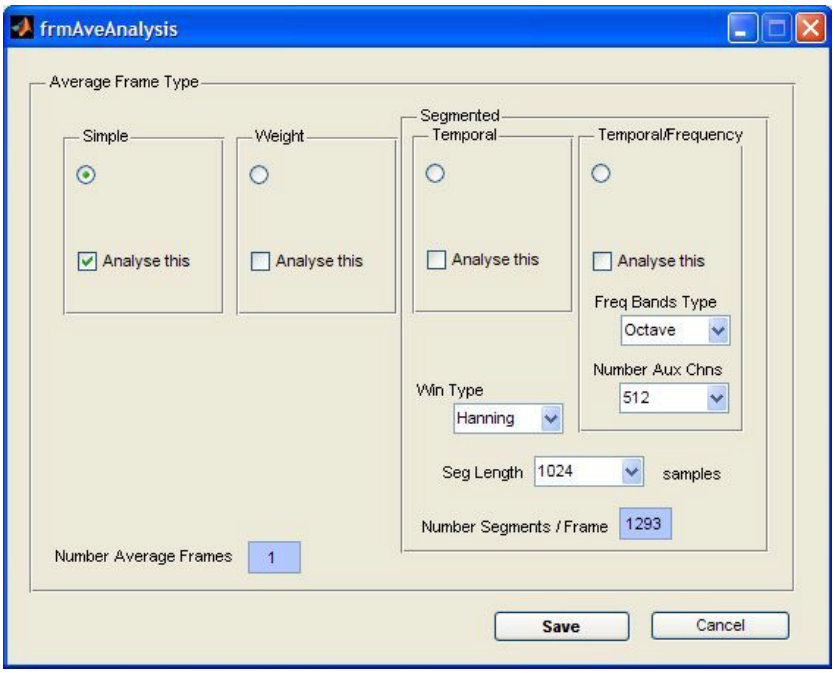
The frames can be segmented before sent to the room and segments with different lengths are possible. Therefore, the techniques based on swept sine grains, developed in Chapter 5, are implemented. Consequently, the segments can be assigned to a tempered musical scale and subsequently, used together with midi files. This configuration has particular interest for creating test signal patterns adapted to music tracks, via midi. The resulting test signal patterns can be used in stand alone or for perceptual masking purposes. The Perceptual model option allows selecting the simultaneously masking effects if noise is added to the swept sine test signal. The masking can be continuous (with no segmentation of the test signal) or by grains.



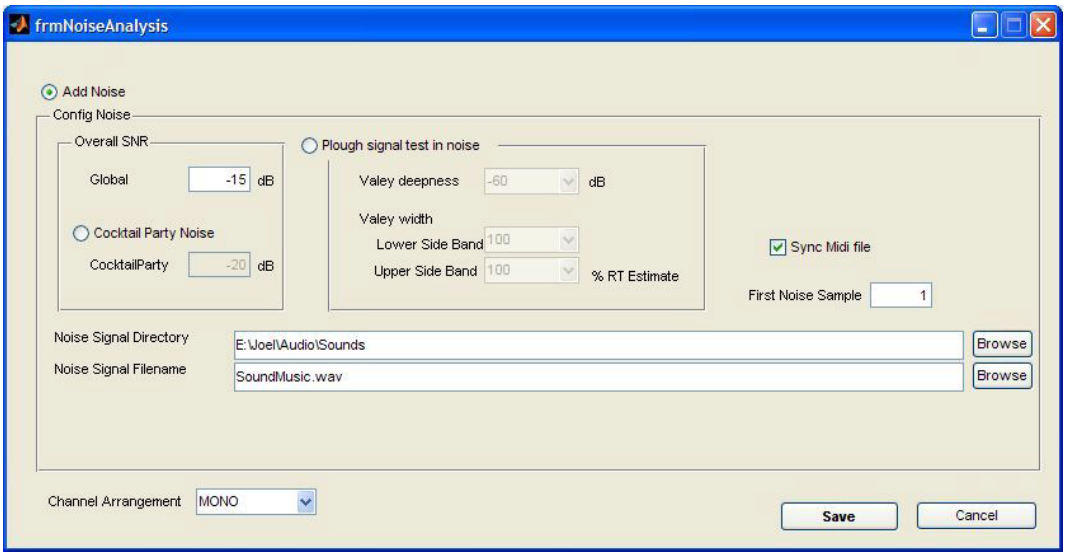
Config Music Track module. This module deals with the parameterization of the midi files. After midi file loading, the table is populated with the different voices (which constitutes the chords of the music). The resulting test signal pattern is built by choosing different combinations of the existing voices building a masker for the test sweep signal. A number of parameters related with the musical notes, such as the *tempo*, maximum number of notes and pitch deviation, are allowed to be adjusted as well as music transpositions.



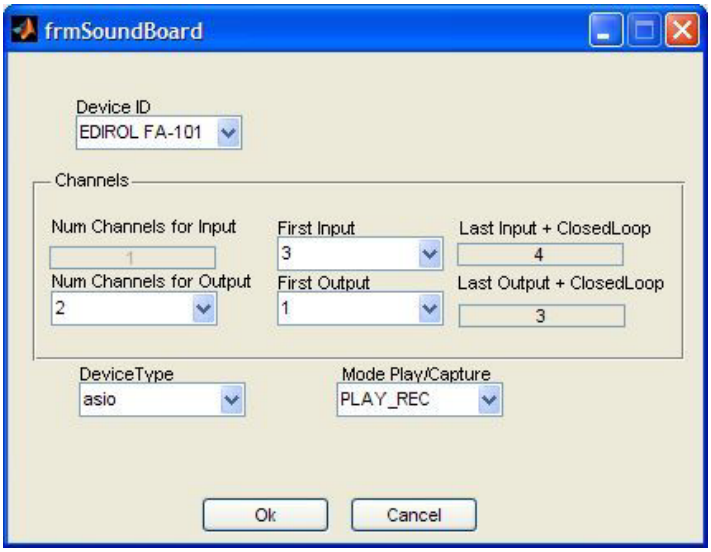
Averaging method selection module. This module implements the averaging techniques mentioned in Chapter 2.5 and 2.6 (Ave and WAve) and developed in Chapter 4.2 and 4.3 (SWave and SSWAve).



Noise signals configuration module. This module permits adding noise to the test signals, before sent to the room. The spectrum of the noise can be modified before added to the test signals selecting the Plough signal test in the noise option. This spectral modification consists of applying a tracking band reject filter to the noise to nest the swept sine signals. This analysis corresponds to the technique TSNM developed in Chapter 6.4, by selecting the Perceptual model option in the **Frame Arrangement Module.**

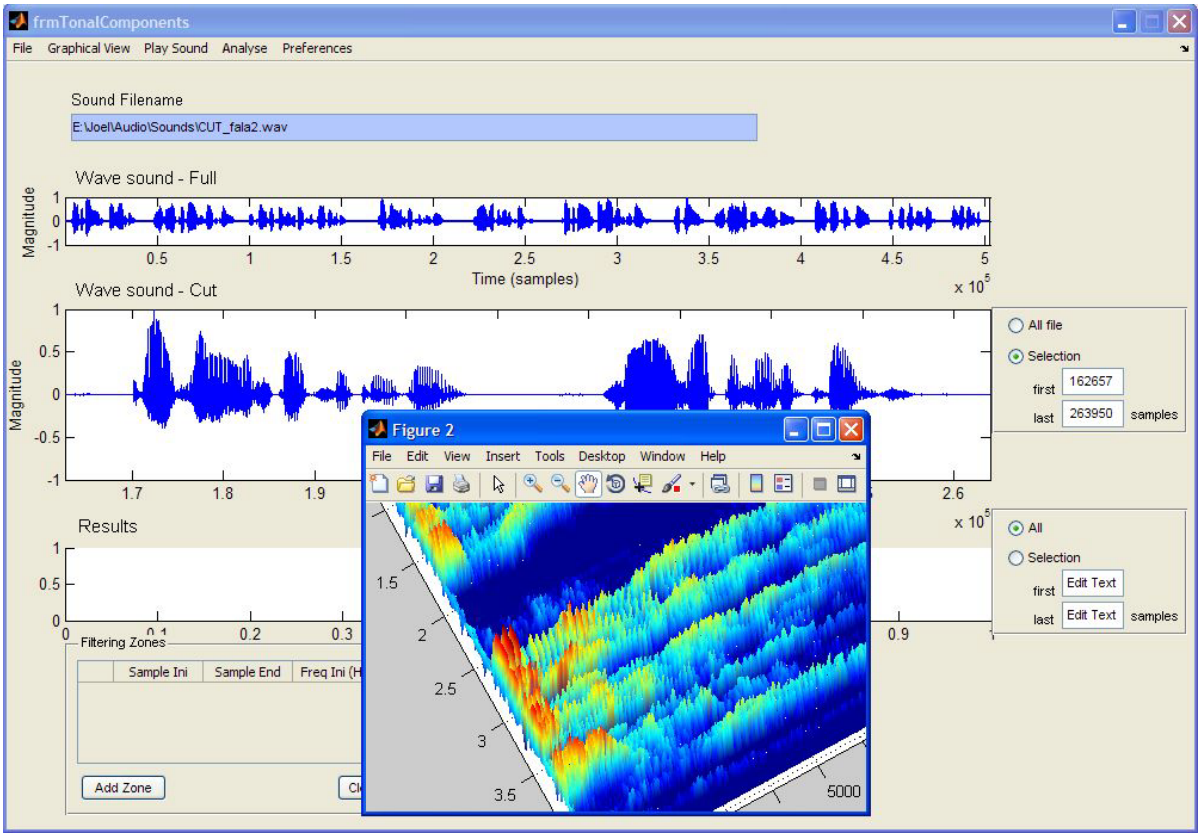


Sound board setup configuration module. Configures the number of input and output audio channels. Selects the sound board and communication audio driver.

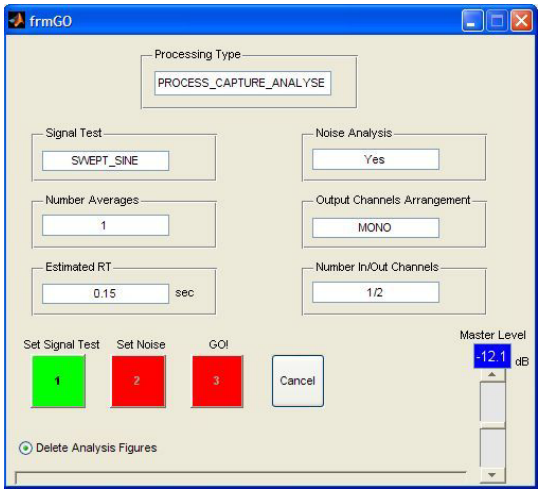


Time/frequency signal analysis module. This module is a sort of audio wave editor. Besides editing the sound waves, in the time and in the frequency domain, this module has some filtering tools used in the perceptual analysis. It possesses a tool for eliminating selected partials components by applying band reject filters. This procedure is simply made by visually selecting an area on the display with the mouse. This procedure is useful to perceptually inspect the degradation applied to the noise (music), especially in the STSNM technique. This module uses also a set of algorithms of analysis/synthesis of the signals and techniques of peak continuation to characterize tonal signals by its basic parameters. Therefore, tonal-like and noise-like parts of the signals are defined. This procedure is important to synthesize some tonal-like parts of music in order to give more room to the swept sine test signals be added to the noise, in the TSNM and STSNM techniques. Thus, less degradation is produced and the SNR can be more efficiently controlled, and therefore, increased.

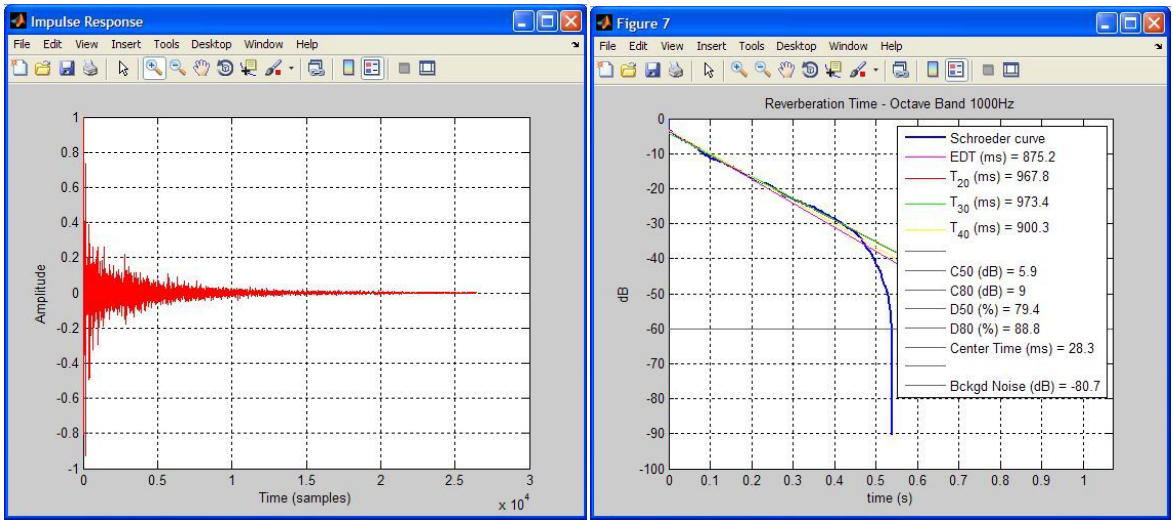
This module is also used to analyse the ultrasound tonal probe technique.



***Resumé* of the measurement setup and start running the analysis**. This is the last window appearing before the acoustical measurement starts. It shows the basic configuration of the session.



Some examples of the results estimated from the Impulse Response (left hand figure) are depicted, such as the Sound Energy Decay curve and the estimated room acoustic parameters (right hand figure).



Note: This set of windows is an overview of the graphical interface of the Room Acoustics Project framework. A User Manual of this platform, describing the procedures for a measurement setup and the techniques implemented, will be designed soon. This tool serves as a framework to demonstrate the acoustical measurement techniques implemented according to my PhD work.