POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Compensation of phase distortion in high-performance audio systems

Tesi di Laurea Magistrale in
Music and Acoustics Engineering - Ingegneria della Musica e dell'Acustica

Author: **Martino Schgor**

Student ID: 945825
Advisor: Prof. Giuseppe Bertuccio
Co-advisors: Marco Tagliaverga
Academic Year: 2020-21

Abstract

Hi-end manufacturers often make use of some quantitative measures such as spectral flatness, harmonic distortion or maximum power as indicative of the quality of their systems. Such features are optimized with extreme care, and usually boasted in advertising as granting the highest possible fidelity in audio reproduction. The phase response of a system (i.e. the distribution of delays which affect each frequency component) is hardly addressed, thought by many to have absolutely no effect on the listening experience. Nonetheless, the scientific literature from the last decades reports several phenomena of audibility of phase distortion, i.e. the waveform distortion due to a non-ideal phase response. Its consequences on the audio experience might be subtle, but cannot be ignored in an environment where the search for high fidelity is pushed to the limit.
The present thesis is meant to analyze the origins of phase distortion in audio systems, together with its perceptual effects. Furthermore, we report some experiments that have been performed in an attempt to correct the phase distortion for higher audio reproduction fidelity, with the use of real-time digital signal processing.

Keywords: Hi-Fi, Loudspeakers, DSP, Audio, Phase distortion

Abstract in lingua italiana

I costruttori di dispositivi Hi-Fi usano spesso delle misure quantitative quali la risposta in frequenza, distorsione armonica o potenza massima, come indici di qualità dei propri prodotti. Queste caratteristiche vengono ottimizzate con cura estrema, e riportate orgogliosamente nelle pubblicità per vantare la miglior fedeltà di riproduzione possibile. La risposta in fase dei sistemi (ovvero la distribuzione dei ritardi a cui è soggetta ciascuna componente spettrale) è pressochè ignorata, poichè molti ritengono che non abbia alcun effetto sull'esperienza di ascolto. Tuttavia, la letteratura scientifica degli ultimi decenni documenta diversi fenomeni di udibilità della distorsione di fase, ossia il cambio nella forma d'onda di un segnale dovuto a una risposta in fase non ideale. Le sue conseguenze sull'ascolto sono marginali, eppure non possono essere ignorate in un ambito dove l'ottimizzazione della fedeltà è portata agli estremi.
La presente tesi si propone come un'analisi delle origini della distorsione di fase in sistemi di riproduzione audio, nonchè dei suoi effetti percettivi. Inoltre, riportiamo degli esperimenti che sono stati eseguiti nel tentativo di correggere la distorsione di fase per una migliore fedeltà di riproduzione audio, usando del processamento di segnali digitali in tempo reale.

Parole chiave: Hi-Fi, Altoparlanti, DSP, Audio, Distorsione di fase

Contents

Acronyms

APF All-Pass Filter. 6–8, 51

APH Alternating Phase. 45

BJT Bipolar Junction Transistor. 11, 12

BM Basilar Membrane. 29

BMFD Binaural Matrix Feature Decoder. 34, 35

BMLD Binaural Masking Level Difference. 33, 34

CPH Cosine Phase. 44

DAC Digital to Analog Converter. 9, 57

DRNL Dual Resonance Non-Linear. 31

DSP Digital Signal Processor. 49, 51, 52, 64

FFT Fast Fourier Transform. 41

FIR Finite Impulse Response. 51, 52

GCC Generalized Cross-Correlation. 59, 60, 62–65

GUI Graphical User Interface. 47, 53, 58, 59

HRTF Head Related Transfer Function. 35

IHC Inner Hair Cells. 29, 31, 44

IIR Infinite Impulse Response. 51, 52, 56, 57

ILD Interaural Level Difference. 33, 35, 36

ITD Interaural Time Difference. 33–36

JFET Junction Field Effect Transistor. 11, 54

LIFO Last In - First Out. 52

LPF Low-Pass Filter. 31, 32

LTI Linear Time-Invariant. 3, 4

OHC Outer Hair Cells. 29

OLA Overlap and Add. 57

OME Outer and Middle Ear. 30, 31

OSD Optimal Source Distribution. 66

PA Power Amplifier. 12, 13

PGA Programmable Gain Amplifier. 53

PWM Pulse Width Modulation. 13

RMS Root Mean Square. 13

SNR Signal to Noise Ratio. 44, 63

SPL Sound Pressure Level. 30, 50

T/S Thiele/Small. 37

TF Transfer Function. 41

THD Total Harmonic Distortion. 23

TM Tympanic Membrane. 27

XTC Cross-Talk Cancellation. 65, 66

Introduction

Aim of the present thesis

This thesis is meant to introduce a scientific approach to the wide and complex topic of phase distortion in high-performance electroacoustic systems. It is a common opinion, and strongly embraced by the author, that the Hi-end world needs more scientific research. It appears in fact, that the people involved in such environment are often easily deceived by subjective bias, commercial advertisement and scientific-sounding statements that are not supported by any quantitative experimental data. Moreover, the concept itself of *high fidelity*, usually regarded as the unique qualitative goal in the design of audio systems, suffers the lack of a clear definition, often leading to the abuse of quality measures.
In particular, phase distortion is generally regarded with extremely variable importance. Some electroacoustic designers say phase distortion is inaudible and thus completely negligible, whereas others pay extreme attention to it, sometimes trading off other objectively important audio quality measures, such as spectral flatness or dynamic range.
The aim of this thesis is to gather information about phase distortion, discussing both its origins and perceptual effects, as well as to offer a solution for the complete control of the phase response of a system. Several models are provided for the definition of the *correct* phase response and their quality is discussed through experimental listening tests.

Applications

The most obvious application for the present thesis is to be a support for electroacoustic engineers, providing data that could be useful in trade-off design choices. Moreover, a precise characterization of phase distortion audibility is strongly needed for the application of linear compression techniques, in which an inaudible phase distortion is purposely introduced to lower the instantaneous energy in peak transients. A great example of such techniques is proposed in [39] where the threshold of audibility is chosen arbitrarily.
The proposed experimental setup can be useful for instant measurement and further experimentation, and could be usefully combined with other audio techniques, such as

audio virtualization systems. Furthermore, it could be used to emulate the phase response of a Hi-end system on a cheaper setup, lowering the (sometimes exceptionally high) prices of such devices or allowing engineers to have more ease in trade-off design processes.

Thesis layout

This thesis is composed of 5 chapters:
The first chapter contains the theoretical background needed for the present discussion, with the necessary mathematical definitions.
The second chapter addresses the causes of phase distortion in audio systems, starting from their working principles and providing some reasonable numerical data obtained by simulation and measurement of real systems.
In the third chapter, a bibliographic research is conducted on the topic of phase audibility, both in monaural (absolute) and binaural (relative) terms. Some psychoacoustical models are explored and presented.
The fourth chapter documents the experimental activities that have been carried out during the development of the present thesis.
Finally, the last chapter reports the experiments that have been performed in the attempt of correcting phase distortion with the use of a digital signal processor.

1 | Theoretical Background

In this chapter, we report some mathematical definitions and properties that are considered necessary for the understanding of the following parts. The author chose to be consistent with the notation from [42].

1.1. Phase distortion and its measures

1.1.1. Linear Time-Invariant systems

Let S be a generic linear, time-invariant (LTI) system with complex frequency response $H(\omega)$ and impulse response $h(t)$. We know from the Fourier Analysis that $H(\omega)$ and $h(t)$ are characterized from each other as follows:

$$H(\omega) = \int_{-\infty}^{\infty} h(t) \cdot e^{-j\omega t} dt \tag{1.1}$$

$$h(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\omega) \cdot e^{j\omega t} d\omega \tag{1.2}$$

Both of these representations carry enough information to describe completely the LTI system. We will focus mainly on the complex frequency response $H(\omega)$, since it is most meaningful for this analysis. As a complex function, it is often expressed in polar form:

$$H(\omega) = |H(\omega)| \cdot e^{\Phi(\omega)} \tag{1.3}$$

Where $|H(\omega)|$ is often referred to as *magnitude response* or *spectrum*, and $\Phi(\omega)$ as phase response.

1.1.2. Distortionless systems

A system is defined *distortionless* when it does not change the waveshape of the signal from its input to its output, introducing at most an amplification factor (which must be

constant with respect to frequency) or a pure time delay. A distortionless system will have an impulse response of the following format:

$$h_{dl}(t) = A \cdot \delta(t - T) \tag{1.4}$$

where $\delta(t)$ is the Dirac pulse distribution, A the amplification factor and T the time delay. In order for the system to be causal (thus practically feasible to implement in real time) T can not be negative. For the scope of this thesis, we will also assume A to be strictly positive. Later in this work, polarity inversion will be taken care of. Substituting (1.4) into (1.1) gives the complex frequency response of a distortionless system:

$$H_{dl}(\omega) = A \cdot e^{-j\omega T} \tag{1.5}$$

The phase response $-j\omega T$ of a distorsionless system must be proportional to the frequency. Any deviation from this linear response is associated to a change in the signal shape and called *phase distortion*[42].

1.1.3. Minimum and Excess Phase

Any LTI system has a minimum amount of phase lag, given its magnitude response, in order to be causal. This component of the phase response is called *minimum phase* and can be extracted from the magnitude response by applying the Hilbert transform relation: (proof in [45])

$$\Phi_m(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\log |H(\omega')|}{\omega' - \omega} d\omega' \tag{1.6}$$

Conversely, there can be a deviation from the minimum phase in the phase response of a system. This is called *excess phase* and defined as:

$$\Phi_x(\omega) = \Phi(\omega) - \Phi_m(\omega); \tag{1.7}$$

The excess phase can be rewritten in the form shown below:

$$\Phi_x(\omega) = \Phi_a(\omega) - \omega T + \Phi_0 \tag{1.8}$$

where Φ_0 is a constant and $\Phi_a(0) = 0$. Φ_0 is associated to a polarity reversal or the presence of a Hilbert transformer. We will exclude the presence of a Hilbert transformer because there is no physical implementation of it in audio systems. For the scope of this

work, Φ_0 is only associated to a polarity reversal and can only assume values of integer multiples of π (in radians). $\Phi_a(\omega)$ is referred to as the *all-pass component* of the excess phase [42].

1.1.4. Phase and group delay

Two useful definitions are provided: namely of the phase delay and group delay.

$$\tau_p(\omega) = -\frac{\Phi(\omega)}{\omega} \tag{1.9}$$

$$\tau_g(\omega) = -\frac{\partial\Phi(\omega)}{\partial\omega} \tag{1.10}$$

It can be easily demonstrated that in order to avoid wave shape distortion, it must be:

$$\tau_p(\omega) = \tau_g(\omega) = T \quad \forall\omega \tag{1.11}$$

Similarly to the phase distortion, these measures are more useful when considered as deviation from the same-delay distortionless system. Hence we define *phase delay distortion* and *group delay distortion* the following:

$$\Delta\tau_p(\omega) = \tau_p(\omega) - T \tag{1.12}$$

$$\Delta\tau_g(\omega) = \tau_g(\omega) - T \tag{1.13}$$

Several sources, among which [15, 27, 35, 42], state that group delay distortion is the most suitable indicator to describe numerically the greatness of phase distortion phenomena. Due to linearity of the differentiation operator, the following equations hold:

$$\tau_g(\omega) = T - \frac{\partial\Phi_m(\omega)}{\partial\omega} - \frac{\partial\Phi_a(\omega)}{\partial\omega} \tag{1.14}$$

$$\tau_g(\omega) = T + \tau_{gm}(\omega) + \tau_{ga}(\omega) \tag{1.15}$$

$$\Delta\tau_g(\omega) = \tau_{gm}(\omega) + \tau_{ga}(\omega) \tag{1.16}$$

where τ_{gm} and τ_{ga} are namely the *minimum group delay* and the *all-pass group delay*, while Phi_m and Phi_a are the *minimum* and *all-pass* phase, defined in equations 1.6 and 1.8.

1.2. All-Pass Filters

1.2.1. Definition

An *All-Pass Filter* (APF) is a linear time-invariant system whose transfer function has constant magnitude at all frequencies, while the phase response may be varying. Its name is chosen in analogy with the more commonly used "Low-Pass" and "High-Pass" filters, and it suggests that no frequency component is dimmed out. By definition, its transfer function is of the form:

$$H_{ap}(\omega) = A \cdot e^{j\Phi(\omega)} \tag{1.17}$$

where A is a constant gain, and $\Phi(\omega)$ the phase response. When the absolute gain of the filter is not interesting for the scope, we can simplify the definition with a *unitary gain* filter and set $A = 1$.

1.2.2. Analog All Pass Filters

In the s complex plane, All-Pass Filters are characterized by complex conjugate zero-pole couples. As shown in [44], the Laplace representation of a first-order APF transfer function is of the form:

$$H_{ap}(s) = e^{j\Phi_0} \cdot \frac{s + p^*}{s - p} \tag{1.18}$$

Where Φ_0 is an arbitrary constant phase shift.
It can be shown that this transfer function has unity gain by evaluating it in $s = j\omega$ as follows:

$$H(j\omega) = e^{j\Phi_0} \cdot \frac{j\omega + p^*}{j\omega - p} = e^{j\Phi_0} \cdot \frac{(-(j\omega - p))^*}{j\omega - p} \tag{1.19}$$

$$|H(j\omega)| = |e^{j\Phi_0}| \cdot |\frac{(j\omega - p)^*}{j\omega - p}| = 1 \cdot 1 \tag{1.20}$$

For *real* first-order APFs, complex poles are in conjugate pairs, so the transfer function can be simplified into:

$$H(s) = \pm\frac{s + p}{s - p} \tag{1.21}$$

More generally, any real arbitrary-order APF can be expressed as a cascade of first-order APFs, thus:

$$H_{ap}(s) = \pm\frac{(s + p_1)(s + p_2) \cdots (s + p_N)}{(s - p_1)(s - p_2) \cdots (s - p_N)} \tag{1.22}$$

Analog all-pass filters are used in electronics (mostly for communications), combined with other filters to linearize their phase response into some group delay specifications.

An example of a circuital implementation is provided is figure 1.1

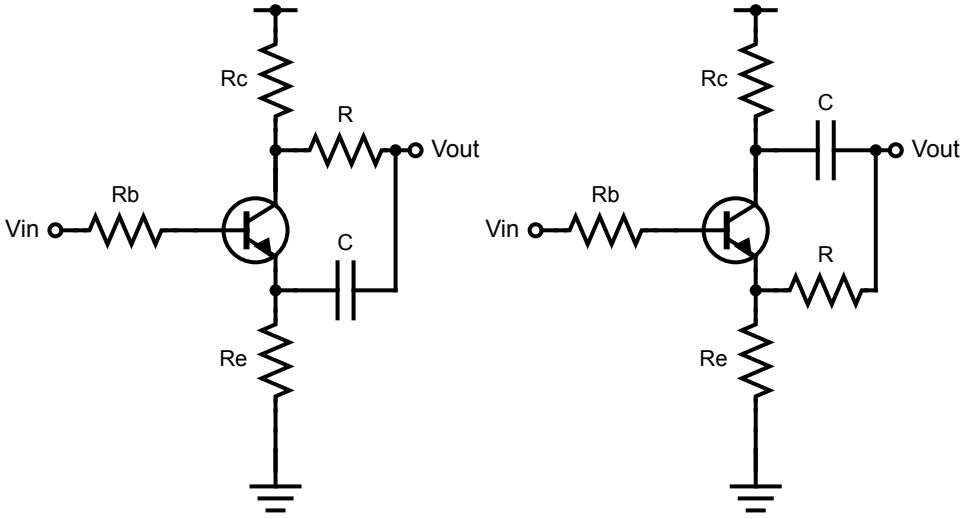

Figure 1.1: Examples of active analog implementations for first-order All-Pass filters, in phase-lead (left) and phase-lag (right) configurations. [49]
Both circuits have a single pole at $f_p = 1/2\pi RC$

.

1.2.3. z-domain representation

The transfer function of the generic APF in z-domain can be obtained by the Laplace domain with a bilinear transform. A first-order APF with a single pole in z_0 will be represented as:

$$H(z) = \frac{z^{-1} - z_0^*}{1 - z_0 z^{-1}} \tag{1.23}$$

It is interesting to notice that the transfer function shows a zero in $1/z_0^*$. It means that a generic APF, which can be considered the cascade of multiple single-pole APFs, features zero-pole pairs in conjugate-reciprocal positions. From a graphical point of view, it is more intuitive to consider that zeros are located in the mirror-image location of their related pole, with respect to the unit circle. An example is provided in figure 1.2.

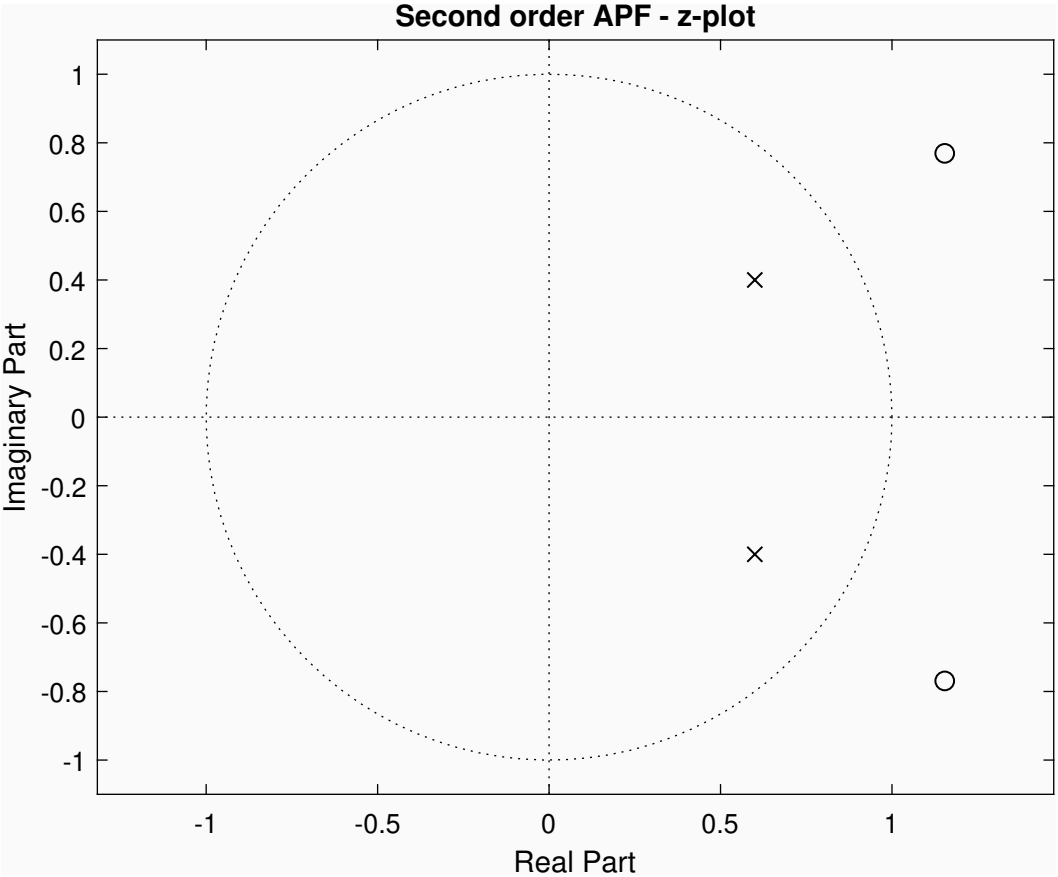

Figure 1.2: The z-plot of a second-order APF. The filter is real because all the poles and zeros appear in complex conjugate pairs. Each zero is located in the mirrored position of a pole with respect to the unitary circle. The filter is stable because all the poles fall inside the unit circle.

2 | Causes of phase distortion

In this chapter, an analysis is performed to explore the origin of phase distortion in electroacoustic systems. The working principles of the most common audio systems are reported, together with some useful modeling methods and a discussion on the non-idealities.

2.1. Working principles of audio systems

Most audio systems start their signal path by fetching a source into an analog electric signal. The signal power is amplified by a pre-amplifier and a power amplifier, then fed to the loudspeakers to produce a controlled air pressure at their diaphragm.

2.1.1. Source

The musical signal must be retrieved from a source. Digital sources include CDs or DVDs, online music services, audio files stored in internal memory or in a USB drive. Other kinds of sources bypass digitalization, by using non-digital storage supports, such as vinyl disks or magnetic tape devices. Most sources output an analog signal, either from a pickup or a DAC (Digital-to-Analog Converter). Non-idealities from this process are usually negligible with respect to the others, thus we will assume the analog signal as perfectly matching the recorded information.

2.1.2. Pre-amplifier

The pre-amplifier is usually an electronic circuit whose aim is to increase the amplitude of a voltage signal, with an ideally linear transfer function and with voltage gain greater than unity. It is relatively easy to obtain a good linearity thanks to the usage of feedback circuits. The most used circuit configurations are described.

Vacuum triodes are the simplest vacuum tubes. Consisting in a glass bulb with a strongly pumped vacuum inside, they feature at least three electrodes: the *cathode* (K) is

the innermost electrode, charged negatively and electrically heated by a tungsten filament (HH). The cathode shows thermoionic effect, by emitting electrons due to the high temperature. The electrons are collected at the outermost electrode, the *anode* (A), usually polarized positively with hundreds of Volts with respect to the cathode. The flow of negative charges from cathode to anode results in an electrical current from anode to cathode, called anodic current and measuring around 1mA for most audio preamplifying triodes. A third electrode called *control grid* (G) has the form of a grid and is placed between the others, near to the cathode. Any difference of potential between the control grid and the cathode, generates an electric field that controls the flux of electrons. In figure 2.1 is reported the most used circuital topology for a triode preamplifier, the *common cathode configuration.*

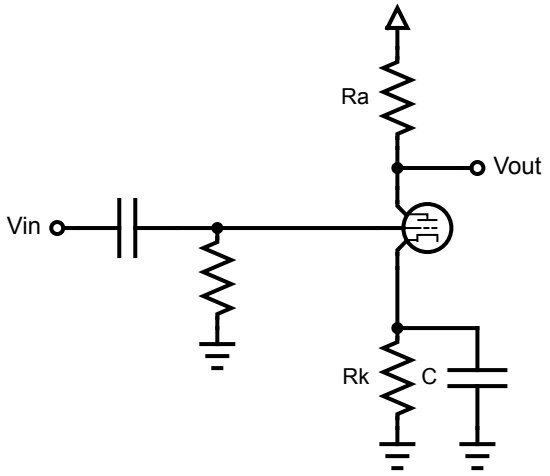

Figure 2.1: Common cathode triode preamplifier

The input is AC-coupled through a suitably designed CR network, resistor R_k and its parallel bypass capacitor are used to bias the triode into an approximately linear working region. Note that the bias potential of the control grid is usually lower than the cathode. The resistor R_a is used to reconvert the anodic current into a voltage signal, thus it is crucial for the gain of the amplifying stage. The output is also usually AC-coupled, given that the anodic voltages are very high.
Vacuum triode amplifiers can reach great linearity and in general good audio quality. Moreover, they feature a very high input impedance, that is useful to avoid voltage partitions between stages, causing an unwanted reduction in gain or a frequency coloration. On the other hand, they feature also a quite high output impedance, making them unsuitable for low-impedance loads, and require a great amount of power for the heater, introducing more requirements on the power supply and the thermal behaviour of the bulb itself.

Solid-state transistors. Invented in 1947 at the Bell Telephone Laboratories, the transistor is one of the most revolutionary devices in the last century. There are several kinds of transistors of which the most used in audio preamplifiers are the JFET (Junction Field Effect Transistor) and the BJT (Bipolar Junction Transistor). They have in common being made on a semiconducting substrate, by doping contiguous regions with opposite polarity. The main channel of a BJT is thus a series of two bipolar junctions, one of which is reversely biased. The most common circuit, in strong analogy with the previous, is reported in figure 2.2. It is called *common emitter* when used with a BJT or *common source* in the case of a Field Effect Transistor (FET).

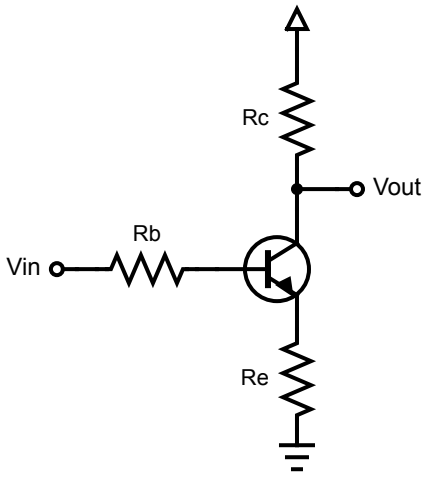

Figure 2.2: Common Emitter Amplifier with a NPN Bipolar-Junction Transistor

In this circuit, the resistance R_e is introduced to generate a negative feedback in order to improve the linearity of the amplifier, at the cost of reducing the overall gain and the output dynamic range.
Transistors have the advantage of working properly with low voltages, furthermore, their power consumption is very low, their package can be small and lightweight, and the common circuits are extremely simple. Their main drawback is the non-linearity of the transfer function, which can be improved by negative feedback loops.

Operational amplifiers are integrated circuits that work as high-gain differential amplifiers when used in open-loop configuration. They are specifically thought to work in a feedback loop with a resistor divider, exploiting its high linearity. Figure 2.3 shows the most common circuits. Op-amps are usually not well seen in the Hi-Fi industry, as they were historically designed with requirements for big volumes of production, low cost, low power consumption, small package, optimization of production. Only recently the big semiconductor industries are designing audiophile-grade operational amplifiers [23, 55]

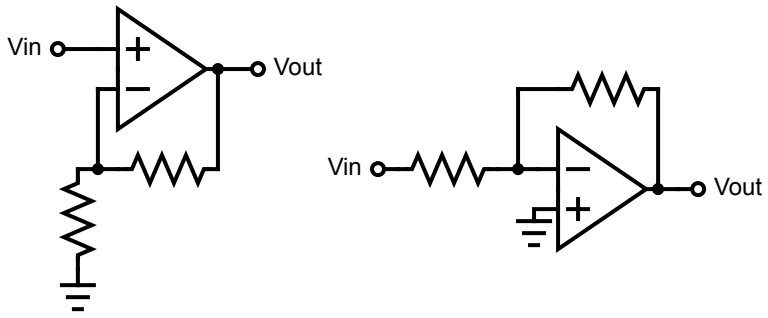

Figure 2.3: The non-inverting (left) and inverting (right) op-amp circuits

2.1.3. Power amplifiers

Power amplifiers, or PAs, are used to feed the loudspeakers with the high current they need to work properly. They are classified into groups based on their functioning principle. [12]

Penthodes are thermoionic tubes like the triodes, but designed for high power. As suggested by the name, penthodes have more electrodes than triodes (namely the *screen* and *suppressor*) to prevent unwanted secondary emission of electrons from the anode back to the cathode. This effect is due to the high current flow, so it is negligible for low-power triodes. The smallest penthodes have nominal anodic current around 10mA, and anode voltage still in the range of hundreds of Volts. Since the output impedance is too high for a typical loudspeaker, a transformer is added between the valve stage and the load.
Penthodes can be used in common cathode configuration, to achieve a small positive voltage gain or as voltage followers. In this latter configuration, the voltage gain is approximately unitary but the amplifier works as an active impedance adapter, supplying enough current to drive the load.
Penthodes share the same advantages and drawbacks as triodes, having a terribly low power efficiency. Moreover, the follower configuration needs a high-voltage input signal and proper polarization.

Mosfet Class A - AB - G Class A stages feature a single transistor (a power Mosfet or BJT) in gate-follower configuration. The stage has voltage gain slightly below unity but can have a huge current gain, so it is suitable for load impedance matching. It is considered the best sounding PA configuration but it has the disadvantage of having to dissipate a considerable part of the supplied power into heat, because of the constant DC current component flowing in the transistor to keep its bias. Not only this affects the power consumption, but also requires to handle the heat dissipation with heavy heat

sinks. More efficient cooling methods involving moving parts (fans, fluid coolers, etc...) are not used in amplifiers because of their acoustic noise.
Class AB PAs feature a couple of power transistors. This configuration has better power efficiency than the class A because the bias current is strongly reduced. However, in practical applications, the thermal issue is still present and worsened by the need of a high supply voltage for the accurate reproduction of high-energy transients.
Class G power stages work by selecting the supply voltage among a set of different rails. Since musical signals have a high crest factor [1], the higher voltage rails are activated only during high-energy transients. This allows to save on power consumption during the majority of time.

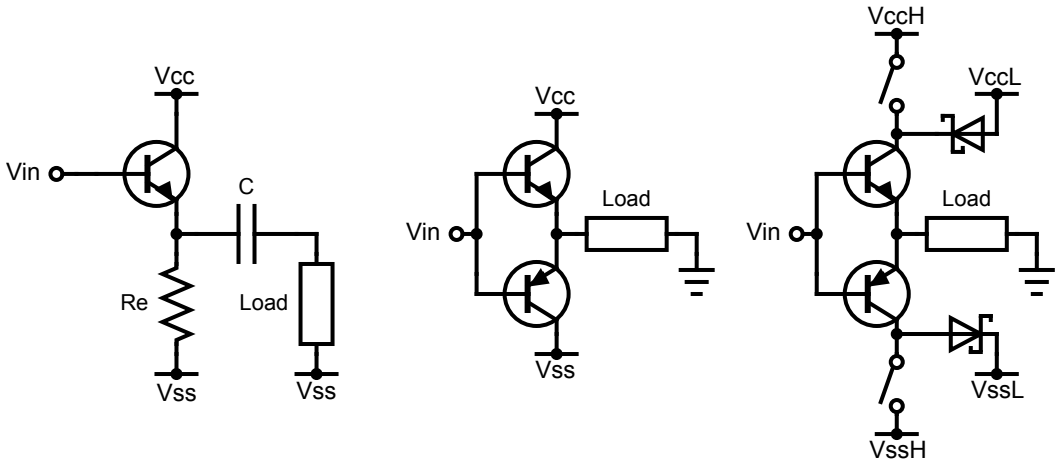

Figure 2.4: Simplified circuits of class A (left), AB (center), G (right), power amplifiers. The switches in class G are actuated by a peak detecting circuit

Class D Class D power amplifiers perform a Pulse Width Modulation (PWM) of a square carrier wave at high frequency (usually 200KHz). The modulated wave is then amplified by a power stage and then demodulated back into audio band by a LC Low-Pass Filter. This configuration allows the active elements of the output stage to work at the extremes of its dynamic range, reaching a high power efficiency. [18]

2.1.4. Crossover filters

Crossover filters split the signal in usually 2 or 3 bands when the loudspeaker system is powered by multiple loudspeakers, working on distinct frequency ranges. Loudspeakers are classified as: subwoofers, woofers, mid-range, tweeters and super-tweeters depending on

[1] The crest factor of a signal is defined as the ratio between the peak amplitude and the RMS value. It is an adimensional number, always greater or equal to 1 and often expressed in dB scale, that provides a relative measure of the strength of the peaks [46].

their optimal working frequency band. It is common to use different kinds of loudspeakers to cover the full range and the crossover is supposed to feed them correctly. Passive crossovers are usually analog high-current filters, mostly Butterworth filters of the first to third order, in Cauer topology[54].
Other auxiliary functions may be inserted in the same circuit, mostly for compensating room acoustics or electrical properties of the driver:

- the crossover may show an L-Pad [2] resistive divider for tweeter attenuation in particularly reverberant rooms, where the high frequency sound components are perceived stronger than the low frequency ones[50].
- a notch LCR filter can be inserted in the circuit to compensate for acoustic resonances in the space[50].
- a Zobel network can be used for equalizing the electrical impedance of a loudspeaker (always inductive because of the voice coil being an inductor)[37].

2.1.5. Loudspeakers

Loudspeakers are responsible for the electro-mechanical and mechano-acoustical transduction of power. They mostly consist in a coil made of electric wire, the *voice coil*, left free to move axially in the air gap of a ferromagnetic nucleus. This nucleus is magnetized by a strong permanent magnet. The build geometry is such that the voice coil is always immersed in a radial magnetic field. Electric current flowing through the voice coil subjects it to the Lorentz force, and this force is mechanically transferred to a diaphragm.
Figure 2.5 shows the working parts of a loudspeaker.

The magnetic circuit is used to maximize the flux of magnetic field at the voice coil. It is made of ferromagnetic material, usually steel, and features a permanent magnet to be always magnetized. Most of the magnetomotive force drops at the inevitable air gap where the coil is allowed to move, since its reluctance is considerably higher than the rest of the ferromagnetic circuit. Some tweeters use *ferrofluid* (a ferromagnetic fluid made with a suspension of iron particles in mineral oil) to fill the air gap and maximize the flux of magnetic field, however, this technique is rejected by Hi-Fi enthusiasts, because it introduces the non-linear viscous friction of oil. Such viscosity rises in time, making the sound inconsistent over the years.

[2] An L-Pad attenuator is the combination of a series resistor and a parallel resistor. It has the duplex effect of lowering the signal voltage, due to the resistive partition, and of raising the series resistance seen at the load.

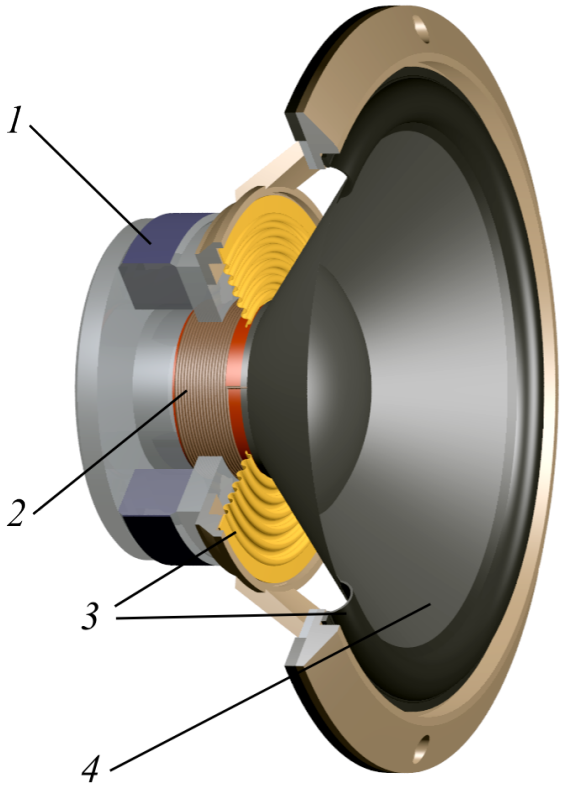

Figure 2.5: Loudspeaker parts:
1) Magnetic circuit
2) Voice coil
3) Suspension
4) Cone
Image by Svjo - Own work, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=30341185

The voice coil is made of copper wire to minimize the electrical resistance. Its thickness is kept as low as possible to allow the paramagnetic air gap to be narrow. For high-power loudspeakers, further attention must be paid to its thermal dissipation since the power absorbed by its parasitic resistance is converted into heat and high temperature could modify important parameters of the loudspeaker or damage it permanently.

The spider is a suspension that forces the movement of the coil to be axial. With too many degrees of freedom, the coil could touch the ferromagnetic nucleus, resulting in (non-linear) sliding friction.

The diaphragm, also called *cone* for its conical shape, provides the surface needed for the interaction with the air. It has the conflicting requirements of being extremely rigid and lightweight, so various materials are used for its construction, from paper to Kevlar or carbon fiber.

2.1.6. Cabinets

Loudspeaker enclosures have multiple functions, from physically sustaining the speaker, to hiding the circuits from sight. But most importantly, they are crucial for the correct acoustic loading of the loudspeakers. Here is given a description of the most common loading techniques.

Open Baffle loudspeaker systems feature only the front face of an enclosure, called *baffle*. This kind of loading allows the parasitic radiation from the inner surface of the loudspeakers to propagate in the listening room. This configuration is highly sensitive to frequency colorations due to the interference of this parasitic source with the main source of acoustic field. Open Baffle systems change sensibly their characteristics with small variations in their location in the room. They also have the drawback of requiring a considerable amount of space in the listening room.

Sealed Box. This is the most common configuration, featuring a sealed box to enclose the parasitic radiation from the inner surface of the diaphragm. This cabinet allows great control on many characteristics of the system, such as frequency response, directional properties (the sealed box is the most similar to the abstract model of a pulsating sphere, thus highly omnidirectional). Being the most insensitive to the presence of other objects around it, the sealed box is suitable for bookshelf positioning. The most intuitive drawback is the creation of an air cavity, that introduces an acoustic elastance in the radiation

impedance of the speaker. High attention must be paid to this kind of reactive components, since they interact with the natural resonances of the driver. In particular, if the box is too small (thus increasing the unwanted elastance) the overall frequency response shows a drop at low frequency.

Vented Box systems make use of a box with one or more openings, called *ports*, to introduce acoustic resonances, in order to compensate for the previously described drop in frequency response. *Reflex* ports, for example, are Helmholtz resonators that can be tuned to the frequency where the magnitude response starts to roll off. Subwoofers working in a limited frequency range can be loaded with a *transmission line port* to achieve a phase reversal in the parasitic radiation. Vented boxes are generally smaller than closed ones, but are more sensitive to the presence of objects near the port and have radiation patterns that depend on the interference between the cone and the port. Moreover, at low frequency, the flow of air could transition from laminar to turbulent, generating white noise.

Passive Radiators are used to substitute acoustic impedances with mechanical ones, requiring less space. They are usually tunable, with the possibility of adding metal washers as mass, and their only drawback with respect to their acoustic counterpart is the cost. Passive radiators solve completely the issue of air turbulence in long reflex ports.

2.2. Equivalent circuit model

Linear electric dipoles are entirely characterized by a relationship between the difference of electric potential, or voltage (v), across their terminals and the intensity of electric current (i) flowing through them. Similarly, linear mechanical entities can be fully described by a relationship between the sum of forces (f) applied to them and their instant velocity (u). Finally, linear acoustic elements are characterized by a relationship between the air pressure at their interface (p) and the volume velocity (U) of the air at their interface. Such characteristic relationships are formalized in linear differential equation, of the first order for our simple case.

2.2.1. Differential Equations

Electrical domain is related to voltage and current in electric dipoles.
Electrical resistance is governed by Ohm's Law of conductors:

$$v = R_e \cdot i \tag{2.1}$$

where the constant R_e represents the electrical resistance of the dipole.
Capacitance is modelled through the following equivalent differential equations:

$$v = \frac{1}{C_e} \cdot \int_0^t i dt \tag{2.2}$$

$$i = C_e \cdot \frac{dv}{dt} \tag{2.3}$$

where C_e is the capacitance value.
Inductance follows the equivalent equations:

$$v = L_e \cdot \frac{di}{dt} \tag{2.4}$$

$$i = \frac{1}{L_e} \cdot \int_0^t v dt \tag{2.5}$$

where L_e is the inductance of the dipole.

In mechanical domain the measures of interest are the force and the instant velocity. The mechanical resistance R_m, related to mechanical damping effects or any linear dissipative component, is formalized as follows:

$$f = R_m \cdot u \tag{2.6}$$

The springs, or more generally, the elastic behaviours, follow the equations:

$$f = \frac{1}{C_m} \cdot \int_0^t u dt \tag{2.7}$$

$$u = C_m \cdot \frac{df}{dt} \tag{2.8}$$

where the quantity C_m is the mechanical compliance, or the inverse of the elastic constant. Mechanical masses (M_m) can be modelled through the following equations:

$$f = M_m \cdot \frac{du}{dt} \tag{2.9}$$

$$u = \frac{1}{M_m} \cdot \int_0^t f dt \tag{2.10}$$

Acoustic domain. The acoustic resistance (R_a) is related to an acoustic power transfer, be it caused by either a dissipative phenomenon or a far-field radiation. This is modelled as:

$$p = R_a \cdot U \tag{2.11}$$

Elastic compression (approximated linear for SPL much lower than the atmospheric pressure, so for any SPL of interest in music reproduction) can be described by the following set of equivalent equations:

$$p = \frac{1}{C_a} \cdot \int_0^t U dt \tag{2.12}$$

$$U = C_a \cdot \frac{dp}{dt} \tag{2.13}$$

where C_a is the value of the acoustic compliance.
Finally, inertia effects are taken into account with the following differential equations:

$$p = M_a \cdot \frac{dU}{dt} \tag{2.14}$$

$$U = \frac{1}{M_a} \cdot \int_0^t p dt \tag{2.15}$$

where M_a represents the acoustic mass of the air system.

Transduction. The interface between the domains works with a set of two equations, each representing a proportionality among the Kirchhoff variables: for the electromechanical transduction holds:

$$v = Bl \cdot u \tag{2.16}$$

$$f = Bl \cdot i \tag{2.17}$$

$$i = \frac{f}{Bl} \tag{2.18}$$

$$u = \frac{v}{Bl} \tag{2.19}$$

where Bl is the force factor of the loudspeaker. Equation 2.16 is related to the Faraday-Lenz law applied to the loudspeaker build, while equation 2.17 comes from the definition of the Lorentz force. The other equations are just the reformulations of these.
Regarding the mechano-acoustic transduction, the equations are:

$$f = S_d \cdot p \tag{2.20}$$

$$U = S_d \cdot u \tag{2.21}$$

$$p = \frac{f}{S_d} \tag{2.22}$$

$$u = \frac{U}{S_D} \tag{2.23}$$

Where S_d is the (effective) surface area of the diaphragm.
These sets of equations allow to model the transduction process with electrical transformers or gyrators, having the primary and secondary sides in distinct domains.

2.2.2. Loudspeaker model

A loudspeaker is undoubtedly a complex machine, having to interact with three distinct domains. However, given all the previous inter-domain similarities, it is possible to model each component in a single domain, highlighting the interaction among the parts. It is a very common practice to represent all the mechanical and acoustic component in an electric circuit as follows:

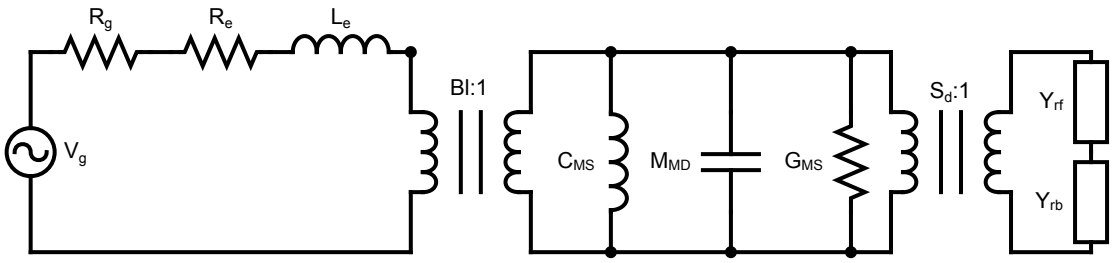

Figure 2.6: *Impedance analogy* configuration

This circuit is extremely useful for the understanding and simulation of the linear behavior of a loudspeaker system. The three main visible sections represent the three physical domains and the interaction between them is correctly modeled. The Power Amplifier is reduced to its Thevenin equivalent circuit and represented by V_g and R_g and the resistive-inductive impedance of the voice coil by the series of R_e and L_e. In the mechanical section, the mass of the moving parts is taken account of via the capacitance M_{MD}, the suspension compliance via the inductance C_{MS} and the mechanical damping conductance

via the resistance G_{MS}. Finally, the acoustic admittance at the front and back surface of the diaphragm are modelled as generic impedance dipoles Y_{rf} and Y_{rb}.
The circuit is then simplified as shown in figure 2.7 by referring all loads of the transformers to their primary side:

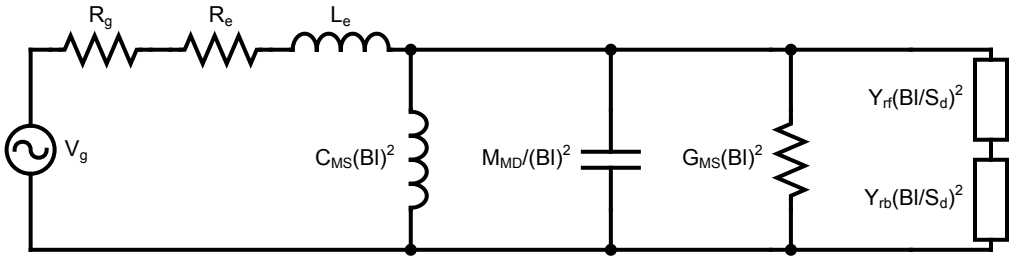

Figure 2.7: Impedance analogy configuration with all the loads referred in electrical domain

2.2.3. Acoustic Impedance types

Small cavities Small cavities (the approximation holds for linear dimensions much smaller than the wavelength of interest) can be considered as a lumped-parameter reactive element in acoustic domain. The admittance of the small cavity can be approximated as a pure acoustic compliance, depending only on the internal volume of the cavity and some constants related to the conditions of the air. The acoustic compliance is:

$$C_a = \frac{V}{\rho_0 c^2} \tag{2.24}$$

with ρ_0 the atmospheric density and c the speed of sound.
The study of cavities is particularly important for the design of sealed enclosures, because such boxes can be considered small cavities for the most of the audible spectrum.

Air masses Short volumes of unconstrained air (i.e. where the particle velocity can be considered constant) can be modeled as lumped acoustic masses. The acoustic inertance of a volume characterized by length L (parallel to the considered direction of motion) and normal area A is:

$$M_a = \frac{\rho L}{A} \tag{2.25}$$

This formulation is important in the design of ported enclosures, as the air in the ports can be approximated as a pure acoustic inertance for low frequencies.

Acoustic resistance Acoustic resistance is related to the acoustic power transfer. Damping elements can be modeled as resistances because they absorb power and convert it into heat, but also the free-space radiation introduces a resistive component in the acoustic impedance seen at the source, because of the power being radiated. Radiation resistance is the parameter we are most interested in, as it plays a crucial role for the correct power transfer from the system.

Piston impedance Particularly interesting for audio application is the study of the *Radiation of a Circular Piston moving axially in an Infinite Rigid Baffle*, as this is a first analytic approximation of the (front surface) radiation of a loudspeaker. The analytic formulation is well explained in [28], here is only reported the resulting formula:

$$Z(ka) = \pi a^2 \rho_0 c[R_1(ka) + jX_1(ka)] \tag{2.26}$$

$$R_1(ka) = 1 - \frac{2J_1(2ka)}{2ka} \tag{2.27}$$

$$X_1(ka) = \frac{2H_1(2ka)}{2ka} \tag{2.28}$$

Where $k = \omega/c$ is the wave propagation constant, a the radius of the piston, J_1 and H_1 namely the Bessel and Struve functions of first order and first kind.

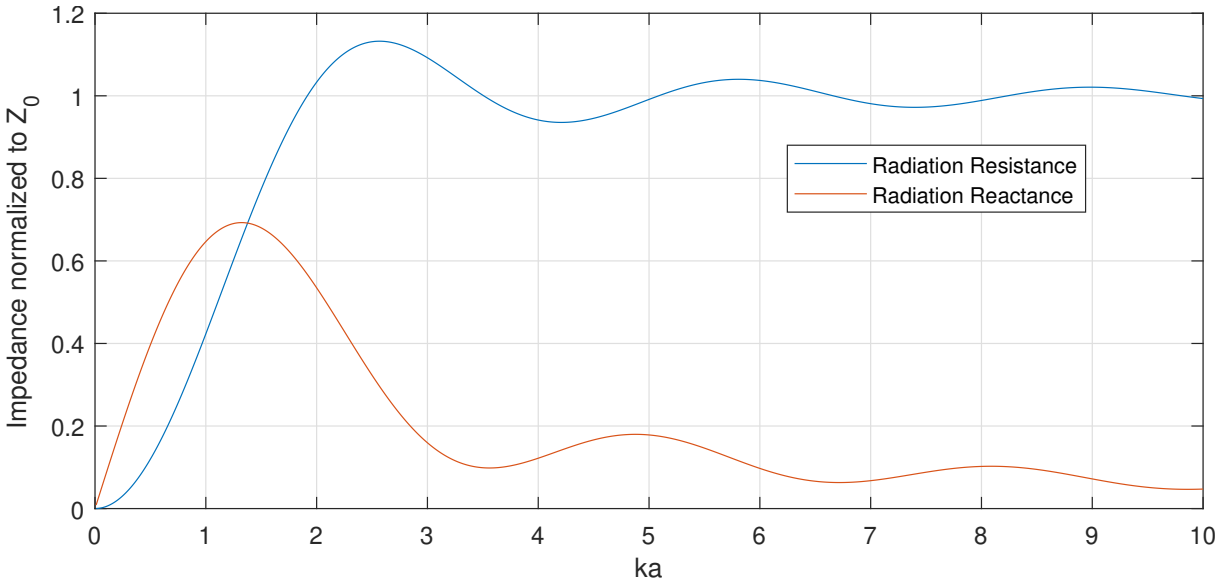

Figure 2.8: Normalized Radiation Impedance of a circular piston radiating from an infinite rigid baffle.

Figure 2.8 shows the variation of R_1 and X_1 with frequency. The values of impedance are normalized to the characteristic acoustic impedance of air $Z_0 = \rho c$. For low frequency

($ka << 1$) the impedance is mostly consisting in a positive reactance (inertance). At high frequency the impedance converges towards being constant and resistive. However, some non negligible oscillations are still present for reasonable values of ka. For practical build reasons, it is extremely rare to find speakers used in $ka > 5$ range.
A lumped-elements model for the approximation of this impedance pattern is proposed by [8].

2.3. Non-idealities

This part addresses the unavoidable deviations from the ideal models that occur in an audio system in practice. [9, 47]

2.3.1. Non-linear distortion

Surely the most offensive non-ideality of an audio system is the distortion introduced by non-linear transfer function (or inaccurate time-invariance) of any block in the signal path. Amplifiers with a single active component suffer from intrinsic non-linearity, that can be strongly reduced by negative feedback loop, but not completely corrected. Tube amplifiers have to deal with the non-linear transfer of the output transformer, that can be reduced by choosing a transformer with a heavy nucleus. Moreover, high-end tube amplifiers use a push-pull configuration to amplify a differential signal and compensate the even-order harmonics.
Other sources of non-linearity are mostly related to the loudspeaker build: the non-linear suspension stiffness, the flux response of the magnetic circuit and the non-uniformity of the magnetic field at the voice coil. Ported cabinets suffer from a fluidodynamic asymmetry, in fact, the air flux is more laminar when exiting the cabinet and more turbulent when entering. [26]
Nonlinear transfer functions introduce harmonic distortion for pure tones, but most importantly intermodulation distortion when the input signal has more frequency components. Both phenomena have the effect of modifying the spectral content of the signal, inserting spectral components that might be absent in the original input.
Studies are being carried on [4] to compensate the non-linear transfer with complementary non-linear control techniques.
In High-End systems, the non-linear distortion is considered acceptable under strict tolerances (THD $< 0.3\%$ at nominal power for most systems), thus for the scope of this thesis we will consider the transfer functions to be linear.

2.3.2. Off-axis variations

Loudspeakers are often approximated as point sources but their geometry forces a non-uniform radiation pattern for off-axis field. Furthermore, multiple-way or ported loudspeaker systems have sources located in distinct points in space, so their interference is dependent from the position of the listening point. Such dependence sets a limit to the design process, as it is impossible to have spectral flatness over a large sweet-spot.

2.3.3. Linear filtering

Mechanical or acoustic resonances, voice coil electrical inductance, impedance mismatches and poor crossover design may introduce linear filtering effects in the transfer function of an audio system. This effect may result in a frequency coloration or a non-linear phase response.

Resonances The interaction between pure reactive components of opposite reactance gives origin to the phenomenon of resonance. Examples of simple physical resonators include the LC circuit, the spring-mass oscillator and the Helmholtz resonator. Resonances have the effect of varying strongly the input impedance of the system, introducing frequency-dependent partitions. For example, the mechanical resonance due to the mass and stiffness of the moving parts in a loudspeaker, causes the input impedance to rise, with a prominent peak. The frequency dependent impedance causes a frequency dependent partitions with the unavoidable $R_e + R_g$ and L_e in electrical domain. This effect is more intense for tube amplifiers, given their relatively high equivalent series resistance (R_g).
It is important to notice that such partition is responsible not only for a spectral coloration, but also for a phase distortion.

Unloading At low frequencies, the radiation resistance of a loudspeaker is low, so little part of the mechanical power from the vibration is radiated correctly. This effect causes the frequency response to roll-off with lowering frequencies. In the zone $ka < 2$ the radiation reactance can be comparable or even higher than the resistance, indicating that a phase shift is present between the volume velocity and the pressure at the diaphragm. Since we are controlling the loudspeaker in voltage, thus in volume velocity [3], and reading

[3]The vast majority of systems uses low-impedance voltage sources, so that the effect of mechanical resonance is negligible. As visible in figure 2.7, if the generator and the coil were ideal (with no impedance), the mechanical elements would have no impact on the transfer function to the acoustic domain. The voltage drive is transduced into a mechanical velocity drive and then in a volume velocity, as seen with equations 2.16 and 2.21. Some research about current driving of loudspeakers can be found in [32].

its output as a pressure, the transfer function of the system suffers from this non-linear phase response.

Break-up The phenomenon of *Break-up* occurs at high frequency where the diaphragm stops acting like a rigid body and moves with non-uniform velocity. It sets an upper limit to the frequency response of a loudspeaker being responsible for its high-frequency roll-off. In fact, break-up lowers the radiation efficiency of the diaphragm. In this condition, the complete frequency response of the loudspeaker shows scarce repeatability and the phase response is practically unpredictable, due to the high sensibility to the mechanical properties of the diaphragm[26].
The cone material is crucial for extending the possibility of trade off between the conflicting requirements of rigidity and low weight.

Porting Apertures in cabinets, or ports, add purposely some acoustic resonances to alter the acoustic impedance. Their main objective is to load correctly the loudspeaker at low frequency without having to build big enclosures, which are impractical and expensive. Ported cabinets suffer from a gross phase distortion because of the interference between the directly radiated acoustic field and the field generated by the port.

3 | Audibility of phase

In this chapter, the effect of phase distortion on sound quality are addressed. Starting with the anatomy of the human auditory system and addressing some psychoacoustical references.

3.1. Human auditory system

To fully understand where perceptive models and experiences come from, it is necessary to introduce quickly the anatomy of the human hearing system. Figure 3.1 shows schematically the parts involved in the process of hearing:

3.1.1. Outer and Middle ear

Pinna and ear canal Consisting mainly in cartilage folds, the pinna has the function of performing an acoustic impedance match with the *ear canal*. This latter works as a transmission line for acoustic waves, terminated with a thin mobile membrane called *tympanum, eardrum* or *tympanic membrane* (TM).

Tympanum Being a diaphragm, the tympanum performs the conversion of acoustic energy into mechanical. It is on average 0.1mm thick, with 9mm radius and weight of 14mg. Its mechanical compliance can be measured via tympanometry and varies for each individual, however keeping the mechanical resonance in the range 800 - 1200 Hz.

Ear bones The TM transfers its motion to a set of bones: the *Malleus, Incus* and *Stapes*, which act as a leverage, again to perform a mechanical impedance match with the *oval window*, a mobile membrane on the surface of the *cochlea*. An increasing body of evidence shows that the leverage ratio is frequency-dependent. The whole mechanism is placed in the *tympanic cavity*, an air cavity whose static pressure is slowly equalized to the atmospheric pressure via the eustachian tubes.

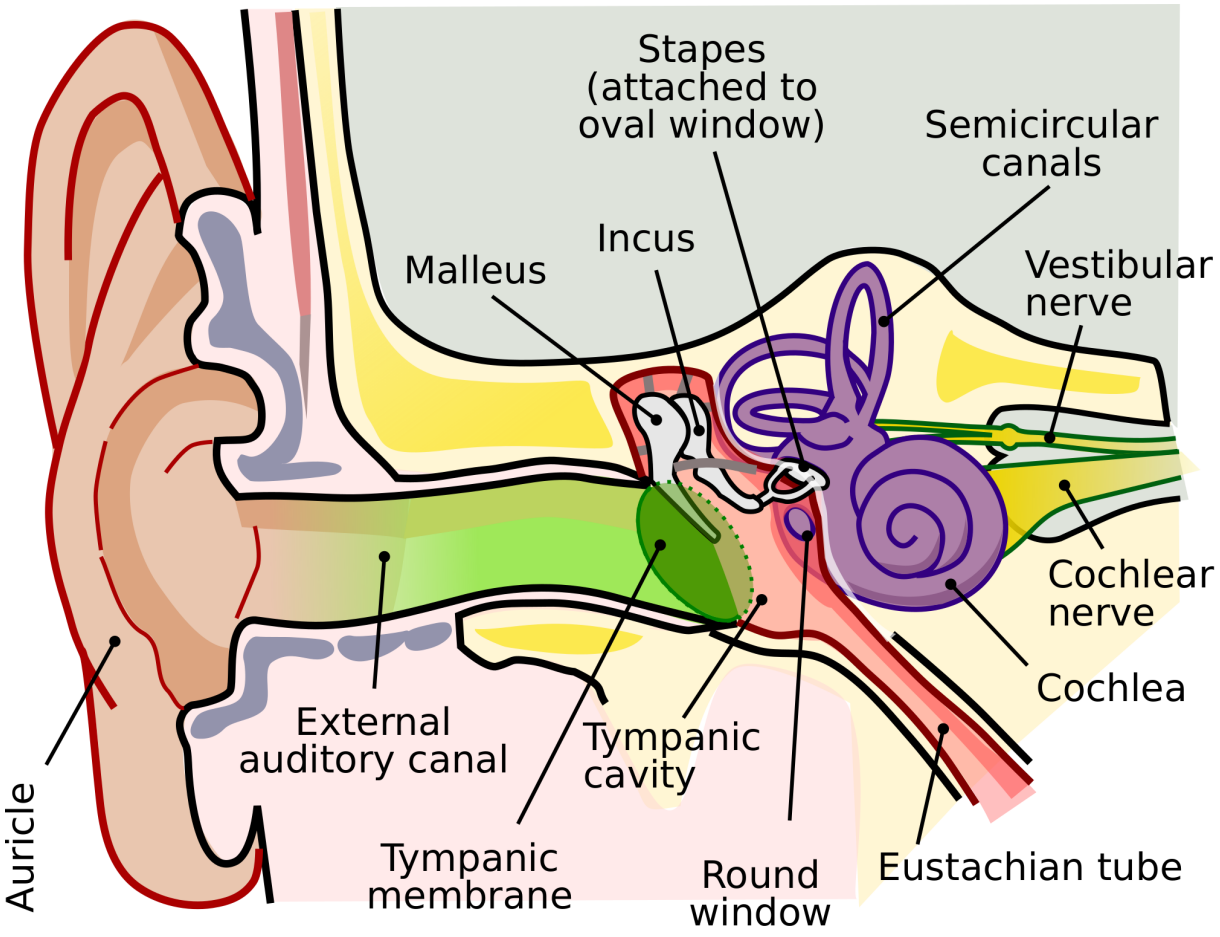

Figure 3.1: Anatomy of the human auditory system. Image by Lars Chittka; Axel Brockmann - Perception Space—The Final Frontier, A PLoS Biology Vol. 3, No. 4, e137 doi:10.1371/journal.pbio.0030137 (Fig. 1A/Large version), vectorised by Inductiveload, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=5957984

3.1.2. Inner ear

Cochlea The cochlea is a spiral-shaped organ filled with liquid (the *perilymph and endolimph*). It consists in 3 chambers: the *vestibular duct*, the *tympanic duct* and the *cochlear duct*. The interface between the cochlear and tympanic ducts is called *Basilar membrane* (BM), and hosts the *organ of Corti* with the mechano-sensitive hair cells. Due to its structure, the cochlea is interested by distributed acoustic resonances, thus splitting the incoming frequency components into vibration of distinct zones of the BM. Counterintuitively, high frequency sounds are mostly converted into vibration at the base of the cochlea, whereas low frequency components at the apex.[21]

Hair cells The Inner Hair Cells (IHC) are responsible for the transduction of mechanical stimuli into neural activation. They use hair-like structures, the *stereocilia* to detect waves with a mechanically gated ion channel. In mammals, the organ of Corti features also the *Outer Hair Cells* (OHC). Recent studies show their effect on dynamic range, as they are thought to non-linearly amplify dim sounds with a positive feedback mechanism. This feedback, known as *electromotility*, leads to the presence of measurable sound emissions from the middle ear, called *otoacoustic emission*. This non-linearity may have an effect on steady-state relative phase perception.

3.1.3. Neural processing

Information is carried from the cochlea to the *pons* (a structure in the brain stem) through the *cochlear nerve*. The terminations of both the cochlear nerves are connected to the *superior olivary nucleus*, where the first processing of binaural information can take place [36].
Then the auditory signal processing path includes other structures from the central neural system, such as the *Lateral lemniscus*, the *inferior colliculus*, the *medial geniculate nucleus* and the *primary auditory cortex*, but their effect is related to higher level sound processing and out of the scope of this work.

3.2. Monaural psychoacoustic models

Many models have been proposed in scientific literature to simulate the perceptive process and find an explanation to the complex phenomena that may occur in listening experiences.

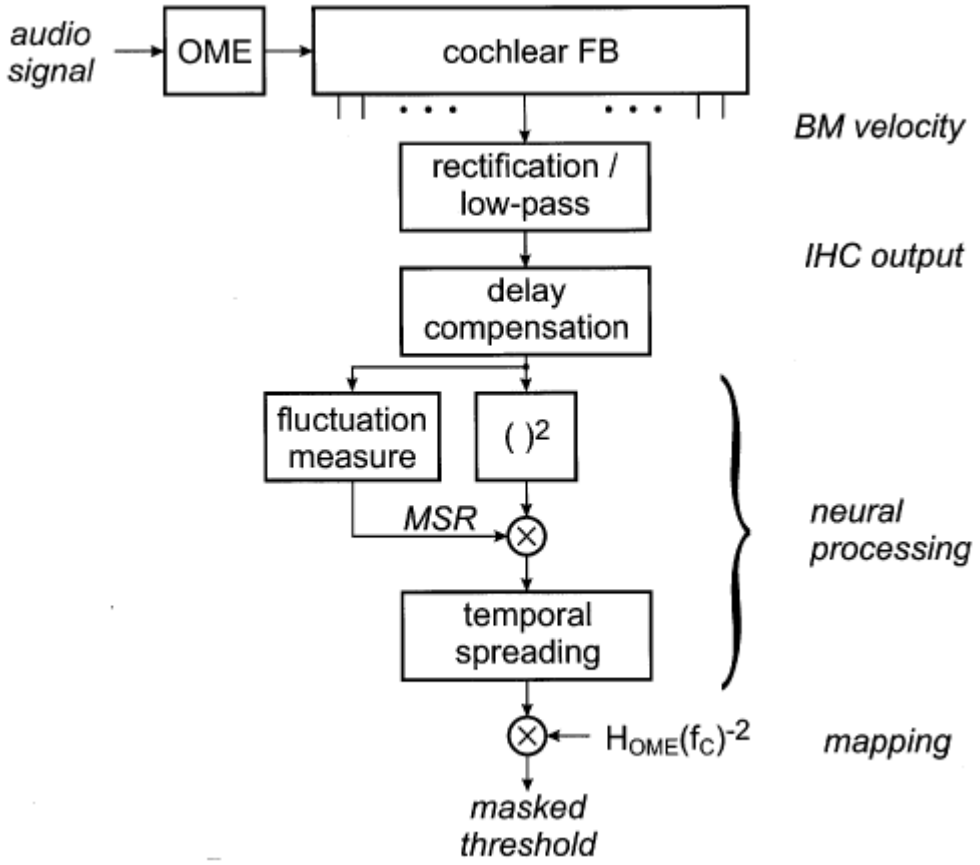

Figure 3.2: Block representation of the simplified psychoacoustic model. Image by [5]

3.2.1. Simplified filter-bank model

The simplest model proposed is the so-called *filter bank model* that simulates the filtering effect of the cochlea with a set of narrow Band-pass filters cascaded by a level detector. Due to its simplicity and effectiveness, it is widely used in perceptual audio encoders [38] and biomedical applications such as hearing aids. An enhanced version has been proposed in 2002 by [5]. This model is strongly anatomy-oriented, as its main core is the behavioural simulation of each physiological system. Figure 3.2 shows the schematic description of the full process, already adapted for audio encoding. However, we are mostly interested in the first blocks, the most related to human physiology.

OME filtering The outer and middle ear perform a strong frequency filtering that is matter of deep studies and has been precisely measured by [20] with laser Doppler vibrometers on the basilar membrane of human corpses. This work is generally accepted in scientific literature, however, in more recent studies [24] the transfer is thought to be non-linear and even fed back with unstable positive phase for low SPL, implementing an

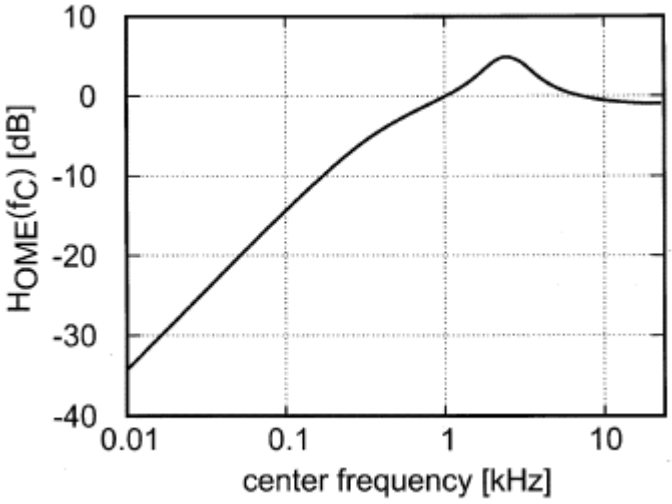

Figure 3.3: Outer and Middle Ear (OME) magnitude response plot, used by [5], implemented with a 5th order IIR filter

active dynamic compression.
The most accepted frequency response results are represented in figure 3.3, where the measurement has been taken directly at the zone of resonance for each frequency. It is fair to notice, however, that the actual OME response lacks a shared formal definition and varies with the zone of measurement on the basilar membrane.

Filter-bank The Inner Hair Cells (IHC) sample the surface of the Basilar membrane specializing in narrow frequency ranges. Together with the cochlea, they act like a filter-bank whose parametric frequency response has been studied deeply. At first it has been approximated with the use of *gammatone filters* whose time-domain impulse response is given by the product of a gamma distribution and a sine tone. Here follows the formal definition of a gammatone:

$$gt(t) = at^{n-1} \cdot exp(-2\pi bt) \cdot cos(2\pi f_c t + \Phi) \quad (t > 0) \tag{3.1}$$

This filtering model has been introduced by [2] in 1980 and is still widely used.
More recent studies [40] propose the use of non-linear filtering as a better approximation of the cochlear hair cells' activity. A good substitute for the gammatone is the Dual-Resonance Non-Linear (DRNL) filter, described in [30].

Rectification and LPF The mechanism of transduction in the hair cells consists in the triggering of mechano-sensitive ion gates due to the deflection of the stereocilia. The gates control the flow of Ca^{++} ions, responsible for the electrical activation of the neural

terminals. The whole process can be modelled as a half-wave rectification, because the ionic flow is proportional to the absolute deflection of the stereocilium from its equilibrium position, mainly in one direction. For this reason, most of the psychoacoustic models feature a rectification block.
The Low-Pass filtering stage is combined with rectification to implement an in-band level detector. Baumgarte [5] used second-order LPFs with the cutoff frequency f_{LP} dependent on the center frequency f_c of each band from the filterbank:

$$f_{LP} = \begin{cases} f_c & f_c \leq 300 \\ 300 \cdot (\frac{f_c}{300})^{0.25} & f_c > 300 \end{cases} \tag{3.2}$$

The envelope detection through Low-Pass Filtering is the block where the information about the relative phase of distinct component in steady-state is lost. A more complex model is needed to explain experiences of phase perception in human hearing.

3.2.2. In-band correlators

An interesting model proposed by [29] theorizes that the human audio information retrieval process features periodicity-sensitive neural autocorrelators for each band at the output of the filter bank. Starting from the separation of the perceptual features of pitch and chroma, this study associates each of them to the filter-bank in frequency domain and the correlation in time domain respectively. Although not perfect (requiring a non trivial redefinition of "pitch"), this model finds a valid explanation for several psychoacoustic phenomena, such as the *missing fundamental reconstruction*, the Miller and Taylor experiment (more about it can be found later in this work) and the phase perception in monaural and binaural experiences.

3.2.3. Effects of phase distortion

Knowing the signal path in the human auditory system, we can subdivide effects of phase perception in 3 main classes:

- **Short-Time Related** From a mathematical point of view, a Dirac Delta and a stationary white noise have the same magnitude spectrum, but they can be easily told apart. This is due to the short-time analysis performed by the combination of filterbank and level detector: the white noise would output a continuous constant neural firing (at all frequency bands) whereas the Delta would produce a firing pattern influenced by the damping of the resonances in the filterbank, combined with the time-domain impulse response of the LPF in the envelope detector. This

trivial case would be explained by any of the cited monaural models.

- **Steady-State** In some steady-state signals, a change in the relative phase between frequency component can be detected. A study by [35] reports that listeners described the detected difference as a change in timbre in the sound, adding more "ringing" when changing synchronization from a cosine phase towards a sine phase. Such effect would be unexplainable from the simple filterbank model, but Licklider's work with the neural correlator explains satisfactorily the phenomenon.
- **Transients** As stated by G.E.Wentworth [53]: "The character and quality of musical tones lie largely in the attack transients [...]". This evidence suggests that great part of our attention is given to the time development of the envelopes resulting as output from the filterbank. A strong phase distortion may alter the synchronization between the neural firing related to different frequencies.

3.3. Binaural psychoacoustic models

Binaural listening allows humans to retrieve much more audio information than monaural listening. An example is the effect of *Binaural Masking Level Differences* (BMLD) i.e. the phenomenon for which a signal, identical at both ears, masked by a noise, also identical at both ears, is made sensibly more detectable if inverted in phase at either ear [19].
Most importantly, it is well known that the relative phase between the audio signals arriving at the ears plays a crucial role in the localization of sound sources. This effect is of particular interest for Hi-Fi application, because an important quality requirement for high-performance audio systems is the construction of a wide and resolute soundstage (i.e. the capability of the listener to retrieve the spatial provenience of distinct sound sources). In this section, a research on the binaural perceptive models is presented.

3.3.1. Binaural cues extraction

Sound cues Intense research has been conducted on the process of extracting cues for localization. The most meaningful relative sound properties are the *Interaural Time Delay* (ITD) and the *Interaural Level Difference* (ILD). The first is mostly used for frequencies below 1KHz while the latter is given more importance for high frequencies. These cues are sufficient to explain the human capability of sound source localization, into the *cones of confusion*, i.e. the 3-dimensional loci in space characterized by the same values of ITD and ILD. Additional information about the source localization is retrieved by the human auditory system via the filtering effect performed by the torso, head and pinnae [17].

Theories for ITD extraction Carr and Konishi [11] observed the neuronal topology in the brainstem of birds, concluding that the ITD extraction is performed by an axonal delay line in the superior olivary nucleus. Coincidence-sensitive neurons are fed with both the monaural channels from the ears, each with a distinct relative time delay. Such mechanism had been theorized in 1948 by Jeffress [25].
Later studies [14, 31] show that such theory is inconsistent with the observations performed on mammals, giving more acceptance to other theories such as the *Opponent-Channel Coding*.

Binaural Matrix Feature Decoder An interesting model has been proposed in scientific literature by [7], pushed mainly towards the explanation of the BMLD effect. This model features a *Binaural Matrix Feature Decoder* (BMFD) i.e. a processing block with 5 fixed (non adaptive) outputs: two channels related to the monaural information retrieved by each ear (BE_R and BE_L) and 3 binaural interaction (BI_L, BI_R and BI_C): two more polarized towards the ears and one with the central mixing of the ears. Figure 3.4 shows the block diagram of this model. As one can see, the Low-Pass Filtering for the envelope detection is performed after the binaural comparison.

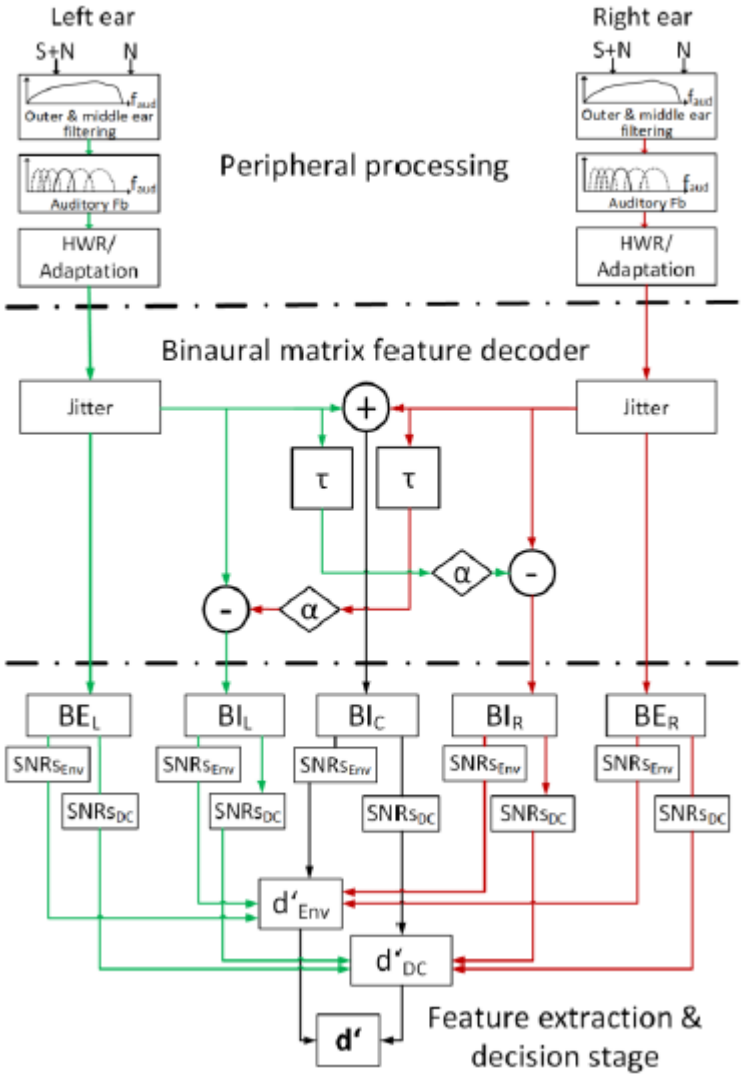

Figure 3.4: Binaural model with BMFD. Image by [7]

3.3.2. Stereo rendering

Headphones Stereo rendering with headphones works by simulating the ILD and ITD, in the stereo channels. More sophisticated algorithms for binaural rendering include the HRTF filtering (usually taken from a dataset of measured transfer functions) and the equalization of effects due to the modified conditions of the ear with respect to free-air reception.

Furthermore, source virtualization techniques take account for the head movements by updating the rendering process with the tracked orientation of the listener's head. Such techniques require a complex setup (with optical sensors for the tracking) and are used mostly for immersive audio. [48].

Loudspeaker array Stereo techniques with loudspeakers focus on the approximate reconstruction of sound field around a narrow *sweet spot*. The most used technique for stereo systems (even in Hi-Fi environment) is simple panning, i.e. modulating the amplitude gain of the signal at each loudspeaker as follows:

$$\frac{g_L}{g_R} = \frac{sin(\Theta_1) - sin(\Theta)}{sin(\Theta_1) + sin(\Theta)} \tag{3.3}$$

Where g_L and g_R are the gains at the left and right channels respectively, Θ_1 the azimutal angle of the sources from the listener's axis and Θ the desired angle of arrival of the sound generated by the virtual source [6].

3.3.3. Effects of phase distortion

In both the cases of loudspeaker and headphones listening experience, phase distortion has effect on sound localization only in differential terms, in fact, common mode phase distortion does not generate any interaural difference. Differential phase distortion in headphones has the unwanted effect of altering the ITD resolution, while the ILD is preserved. For wide-band musical signals the damage on the stereo image is limited, as the localization is mostly dependent on the ILD in case of mismatch. A comparison of ITD and ILD deviations, made with a snare drum sound, can be found at [10].
A differential phase distortion in loudspeakers may affect the interference pattern in the listening space, with huge perceptual loss in the construction of a clear and wide soundstage. Fortunately, the stereo loudspeaker systems can rely on a good level of repeatability, so we can assume that identical systems have identical phase response for most of their frequency range. As already mentioned in chapter 2, the phase distortion becomes uncontrollable in conditions of break-up. The consequent damage on the soundstage clarity is mitigated by the use of multi-way systems, so that the break-up zone of each driver is always avoided and filtered out.

4 | Preliminary activities

This chapter reports some quick experiments or builds that have been carried out during the development of the present work. All the activities have been performed at TagMa S.r.l.s. in Milan, Italy, as part of an internship.

4.1. Build and measurement of an audio system

A simple loudspeaker system has been built to validate and measure quantitatively the effects of phase distortion introduced by each part. The same system will be used for later experiments in the present thesis, so it is already designed to be as simple as possible and to have a low distortion. It features a wide-band loudspeaker in a sealed enclosure, with no crossover filters or additional circuits. The project is called R2c from TMAUDIO.

4.1.1. Circuit simulation

An article from Micka [33] focuses on Spice simulations of the loudspeaker equivalent circuits; it shows how to perform a correct simulation and proposes a precompiled LTspice file with the possibility of inserting the T/S parameters. The acoustic impedance is estimated with a first order approximation.
Figure 4.1 shows the complete circuit in the LTspice environment.
All the values in electro-mechanical domain have been derived from the T/S parameters of the chosen loudspeaker and are reported in table 4.1.1:

Component	Expression	Value
R_g	$\frac{dV_g}{dI_g}$	0.022Ω
R_e	DCR	6.6Ω
L_e	L_{evc}	$0.2mH$
C_{MM}	$\frac{M_{MD}}{(Bl)^2}$	$132\mu F$
R_{MS}	$(Bl)^2 G_{MS}$	63Ω
L_{MS}	$(Bl)^2 G_{MS}$	$17mH$

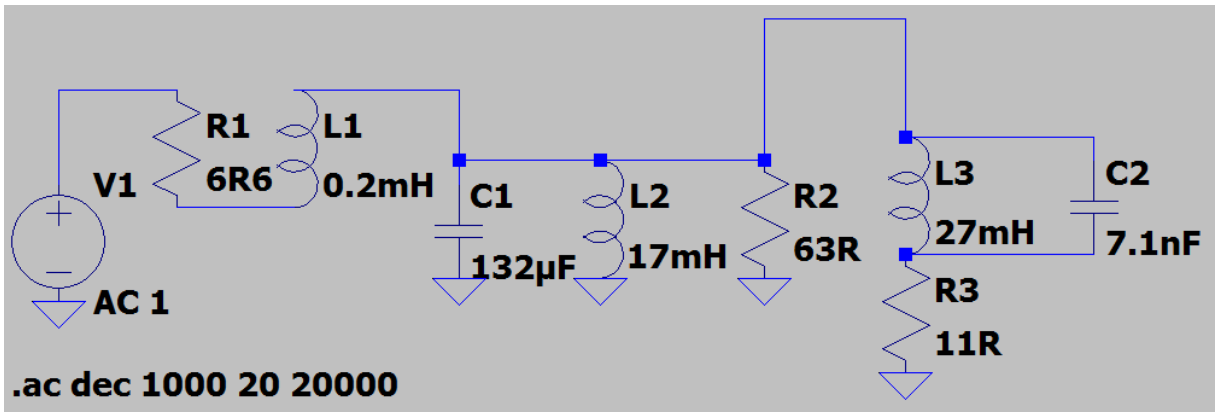

Figure 4.1: The complete LTspice circuit for the simulation of R2c

The values for the components in acoustic domain have been evaluated with the Micka online simulator [33].

Component	Value
R_{ar}	11Ω
L_a	$27mH$
C_{am}	$7.1nF$

The simulation has been run with sine sweeps over the whole audio range (20-20000 Hz) with a logarithmic resolution of 1000 points per decade. The output is taken as the current flowing in R3, in analogy with the pressure exerted by the cone on the radiation resistance. The simulated transfer function is reported in figure 4.2.

4.1.2. Simple loudspeaker build

The described loudspeaker is supposed to be built out of lamellar beech wood. A prototype has been realized first, for testing purposes, with 3D printing technology.
The walls of the cabinet are designed 9mm thick, with 3mm border and honeycomb-shaped infill. The total internal air volume is 2.41 litres, as in the technical specifications. Threaded inserts and bolts have been used to mount the loudspeaker, the electrical connectors and the bottom face.

4.1.3. Measurement

There are many techniques for measuring a loudspeaker frequency response. The most suitable for the case is the near-field recording of white noise, because this technique does not require an anechoic room [52]. In contrast, it is not effective for multi-way or ported

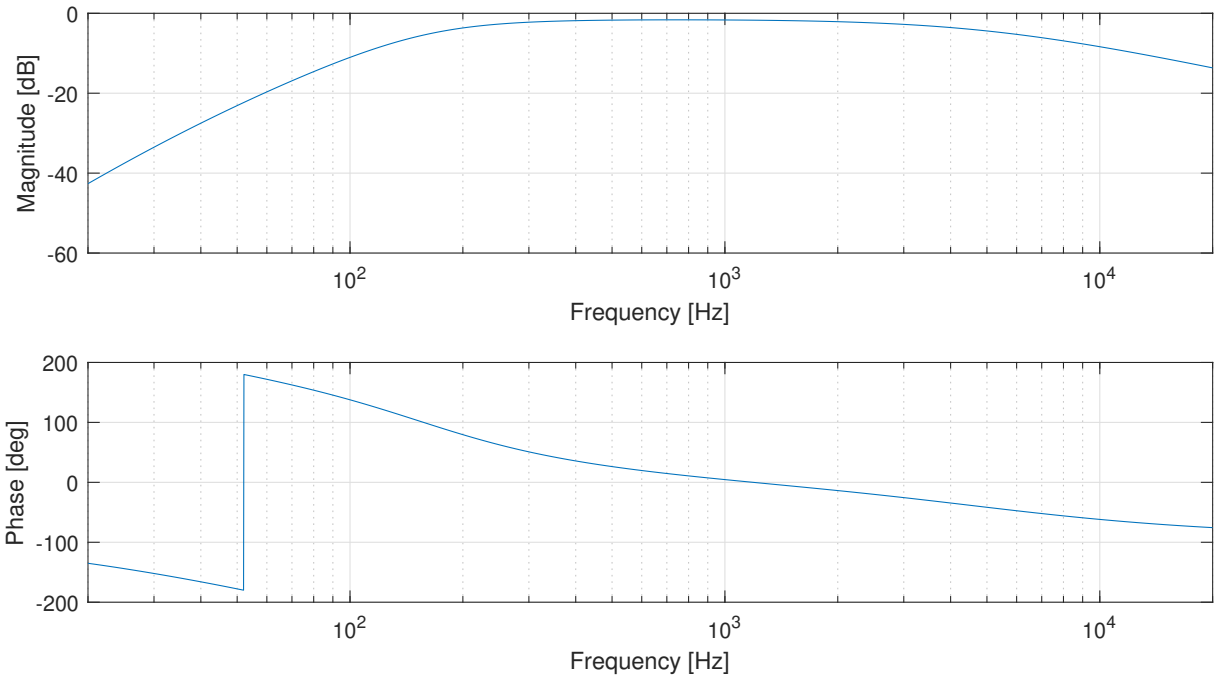

Figure 4.2: The simulation results of the circuit in figure 4.1 The graphs show the magnitude and phase of the transfer function from the voltage over V1 to the current flowing in R3. The magnitude scale is relative, the 0dB line has no physical meaning.

systems, but the device under test is a sealed single-way speaker system.

Amplifier transfer function A 100W class AB module (MX50SE stereo amplifier kit) was used as amplifier for this test. Its frequency response has been measured with a digital oscilloscope (Analog Discovery 2), while loaded with a TMAUDIO R2c loudspeaker system. The scope has a built-in function generator that can produce stepped sine sweeps in the chosen frequency range. Figure 4.4 shows the measured frequency response.
Its flatness over the audible range in loaded conditions suggests that the circuit does not introduce any appreciable phase distortion (nor magnitude filtering) and the impedance loading is correct, i.e. $R_g << R_e$. A pure time delay of less than 1ms has been recorded, but it may be due to the synchronization of the measurement setup. The peak group delay is $0.087ms$, certainly acceptable for audio application and negligible with respect to the usual group delays introduced by loudspeakers, in the order of several milliseconds.

Figure 4.3: Near-field Recording of a R2c loudspeaker box

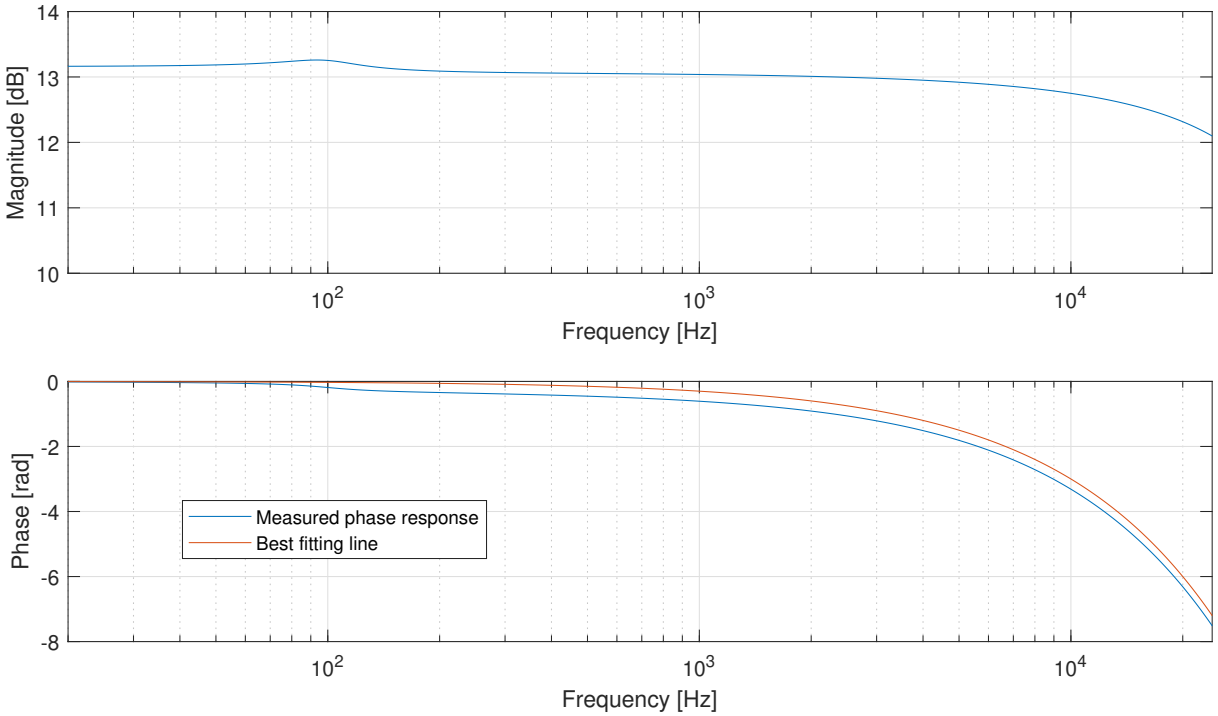

Figure 4.4: Complex Frequency response of the amplifier used to measure the frequency response of TMAUDIO R2c. The red line is the best fitting pure delay.

R2c near-field Measurement A R2c loudspeaker system was placed in a quiet room, on a loudspeaker stand, with a professional measurement microphone (OmniMic v2) as close as possible to the center of its diaphragm, as shown in figure 4.3. A 30-second white noise has been generated with the audio editor Audacity and played through an external audio interface (Steinberg UR12) at 48KHz sample frequency. One of its output channels has been fed into an input channel as a control. The played signal and the recording have been analyzed through 65536-point FFTs and their comparisons windows have been averaged. Figure 4.5 shows the measured TF in comparison with the data previously obtained by simulation (subsection 4.1.1):

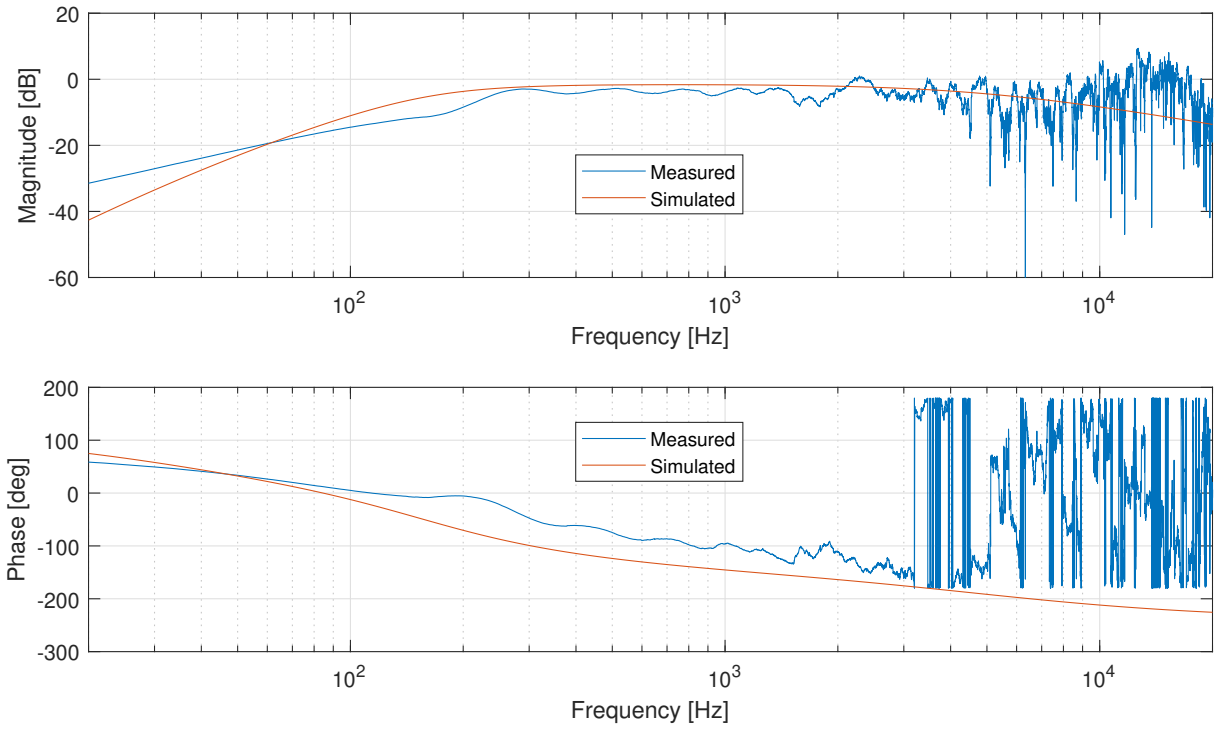

Figure 4.5: Measured and Simulated Transfer Functions of R2c loudspeaker system

The magnitude is relative and has been translated "by hand" to match for graphical reasons. The phase plot has been unwrapped and inverted in polarity (probably consequently to a wiring mistake). There is no pure delay because the recorded track has been resynchronized in the audio editor.
The noise at high frequency is due to break-up effects (whose poor repeatability affects the averaging process) and to the approximation used to esteem the radiation impedance. The zone over 2KHz cannot be easily unwrapped, it is shown in wrapped form for graphical reasons.
The group delay peaks to $5.2ms$, it has been evaluated for all frequencies by differentiating the simulated phase response, and it is shown in figure 4.6.

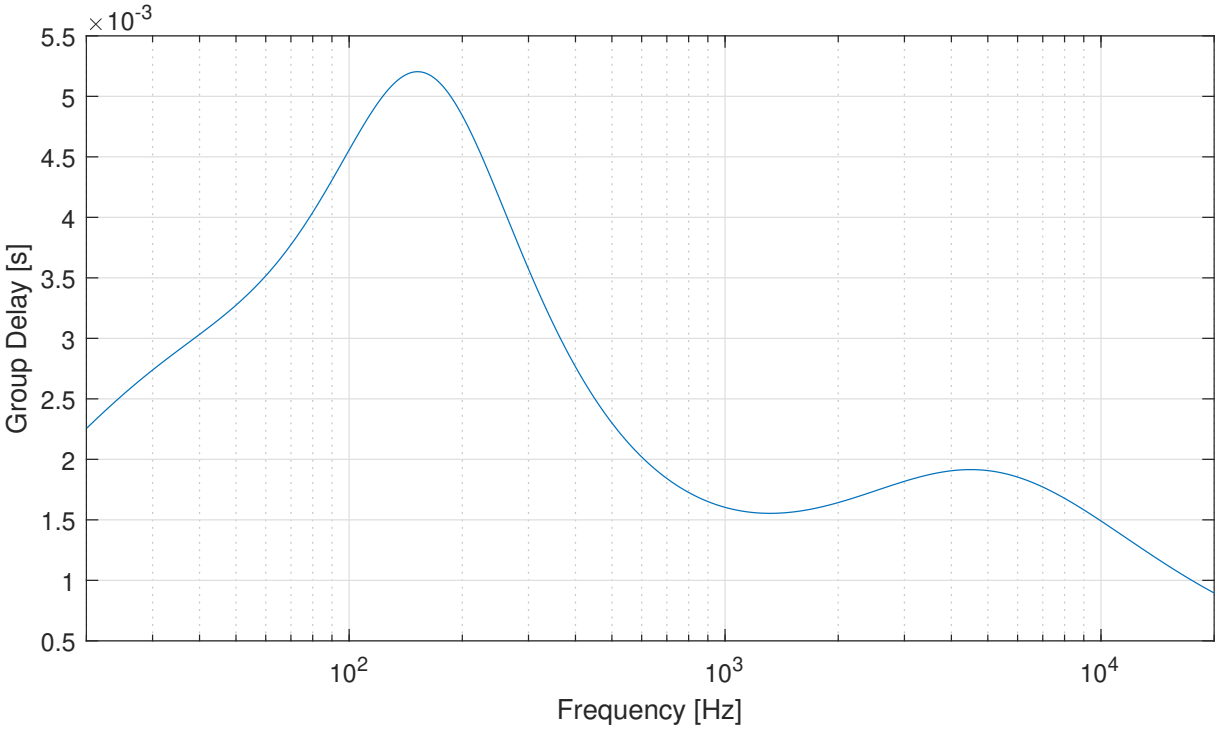

Figure 4.6: Simulated Group Delay for R2c loudspeakers

4.2. Audibility experiences

The audibility of phase has been topic of discussion for decades. Helmholtz first stated that human beings are "phase deaf", as the human auditory systems is mainly built on a filter bank, for frequency detection. However, his experiments were limited and executed with poor reliability [29]. With the modern capability of computerized signal generation, it is relatively easy to generate signals and simulate phase distortion on them. Here are proposed some experiments that demonstrate and quantify the audibility of phase responses.

4.2.1. Licklider's model validation

We recreated two quick experiments, to validate the presence of time-domain periodicity sensors in our auditory system.

Miller and Taylor's experience This experiments consists in the generation of a periodic sequence of bursts of white noise. Such signal is equivalent to a white noise modulated in amplitude by a square wave, the plot of an example signal is provided

in figure 4.7. From Fourier's analysis we cannot extract any peaking component at the fundamental frequency of the modulating wave, nonetheless, it is clearly audible as a peak tone. This result gives credit to the chosen model, because the perceived tone comes from the periodicity of the activation at the output of the IHC, and not from a cochlear resonance.

We modified slightly this experience, described in [34] by adding a masking noise with

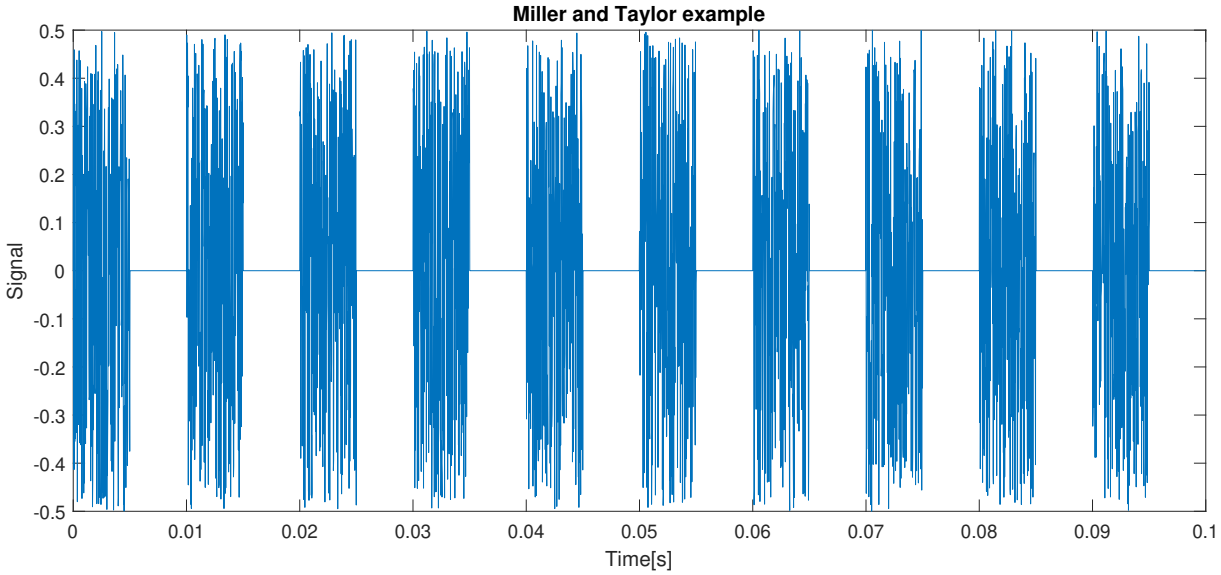

Figure 4.7: Example of signal for Miller and Taylor's experiment recreation. This signal has been generated on MATLAB (code in Appendix A)

adjustable SNR. For the author of this thesis (male, age 26, normal hearing), the tone was detectable in the frequency range 50-250Hz, with a (negative) peak of -8dB SNR at 200Hz. The MATLAB code for signal generation can be found in Appendix A.

Schouten's experience [43] A signal is generated by summing high-frequency tone components, equidistant from one another of a constant amount. This latter frequency is perceived in the resulting mix. Again it is originated by a phenomenon of periodicity, instead of frequency content, with the difference that this time we force the spectrum of our signal to have zero magnitude at the frequency of interest.
The code can be found in Appendix A.

4.2.2. Phase detection

An interesting study from Patterson et.al. [40] shows how it is possible to detect relative phase between the harmonic components of a periodic signal. For this experiment, a signal with 31 harmonics has been synthetized, both in Cosine Phase (CPH) and in Alternating

Phase (APH). Such signals have the following formulation:

$$CPH = \sum_{n=1}^{31} cos(n\omega t) \tag{4.1}$$

$$APH = \sum_{n=1}^{31} cos(n\omega t + \phi_n) \tag{4.2}$$

where $\phi_n = 0$ for odd values of n and a constant for even values of n. Patterson measured the human capability of telling these signals apart, changing the fundamental frequency, the number of harmonics and the phase lag.
Using computers, it is possible to generate even simpler signals, with just 2 harmonics, with variable phase lag between one another. A MATLAB code for this experiment has been reported in Appendix A, an equivalent C++ implementation has been done on a Bela board (more about it can be found later in this work) for usability reasons. I recreated informally this test, with 10 fundamental frequencies. The results of this experiments are reported in figure 4.8.

The threshold of audibility follows an irregular curve, suggesting that none of the proposed measures alone (i.e. phase response, group delay, phase delay) are suitable to quantify the perceptual effects of phase distortion.
The plot for the previously described parametric signal has been converging after hundreds of listening test. Thus, it is important to notice that the audibility of phase can be trained significantly. A high level of attention is required to detect differences in the first listening tests, resulting in the process of learning a new sound feature.

4.2.3. Audibility measurement - steady state

An experiment has been conducted to extract the average of the audibility results from a large sample of listeners. Due to the COVID-19 pandemic that makes it incautios to meet many people in presence, we chose to implement the experimental setup on a web application, hosted on GitHub Pages and accessible from the Internet. Given the previous considerations about the learning curve in phase detection, this test is expected to return laxer thresholds, as it is focused on the first experiences for the listeners. All the subjects were volunteers, so they could not be required to perform a large number of listening tests.

Experimental Setup The web app was developed in HTML, CSS and JavaScript; it consists in a graphical user interface, shown in figure 4.9, as well as a sound generation

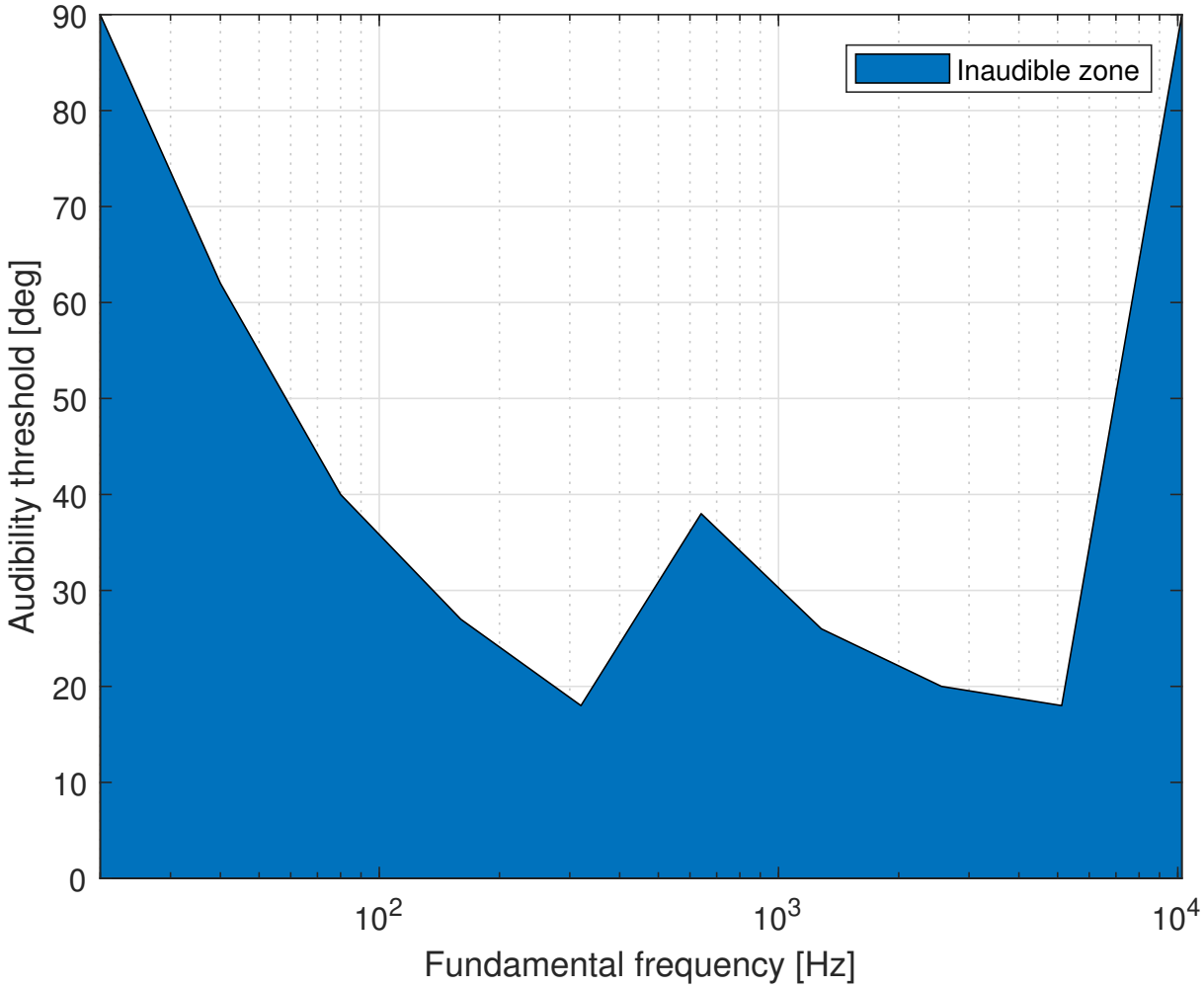

Figure 4.8: Audible and Inaudible zones for the 2-component experiment, after hundreds of listening tests

motor. The users are given a monophonic audio signal to listen to through headphones, consisting in 5 equi-amplitude harmonics with random fundamental frequency in the range [40 - 400] Hz with logarithmically uniform probability. The GUI has a slider, whose effect is the regulation of the phase lag between the even and odd harmonics. Moreover, a clickable button toggles the effect of phase distortion. The users are asked to set the slider at the exact threshold where the button has no audible effect on the sound. For each test taken, a picture of a smiling star is shown on the screen, as an incentive to take multiple tests. The results are saved on a real-time database and classified by user ID. Anonimity is guaranteed in the whole process.
A private Google Colab python sheet is used to poll the database and extract the data.

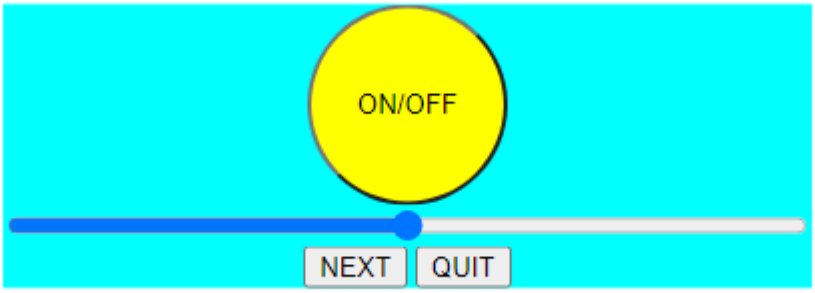

Figure 4.9: Graphical User Interface (GUI) of the online web app for measuring phase audibility.

Results A total of 24 participants, mainly among friends of the author, volunteered online. The average number of listening tests per participant was about 9. The outcome of this experiment was pretty unexpected and delusional, some subjects commented that the test was too difficult. Most of the plots of the *audibility thresholds* (i.e. the plot of the minimum audible phase difference over frequency) are pretty randomic and meaningless. One listener took 18 listening test (the highest number) and took the test the most seriously, showing a downwards behaviour in the plot of audibility threshold. This gives credit to the learning process theory. His/her results are reported in figure 4.10.
Unfortunately, there was scarce control of the experiment; non-linearities in the reproduction system or a personal bias could affect the results, leading to a bad reliability of the retrieved data.

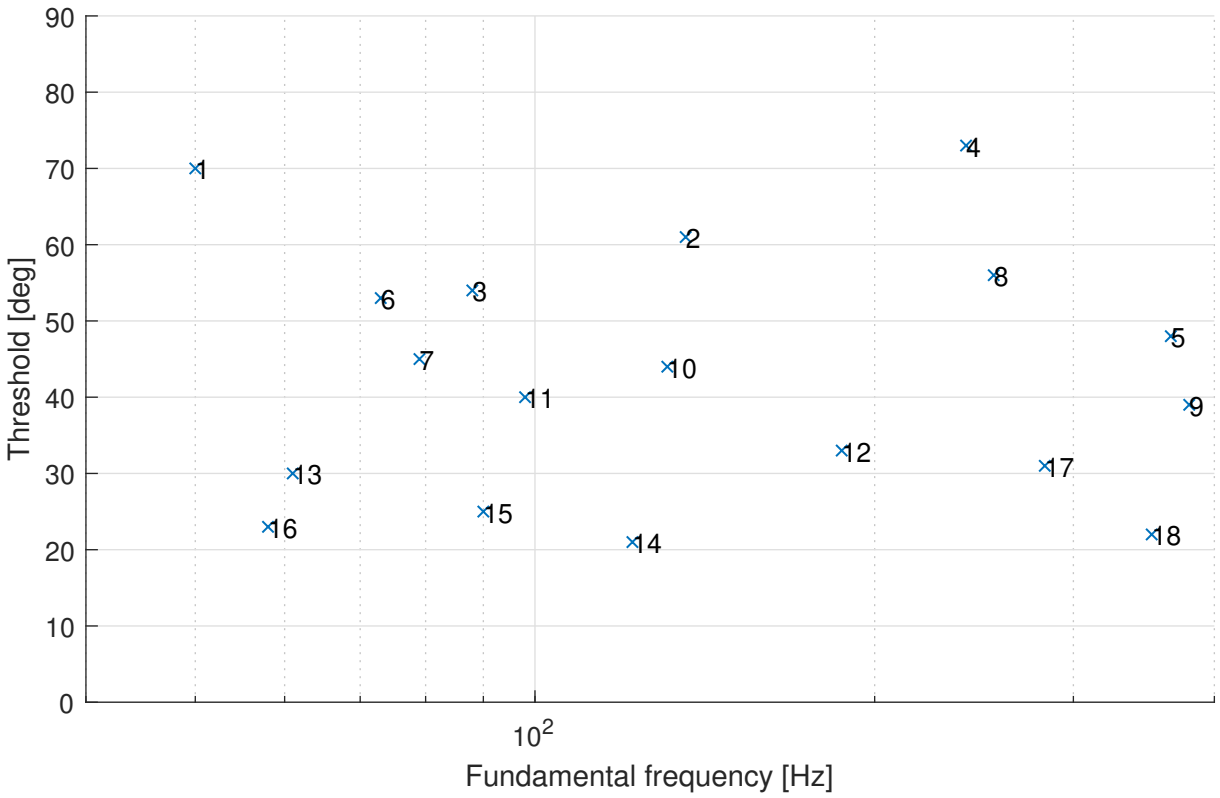

Figure 4.10: Progress in audibility shown from an unknown listener. During the session, the subject trained his/her audibility thresholds of 40-50 degrees over the frequency range of interest. The spots are numbered in chronological order.

5 | Phase distortion compensation

In this chapter, we describe some experiments that have been carried out with the aim of compensating phase distortion. Since there is no clear definition of the *correct* phase response of a system, four models are proposed and tested out with the help of a digital DSP.

5.1. Models

5.1.1. Headphones listening model

When using headphones, the sound pressure at the diaphragm of the loudspeakers and at the listener's ears coincide. It it thus intuitive to choose that pressure signal and compare it with the stereo audio signal stored in the source file. We will consider the phase corrected when the phase response of the system, from the digital audio to the pressure at the ears, has flat or linear phase response. Since headphones are stereophonic, it must be considered that the time delay introduced in the two channels must be the same. So:

$$T_r = T_l \tag{5.1}$$

In fact, given the human sensibility to interaural time delay described in chapter 3, we cannot accept to introduce a different delay between the channels of the ideal system transfer function.

5.1.2. Loudspeaker listening model

The use of stereo loudspeaker systems, placed at several meters of distance from the listener, is considered the most pleasing listening experience in the HI-FI environment. Yet it is a hard situation to model from an engineering point of view because of the acoustic mixing, happening between loudspeakers and listeners even in perfectly anechoic rooms. As previously stated, this mixing effect introduces not only a strong phase distortion, but also a frequency coloration. We will consider as correct the phase transfer between

the source signal and the sound pressure evaluated at the mouth of the loudspeaker. This model allows to measure easily the sound pressure generated by the loudspeakers independently. It is sufficient to place the microphones very near to each speaker to isolate the effect of the other, in fact the SPLs are not comparable and the cross-talk has extremely low impact on the phase measurement.
It is demonstrated by Stroh [47] that the on-axis phase response $\Phi(\omega)$ is invariant with the distance D between the loudspeaker and the microphone. In fact, the on-axis sound pressure generated in theory by a circular piston of radius a pulsating in an infinite rigid baffle with velocity $U_0 e^{j\omega t}$ has the following expression:

$$p(D) = 2U_0\rho c \sin[\frac{k}{2}(\sqrt{D^2+a^2} - D)] \cdot e^{j[\omega t + (\pi/2) - (\omega D/2c)(\sqrt{1+D^2/a^2}+1)]} \tag{5.2}$$

where the phase terms are either constant ($e^{j\pi/2}$) or proportional to frequency, thus relatable to a pure time delay.
However, this model does not take account for the room reverberations, nor their effect on phase response.

5.1.3. Loudspeaker model with room effects

Most listening rooms are not anechoic. Their acoustic response may be treated with absorbent panels but the presence of physical reverberation is related to a more natural sound, thus highly appreciated. A way to include the room effect would be to perform the measurement of pressure at the approximate position of the listener. However, in this placement, the acoustical channel mixing occurs, making it impossible to measure the effect of each loudspeaker independently. Speakers can be measured singularly to tune the phase response and then corrected for a single listening spot.

5.1.4. Acoustic correction model

This last model focuses of the global transfer from the source signal to the sound pressure at each of the listener's ears. It is indeed the most complex because it features the compensation of the phase distortion introduced by the acoustic channel mixing, evaluated singularly at very close points in space.

5.2. Previous experimental literature

We can find in scientific literature some attempts at the linearization of the phase response of audio systems:

5.2.1. Reversed-time APFs

Some studies aim at the compensation of phase distortion by applying to the input signal the inverse filter with respect to the APF component of the transfer function. Unfortunately, All-Pass Filters are notoriously difficult to invert. In fact, their transfer function in z-transform domain shows zero-pole pairs mirrored over the unitary circle. The practice of simply swapping zeros and poles would lead to poles outside the circle, thus an unstable IIR filter. The approximated approach with FIR filters guarantees stability at the cost of requiring more computational resources. An approach for inverting APFs is the convolution with the reverse-time impulse response of the original filter. In fact, let $H(\omega)$ be the original transfer function, related to the impulse response $h(t)$. Its reverse-time IR is $h(-t)$, we can extract the Fourier Transform with the definition:

$$H_{inv}(\omega) = \int_{-\infty}^{\infty} h(-t) \cdot e^{-j\omega t} dt = \int_{-\infty}^{\infty} h(t) \cdot e^{j\omega t} dt \tag{5.3}$$

Recalling that:

$$H(\omega) = \int_{-\infty}^{\infty} h(t) \cdot e^{-j\omega t} dt \tag{5.4}$$

We can substitute equation 5.4 in 5.3 and get:

$$H_{inv}(\omega) = H^*(\omega) \tag{5.5}$$

$$H(\omega) \cdot H_{inv}(\omega) = |H(\omega)| e^{Phi(\omega)} \cdot |H(\omega)| e^{-Phi(\omega)} = |H(\omega)| \cdot |H(\omega)| \tag{5.6}$$

Since $H(\omega)$ is the transfer function of an all-pass filter, its magnitude response is flat and unitary: $|H(\omega)| = 1$ Thus:

$$H(\omega) \cdot H_{inv}(\omega) = 1 \tag{5.7}$$

To demonstrate that $H_{inv}(\omega)$ is the actual inverse transfer function.
A study conducted in 1991 [41], proposes a way to correct phase distortion with the use of a simple digital DSP. The method (adapted for our case) consists in:

- Extraction of APF filter parameters via spice simulation of loudspeaker and crossover

- Buffering of source signal in a Last-In-First-Out (LIFO) structure
- Filtering of reversed-time signal
- Time reversal of the output in another LIFO buffer

Such system implements an approximated approach when used in real time, in fact, the signal is windowed at each loading in the LIFO. The presence of an IIR reverse-time filter, leaves in every window unwanted residues from the processing of the previous windows. This method has been succesfully implemented and tested on loudspeakers (with particular regard to crossover phase distortion) in 2007 by Adam and Benz [1]. They managed to compensate for the system's phase distortion, and the main advantage of this implementation is a highly reduced computational complexity with respect to the complete FIR approach. Drawbacks include the need for a simulation-driven filter parametrization and the artifacts derived from the approximation.

5.2.2. Open-loop phase correctors

Digital systems always require an analog anti-aliasing low-pass filter, usually tuned in the margin between the maximum frequency of interest (20KHz for high quality audio signals) and the Nyquist frequency. This leads to a design trade-off between alias filtering and phase distortion in the high frequency range, clearly audible by well trained listeners who can recognize the presence of a DSP in the signal path. As highlighted by Greenspun [22], raising the sampling frequency allows more space for the quality trade-off, but in order to have both a high quality alias filtering and a flat phase response, the Nyquist frequency should be much greater than the maximum frequency of interest, with enormous weight on the computational and data storage cost. Commercial companies produce phase correctors featuring all-pass filters, but their measurable effect (shown with an oscilloscope) is very subtle. Greenspun states that their benefit is not enough to justify the insertion of a new block in the signal path, because of the degradation introduced by side-effect non-idealities.

5.3. Bela platform

For the experimental part of this thesis, a powerful DSP was required. This thesis is being written during the COVID-19 pandemic and among the consequences of the virus there is the lack of availability in silicon-based electronic devices. In particular, audio DSPs have generally low volumes of production, thus they are unavailable for purchase, or shipped with several months or years of delay.

However, even in these conditions, the author could use a computing board, called Bela, engineered by the English company Augmented Instruments Ltd.

Hardware Bela is a computing platform that features a 1GHz ARM Cortex-A8 processor with 512MB of RAM, in a single-board computer from the BeagleBone Black family. It is equipped with 8X 16-bit analog I/O channels, 8X digital I/O channels and stereo audio channels as well as multiple useful hardware features for connectivity, such as Ethernet and USB host ports, and audio devices, as on-board input programmable gain amplifiers (PGAs) and output power amplifiers.

Software Bela's program memory is flashed with a Linux distribution called Xenomai, optimized for real time audio management. The Bela software features a browser-based IDE with the possibility of using several programming languages. The author chose to program in C++, thought to be the most suitable for audio management. Bela offers the possibility to design interactive GUIs with the JavaScript library P5, and most importantly, the *extremely* useful feature of the virtual oscilloscope. In debugging phase, the user can plot any signal on the pc screen in real time. Plotted signals can also be exported as csv files for further processing.

Preliminary testing As already mentioned in chapter 2, digital audio systems may suffer from the phase distortion introduced by the analog anti-aliasing filters. A quick measurement has been performed on the Bela board, with the use of a digital signal oscilloscope (Digilent - Analog Discovery 2). The instrument features a built-in function generator and can be automated to perform stepped frequency swipes while measuring the net response in magnitude and phase. The board has been programmed as "pass-trough", i.e. it samples and reconstructs a signal with no further processing. In figure 5.1 is reported the plot of the complete response.
The anti-aliasing effect is evident with the steep magnitude roll-off at 20KHz, nevertheless, the phase response is flat over the whole audible range. The artifacts at higher frequency are related to interaction with aliases and poor measurements of dim signals, thus are not meaningful for our scope.
The trade-off, addressed by Greenspun [22], between linear phase and alias filtering is visible: the Bela board features a filter with an excellent phase response up to the Nyquist frequency ($22.1KHz$) but the alias filtering is not very effective, recording only $20dB$ of attenuation at Nyquist frequency.

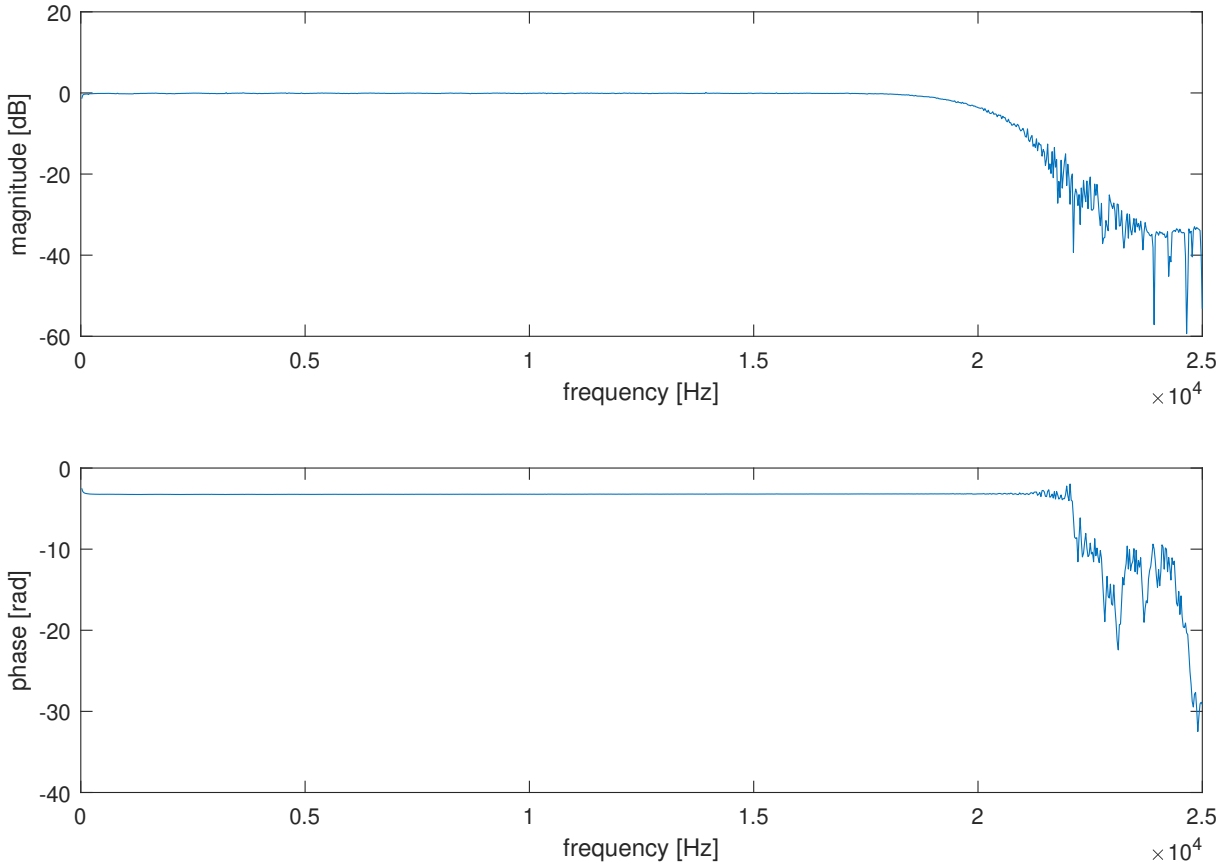

Figure 5.1: The measured frequency response of the "pass-trough" configuration of the Bela board

5.4. FFT-powered real time correction

5.4.1. Hardware setup

Microphones. The microphones used are four CMA-6542TF omnidirectional electret capsules. A preliminary tuning experiment has been performed to calibrate the capsules individually. Each capsule was placed near to the Omnimic V3 and loud signals (white noise and frequency sweeps) were played by a speaker. Both recordings have then been analysed in frequency domain to show unrepeatable oscillations over the whole audible range of $\pm 4dB$ in magnitude and $\pm 8°$ in phase. Being the error so low and unrepeatable, the author chose to neglect microphone calibration. However, the electret capsules showed a polarity reversal, probably due to their internal JFET pre-amplifier. It cannot be corrected by swapping the wires for polarization reasons, thus this issue is taken account for in software.

All microphones have been biased at $2.4V$, supplied by the $3.3V$ working voltage of Bela,

and loaded with $10k\Omega$ gain resistance [1].

Supports and cables. Two microphones are mounted at the endings of two 1.5m low-noise coaxial cables, designed for microphone signals. Nothing has been added at their base, in order to minimize any scattering or interference with the acoustic field. These capsules have then a reach of 3m distance from one another and are suitable for fixed positioning in the listening room.
The other two microphones have been placed on a wearable headset, as close as possible to the position of the listener's ears. the connection cable is a balanced screened cable, used in this case as a stereo screened cable. Less attention has been paid on the noise performance of this cable, with respect to the previous, because it is shorter and designed for more comfortable wearability.
A quick listening test confirmed that the noise introduced by the cable is acceptably low. The gain resistors are enclosed in small 3D-printed containers and the connection to Bela is achieved by the use of jumper wires.

Output For the headphones experiment, a pair of commercial gaming headphones has been used (OMEN hp 800). The choice for this set was justified by wearability reasons: the used pair of headphones features spacious and soft pavillions, so it can be worn over the microphone headset without pressing the capsules, that would be uncomfortable for the listeners.
The loudspeaker used were Hi-end 2-way ported loudspeaker systems (TMAUDIO GEM46) driven by a class AB power amplifier (MX50SE stereo amplifier kit).

5.4.2. Simplified version

The block scheme in figure 5.2 shows the implementation of our first attempt at the correction of phase distortion without crosstalk (the fourth proposed model is excluded for now). The approach is clearly closed-loop and computable, the latched blocks are indicated on the scheme with a red frame.

[1] Most electret capsules have to be biased through a resistance, usually called "gain resistance". Their output is driven in current, so the voltage amplitude is proportional to the value of the gain resistance.

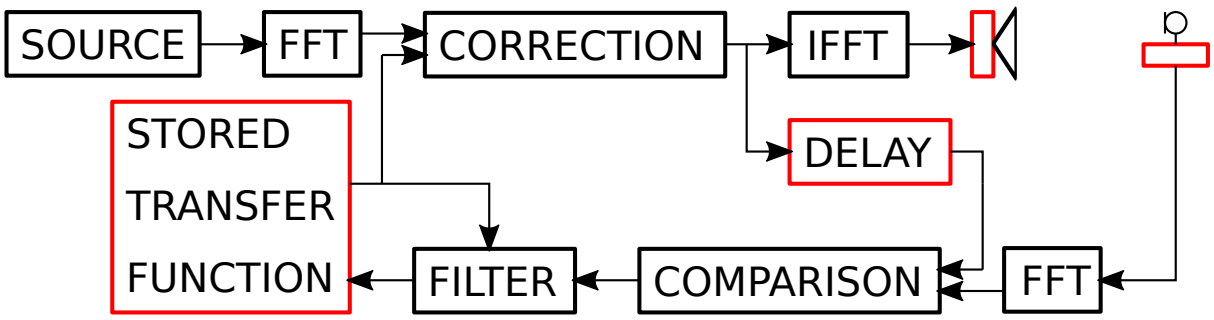

Figure 5.2: Block implementation of the first Real-Time version of the software

Description Every 186ms, windows of 8192 samples are fetched from the music source file, stored in the internal memory. Its FFT is performed, since the system works mainly in frequency domain, then its phase is adjusted by point-wise complex division with a stored transfer function, initialized as a sequence of complex ones. The corrected signal is stored in a delay buffer and retransformed in time domain for audio playout. The delay is inserted to compensate for the latency introduced by the real-time environment. Finally, the played signal (suitably resynchronized) and the input from the microphone are compared to evaluate the transfer function from the output to the microphone. The obtained transfer function may suffer from random fluctuations due to model inaccuracy or injected external noise, so it is low-pass filtered with a first order IIR filter. The parameter of the IIR filter is calculated from the power spectrum of the source signal, so that faster adaptation is given to more reliable components.
The processing is performed for both the stereo channels in a completely independent fashion.

Performance The described software works with an average consumption of 60% CPU time. It is particularly suitable for steady-state signals, in which it is clear to see the phase compensation. Unfortunately, this approach suffers the disadvantage of working with rectangular windows, thus inserting phase shifts instantaneously between one window and another. This effect is annoyingly perceived as a sequence of clicks. Moreover, the time delay due to the acoustic channel inserts a shift in the windows, this shift contains audio information belonging to different windows (along with the clicks and their echoes) that introduces a spurious component in the comparison. For steady-state signals, the frequency content is the same in every window, so the system converges with the only drawback of the annoying clicks. For music signals, placing the microphone far from the speakers may prevent the system from reaching convergence.

5.4.3. Overlap and Add version

To correct the drawbacks of the previous implementation, an Overlap-and-Add (OLA) technique has been proposed, schematized in the following picture.

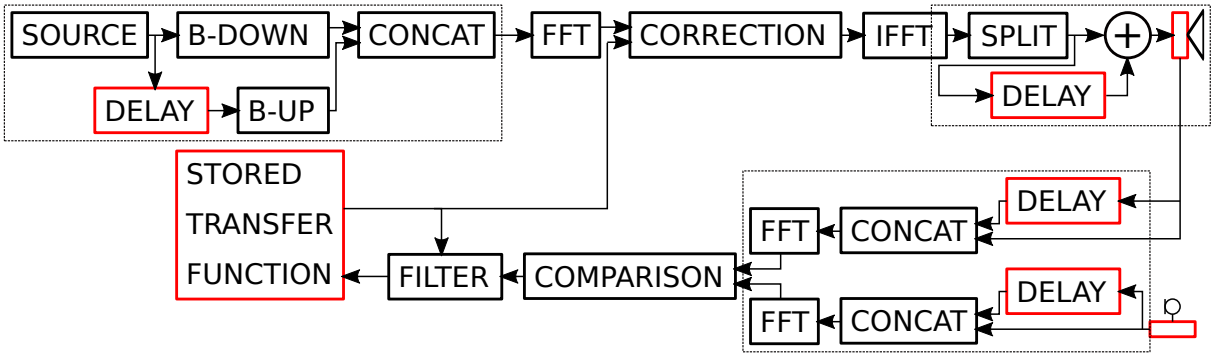

Figure 5.3: Block scheme of the OLA version. The blocks B-UP and B-DOWN represent the multiplication with a linear signal, growing upwards and downwards respectively.

The algorithm complexity is mainly due to the real-time organization and the optimization in terms of latency, in fact, this software layout works with processing blocks called at half the length of a window.

Description The working principle is the same as the previous implementation, apart from the blocks drawn in the dashed lines: the transfer function from the output to the microphone is evaluated and used to update the stored transfer function via an IIR filter. This time the correction is performed on overlapping Bartlett windows. Any change in the phase response is corrected continuously over the time of a half window, so no clicks are generated. Moreover, the spurious correlation between audio contained in distinct windows is given less importance, because it is located at the extremities of the Bartlett window, where the gain is minimum.

Clip detector The limited dynamic range of the output DAC results in the saturation of signals exceeding the range $[-1, 1]$. As already mentioned, a system with non-linear phase response could lead to a change in the crest factor of thee signal going through it, so it is possible that the compensation mechanism generates higher peaks than the saturation threshold. For this reason, an attenuation stage a clip detector has been implemented. It simply prints the warning "Clip" on the user's console if an output signal exceeds the

maximum values. The user can then reset the attenuation factor to a lower value until the full song can be played with no clipping detected.

Graphical User Interface Finally, the new version has been added a Graphical User Interface (GUI) for testing purposes. It simply consists on a clickable button on the computer screen, whose effect is to bypass the correction block. The button switches colour with every click, but no intuitive information is given regarding the status of the machine. The listener can thus know to be in "red" or "blue" configuration, but does not know which colour is associated to the system being switched on. This is purposely made to avoid any bias.

Performance At the cost of almost doubling the computational complexity (the processing block has to be called twice per window, so at double frequency), this scheme successfully solves the problems of the first one. It runs with an average of 85% CPU usage and a peak of 93%. Debugging is made difficult by the fact that invoking the virtual oscilloscope would cause the processor to saturate and drop out audio samples.

5.4.4. Results

The system has been tested by a total of 12 listeners (age 23-60, mixed gender, all normal hearing, all able to play a musical instrument at least and all used to active listening experience) who volunteered for taking part in the experiment. The tests were performed in a non-anechoic, lightly treated listening room.
One of the following musical pieces has been used, depending on the listener's taste:

- Joss Stone - Don't start lying to me now
- Frank Sinatra - New York, New York (remastered 2008)
- Diana Krall - Temptation
- Stevie Wonder - Isn't she lovely

Headphones model The volunteers were asked to wear the microphone headset and a pair of headphones (OMEN hp 800). They had access to the GUI and could comment qualitatively on the sound experience.
100% of them reported the phase compensation to have a negative effect on the sound quality, mostly related to the perceived effect of the "sound coming from inside the head".

Loudspeaker model For the second experiment, a measurement of the phase response was taken, of each loudspeaker singularly, at about 3m distance (the approximate distance of the couch from the system). Moreover, the GUI was modified to have three possible states: no effect, real-time closed-loop phase compensation, and compensation obtained by the pre-recorded phase response. The microphone capsules were placed as near to the loudspeakers as possible, with the aid of microphone stands, about half way between tweeter and woofer. High attention has been paid to the symmetry of the setup.
Listeners were asked again to comment on the sound. While the unanimity was recognized in asserting that "no effect" is still the best, most natural and most satisfying option, a few disagreements were recorded regarding the two processes of phase correction. The majority (5 out of 12) preferred the real-time correction, describing the sound more "dry, resolute", 1 subject preferred the "fluid, warm" sound of the open loop approach, 6 listeners declared not perceiving any difference. Apart from the personal taste in music listening, most of the subjects (who declared perceiving the difference in the correction models) agreed on one another's qualitative comments, presented to them after the experiments.

5.5. A GCC-based microphone localization technique

In a listening room with loudspeakers, complex interference patterns arise along the whole space. In order to attempt any kind of correction involving binaural audio it is necessary to track the listener's ears and adjust the generation of the acoustic field depending on the instantaneous position in the room.

5.5.1. GCC for source localization

Generalized Cross-Correlation (GCC) is a technique used mainly for source localization with an array of microphones. It consists in evaluating the correlation between the microphone signals. Since the signals are expected to be delayed versions of the same wave, the correlation will show a peak in correspondence of the TDOA (Time Delay Of Arrival).
Musical signals often show periodicity, that introduces parasitic peaks in the correlation function, in correspondence of the period and its integer multiples. This unwanted effect can be tampered by the introduction of a whitening filter.
For performance reasons, the correlation is performed in frequency domain, thanks to the

following property: [16]

$$x_{ab}(t) = \int_{-\infty}^{\infty} a(t) \cdot b(t+\tau) d\tau \tag{5.8}$$

$$... \tag{5.9}$$

$$X_{ab}(f) = A(f) \cdot B^*(f) \tag{5.10}$$

The whole algorithm works as follows: first, the signals A and B are Fourier-transformed into frequency domain. Then, the whitening factor (called *PHAT* as *PHAse Transform*) is computed as follows:

$$PHAT(f) = \frac{1}{|A(f)| \cdot |B(f)|} \tag{5.11}$$

Note that $PHAT(f)$ is always real. Its introduction in the cross-correlation has no effect on the relative phase of the signals. In practice, this quantity should be clipped to avoid divide-by-zero issues.
Finally, the actual GCC is computed as follows:

$$GCC(t) = \mathcal{F}^{-1}\{A(f) \cdot B(f) \cdot PHAT(f)\} \tag{5.12}$$

5.5.2. Proposed technique for microphone localization

In this section, a technique is proposed by the author for the task of microphone localization in a listening room with two loudspeakers playing music. The objective of this task is to return the distances of the microphone from each speaker, given the stereo signal fed to the system and the input captured by the microphone. This technique is thought to be implemented in real time, nevertheless, a MATLAB prototype has been prepared first.

The following assumptions are taken for the correct functioning of the proposed GCC technique:

- The room is anechoic, or features low reverberation.
- The signal played is a stereo musical signal, forcedly not monophonic.
- The difference between the stereo channels is uncorrelated enough from the signal.

The figure 5.4 shows the proposed setup.

Let x_1 and x_2 be the sound pressure signals reproduced by the loudspeakers, X_1 and X_2 their Fourier representation in frequency domain. Similarly G_1 and G_2 are the transfer functions from each loudspeaker recorded at the microphone. y is the mixed signal

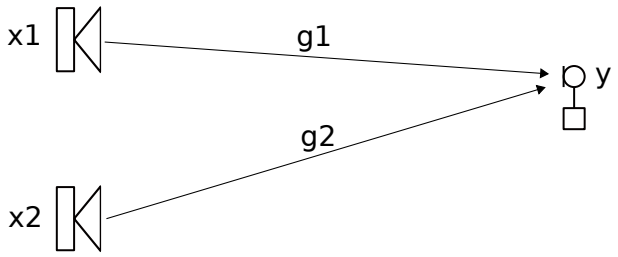

Figure 5.4

captured by the microphone when both loudspeakers are playing, Y its representation in frequency domain.
The transfer functions G are expected to be in the following form:

$$G_n(\omega) = A_n e^{-j\omega \cdot \Delta T_n} + n(\omega) \tag{5.13}$$

$$\Delta T_n = c \cdot d_n \tag{5.14}$$

Where c is the speed of sound in air, d_n the distance of the microphone from the n-th loudspeaker, ΔT and A_n the acoustic delay and attenuation due to the sound propagation. n is a term that takes account of additive disturbances to the model, mostly due to unwanted reverberation and external noise sources in the room. In ideal conditions (i.e. anechoic and silent room) the transfer function simplifies as:

$$G_n(\omega) = A_n e^{-j\omega \cdot \Delta T_n} \tag{5.15}$$

Under the previously stated assumptions, we will neglect the noise components and use this latter equation.
The microphone records the linear combination of the stereo signals, each weighted to its own transfer function. So:

$$Y(\omega) = G_1(\omega) \cdot X_1(\omega) + G_2(\omega) \cdot X_2(\omega); \tag{5.16}$$

Now, we consider the half sum and half difference of the stereo signals. Let them be called C as "Common" and D as "Differential". The sources will emit:

$$x_1(t) = c(t) - d(t) \tag{5.17}$$

$$x_2(t) = c(t) + d(t) \tag{5.18}$$

With the same notation, in frequency domain:

$$X_1(\omega) = C(\omega) - D(\omega) \tag{5.19}$$

$$X_2(\omega) = C(\omega) + D(\omega) \tag{5.20}$$

Substituting these equations in 5.16 we get:

$$Y(\omega) = G_2 \cdot [C(\omega) + D(\omega)] + G_1 \cdot [C(\omega) - D(\omega)] \tag{5.21}$$

Consider the cross-correlation index A between the signals c and d. By definition:

$$A = \frac{R_{CD}}{R_{CC}} = \frac{E[c(t) \cdot d(t)]}{E[c(t) \cdot c(t)]} \tag{5.22}$$

Intuitively, the cross-correlation index is used to perform a base change to represent the same information as the original stereo channels, with uncorrelated signals. We define a new signal b as:

$$b(t) = d(t) - A \cdot c(t) \tag{5.23}$$

Given that all the signals are zero-mean (all audio signals filter out any DC component), it is possible to demonstrate (Proof in appendix B) that b and c are uncorrelated.
We will use the inverse form of the previous equation:

$$d(t) = A \cdot c(t) + b(t) \tag{5.24}$$

$$D(\omega) = A \cdot C(\omega) + B(\omega) \tag{5.25}$$

Substituting equation 5.25 into 5.21 we obtain:

$$Y(\omega) = [(A+1) \cdot C(\omega) + B(\omega)]G_2(\omega) + [(1-A) \cdot C(\omega) - B]G_1(\omega) \tag{5.26}$$

$$Y(\omega) = B(\omega) \cdot [G_1(\omega) - G_2(\omega)] + C(\omega) \cdot [(A+1) \cdot G_2(\omega) + (1-A) \cdot G_1(\omega)] \tag{5.27}$$

The recorded signal has been rewritten as a combination of two uncorrelated signals (B and C). So, by performing GCC method between Y and B, we expect to extract an estimate of the term $G_1(\omega) - G_2(\omega)$, because the second term of the equation is proportional to C, thus expected to have no effect in the GCC of Y. The time-domain $g_1(t) - g_2(t)$ can be then evaluated through an Inverse Fourier Transform.

g_1 and g_2 represent the impulse response related to the transfer from namely speakers 1 and 2 to the microphone. They are expected to show a dirac pulse in correspondence of their direct time delay and low-energy disturbance components. So, the signal $g_1 - g_2$ should show a positive peak and a negative peak. If the SNR of the whole system is reliable enough, we can extract the time delays as indexes of the maximum and minimum of this signal. The SNR can be improved sensibly by processing longer frames of audio signal.
A more sophisticated version features the redefinition of B as the uncorrelated residue between C and D and the same steps to obtain $g_1 + g_2$. Both paths can be combined to extract singularly the impulse responses. Then again, we are interested in the maximum of each of them, since it is expected to happen at $t = \Delta T$. This new version improves the SNR (and thus the reliability of the detection) at the cost of doubling the computational load.

This GCC technique was tested with simulated signals first and then in an experimental situation. The MATLAB code for the processing is reported in appendix A. The experiment was conducted in a non acoustically treated living room, with a stereo player (Logitech stereo system 2320), a carpet, a sofa and several pieces of furniture. An omnidirectional measurement microphone (OmniMic V3) was placed over the sofa at head height and precise measurement of the distances has been performed with a metre rule. Several pieces of music or audio signals have been played and recorded.

5.5.3. Performance and Results

This algorithm has been tested in a MATLAB simulated environment first, for debugging purposes, then it was made to run on the Bela board in some real-world situations, i.e. two living rooms with a stereo setup and a highly reverberant bathroom. The first living room had a cheap set of loudspeakers and the second was equipped with TMAUDIO R^2c. A capsule was placed on a microphone stand in the middle of the room, and the stand was able to rotate along a fixed axis. A piece of music with average correlation between the stereo channels was played (Frank Sinatra - New York New York). Every 10 seconds, the stand was rotated in a random direction. The localization data were processed in MATLAB and plotted to check if the localization results stayed on a circle.
Unexpectedly, the real-time processing required low hardware resources, with a surprising 19% of CPU usage for a single microphone and around 24% for the stereo version, running at 8192 samples per window, which gives a more than acceptable reliability. As a comparison, a program doing absolutely nothing stabilizes its CPU absorption to 12%,

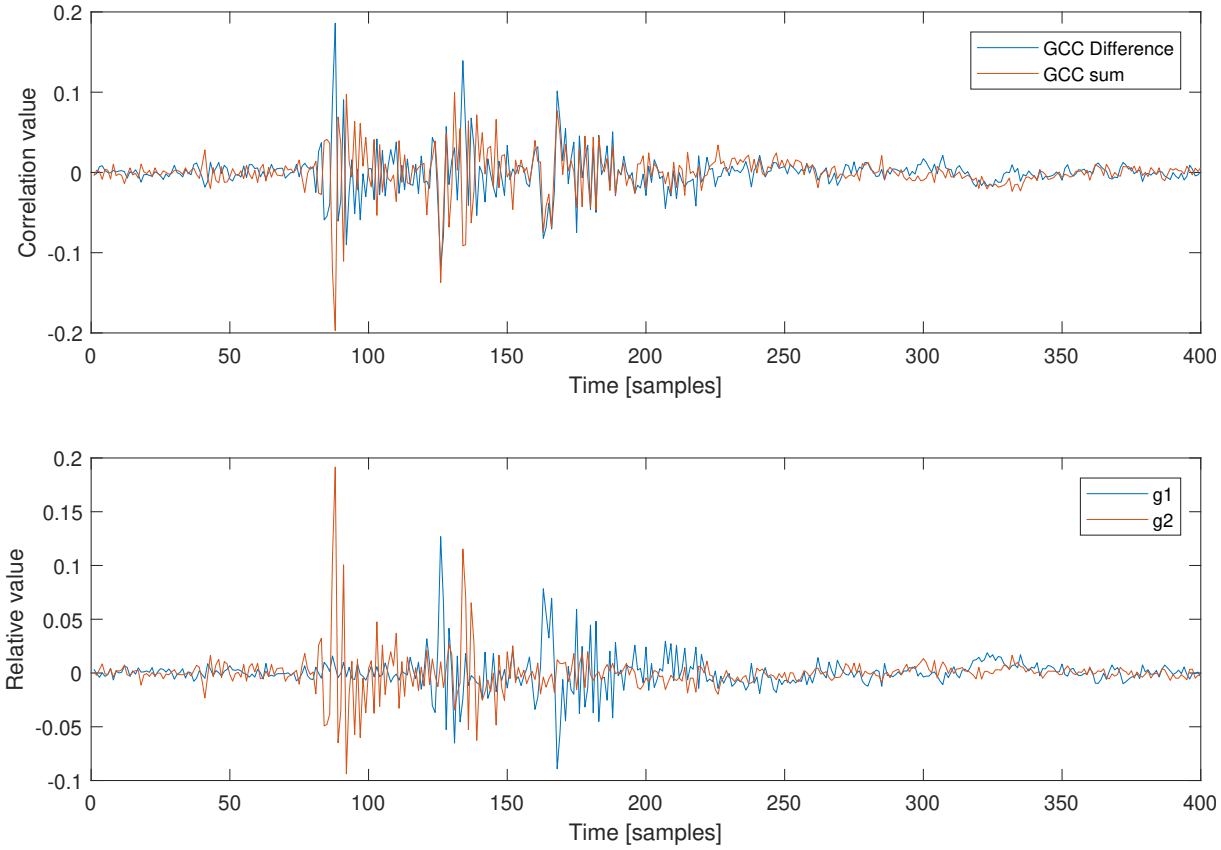

Figure 5.5: Example of GCC from real-world situations.

just for running the operating system. The image 5.5 shows an example of the GCC. A max and min functions on the "GCC Difference" signal are sufficient to detect the peaks. However, using the refined version of the software, we can see from the picture that the peaks of g1 and g2 lie in the exact same positions, with a better crest factor.

The plot in figure 5.6 represents the detected position of the microphone in the room. It is clear that the spots lie on a circle. During the transients, when the microphone was used, the system slowly dims out the old peaks, due to the low-pass filtering, and forms peaks related to the new position. This approach loses continuity in the detection and presents some spurious maxima when the amplitude of the peaks is low. However, all of these false detections were several tenths of meters far from the previous, so, a reliability index could be easily implemented and could be helpful to avoid false detections.
The code used to generate this plot can be found in Appendix A.

The cheap set contained a DSP, its presence introduces a considerable latency (about 10ms) and a phase distortion, with t_g over 4ms. The first makes the system overesteem the distances by tenths of meters, but the delay can be easily compensated, and the detection still works. Instead, the latter has the effect of lowering the amplitude and resolution of the peak, resulting in less reliability for the detection. The system reaches convergence

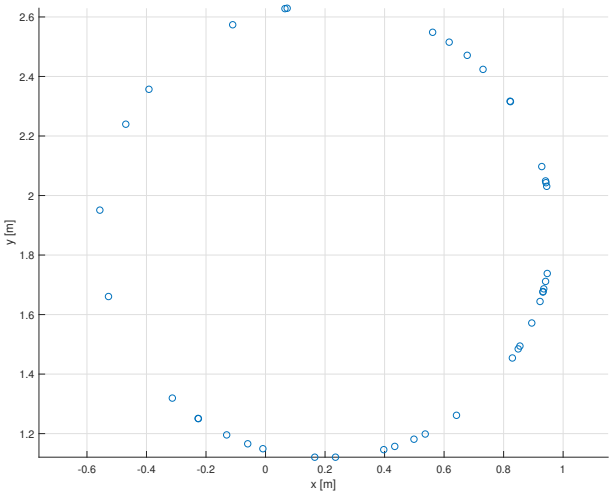

Figure 5.6: Detected position of the microphone, rotated around the axis with a microphone stand. Apart from the false detections, all the reasonable data appear to lie in a circle

but returns many false detections and shows low robustness to transients.
A highly reverberant room shows a strong degradation in the quality of GCC, as many peaks arise, related to the acoustic reflections on the walls. However, the peak related to the direct sound is stronger, so the reverb has again little effect on the localization of a fixed microphone, instead affecting the speed of adaptation to transients.
In conclusion, even though further signal processing could be useful, especially for transients in microphone placement, the software has showed to be working with acceptable speed, reliability and precision for non-treated (but reasonably damped) listening rooms with high quality sound systems.

5.6. Binaural phase equalization

The last listening model requires the most complex correction algorithm, as it has to take account for the acoustic mixing between the stereo channels happening at both ears. The same problem has been encountered in scientific literature by studies regarding the cross-talk cancellation (XTC) techniques.

5.6.1. Dual-channel XTC techniques

When listening with a 2-speaker system, the transfer function between each source and the contralateral ear is called *cross-talk*. It is responsible for the most of the difference between listening through headphones and through a stereo loudspeaker setup.

Several studies aim at finding the correct change of coordinates in the stereo signal, so that full control is achieved of the sound pressure at the ears. As higlighted by Choueiri [13], such techniques are easy to implement in theory, but extremely sensitive to reverberations and non-idealities, as well as small movement of the listener's head.
Famous XTC implementations include *Optimal Source Distribution* (OSD) [51] and the *BACCH* algorithm [13].

5.6.2. Implementation

The following implementation has been inspired from the work of Anushiravani [3]. We can model the mixing of ipsilateral signals and crosstalk as follows:

$$Y(\omega) = \begin{bmatrix} G_{11}(\omega) & G_{12}(\omega) \\ G_{21}(\omega) & G_{22}(\omega) \end{bmatrix} X(\omega) \tag{5.28}$$

Where G_{nn} are extracted through the microphone localization process. In fact:

$$G_{nn}(\omega) = |G_{nn}| \cdot e^{-j\omega\Delta t_{nn}} \tag{5.29}$$

Neglecting noise and reflections, $|G_{nn}|$ can be evaluated in relative terms as the peak value of the estimated g functions.
In theory, it would be sufficient to filter the signals with the inverse matrix of $G(\omega)$.

$$G^{-1}(\omega) = \frac{1}{G_{11}(\omega)G_{22}(\omega) - G_{12}(\omega)G_{21}(\omega)} \begin{bmatrix} G_{22}(\omega) & -G_{12}(\omega) \\ -G_{21}(\omega) & G_{11}(\omega) \end{bmatrix} \tag{5.30}$$

Such operation would force the sound pressure at the ears to match perfectly the source signal, correcting all the transfer from stereo source to ears.
The algorithm has been tested on MATLAB simulation and on Bela.

5.6.3. Results

While the simulation offered good results, the real-world real-time implementation on Bela failed. The system requires much more dynamic range than the one supplied by the loudspeaker, triggering annoying reverberations and easily reaching the clipping threshold.

Conclusions and future work

This thesis highlighted that phase distortion happens in almost every kind of audio systems, even the most expensive. Its perceptual effects are way more subtle with respect to other waveform distortions, such as nonlinear distortion or frequency filtering, nevertheless, we can state that monaural phase is audible, and a reasonable psychoacoustic model can give an explanation to this phenomenon. The measuring for monaural phase perception is hard to perform, the measures that were proposed in scientific literature cannot be considered absolute and the perception requires a certain degree of attention from the listener.

Binaural differential phase distortion is more easily perceivable, not related to a timbric change, but rather responsible for the generation of the soundstage. However, such effect can be easily avoided by respecting the symmetry of the audio system build.

The last experiment has been a failure because the setup was extremely sensitive to model non-idealities. The experience should be performed in a quiet anechoic room but since we had no access to such an environment, the experiment is left for later developments.

The first compensation experiments, have been carried out with the most interesting results. The correct functioning of the setup was guaranteed by the virtual oscilloscope, but the overall listening experience was considered worse by the majority of the listeners who declared themselves able to spot the difference. A possible reason could be that humans are so used to listening to a certain pattern of phase distortion that the compensated version may sound unnatural. In any case, further research will be needed to find a plausible explanation. We cannot completely reject the idea that phase distortion should not be regarded as a dangerous non-ideality, but rather as a parameter that can be artfully mastered by electroacoustic engineers. Such statement clashes with the obsessive search for "fidelty" often shown by audiophiles, but there might be a point where the pursuit of perfection gives way to the more meaningful mastery of the good sounding imperfections.

Bibliography

[1] V. Adam and S. Benz. Correction of crossover phase distortion using reversed time all-pass iir filter. *Audio Engineering Society - 122nd Audio Engineering Society Convention 2007*, 2:585–590, 01 2007.

[2] A. Aertsen and P. Johannesma. Spectro-temporal receptive fields of auditory neurons in the grassfrog - i. characterization of tonal and natural stimuli. *Biological Cybernetics*, 38:223–234, 11 1980. doi: 10.1007/BF00337015.

[3] R. Anushiravani. *3D Audio Playback through two Loudspeakers*. PhD thesis, University of Illinois, January 2014.

[4] M. Arvidsson and D. Karlsson. Attenuation of harmonic distortion using nonlinear control. Master's thesis, Linköpings universitet, 2012.

[5] F. Baumgarte. Improved audio coding using a psychoacoustic model based on a cochlear filter bank. *IEEE Transactions on Speech and Audio Processing*, 10(7): 495–503, 2002. doi: 10.1109/TSA.2002.804536.

[6] A. Bernardini, L. Bianchi, and A. Sarti. Spatial sound with loudspeaker - stereophony and panning. Published as lesson slideset for the course of Sound Analysis Synthesis and Processing at Politecnico di Milano, 5 2020.

[7] T. Biberger and S. Ewert. *Towards a Generalized Monaural and Binaural Auditory Model for Psychoacoustics and Speech Intelligibility*. Acta Acustica, 06 2021.

[8] A. Bozkurt. A lumped-circuit model for the radiation impedance of a circular piston in a rigid baffle. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 55(9):2046–2052, 2008. doi: 10.1109/TUFFC.896.

[9] S. Brown. Linear and nonlinear loudspeaker characterization. Technical report, Worcester Polytechnic Institute, 01 2006.

[10] H. Cardew. Ild vs itd comparison (wear headphones), 2016. URL `https://www.youtube.com/watch?v=K1AoWfAKCqw`.

[11] C. Carr and M. Konishi. Axonal delay lines for time measurement in the owl's brainstem. *Proc. Nadl. Acad. Sci. USA*, (85):8311–8315, 1988. doi: 10.1073/pnas.85.21.8311.

[12] D. Cartasegna. *Study, Modeling and Realization of an Audio Class-D Power Amplifier in 0.18µm CMOS Technology.* PhD thesis, Università degli Studi di Pavia, 2011.

[13] E. Y. Choueiri. Optimal crosstalk cancellation for binaural audio with two loudspeakers. Self-published, URL `http://www.princeton.edu/3D3A/Publications/Choueiri_3D3A_OptimalXTC.html`, 2010.

[14] J. Encke and W. Hemmert. Extraction of inter-aural time differences using a spiking neuron network model of the medial superior olive. *Front Neuroscience*, pages 12–140, 03 2018. doi: 10.3389/fnins.2018.00140.

[15] J. Fagerström. Phase equalizers, overview and experiments in audio processing. Paper published for Master's Programme CCIS / AAT.

[16] B. J. Fischer, G. B. Christianson, and P. J. Luis. The cross-correlation and wiener-khinchin theorems. *Journal of Neuroscience*, pages 8107–8115, 2008.

[17] A. Floros and N. A. Tatlas. Spatial enhancement for immersive stereo audio applications. In *2011 17th International Conference on Digital Signal Processing (DSP)*, pages 1–7, 2011. doi: 10.1109/ICDSP.2011.6005001.

[18] E. Gaalaas. Class d audio amplifiers: What, why, and how. *AnalogDialogue*, 06 2006.

[19] H. Gilbert, T. Shackleton, K. Krumbholz, and A. Palmer. The neural substrate for binaural masking level differences in the auditory cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 35:209–220, 01 2015. doi: 10.1523/JNEUROSCI.1131-14.2015.

[20] R. L. Goode, M. C. Killion, K. Nakamura, and S. Nishihara. New knowledge about the function of the human middle ear: development of an improved analog model. *The American journal of otology*, 15 2:145–154, 1994.

[21] L. Gray. Auditory system: Structure and function, 2020. URL `https://nba.uth.tmc.edu/neuroscience/m/s2/chapter12.html`.

[22] J. Greenspun. Audio analysis i: Phase correction for digital systems. *Computer Music Journal*, pages 13–19, 1984.

[23] R. Harley. New op-amps promise high-end sound, 2008. URL `https://www.theabsolutesound.com/articles/new-op-amps-promise-high-end-sound`.

[24] A. Hudspeth. Making an effort to listen: Mechanical amplification in the ear. *Neuron*, 59:530–45, 09 2008. doi: 10.1016/j.neuron.2008.07.012.

[25] L. Jeffress. A place theory of sound localization. *J. Comp. Physiol. Psychol*, (41): 35–39, 1948. doi: 10.1037/h0061495.

[26] W. J. Klippel. Loudspeaker nonlinearities – causes, parameters, symptoms. *Journal of The Audio Engineering Society*, 2005.

[27] D. Koya. Aural phase distortion detection. Master's thesis, 2018.

[28] H. Köymen. Radiation from a piston: radiation impedance, radiation pattern and diffraction constant. Master's thesis, Bilkent University.

[29] J. Licklider. A duplex theory of pitch perception. *Experientia*, 7:128–134, 1951.

[30] E. Lopez-Poveda and R. Meddis. A human nonlinear cochlear filterbank. *The Journal of the Acoustical Society of America*, 110:3107–18, 01 2002. doi: 10.1121/1.1416197.

[31] D. Magezi and K. Krumbholz. Evidence for opponent-channel coding of interaural time differences in human auditory cortex. *Journal of neurophysiology*, 104:1997–2007, 10 2010. doi: 10.1152/jn.00424.2009.

[32] E. Merilainen. Current drive - the natural way of loudspeaker operation, 2013. URL `https://www.current-drive.info/`.

[33] J. Micka. Spice simulation of loudspeaker using thiele small parameters, 2009. URL `https://www.micka.de/en/download/spice-tsp_en.pdf`.

[34] G. A. Miller and W. G. Taylor. The perception of repeated bursts of noise. *The Journal of the Acoustical Society of America*, 20(2):171–182, 1948. doi: 10.1121/1.1906360. URL `https://doi.org/10.1121/1.1906360`.

[35] H. Møller, P. Minnaar, S. K. Olesen, F. Christensen, and J. Plogsties. On the audibility of all-pass phase in electroacoustical transfer functions. *Journal of The Audio Engineering Society*, 55:115–134, 2007.

[36] J. Moore. Organization of the human superior olivary complex. *Microscopy research and technique*, 51:403–12, 11 2000. doi: 10.1002/1097-0029(20001115)51:4<403::AID-JEMT8>3.0.CO;2-Q.

[37] J. Murphy. Neutralizing l(e) with a zobel, 2017. URL `https://trueaudio.com/st_zobel.htm`.

[38] D. Pan. A tutorial on mpeg/audio compression. *IEEE MultiMedia*, 2(2):60–74, 1995. doi: 10.1109/93.388209.

[39] J. Parker and V. Välimäki. Linear dynamic range reduction of musical audio using an allpass filter chain. *IEEE Signal Processing Letters*, 20(7):669–672, 2013. doi: 10.1109/LSP.2013.2263136.

[40] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, Z. C.Q., and A. M. Complex sounds and auditory images. Published in: Auditory physiology and perception, Proc. 9th International Symposium on Hearing, Eds: Y Cazals, L. Demany, and K. Horner. Pergamon, Oxford, 429-446., 1992.

[41] S. Powell and P. Chau. A technique for realizing linear phase iir filters. *IEEE Transactions on Signal Processing*, 39(11):2425–2435, 1991. doi: 10.1109/78.97998.

[42] D. Preis. Measures and perception of phase distortion in electroacoustical systems. In *ICASSP '80. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 490–493, 1980. doi: 10.1109/ICASSP.1980.1170917.

[43] J. F. Schouten, R. J. Ritsma, and B. L. Cardozo. Pitch of the residue. *The Journal of the Acoustical Society of America*, 34(9B):1418–1424, 1962. doi: 10.1121/1.1918360. URL `https://doi.org/10.1121/1.1918360`.

[44] J. Smith. Introduction to digital filters with audio applications, 2007. URL `https://ccrma.stanford.edu/~jos/fp/Analog_Allpass_Filters.html`.

[45] J. Smith. Spectral audio signal processing, 2011. URL `https://ccrma.stanford.edu/~jos/sasp/Minimum_Phase_Filter_Design.html`.

[46] I. Stewart. What is crest factor and why is it important?, 2020. URL `https://www.izotope.com/en/learn/what-is-crest-factor.html`.

[47] R. Stroh. Phase shift in loudspeakers. *IRE Transactions on Audio*, AU-7(5):120–124, 1959. doi: 10.1109/TAU.1959.1166212.

[48] K. Sunder, J. He, E.-L. Tan, and W.-S. Gan. Natural sound rendering for headphones: Integration of signal processing techniques. *IEEE Signal Processing Magazine*, 32: 100–113, 02 2015. doi: 10.1109/MSP.2014.2372062.

[49] H. Suzuki, S. Morita, and T. Shindo. On the perception of phase distortion. *Journal of The Audio Engineering Society*, 28:570–574, 1980.

[50] M. Tagliaverga and A. Ragni. private communication, 2021.

[51] T. Takeuchi and P. Nelson. Optimal source distribution for binaural synthesis over loudspeakers. *JAES*, pages 2786–2797, 2002.

[52] Unknown. Measure speaker frequency response – different techniques, 2016. URL `https://audiojudgement.com/measure-speaker-frequency-response/`.

[53] G. Wentworth. On the audibility of amplifier phase distortion. *IEEE Transactions on Audio*, AU-13(4):99–99, 1965. doi: 10.1109/TAU.1965.1161805.

[54] Wikipedia. Elliptic filter, 2021. URL `https://en.wikipedia.org/wiki/Elliptic_filter`.

[55] W. Young. Notes on audio op-amps, 2016. URL `https://tangentsoft.net/audio/opamps.html`.

A | MATLAB Codes

Patterson's experience

```
%% Patterson's experience
% Simplified to 2 component audio signals
%Choose the fundamental frequency and the phase in degrees
freq = 200;
phi_deg = 45;

Fs = 44100;
Ts = 1/Fs;
omega = 2*pi*freq;
phi = phi_deg*pi/180;
time = 0:Ts:10-Ts;
sine_L = cos(omega*time);                    %generate lower tone
signal_u = sine_L + cos(2*omega*time);       %sum higher tone
signal_d = sine_L + cos(2*omega*time + phi);%sum delayed higher tone

soundsc(signal_u,Fs);                        %audio playout
pause();
soundsc(signal_d,Fs);
```

Miller and Taylor's experience

```
%% Miller and Taylor experience
%Choose the frequency and the SNR in dB
freq = 100;
SNR_db = -25;

Fs = 44100;
Ts = 1/Fs;
time = 0:Ts:10-Ts;                  %10s of time axis
```

```
signal = rand(1,10*Fs)-0.5;          %10s of noise
sq_wave = sign(sin(2*pi*freq*time));%modulating wave
signal(sq_wave<=0) = 0;              %chop the noise
mask = rand(1,10*Fs)-0.5;            %10s of noise
SNR = 10^(SNR_db/20);                %linear SNR value
out = SNR*signal + mask;             %mix signal and noise
out = out/max(abs(out));             %normalization

sound(out,Fs);                       %audio playout
```

Shouten's experience

```
%% Schouten experience
%Choose the frequency
freq = 100;

Fs = 44100;
Ts = 1/Fs;
time = 0:Ts:10-Ts;                   %10s of time axis
signal = zeros(1,10*Fs);             %allocate signal
for ii=0:10                          %10 sinusoidal    components
    this_freq = 40*freq + ii*freq;
    sine = sin(2*pi*this_freq*time);
    signal = signal + sine;
end

soundsc(signal,Fs);                  %audio playout
```

GCC microphone localization

```
%%Microphone Localization demo

time = 1; %seconds
iteration = 1;

%% ORIGINAL SIGNAL PROCESSING
[y,Fs] = audioread('Esperimenti_rifatti/frank_sinatra.wav'); %import ...
    audio
```

```
8  y = y(hop*Fs*iteration+1:round(time*Fs) + hop*Fs*iteration,:); ...
       %select a window
9
10 C = (y(:,1) + y(:,2))/2; % common mode
11 D = (y(:,1) - y(:,2))/2; % differential mode
12
13 X_cd = mean(C.*D); % cross-correlation C-D
14 X_cc = mean(C.^2); % autocorrelation C
15 A = X_cd/X_cc; % coefficient of correlation C-D
16 B = D - A.*C; % uncorrelated residue B
17
18 X_dc = mean(D.*C); % cross-correlation D-C
19 X_dd = mean(D.^2); % autocorrelation D
20 E = X_dc/X_dd; % coefficient of correlation C-D
21 F = C - E.*D; % uncorrelated residue F
22
23 %% DATA INPUT
24 mic_in = audioread('Esperimenti_rifatti/record3.wav');
25 Y1 = mic_in(hop*Fs*iteration+1:round(Fs*time) + hop*Fs*iteration,:); ...
       % a few seconds
26
27 %% GENERALIZED CROSS-CORRELATION
28 X11 = fft(Y1); % evaluate GCC in frequency domain
29 X2 = fft(B);
30 PHAT = 1./abs(X11.*conj(X2)); %PHAse Transform (whitening filter)
31 GCC1 = ifft(PHAT .* X11 .* conj(X2));
32
33 X2 = fft(F);
34 PHAT = 1./abs(X11.*conj(X2)); %PHAse Transform (whitening filter)
35 GCC2 = ifft(PHAT .* X11 .* conj(X2));
36
37 sum = -GCC2;
38 diff = -GCC1;
39 G1 = (sum + diff)/2;
40 G2 = (sum - diff)/2;
41
42 range = 1:400; %DBG PLOT
43 figure(1);
44 subplot(2,1,1);
45 plot(GCC1(range));
46 hold on%figure(2);
47 plot(GCC2(range));
48 legend('GCC Difference','GCC sum');
49 xlabel('Time [samples]');
```

```
ylabel('Correlation value');

subplot(2,1,2);
plot(G1(range));
hold on
plot(G2(range));
legend('g1','g2');
xlabel('Time [samples]');
ylabel('Relative value');

%%FIND MAXIMA
[gain1,d1] = max(G1(10:min(Fs,numel(G2))));
[gain2,d2] = max(G2(10:min(Fs,numel(G2))));

d1 = d1-1; % Off-by-1 error... MATLAB indexing...
d2 = d2-1;

dly = d1-d2;

disp("iteration:  " + iteration + ", delay: " + dly);
```

Position plotter

```
%% Compute and Plot position from delays

c = 343;                                %sound speed [m/s]
L = 2.68;                               %distance between speakers [m]
Fs = 44100;                             %sampling frequency
delays = load('detection_delay.mat'); %load data
d = delays * c/Fs;                      %convert delay into space
X = (d(1,:).^2 - d(2,:).^2)./(2*L);    %compute X
Y = sqrt(d(1,:).^2 - (X + ones(size(X))*L/2).^2);    %compute Y

scatter(X,Y);                           %plot
```

B | Theorems

Uncorrelated Residue

Let $c(t)$ and $d(t)$ be real zero-mean signals.
Let A be their correlation index, so, by definition:

$$A = \frac{R_{CD}}{R_{CC}} = \frac{E[c(t) \cdot d^*(t)]}{E[c(t) \cdot c^*(t)]} \tag{B.1}$$

Where $E[*]$ denotes the expected value and R the cross-correlation. Please note that the components related to the average amplitude of the signals have been neglected because the signals are zero-mean by hypothesis.
$c(t)$ and $d(t)$ are real, so the complex conjugation operator is not needed. We can simplify equation B.2 as:

$$A = \frac{R_{CD}}{R_{CC}} = \frac{E[c(t) \cdot d(t)]}{E[c(t) \cdot c(t)]} \tag{B.2}$$

Let $b(t)$ be:

$$b(t) = d(t) - A \cdot c(t) \tag{B.3}$$

b is also zero-mean, in fact:

$$E[b(t)] = E[d(t)] - E[A \cdot c(t)] \tag{B.4}$$
$$E[b(t)] = E[d(t)] - A \cdot E[c(t)] \tag{B.5}$$
$$E[b(t)] = 0 - A \cdot 0 = 0 \tag{B.6}$$

We can calculate the cross-correlation between b and c as:

$$R_{BC} = E[b(t) \cdot c(t)] \tag{B.7}$$

Now we can substitute equation B.3 into B.7 and we obtain:

$$R_{BC} = E[(d(t) - A \cdot c(t)) \cdot c(t)] \tag{B.8}$$
$$R_{BC} = E[d(t) \cdot c(t)] - E[A \cdot c(t) \cdot c(t)] \tag{B.9}$$
$$R_{BC} = E[d(t) \cdot c(t)] - A \cdot E[c(t) \cdot c(t)] \tag{B.10}$$
$$R_{BC} = R_{DC} - A \cdot R_{CC} \tag{B.11}$$
$$\tag{B.12}$$

Substituting equation B.2 into B.11, we finally get:

$$R_{BC} = R_{DC} - \frac{R_{DC}}{R_{CC}} \cdot R_{CC} \tag{B.13}$$
$$R_{BC} = R_{DC} - R_{DC} = 0 \tag{B.14}$$

Showing that $b(t)$ and $c(t)$ are uncorrelated.

Acknowledgements

Grazie...

Ai compagni del calcizzu
Ai colleghi di HRE
Alle antenne di Silvano
Alle scatole di Ste

All'affetto della mamma
Al coraggio di papà
Alle belle canterine:
Anna, Miri, Bea e Fra

A Jack Longo e alle montagne
Alla Ceci e al suo gattino
Agli aerei di Orlandone
Alla nonna e al suo Crodino

Ai pinguini di Lou lou
A zio Bos ed a zio Pidi
Ad un Merlo cantastorie
Ai gelati della Giudi

Alla macchina di Pera
A zio Kave, Pavo e Ferdone
Alle birre col Mariolo
Alle casse di Simone

Ai pranzoni con i nonni
Ai balletti con Doriana
Alle maschere di Chiara
Alla voce di Andriana

Alla musica, ai sorrisi
ed al coro degli alpini
Alle mie care sorelle
Ai cognati e nipotini

Ai miei tutor Marco e Andrea
Ai colleghi fulminati
Alle valvole di Pino
Alle cuge Luci e Mati