MODELING OF GESTURE-SOUND RELATIONSHIP IN RECORDER PLAYING: A STUDY OF BLOWING PRESSURE

LENY VINCESLAS

MASTER THESIS UPF / 2010

Master in Sound and Music Computing

Master thesis supervisor:
Esteban Maestre
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona

ABSTRACT

This work deals with instrumental gestures and their relation to sound in recorder playing. In particular, it presents a study of the relationship between the blowing pressure and a number of timbre characteristics of produced sound.

Blowing pressure is acquired by means of a pressure transducer mounted on the mouthpiece of the recorder, while sound is acquired using a microphone and analyzed in the spectral domain via SMS in order to extract spectral shape descriptors corresponding to odd harmonics, even harmonics and residual components. A performance recording is carried out so that a number of representative playing contexts are covered. Collected data (blowing pressure, produced sound, and spectral analyses) are segmented and used to assemble a multi-modal performance database.

Starting from the acquired data (blowing pressure and spectral descriptors), supervised machine learning techniques are used to construct and evaluate two types of models. The first model takes spectral descriptors as input and produces an estimation of the blowing pressure, in a frame-by-frame fashion. The second model takes the blowing pressure as input and produces an estimation of the spectral descriptors, also in a frame-by-frame fashion. While the former lays in the domain of indirect acquisition of instrumental gestures, the application context of the latter corresponds to spectral-domain timbre synthesis. Both models are evaluated via cross-validation, while the synthesis-oriented model is tested by means of additive synthesis via SMS.

Key words: instrumental gestures, recorder, blowing pressure, spectral descriptors, SMS, spectral modeling synthesis, additive synthesis, indirect acquisition, machine learning, artificial neural network.

ACKNOWLEDGEMENTS

I would like to thank Xavier Serra who gave me the great opportunity to study and to research combining two of my passions: music and technology.

I would like to thank as well all my classmates from the Master *Sound and Music Computing* and chiefly people from the room 312 who accompanied me during this year of studies in Barcelona. In addition a general thank goes to the whole MTG crew for their kindness and their availability.

I warmly thank Alfonso Perez for his valuable advices and friendly help on Artificial Neural Network.

I greatly appreciated to meet Joan Vives, a music history teacher and recorder player who was our first contact. He provided us with a lot of information and introduced us to remarkable people who were extremely valuable for the project.

I am extremely grateful to Juan Izquierdo, professional recorder player, for his voluntary work and his patience, which made possible the recording of the database.

My deepest gratitude goes to Josep Tobau, experimental flute maker, who took part of the project as a real member by sharing his brilliant technical points of view, imagining and constructing an unexpected prototype of recorder. Moreover his perpetual brain storming allowed remarkable progresses on the project.

During this work I collaborated with Francisco Garcia Diaz-Castroverde who has been working on his Master thesis in a complementary topic and whom I would like to sincerely thank for his team spirit.

I especially want to thank my supervisor, Esteban Maestre, for his guidance and his support during my study at the *Universitat Pompeu Fabra*. His perpetual energy for research kept me motivated throughout the year and furthermore, he was always available and enthusiastic to help his students with their research. As a result, this peaceful and rewarding working environment, gave me the will to pursue my career in that direction.

CONTENTS

LIST OF FIGURES

1. INTRODUCTION

The sound produced by recorders depends not only on the physical characteristics of the instrument but also on the gestures of the musician. In the case of the recorder, the main parameter which can be controlled by the performer is the blowing pressure. The next paragraph explains how the blowing pressure has an incidence on the timbre of the recorder. Several previous researches let us imagine that it could be possible to estimate the blowing through indirect acquisition and in the opposite way to reconstruct the timbre from the blowing pressure.

1.1. MOTIVATION

Nowadays, the most ambitious challenge of the digital music field is to faithfully synthesize the sound of musical instruments. Over the past years the synthesizing techniques have been improved to lead to two main techniques: physical models and spectral modeling fed by pre-recorded samples.

Currently most of the synthesizers use pre-recorded samples with spectral modeling. This technique requires a large bank of samples, which take up much storage space and needs to be recorded before building the synthesizer. There are several implementations of this technique where spectral modeling allows applying changes to the recorded samples, but the final sound will still be strongly dependent on the used samples.

Physical models achieve good results for impulsively excited instruments (as percussive instruments), but become very complicated and not very realistic for continuously exited instruments such as wind instruments or bowed string. Physical models of those instruments suffer from a lack of appropriate input control parameters (to be developed).

1.2. OBJECTIVES

The goal of the present field of research is to provide a better understanding of processes involved in sound production of recorder like instruments. Whereas the physique of the instrument starts to be very well understood and allows the creation of physical models, the instrumental gestures which produce the sound has been less studied. The purpose of this thesis is to establish a relationship between the timbre of the generated sound and the instrumental gestures applied by the recorder player.

This research is carried out following the given methodology: In a first time a prototype of recorder capable of measuring the blowing pressure is built. It is then possible to record sound from the instrument and the corresponding blowing pressure. In a second time a set of scores covering the main instrumental contexts is defined, and then played by a recorder player while audio and blowing pressure are recorded. The next step is to analyze the recordings of sound

and blowing pressure to extract spectral descriptors describing the evolution of the timbre according to the blowing pressure. These descriptors, the blowing pressure and the sound are then segmented and assembled in order to build a multi-modal performance database. Thanks to this database, artificial neural networks are trained with the aim of building models capable of estimating the spectral shape of the recorder sound from the blowing pressure or estimating the blowing pressure from the audio spectrum. Once the accuracy of the models is sufficient, a prototype of sound synthesizer can be constructed as well as a prototype of indirect blowing pressure acquisition.

The general objective of this study is to demonstrate the possibility of modeling the timbre of the recorder using the previous method. Going through these steps we can highlight several sub-goals:

- Study of the blowing technics in recorder performance: for a basic set of techniques using blowing pressure as the main instrumental control parameter, and define the typical profiles of the blowing pressure applied by the performer.

- Study of the timbre representation: how to model the timbre of the recorder in an accurate and efficient way. The representation should be close to the real timbre using only few numbers of parameters.

- Study of a relationship between audio features and bowing pressure: it is obvious that the timbre of the sound generated by the recorder relies on the blowing pressure; we can then establish a correlation.

- Development of a low-intrusive technique for blowing pressure acquisition: since there is a relationship between the timbre of the recorder and the blowing pressure applied by the flutist, we can estimate the blowing pressure directly from the sound without bothering the player with sensors on the recorder.

1.3. OUTLINE

This thesis is structured in 5 chapters. The two first are introductory. The first chapter introduces the motivation of this work, the scope, the main goals and the sub-objectives. The second chapter provides the general background of the present research. Particularly it gives an introduction on sound production in the recorder followed by a literature review going through instrumental gesture parameters and techniques for acquisition of instrumental gestures. Finally, Chapter 2 gives some examples of commercial applications of controller and synthesizer.

Chapter 3 describes the techniques developed to acquire sound and gestures from real performances. The procedure applied is detailed in three steps: firstly scores are created to cover a representative recorder playing context. Then, a prototype of recorder is used to measure the blowing pressure. The recording session is carried out by using an adapted installation. All collected data is finally preprocessed to build a database gathering sound and gesture.

The following two chapters (4 and 5), contain the main contributions of the thesis. Chapter 4 presents the spectral analysis algorithm used to extract spectral parameters and describes the spectral evolution of the recorder timbre according to the blowing. Different methods for timbre representation and spectral reconstruction are exposed and the algorithm developed for re-synthesis is detailed. Chapter 5 starts by giving an overview on the method used for sound synthesis and indirect acquisition. In this chapter, the four models listed predict timbre or blowing pressure. Based on these models, predictions are discussed. This chapter concludes with a presentation of the algorithm developed for sound synthesis.

2. BACKGROUND

In this chapter, a brief review of the work related to the topic of this master thesis is presented. The first part of the chapter introduces the scientific context of the researches: the sound production in the recorder. The second part of the chapter is dedicated to the previous researches of the discipline. It firstly exposes the instrumental gesture parameters and then discusses techniques for acquisition of instrumental gestures. Finally, the chapter gives some examples of commercial applications of controller and synthesizer.

2.1. SOUND PRODUCTION IN RECORDER PLAYING: AN OVERVIEW

2.1.1. SOUND PRODUCTION

Sound is a travelling wave made of oscillations of pressure transmitted through a solid, liquid, or gas. With the intention to produce a sound, an instrument has to transmit pressure oscillations to a medium. Wind instruments are regarded as continuously exited instruments (opposing with impulsively excited instruments), in the sense that the performer has to constantly interact and supply energy to the instrument in order to generate a sound. In the case of the recorder, the provided energy comes from the flutist who blows a rapid jet of air across the embouchure hole. This air jet is a continuous source of energy which the flute will make oscillate. Regarding this conversion phenomenon, the recorder can be divided in two main parts: The generator of oscillation which converts the continuous flow in oscillations and the resonator which set the frequency of the previous oscillation.

2.1.1.1. The generator

The generator corresponds to the head of the recorder. It is composed of the mouthpiece and the windows.

The mouthpiece includes two subparts: The windway and the block. The windway is the small canal (~1.5mm) through which the airstream goes. The windway is cut in the block. The block is in fact the base of the windway, it is a wooden piece usually made with a softer wood than the rest of the recorder in order to avoid possible crack of the instrument while the human blow introduces a large amount of humidity.

The window is composed of three critical parts: the labium, the air jet, and the distance between the labium and the windway: The labium can be considered as the master piece of the flute. The labium constitutes the edge on which the air flow is directed. This very thin and delicate edge makes the air flow oscillate around its both sides (inside or outside the recorder). It is this oscillation which allows to generate a tone and then to interacte with the air column in the resonator of the recorder. The distance between the labium and the end of the windway, and the offset of the labium from the center of the jet are parameters generally used by makers to ensure a prompt attack for a suitable range of blowing pressure. The action of setting these

parameters is called “the voicing”. When the air jet leaves the windway, it spreads and slowdowns because of frictions with the surrounding static air. When the air jet meets the labium, it starts to oscillate around the latter. While the air jet is providing the whole oscillatory system with energy, it is this system which defines the frequency of the jet air movements. The next picture represents the two farthest states in the oscillation of the air jet.

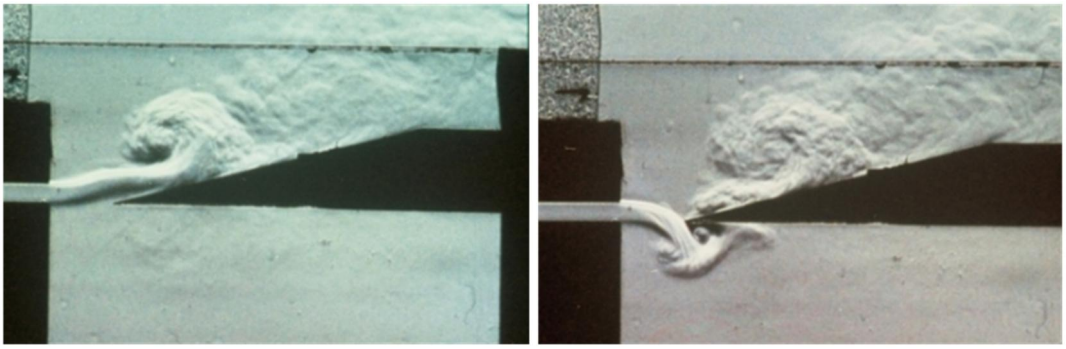

Figure 1: The air jet oscillating around the edge of the labium.
(University of Eindhoven, Netherlands)

2.1.1.2. The resonator:

The resonator corresponds to the bore, the middle piece of the recorder. The acoustical behavior of the recorder depends strongly on the resonance patterns of the bore and on the position of the holes. Assuming the bore can be modeled as a two ends open pipe, and assuming a pressure excitation at one end of the bore we are able to predict the air motion occurring into the tube. In a first time the pressure pulse produced by the excitation start propagating from one end of the bore until it reaches the second open end. At the open end a difference of pressure (acoustic pressure traveling into the recorder and atmospheric pressure) act as a difference of impedance and reflect part of the longitudinal waves. Since the reflected wave is propagating in the opposite direction that the incident wave, a stationary wave result of the addition of the both waves. Note that a complete cycle of vibration is the time taken for the pulse to travel twice the length L of the tube (once in each direction). Given that the pulse travels at the speed of sound c, so the cycle would repeat at a frequency of $c/2L$. Since the two end are open, the acoustic pressure at these points is null (atmospheric pressure in fact), they are called pressure nodes. Inside the tube for the first mode of vibration (lowest frequency) there is one point having a maximum pressure variation higher than in the whole tube. This point is called pressure anti-node and occurs at the middle. The figure 2 shows in red the stationary wave in different modes of vibration. The blue line represents the variation in the displacement of the air molecules. The latter curve has anti-nodes at the ends: air molecules are free to move in and out at the open ends.

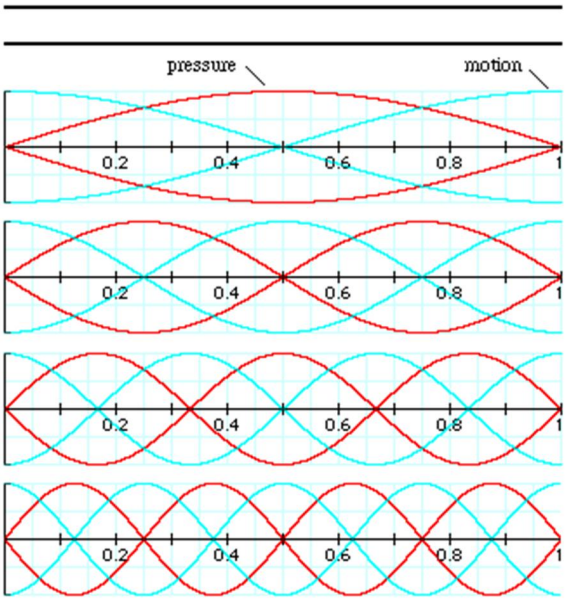

Figure 2: Stationary wave in an open tube [8]

The modern recorder has a total of 8 holes. The positions of these holes are very carefully chosen since there are the only ways to tune the instrument and they cannot easily be modified after the fabrication. The holes are not equally spread over the bore, and they can have a different size. The aim of holes is to control the length of the air column, and then the frequency of its oscillation.

Assuming all holes are closes, the instrument uses its whole air column length. If starting from the far end the holes are opened (we move the point where the pressure must be null), the pressure node move up, the air column gets shorter, and then the frequency of the oscillation increase. The pitch gets higher. The fundamental mode of the recorder involve on node at the middle of the tube, and an antinode at each ends. The figure 3 shows that opening a hole on the middle, a node is created at the same position and the oscillation frequency is the multiplied by two.

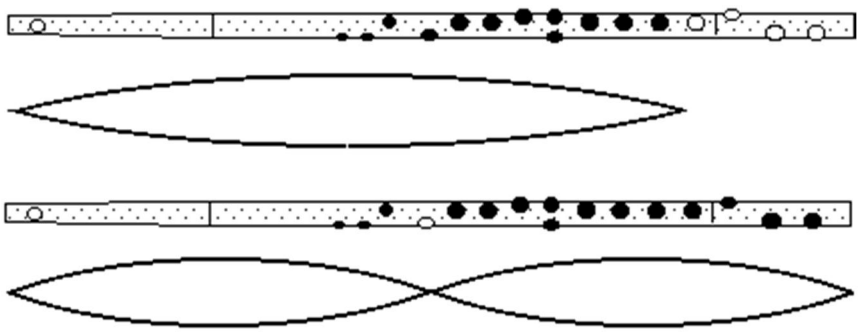

Figure 3: Oscillating wave length according to opening tone holes of flute[8]

Measuring the active length of the bore and ignoring the end correction, we are now able to predict the fundamental frequency of an oscillation.

$$f0=\omega 0/2\pi=c/2L$$

Where *f0* is the fundamental frequency, *c* the celerity of sound in air and *L* the length of the tube.

The spectral properties of the sound of the recorder are strongly dependent on the facture of the flute and on the blowing pressure applied. Indeed, while the odd harmonics are more present in the spectrum, figure 4 shows that their evolution is not coupled with even harmonics. The amplitudes of the even harmonics and of the second harmonic depend quite critically on voicing adjustments of the recorders and of the blowing pressure. These behaviors should be a consequence of the non-linearity of the oscillating system [22] [2].

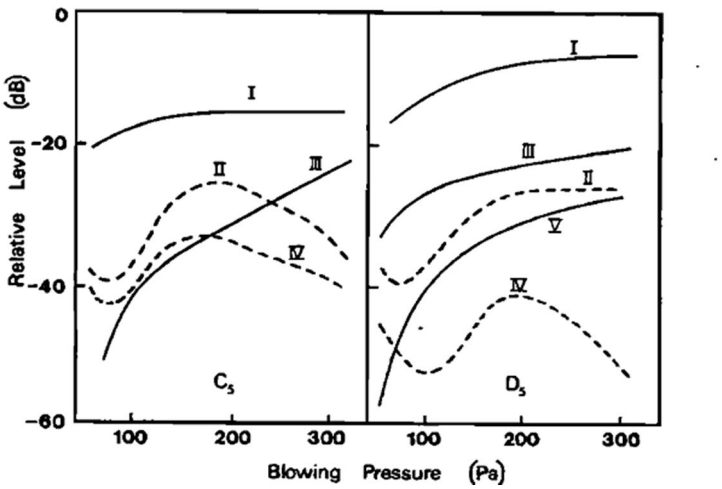

Figure 4: Measured relative sound pressure levels of the harmonics for the notes C5 and D5, as a function of blowing pressure, played on an alto recorder [22].

2.1.1.3. Interaction between the Generator and the resonator

As we have seen before, the generator is able to convert a linear air flow in oscillations, which interact with the air column contained in the bore. When the air jet goes outside the recorder (over the labium or north position) the pressure in the air column decreases and the pressure pulse goes from the end of the recorder to the top. When the air jet goes into the recorder (under the labium or south position), the pressure increases and the wave goes from the top to the end of the recorder. Therefore this air column is vibrating lengthwise according the length of the tube which is defined by the holes positions.

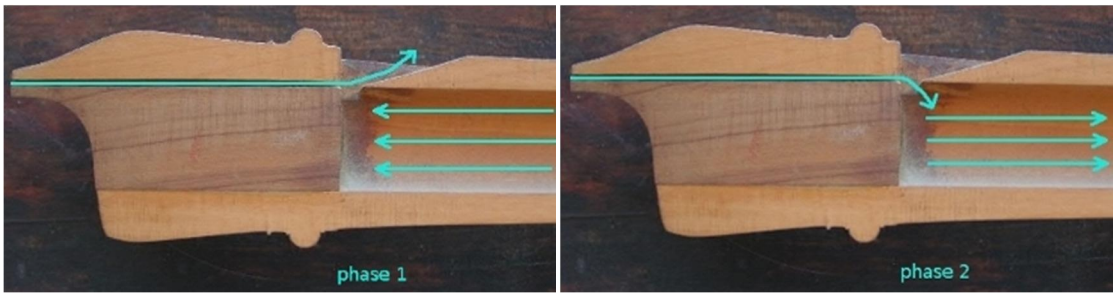

Figure 5: The interaction between the windway and the recorder's air column [7]

The working of the recorder can be simplified by a looped system where the air jet propagates to the labium which creates the oscillation that then resonates in the bore (*Figure 6*). Taking into account the phenomenon of interaction where the vibrations of the air column affect the formation of the air jet, it can be assumed that there is a feedback from the resonator to the air jet(*Figure 6*). Physical models of flute-like instruments can be based on this modeling which produces a quiet realistic sound but generates non-convincing transients.

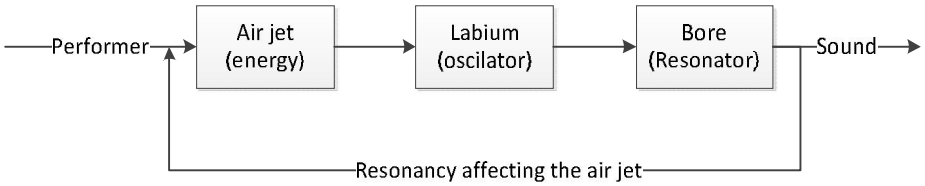

Figure 6: Simplified representation of the interactions of energy

2.1.2. PERFORMANCE TECHNIQUES

Despite the recorder appears as a simple instrument, it has a large range of possible playing techniques. A skilled player can play over two octaves; the use of the third octave is more applied to modern compositions. To reach the notes of the second and third octave, the player must blow harder in order to make the air column vibrating.

2.1.2.1. Articulation

If a note is started with the same blowing pressure as the final pressure, the initial frequency would be the resonance frequency of the bore. In other terms the beginning of the note would not be in tune. The recorder player is able to control the beginning (attack) of a note and its end. In order to immediately reach the right tone the player must blow harder while starting the note. With the aim to increase the pressure and accelerate the first oscillations, the player uses a technique called tonguing. The tonguing is a part of the articulation technique and according to the wanted dynamic there are several possible articulations. The most used articulations can be split into two groups: the simple articulations (T-D-L) and the compound articulations (K-G-R). Those articulations could also be classified from the briefest staccato (T-K-G) to the broadest legato (D-L-R). A group of articulation is called a syllable (du-ru-du-ru or tu-ku-tu-ku). The use of articulation is very important in order to give the perception of subsequent notes.

Discussing about articulation during the recording session, the flutist Juan Izquierdo said" In music, silences are more important than the notes themselves".

2.1.2.2. Vibrato

The vibrato is a technique used to add expression to the notes played. Increasing and decreasing the blowing pressure (around 5Hz) the vibrato produces a change of pitch, loudness

and timbre. A similar effect can be achieved using a “finger vibrato” or “flattement” in which a trill[1] is executed by covering or discovering a hole. Regarding these both methods, the compromise is that the finger vibrato can be done faster, while the breath vibrato can be performed independently of the fingers positions.

2.1.2.3. Multiphonics

In the recorder, when changing the blowing pressure, two or three different notes can be played using the same fingering. Most of the time there is a well-defined transition from one note to another. But for some fingering there is a wide range of pressures where both notes sound. This phenomenon is called multiphonics. Another method to play a multiphonic sound is to use the articulation “R” which produces a rapid alternation of the blow pressure.

[1] A trill is a musical ornament consisting of a rapid alternation between two adjacent notes.

2.2. LITERATURE REVIEW

2.2.1. INSTRUMENTAL GESTURE PARAMETERS AND INFLUENCE OVER THE RECORDER TIMBER

Traditional musical instruments are usually performed by a musician interacting with a control surface made of keys (piano, clarinet), strings (violin), mouthpieces (trumpet), reeds (oboe), etc. Instrumental gesture is the actual instrument manipulation and playing technique on an instrument [14]. From a functional point of view, the gesture is necessary to mechanically produce a sound (like blowing in a flute, bowing on a string or pressing a key of a piano). It is so called instrumental gesture parameters the parameters characterizing the components of the instrumental gesture.

The recorder can be considered as one of the simplest instrument of the woodwind family. While a clarinet, an oboe or a flute allow performers a wide control over tones, dynamics and timbre by changing the position of the mouth, in the recorder, the distance between the windway and the labium is fixed by the maker. Therefore we can consider that only a few numbers of instrumental gesture parameters are able to influence the sound of the recorder. We mainly define three control parameters: The speed of the air jet (blowing pressure), the volume of the mouth (mouth shape) and the way of opening or closing a hole with fingers (fingering). The variation of these parameters affects the timbre and is usually clearly perceived by a trained listener [15].

2.2.1.1. Mouth shape and vocal tract

There is currently no consortium on this point and the effects of the shape of the player's mouth on the sound of the recorder generate many disagreements. If we can hear an alteration of timbre by changing the shape of the mouth, it is not really obvious and most of the people do not notice it. The mouth and the windway form together a Helmholtz resonator. By varying the volume of the mouth the Helmholtz resonator can reach a resonant frequency close to the resonant frequency of the bore and in this case, it can change the resonance of the whole instrument. This resonance modification alters mainly the harmonic content of some notes. In order to achieve this change of timbre the sound of the recorder must travel back up to the mouth through the windway (Helmholtz resonator), resonate, and then travel back to the recorder and alter the timbre [4].

Skilled performers regularly report that manipulating their vocal tract configuration can change the timbre, pitch of the recorder sound. Measuring the acoustic impedance of the vocal tract during a performance has allowed investigating different vocal tract configurations: low palate and high palate. The two performance gestures studied gave significantly different spectra, consistent with a substantial change in the constriction at the back of the tongue. A systematic difference in harmonic structure was observed between the two gestures, and the

more compressed the 'low palate' gesture is, the more a boost in broadband signal at around 6-8 kHz is observe[16].

In the previous study they noted that harmonic structures of recorder notes played using the two throat configurations are similar, with pitch typically differing by not more than five cents. Using Long-term average spectral, the figure 16 shows a comparison of the harmonic contents of notes F4-E6 (left) and the noise content (right). We see the presence of stronger harmonics in the 1.7 to 3.7 kHz region for "low palate" than for "high palate". We note also that from 6 to 8 kHz, the "low palate" broadband signal is typically 3-6 dB stronger than that of the "high palate", for sounds having similar volume flow velocity [16][17].

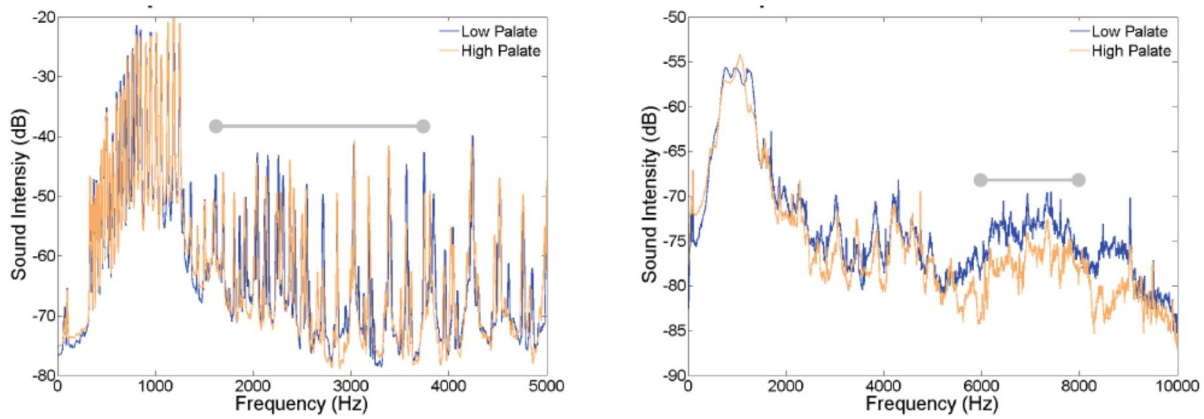

Figure 7: Long-term average spectra of recorder harmonic contour (left) and noise floor (right) [16].

2.2.1.2. The fingering

Regarding different style (Spanish, Dutch or French) the fingering is more or less active on the sound production process. For example the gesture of the French style influences a lot the sound of the recorder since fingers are moved (flattement) during a note. In other styles, fingers are less used to shape the sound, but they are still playing an important part when performers use legato articulation.

2.2.1.3. The blowing pressure

The blowing pressure is in fact the gesture which has the strongest contribution on the sound production process. Actually the main playing techniques act only on the blowing pressure. Changing the pressure of the air flow allows varying the loudness of the notes, the tone or the timbre. Blowing harder the performer is also able to reach different notes keeping the same fingering.

Different studies allow to make a correlation between the frequency of the played note and the strength of the blow. For the transverse flute the blowing pressure is proportional to the frequency. Other researches on the flute have measured a growth of harmonics when the blowing pressure is increasing. The whole spectrum gets shifted to the high frequency.

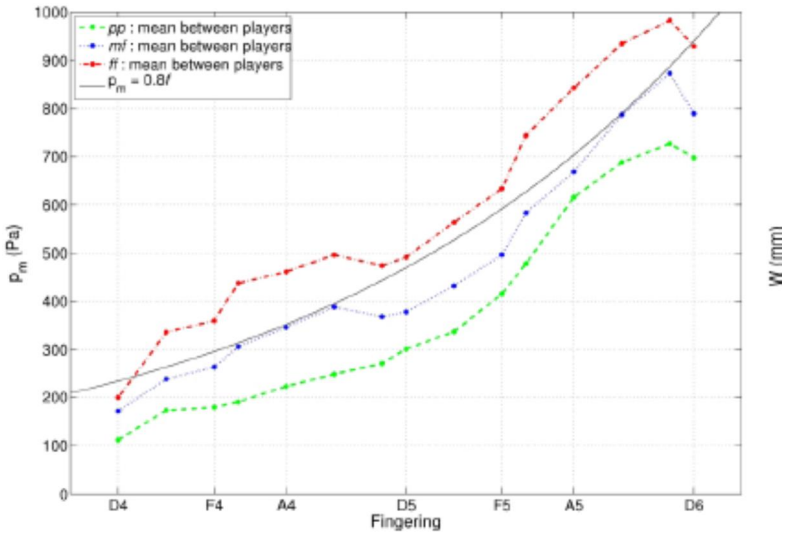

Figure 8: Representative blowing pressure for different notes played with different dynamics. [5]

Figure 8: Twelve flutists were asked to play ascending diatonic D scale with different dynamics (pp, mf and ff). The curves drawn are the mean between the players regarding the note played for several dynamics. Differences between pp, mf and ff are consistent regarding the played note. When the flutist wants to play louder he blows with roughly 100Pa more and when he wants to playing softer, he blows with 100Pa less. It is also interesting to notice that the pressure range is growing with the frequency of the played note.

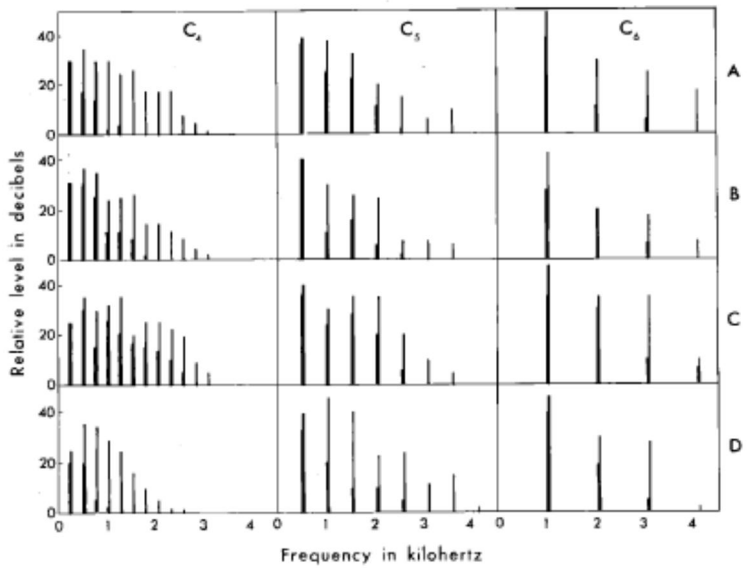

Figure 9: Harmonic analysis of forte and piano notes played by A, B, C, and D. [1]

Figure 9: In the lowest octave of the flute, and for loud playing, the fundamental is lower in level than either the second or third harmonics and lower than the fourth and fifth harmonics as well. When the playing is soft in this octave, the level of the fundamental is the same as for loud playing but the relative levels of all higher harmonics are decreased. This decrease is much more pronounced for players A and D, who reduce wind pressure for soft playing, than for B and C, who use a constant wind pressure. For the middle octave of the flute, the fundamental becomes

the dominant partial for both loud and soft playing, though second and third harmonics are within 10 dB of it in relative level. The sound-pressure level of the fundamental changes little with dynamic level and most of the change is represented by changes in the upper partials. In the third octave the fundamental is clearly dominant and all upper partials are more than 10 dB below it in relative level. The fundamental changes considerably with dynamic level, though still not as much as do the upper partials.

It has also been proved that blowing pressure contains significant frequency content up to 1 kHz and beyond [18]. The highest-frequency components result from vibrations of the vocal folds, most typically at a periodic rate with associated harmonics. These "vocalizations" subsequently modulate the oscillations of the air column under playing conditions. In order to watch those high frequency components it could be possible to couple a traditional breath sensor (limited bandwidth) with an internal microphone [18].

Since the blowing pressure affects the timbre features of the recorder sound, many different spectral features are influenced by its changes (spectral centre, spectral flux, transients, density)

Figure 10 illustrates the work done on a very well-known behaviour of the flute-like instruments: on a same fingering, the pitch of the sound increases as the blowing pressure gets stronger. The present study [26] highlights the existence of a hysteresis when the instrument changes its mode of vibrations. It is noticeable that the threshold of the pitch jump does not happen for the same blowing pressure according that the strength of the air jet is increasing or decreasing.

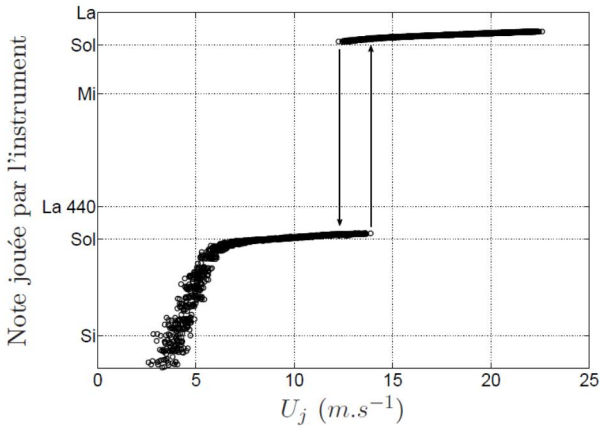

Figure 10: Fundamental frequency of recorder's sound according to the speed of air jet [27]

2.2.1.4. Non-obvious Performer Gestures

We can consider non-obvious gestures as a class of wind instrument gestures that are produced by means of moving the instrument during the performance (lifting it up/putting it down, to one side or another, fast tilt-like gestures, etc)[13].

It has been verified by measurements and simulation that the variations of the direct sound and floor reflection attributes for a specific movement cause sinusoidal partial amplitude

modulations that play an important role in the resulting sound for a particular microphone position. As the radiated spectrum can be different regarding the listening point, moving the sound emitting point from the auditory or the microphone can also change the perceived timbre.

As it follows *Figure 11* presents the partial amplitude modulation of a D3 played by a clarinet and recorded in a reverberant auditorium by a microphone (two meter in front of the instrument). In this situation the performer was playing normally and so he was moving. The *Figure 12* shows the note recorded in the same auditorium with a clarinet immobilized by a mechanical apparatus. It is clearly noticeable that both situations do not produce the same spectral variations.

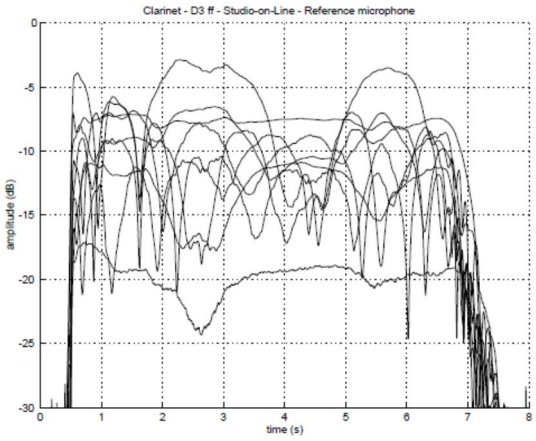

Figure 11: Clarinet recordings in a reverberant auditorium - D3 recorded, standard playing microphone 2 meters in front of the instrument [13]

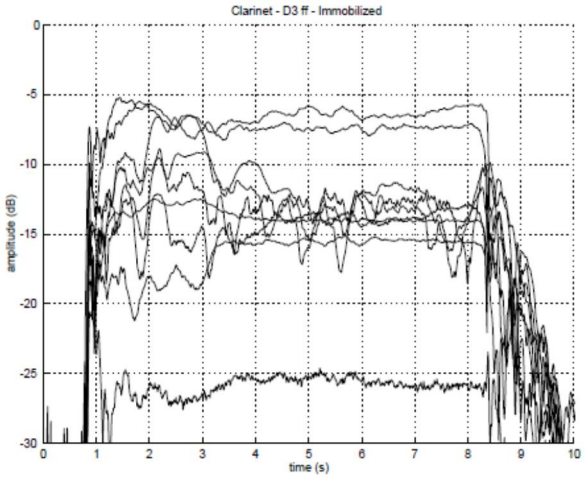

Figure 12: Clarinet recordings - D3 recorded in a reverberant auditorium – clarinet immobilized by mechanical apparatus [13]

2.2.2. ACQUISITION OF INSTRUMENTAL GESTURES

Nowadays there are two common approaches for acquiring data: Direct acquisition which measures the physical event using a dedicated sensor and indirect acquisition which tries to measure the repercussions of the event over its environment. Applied to the measurement of instrumental gesture, each technique has its own strength and drawback.

2.2.2.1. Direct acquisition

Direct acquisition of gesture will provide a hybrid instrument requiring modifications, which is frequently undesirable for the sound quality of the instrument and for the conveniences of the performer. Acoustical instruments can be modified with force sensor, accelerometers or other kind of sensors. The purpose of these sensors is to acquire directly the gestures of the performers interacting with their instruments. One problem of these "augmented" instruments is that there are usually only a few versions, and the builder is the only one able to manage the data coming from the used sensors. Another problem is that it can be expensive to purchase the electronic parts. However they provide relatively straightforward and reliable measuring capability.

2.2.2.2. Indirect acquisition

Indirect Acquisition of gesture requires only the measurement of the acoustic signal. This signal is provided using noninvasive sensors such as microphones. If the pickup process seems very simple the recorded data requires a sophisticated and a computationally expensive signal processing in order to extract the relevant information from the audio signals. One common example is the use of automatic pitch detectors to turn monophonic acoustic instruments into MIDI.

The following *Figure 13* schematizes the implementation of an indirect acquisition for gesture of classical guitar extraction.

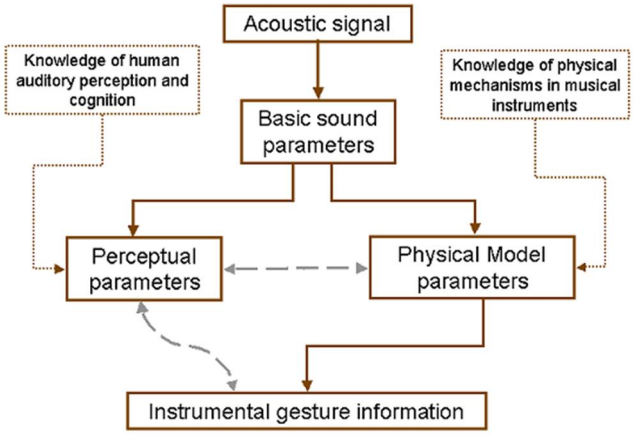

Figure 13: From acoustic signal to instrumental gesture information [15]

In this model the first stage is to extract basic sound parameters from the acoustic signal. These low-level parameters include the short-time energy, fundamental frequency, spectral envelope, amplitudes, frequencies and phases of sound partials, and power spectral density. Knowing the physical mechanisms occurring in musical instruments, it is possible to establish a correlation between the low level parameters which have been extracted and the gesture which causes those parameters. In this particular model perceptual high-level parameters are also used. These perceptual measures are also derived from basic sound parameters and they include spectral centroid, spectral irregularity, odd/even harmonic ratio, low/high harmonic ratio, and log-rise time. These parameters are interesting because they are correlated to perceptual attributes such as brightness, metallic quality and nasality. A correlation can be found between perceptual attributes and instrumental gesture parameters. For example, plucking a string closer to the bridge increases brightness. Modifying the angle of the air jet on the mouthpiece edge of a transverse flute affects brightness as well [15].

Using indirect gesture acquisition, it is very important to define whether or not the non-obvious gestures should be considered. Indeed if in the wanted model there are not taken into account the microphone should have a relative position stationary regarding the instrument.

2.2.2.3. Learning indirect acquisition

One way to solve the disadvantages of both methods is to implement a Learning indirect acquisition. At the beginning the direct and the indirect acquisition are used at the mean time in order to acquire data from the performer's gestures and to record the resulting sound. Then the recorded sound has to be annotated by hand, or by an automatic annotation algorithm extracting relevant spectral features. The next step is to train a model using the two correlated set of data obtained by direct and indirect acquisition. Once the model is trained, it will be able to interpret the data coming from the indirect acquisition according to the gesture parameters which have been used for the training. The model becomes a virtual sensor output.

This method has been successfully implemented for violin in the work [11] [19] where the performers actions and audio are captured with a sensing system and are used to train a model based on neural networks. The trained model is able to predict a sequence of spectral envelopes corresponding to a sequence of input actions (generative timbre model [21]). It is used for sound synthesis, either alone as a pure spectral model or integrated within a concatenative synthesizer. If used alone, the predicted envelopes are filled with harmonic and noisy components. As input the model receives sound and Performance Controls parameter such as the bowing controls (bow-bridge distance, bow force and bow velocity. Then a timbre Representation is computed by the mean of harmonics and residual components envelope. The whole representation takes advantage of a perceptual model such as the Mel scale.

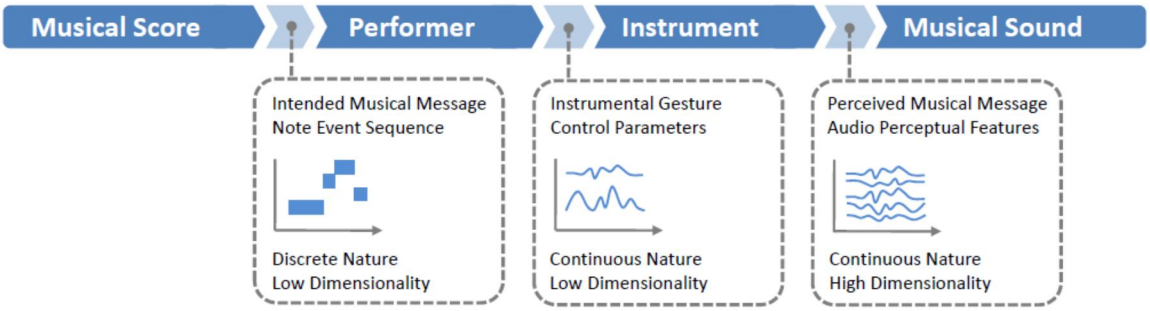

Figure 14: Acquisition of musical score, instrumental gestures, and produced sound [11]

The *Figure 14* shows the acquisition process of music performance from the score to the final sound. The intended musical message is represented on a score which contains low dimensionality information. This information is stored by the system and will be used by the model. Then reading the score the performer converts low dimensionality information in relevant instrumental gestures which produce the desired musical sound. The next set of acquired data provides more explicit evidence of instrumental gestures than the musical score. As a result of instrumental gestures applied by the performer, the instrument produces the sound perceived by the listeners. At this point, acquiring the audio signal and computing audio perceptual futures would give a higher dimensional set of data. This whole process allows to build the data base needed to train the model.

2.2.3. COMMERCIAL APPLICATIONS OF CONTROLLER AND SYNTHESIZER

2.2.3.1. Digital controllers (pseudo-instruments)

A digital controller can be defined as an electronic instrument. Digital wind controllers are, for no logical reason, mostly played and fingered like a saxophone or clarinet. A digital wind controller might convert fingering, breath pressure, bite pressure, finger pressure, and other sensors into control signals which are then used to control internal or external devices such as MIDI synthesizers. Generally these controllers do not produce sound although a few have a built-in synthesizer. They are usually connected to a MIDI synthesizer which is then connected to an amplifier. For this reason, digital controllers can sound like any instrument and is only limited by its measuring capabilities. The fingering and shape of the controller are not related to how the wind controller sounds. Some examples of famous wind controllers are the Yamaha WX5 or the Akai EWI.

2.2.3.2. Augmented instruments

Analog controllers are in fact augmented acoustic instruments. The main difference with digital controllers is that they can produce sound without using any electronic devices. The purpose of analog controllers is more oriented toward research, where there is the need to understand the real instrumental gestures. These controllers are usually mapped to physical

model in order generate a synthesized sound. Here we have the examples of the augmented violin of the Music Technology Group of the university Pompeu Fabra, or the one of the IRCAM.

2.2.3.3. Concatenative and physical modeling synthesis:

Concatenative synthesis methods use a large database of source sounds, segmented into samples, and an algorithm that finds the sequence of samples that match best the target sound. The selection is performed according to the descriptors of each sample, which are extracted from the source sounds. The selected samples can then be transformed to fully match the target specification, and are concatenated. However, if the database is large enough, the probability is high that a matching sample will be found, so the need to apply transformations is reduced. Nowadays most of the commercial synthesizers are based on sample concatenation where, while these systems provide recording sound quality, concatenations can be often noticed and they are not as flexible and controllable as physical models. Sample based synthesizers offer by far the best sound quality [19] [20].

On the other hand, physical models are creating and controlling a physical process that produces the sound. Defining the physical process which models an instrument, and then specifying some of the input parameters, the model will generate the sounds. A popular technique called waveguides allows implementing physical modeling synthesis in an easier way that using a mathematic approach. Instead of solving again and again a sets of equations to produce a sound, waveguides simulate the operation of the instrument. Once the equations have been solved, there are used in the designing process of filters which come up with the virtual instrument. Physical modeling is quite flexible and a single algorithm can achieve a variety of sounds, without requiring more memory.

In the study [10] a physical model of flute is built based on its mathematical representation. The equations model the windway, the blown air forming the jet, the oscillation due to the instability at the output of a channel and the coupled acoustic resonator (*Figure 15*). In this study the main progress comes from the modeling of control of the performer and its including in the final model (*Figure 16*). This model of flute player is based on measurements carried on instrumentalists playing on a transverse flute. The model is generating the basic features of the instrument control in order to produce given pitches and dynamics. The coupling with a flute synthesis algorithm by physical modeling enhances the quality of the control of the virtual instrument.

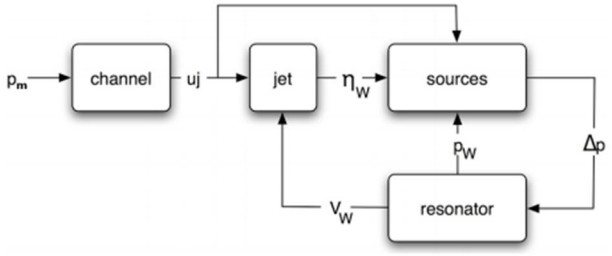

Figure 15: Diagram of the physical model of flute [10]

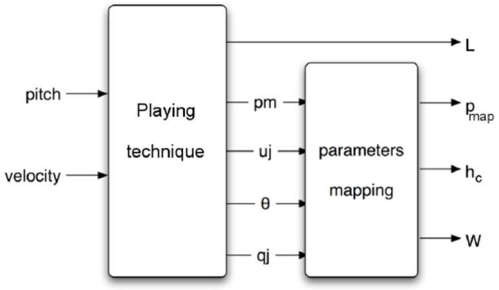

Figure 16: Flutist modeling [10]

2.2.3.4. Gesture acquisition in live performance

In order to generate in real time relevant input parameters, it is possible to map a Human-computer interface to the physical model of music instruments. Typically performers use classic MIDI interfaces such as piano keyboard or electrical sensors where the analog signal is converted into MIDI to control synthesizers. As both instruments are played with different gestures, controlling a violin model with a keyboard would give a poor sound with dynamic close to the dynamic of the piano. Consequently, we attend today to an adaptation and improvement of real-time controllers where the design tends to take into account the real gestures needed to play a virtual instrument. In live performances, electronic devices have to consider the delay that they introduce in the performance. A study has settled that while a large delay (more than 50ms) would make impossible the performance, a too short delay does not seem natural to the performers [23]. In fact the performer is capable to adaptation over the delay and it appears that the most important in live performance is to provide a proper feedback of its own instrument to the performer.

3. DATA ACQUISITION

This chapter describes the techniques developed to acquire sound and gestures from real performances. First of all a set of musical scores is designed with the intention to cover a representative recorder playing context. Then the second section presents the prototype of recorder used to measure the blowing pressure. The third section is dedicated to the explanation of the recording settings and the last part describes the preprocessing process required to build a database gathering sound and gesture. These 2 last sections were the work of Francisco Garcia Diaz-Castroverde but are presented in this thesis in order to provide an overview of the whole project and a better understanding of database construction process. In the frame work of his thesis Francisco Garcia Diaz-Castroverde also developed a preliminary prototype of recorder which will not be described here.

3.1. SCRIPTS

With the purpose of acquiring sound and blowing pressure related to the normal playing situations we designed a script containing principally 4 exercises. Each exercise was meant to study at least one behavior of the playing style so that the set of exercises covers a representative playing context. We defined 4 main exercises:

- The main goal of the first exercise was to study the shape of the blowing pressure according to the variation of the principal playing parameters. For each fingering (a total of 16) the flautist performs a phrase (*Figure 17: Score for the E6* of eight quarter notes, eight eighth notes and sixteen sixteenth notes using different dynamics, articulations and tempo. Finally for each fingering 24 different phrases are obtained *Figure 18: Structure of the scores played for each noteFigure 18*).

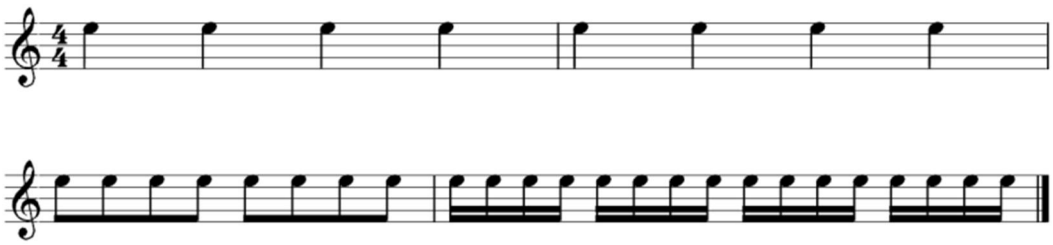

Figure 17: Score for the E6 repetition exercise

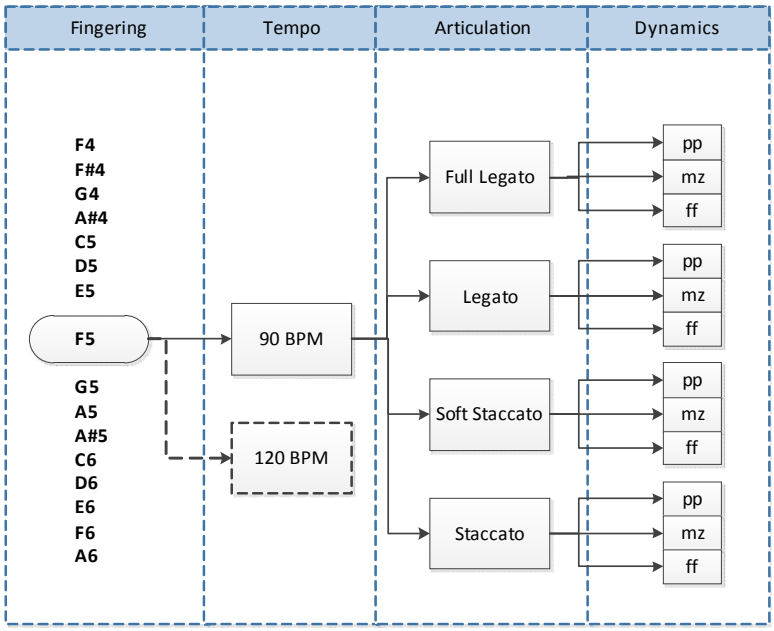

Figure 18: Structure of the scores played for each note

- A second exercise called "scales" aims at covering the tessitura of the instrument with different intervals between the notes by means of musical phrases (*Figure 19*). The data recorded during this experiment allows to study the implication of the pitch interval of consecutive notes on the actual blowing pressure. This musical scale is also played using different dynamics, articulations and tempo. Finally for each fingering, the scale is played in 24 different ways.

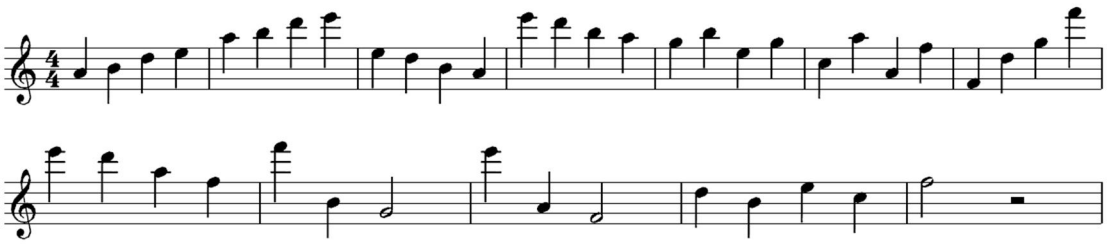

Figure 19: Score for the scale exercise

- A third exercise is to perform a crescendo for each fingering. Increasing the blowing pressure by covering the largest range of pressure allows to observe the changes of timbre and pitch accordingly to the blowing pressure applied by the performer. This exercise is mostly important for the understanding and the modeling of the timbre.

- A last exercise is to play a set of 5 musical pieces (ANNEXES: Musical piecesC) in order to measure sound and blowing pressure in real situation. In a first time we define the articulation and the dynamic .In a second time the performing style was chosen by the player and explained after playing. For this exercise each piece is performed with 10 different styles. 0ne interpretation is free to the player.

3.2. PROTOTYPE OF RECORDER

With the aim of measuring coherent parameters we had to deal with the following compromise: Should we give priority to the accuracy of the measurements or weigh the low-intrusiveness of the measuring method. The precision of the acquisition can be improved by measuring as close as possible from the source. In our case it means setting the sensor as closest as possible to the mouth, avoiding every kind of tubes in between. Decreasing intrusiveness can be achieved by placing the sensors without modifying the instrument neither the way to play it. On musical instruments the accuracy of the measurements is strongly correlated to the intrusiveness of the sensors. Indeed too intrusive measures modify the way of performing and can even affect the acoustical behaviors of the instrument. Therefore, intrusive methods are not able to provide realistic measures of instrumental gestures.

The recorder used for the recordings was a baroque treble recorder in F. The instrument was built by Josep Tubau who took as example the model BRESSAN from the flute maker Pierre Jaillard. The recorder has been designed to acquire the blowing pressure and the internal pressure using two ASCX pressure sensors. The measure of the blowing pressure is achieved by directly acquiring the pressure in the mouth of the player. For reasons of intrusiveness the air flow is carried to the sensor through a small canal in the bloc of the recorder. A second canal allows to measure the internal pressure by carrying the oscillations of the air column from the end of the bloc to the sensor (*Figure 20*). This method of measure does not modify a lot the instrument and does not bother the performer.

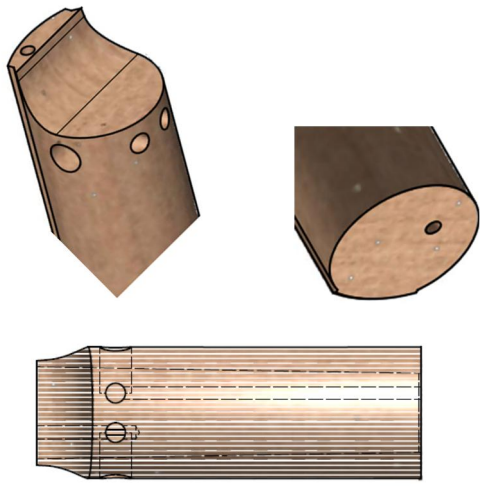

Figure 20: CAD of block of the recorder

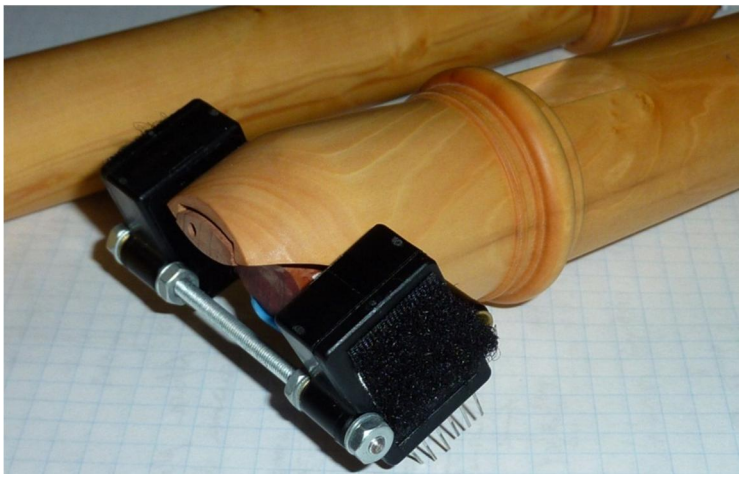

Figure 21: Prototype of recorder with pressure sensors mounted

The prototype of recorder (*Figure 21*) was played by Mr. Joan Izquierdo, a professional recorder player and recorder teacher at the Conservatory of Barcelona.

3.3. RECORDINGS SETTINGS

The recording sessions are carried out using two analog to digital interfaces: one interface is used for sound conversion and the second one for pressure conversion. Two acquisition cards are required since the measure of the pressure provide a signal containing a continuous component which cannot be acquired using the audio mixing desk.

A DAQ NI USB-6009 from National Instrument is used to acquire the analog signals coming from the two pressure sensors mounted on the recorder. The sound is recorded thanks a clip microphone which is mounted on the recorder and pointing at the wind-way of the instrument. A classic *Schoeps* microphone (MK4 capsule + CMC6 amplifier) is also used in order to get a less precise but more musical sound. Both of them are connected to a *Yamaha* mixing console.

The acquisition process (*Figure 22*) using several kinds of interfaces has one important drawback. Since the interfaces are using their own clocks which are not synchronizable, a significant lag appears between the recordings of sound and pressure. This problem can be avoided by using two high quality acquisition devices with the same clock for both. In our case we minimized this issue by recording with each interface a synchronizing signal from a metronome. Assuming that this signal should appear at the same time in both recordings, by using a simple preprocessing algorithm we are able to resynchronize all the recorded tracks.

Acquisition cards are driven by a LabView script. The purpose of the script is to define the sampling rates, the resolution, to start and stop the interfaces at the mean time and to store the recorded sound and pressure in wav or text files.

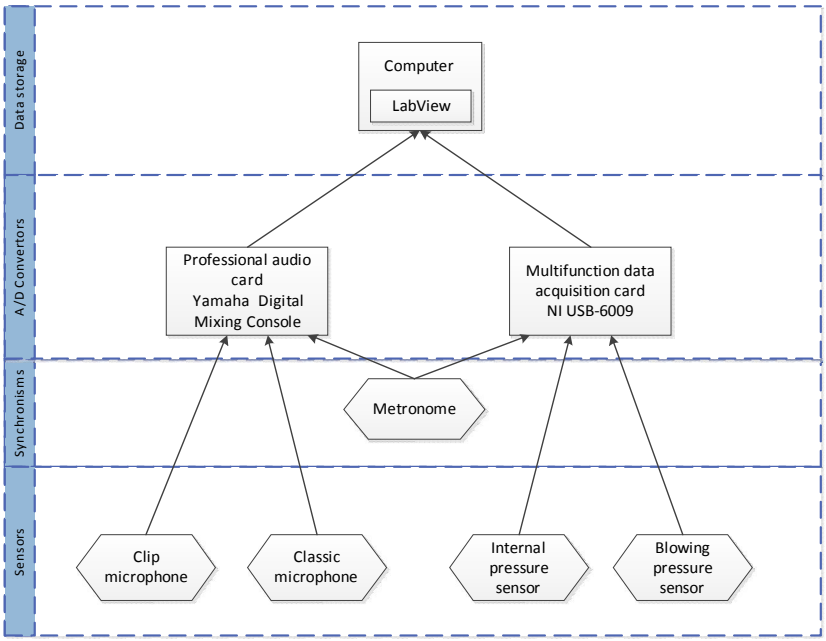

Figure 22: Recordings settings

Figure 23: Juan Izquierdo during the recordings session

3.4. PREPROCESSING

Using the previously described setting to record sound and pressure from a performance provides us with data acquired through different sampling rate. Since the digital mixing console acquire the sound at 44.1 KHz and the multifunction card acquire at 24 Khz, one them has to be re-sampled in order to work later with the same number of points on every recordings. In our case we chose to interpolate the signals obtained through the multifunction card from 24,000 points per second to 44,100 points per second or 44.1 KHz.

At this step we are still dealing with non-synchronous sound and pressure. So the data coming from different interfaces is then re-synchronized by using a MatLab script which looking at the recordings of the metronome (impulse recorded by each interface) defines the time lag between each impulse and consequently adjusts the pressures signals.

The last step of the preprocessing is to segment the recorded tracks regarding the predefine structure of the recordings. After the segmentation process we basically get one folder for each fingering where each sound and pressure track is annotated with the corresponding tempo, articulation and dynamic.

4. SOUND ANALYSIS

In the chapter 4 we will firstly present the spectral analysis algorithm used to extract spectral parameters. In a second time we will describe the spectral evolution of the recorder timbre according to the blowing pressure and we will see that in some cases the spectral evolution is not the one expected. Then we will give an explanation on different methods for timbre representation and spectral reconstruction. The last section will present the algorithm for re-synthesis.

4.1. SPECTRAL ANALYSIS ADAPTED TO RECORDER INSTRUMENTAL CHARACTERISTICS

This section describes the method used and their adaptations to perform an appropriate analysis of the recorder timbre. An explanation of the spectral evolution of the recorder timbre is also provided.

4.1.1. SPECTRAL ANALYSIS ORIENTED TOWARD SPECTRAL RECONSTRUCTION

The method used to carry out the analysis of the timbre of the recorder is directly related to the will to use statistical models to establish the relationship between the instrumental gesture and the sound of the recorder.

Timbre estimation by mean of statistical models can provide acceptable results only using few parameters to predict. Consequently we can make the supposition that lower the number of predicted variables is, higher is the prediction accuracy. The timbre description taking place in the spectral analysis must also take into account that the number of descriptors is directly responsible of the spectral reconstruction quality (section 4.2). We reach here a fundamental compromise between the losses allowed during the spectral reconstruction and the accuracy of timber prediction.

4.1.2. ALGORITHM FOR SPECTRAL FEATURES ACQUISITION

The extraction of features from the sound of the recorder is based on a MatLab SMS (Spectral Modeling Synthesis) algorithm [24] modified to provide an analysis which fits the best as possible the specificities of the recorder timbre.

Algorithm description:

- Harmonics and residual extraction by adapted SMS algorithm: In this analysis section, only the SMS script for spectral modeling is used (no synthesis). Modifications allow the algorithm to apply the window on both the blowing pressure and sound. SMS considers sounds as an addition of harmonic and noise content. After computing the Short Time Fourier Transform the harmonic content is identified by picking the peaks in the

frequency spectrum of the signal. By subtracting the harmonic component to original spectrum, we get the residual component, which can be modeled by white noise. The outputs of the script are thus the locations (frequency + level) of the peaks detected in the harmonic component and the locations of the detected peaks in the residual component.

- Pitch extraction: The pitch extraction is carried out by a modified two ways mismatch (TWM) algorithm. Since the pitch aim at reconstructing the harmonic series, it must be equal to frequency of the first harmonic. Another modification set the pitch of the frame to zero when a non-harmonic series is detected or when the fundamental frequency does not match the expecting pitch (according the fingering played).

- Splitting harmonics according odd and even spectral locations: As explained in the paragraph 6.1, odd and even harmonics are considered distinctly. This functions of the algorithm aim at splitting the harmonic components provided by SMS into odd and even series.

- Interpolating odd and even harmonic locations: Once the harmonics are divided into odd and even harmonic series, the envelopes of both harmonic series are computed. Each point of the series is interpolated by piecewise cubic Hermite interpolation ("pchip" MatLab Function). Several interpolation methods have been tried but cubic Hermite interpolation provided more natural envelopes.

  The *Figure 24* shows the envelope of the even harmonic series using different computation methods. Plotted in red we can observe that the "spline" function creates overshoots with amplitudes higher than the real harmonics. This issue can introduce errors during the spectral reconstruction. In the case of the envelope computed using the linear method (green), it is noticeable that the shape is not smooth and consequently not natural. On the last blue plot we computed the envelope with the "pchip" method which provides a result more acceptable than the two first methods.

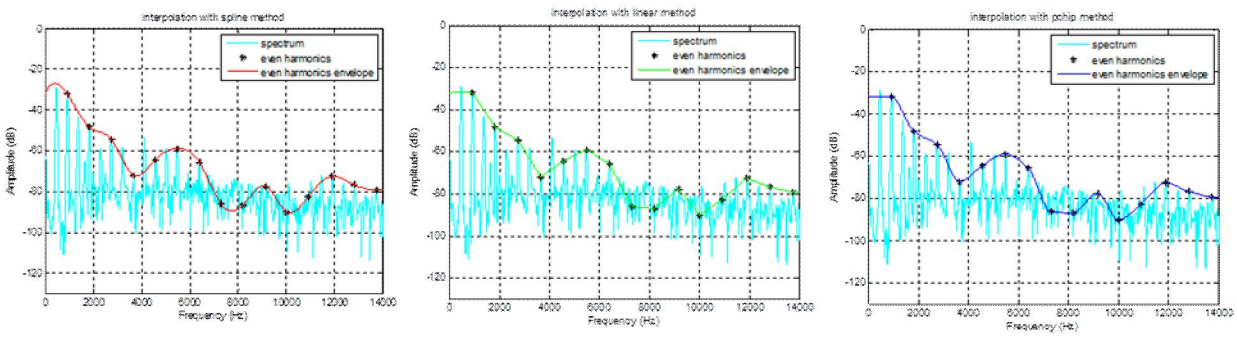

Figure 24: Comparison between three interpolation methods: spline(red), linear(green) and pchip(blue)

- Interpolating residual peaks plus low pass filtering: The residual could have been modeled just by low pass filtering the remaining noise but this technique make a non-realistic approximation of the envelope by flattening the formant frequencies. Instead of averaging the residual spectrum, we pick up the highest peaks which represent in a

better way the maximum amplitude of formant frequencies. Then the resulting locations are interpolated and low pass filtered in order to simplify the general shape.

The *Figure 25* shows the difference between the low pass filtered residual (green) and the peaks interpolation which is closer to the residual envelope. In the following chapters of the thesis we will work only with the peaks interpolation method.

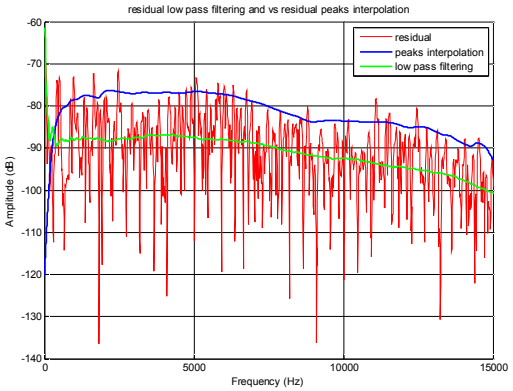

Figure 25: Residual envelope obtained by LP filtering (green) and peaks interpolation (blue)

- Computing the odd, even and residual representative coefficient: The last step of the algorithm is to condition the three envelopes in form of representative coefficients. Since one envelope is composed by a lot of points, it is not possible to use this representation to train the model. The purpose of conditioning the envelopes into another form is to decrease the number of parameters (points) that will train the statistical models. In our experiment two different conditioning methods are tested: MFCC (Mel-frequency cepstral coefficients) and linear cepstral coefficients. The comparison and the choice of the methods are discussed in the section 4.2.

A flowchart of the whole algorithm previously described is presented below in the (*Figure 26*).

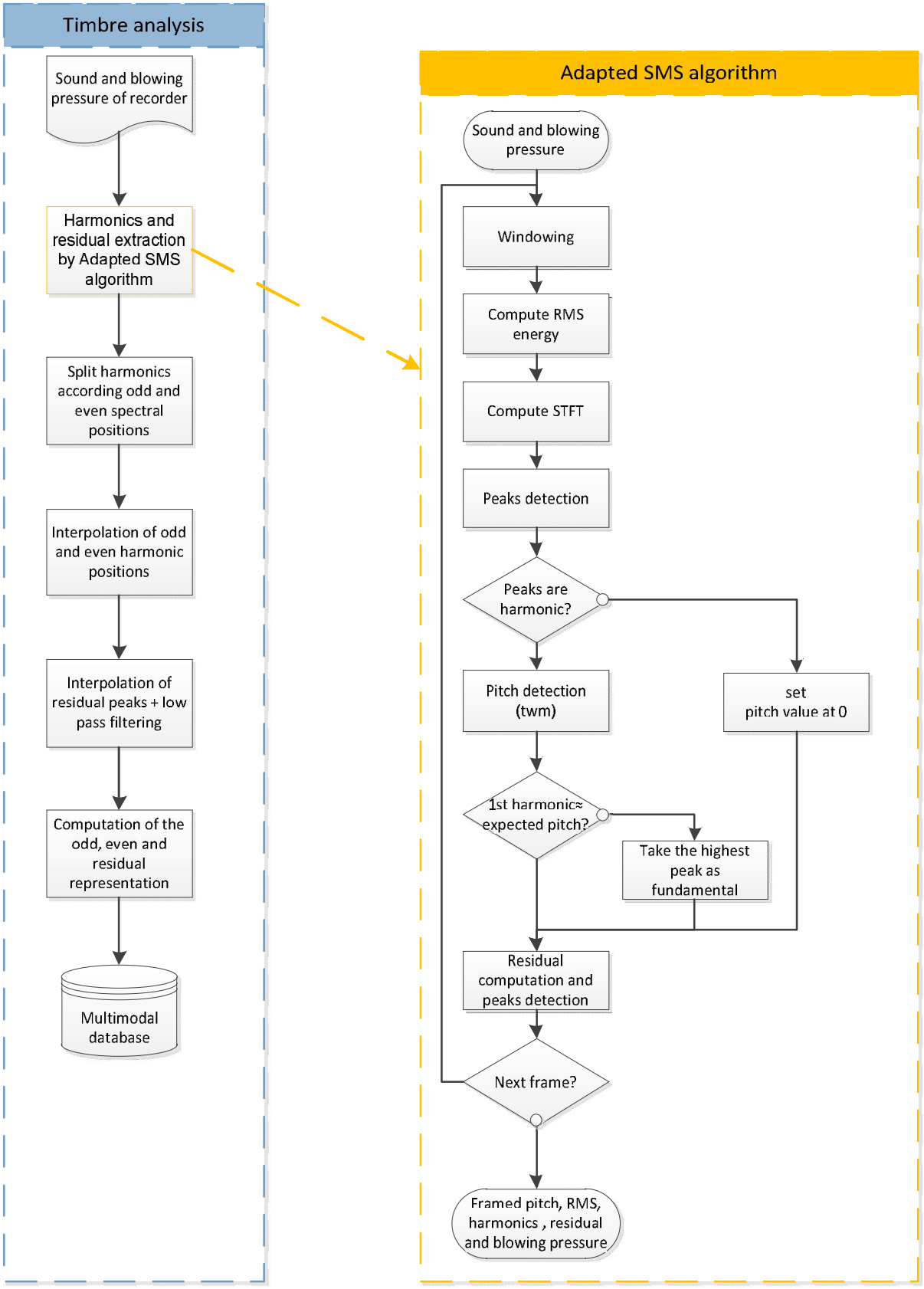

Figure 26: Algorithm for spectral features acquisition

4.1.3. SPECTRAL EVOLUTION

In the section 3.1 we describe the recording protocol where the crescendo exercise is planned. During this exercise the performer increases slowly the blowing pressure until the instrument reach its last mode of oscillation. This work allows us to study and to analyze the evolution of the spectrum regarding the blowing pressure applied. Basically it can be observed that for certain fingering, the recorder can have 1, 2 or 3 modes of oscillation. We call "mode transition" the transient, where the higher harmonic of the spectrum starts decreasing, letting rising above the second harmonic. Later on, when the second harmonic gets higher than the first harmonic the instrument will then enter in its second mode of oscillation. It is also noticeable that in the first and second mode, the pitch increase linearly with the blowing pressure, but since the second harmonic is equal to twice the frequency of the fundamental, when the instrument reaches the transient period, the pitch jump from an octave to the next one (at t=3s in the *Figure 27*).

The *Figure 27* represents in the first and second graph the audio signal and blowing pressure of an E5 fingering during the crescendo exercise. The third and fourth graph shows the pitch, and the spectrogram. We can observe that the evolution of the pitch is not smooth and jump from an octave to the next one. This perception of the pitch can be supported by the spectrogram which shows that increasing the blowing pressure the first harmonic disappears when the second harmonic is growing up.

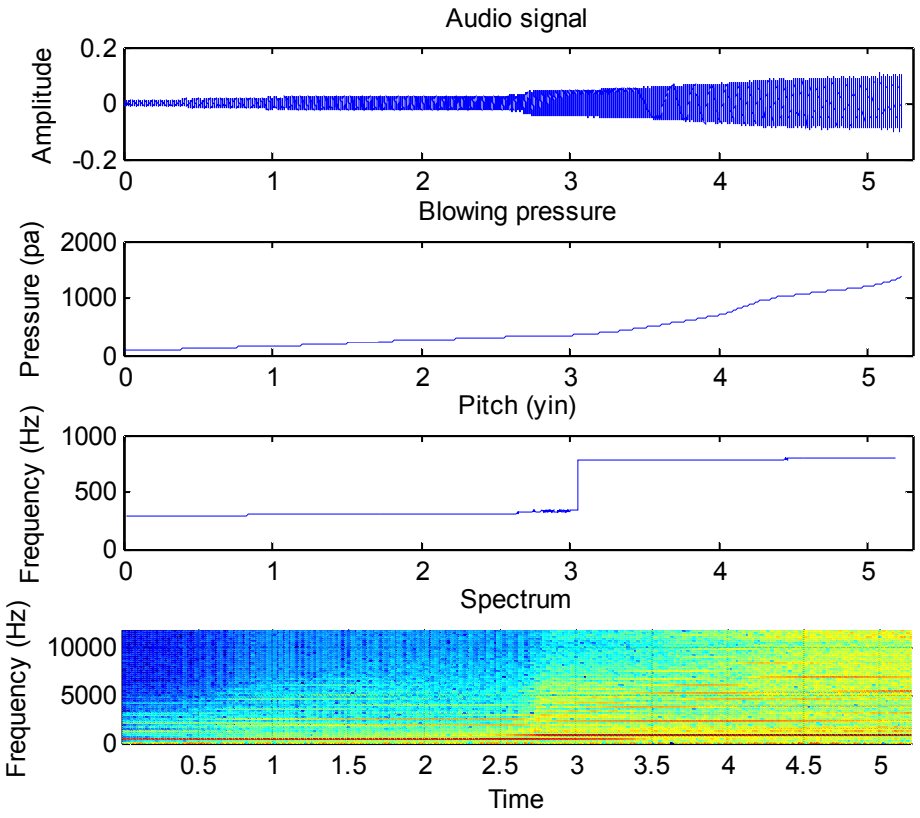

Figure 27: Spectral evolution of an E5 crescendo with its pitch and plowing pressure

Observing the spectral evolution of an E5 crescendo it is possible to highlight three phases in the progression:

-phase 1: as shown in the *Figure 28*, the first phase represent when the first harmonic contains the most of the spectral energy. Therefore it is the fundamental frequency of the sound. In this first case the envelope of odd harmonics (blue envelope) is above the even harmonics' envelope.

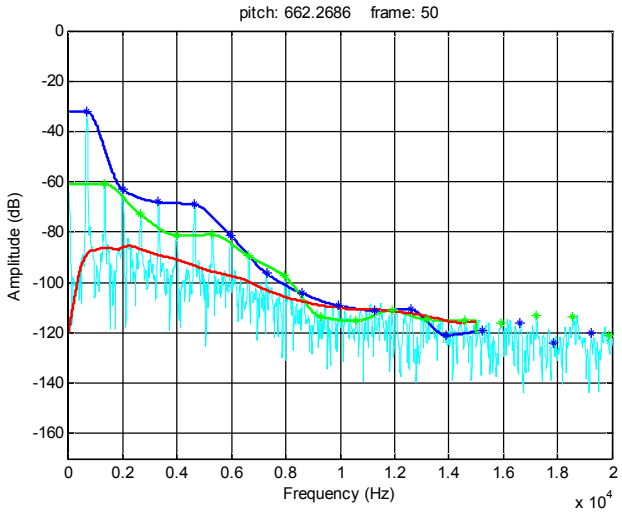

Figure 28: First phase, the odd harmonics (blue envelope) contain most of the energy, even harmonics (green envelope) are lower and the residual (red envelope) has a low level. The full spectrum is represented in cyan.

-Phase 2: keeping the blowing pressure increasing the second harmonic and more generally the even harmonics get more powerful. The even harmonic series crows until the second harmonic become equally powerful than the first harmonic (*Figure 29*). In this situation there is no fundamental anymore and the perceived sound appears like containing two notes. This phenomenon is called multiphonic. This phase of transition happens generally very quickly and does not let time to be heard but for some fingerings this transition occurs on a larger range of pressure.

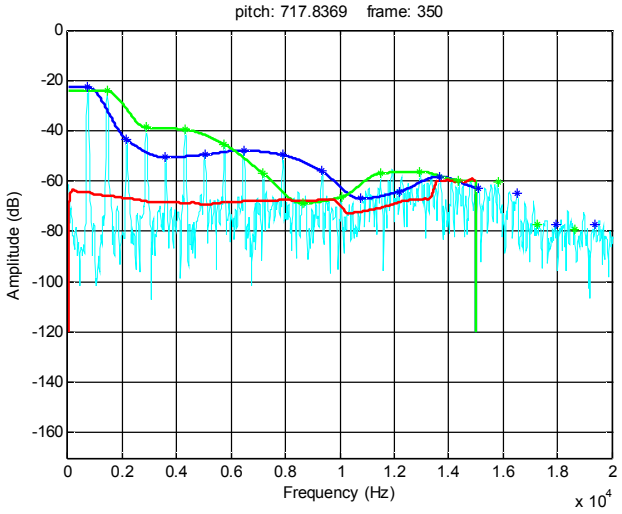

Figure 29: Second phase, the odd harmonics (blue envelope) contain the same amount of energy than the even harmonics (green envelope) and the level of the residual (red envelope) is increasing. The full spectrum is represented in cyan.

- Phase 3: The transition finally ends up when the second harmonic becomes the fundamental. The second harmonic turns out to be the most powerful harmonic and mostly all even harmonics (green envelope) get higher than the odd harmonic series (*Figure 30*). This change of fundamental corresponds to the change of mode of oscillation of the air column. Since the air column oscillates twice faster, the generated sound presents a pitch twice higher.

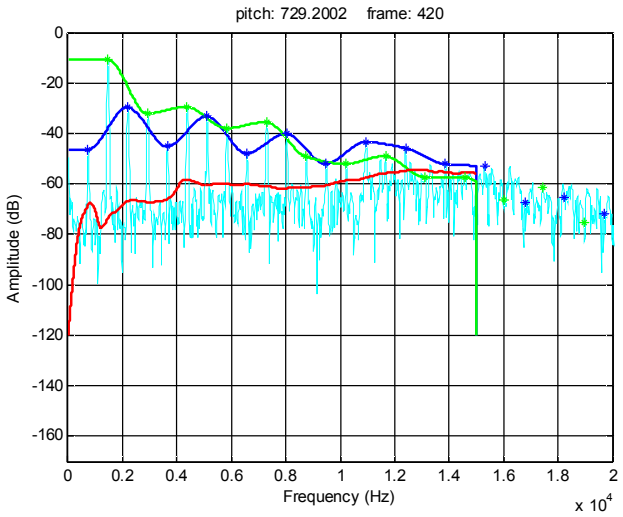

Figure 30: Third phase, the odd harmonics (blue envelope) are lower than the even harmonics (green envelope) and the level of the residual (red envelope) is still increasing. The full spectrum is represented in cyan.

4.1.4. NON-HARMONIC CASE

After having achieved the spectral analysis of several crescendo exercises, we could not extend the previous harmonic evolution rule to all fingerings. In fact for some fingerings it appears that the second mode of oscillation comes up with a new harmonic series that the algorithm for spectral feature acquisition cannot extract.

Observing the spectral evolution of an E6 crescendo highlight the fact that the progression is not harmonic:

-Phase 1: First mode of oscillation, odd harmonics are higher and even harmonics are present (*Figure 31*).

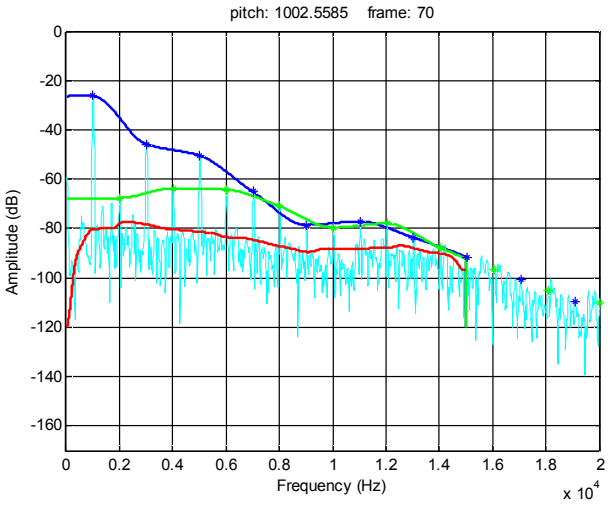

Figure 31: First phase, the odd harmonics (blue envelope) contain most of the energy, even harmonics (green envelope) are lower and the residual (red envelope) has a low level. The full spectrum is represented in cyan.

-Phase 2: The blowing pressure is increasing and a new harmonic series is growing. The odd and even harmonics present in the first mode are still existing (*Figure 32*). In this phase we cannot clearly define the residual since the third harmonic series is considered has noise by the algorithm, therefore, the residual is growing with the growth of the new harmonic series.

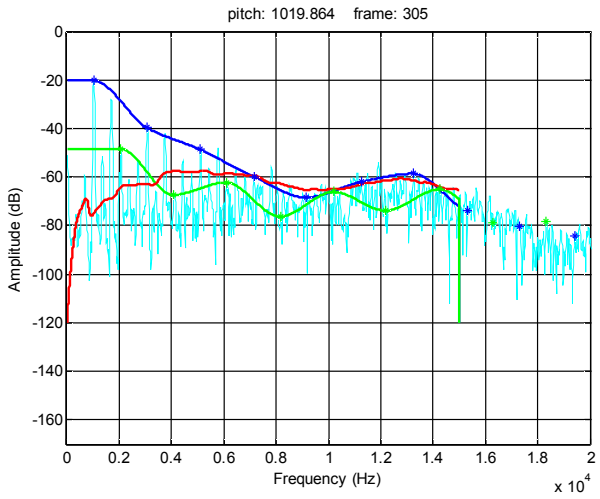

Figure 32: Second phase, the odd harmonics (blue envelope) are higher, even harmonics (green envelope) are lower and a new harmonic series is appearing.

-Phase 3: The whole energy is concentrated on the new harmonic series. The first harmonic series is not present in this second mode of oscillation; consequently the algorithm cannot represent the odd and even envelopes (*Figure 33*).

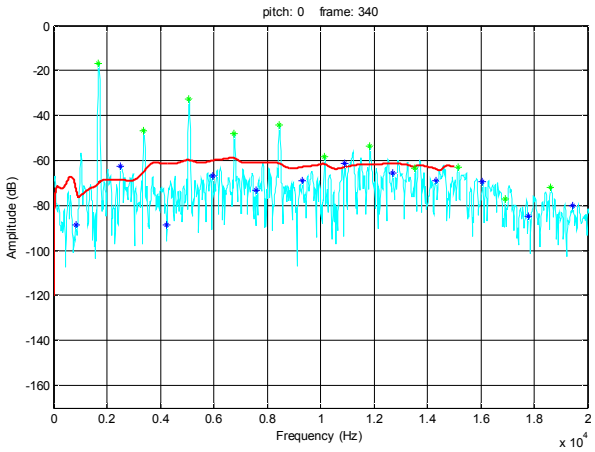

Figure 33: Third phase, both odd and even harmonics (blue and green envelopes) are absent, only the new harmonic series remains.

4.2. TIMBRE REPRESENTATION AND SPECTRAL RECONSTRUCTION

In order to improve the ANNs training process, the number of spectral parameters must be minimized. This data compression can be achieved by representing the spectral envelope with a small set of coefficients. This section presents two different methods for timbre representation and spectral reconstruction. An experiment gathering cepstrum and MFC representation is carried out to determine which technique is most suitable for recorder timbre representation and reconstruction.

4.2.1. REPRESENTATION BY LINEAR BANDS

One way to represent an envelope with few parameters is to use a bank of triangular windows equally spaced in the frequency domain with an overlapping factor of fifty per cent. This bank of windows constitutes a filter (*Figure 34*). Once the filter is built, it is applied by multiplication to the spectral envelope and provides for each window a coefficient. Depending on the number of windows used to construct the filter it will provide a different number of coefficients affecting the accuracy of the spectral reconstruction. With this type of linear bands we assume that humans are equally sensitive to each frequency.

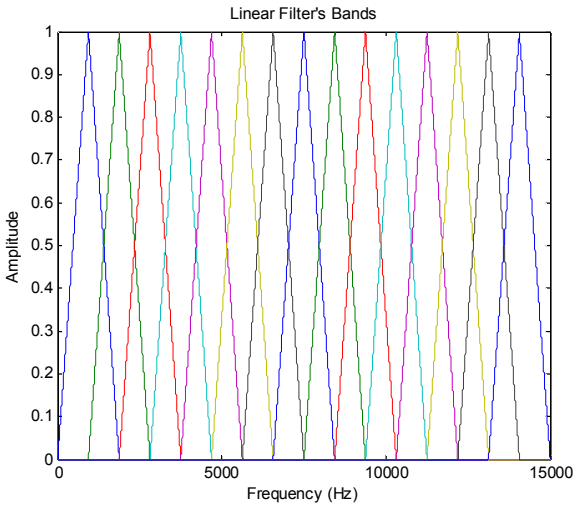

Figure 34: Linear filter's bands for cepstral representation

The computation of linear representation is implemented as follows:

1. Generating the filter by constructing a matrix of triangular windows equally spaced in the frequency scale.
2. Apply the filter bands to the spectral envelope.
3. Compute de DCT (Direct Cosine Transform) of each coefficient.

4.2.2. REPRESENTATION BY MFCCS

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively construct a mel-frequency cepstrum (MFC) and are obtained from cepstral representation of sound. The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are not equally spaced in the frequency domain (*Figure 36*). MFC representation use the mel scale [2], which approximates the human auditory system response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping allows to gain precision in audio signal compression.

$$m = 2595 \log_{10}\left(\frac{f}{700} + 1\right) = 1127 \log_{e}\left(\frac{f}{700} + 1\right)$$

Figure 35: Relation to convert Hertz to Mel scale

[2] Mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another.

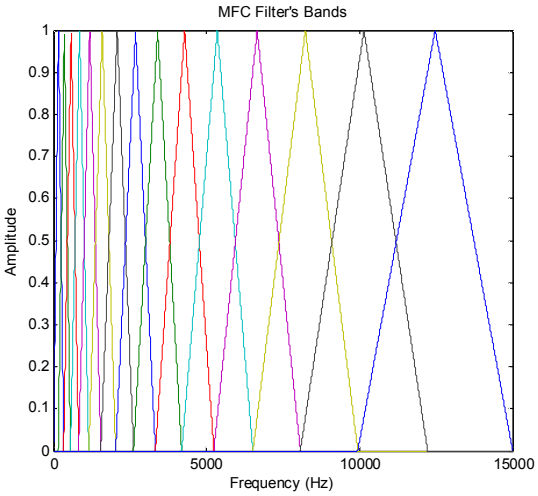

Figure 36: MFC filter's bands for MFCCs representation

The computation of MFCCs is implemented as follows:

4. Generating the filter by constructing a matrix of triangular windows equally spaced in the mel scale.
5. Applying the mel filter to the envelope.
6. Computing the DCT (Direct Cosine Transform) of each coefficient.

4.2.3. SPECTRAL RECONSTRUCTION

Once the odd, even and residual envelopes are compressed into coefficients, it is then possible to use those coefficients to train the ANN in order to build the timber models. Training the ANN with the MFCCs will only let the model providing an estimation of the coefficients and consequently it will be necessary to reconstruct the envelopes from the estimated MFCCs. In order to get back a vector of numbers describing the envelopes the method used before is reversed on the generated MFCCs:

1. Generating the filter with triangular bands equally spaced in the frequency domain or in the mel scale.
2. Computing de IDCT (Inverse Direct Cosine Transform) of each coefficient.
3. Multiplying the Coefficients by the linear filter or the mel filter.
4. Smoothing the obtained vectors

4.2.4. RESULTS OF SPECTRAL RECONSTRUCTION

In order to compare the result obtained with each method we apply both cepstrum and MFC filters to an envelope. For each trial the number of windows used in the filters is changes. The *Figure 37* shows the spectral reconstruction of an original envelope (blue) using the MFC

representation (green) and the cepstrum representation (blue). The experiment is reproduced three times with 5, 15 and 30 windows. It is obvious that cepstrum representation has a lack of precision in the low frequency content (from 0 to 2500Hz), where most of the fundamental frequencies of the recorder are located. It is also clear that increasing the number of windows both representations gain precision and get closer to the original envelope. For the three cases the MFC representation gets closer to the original envelope.

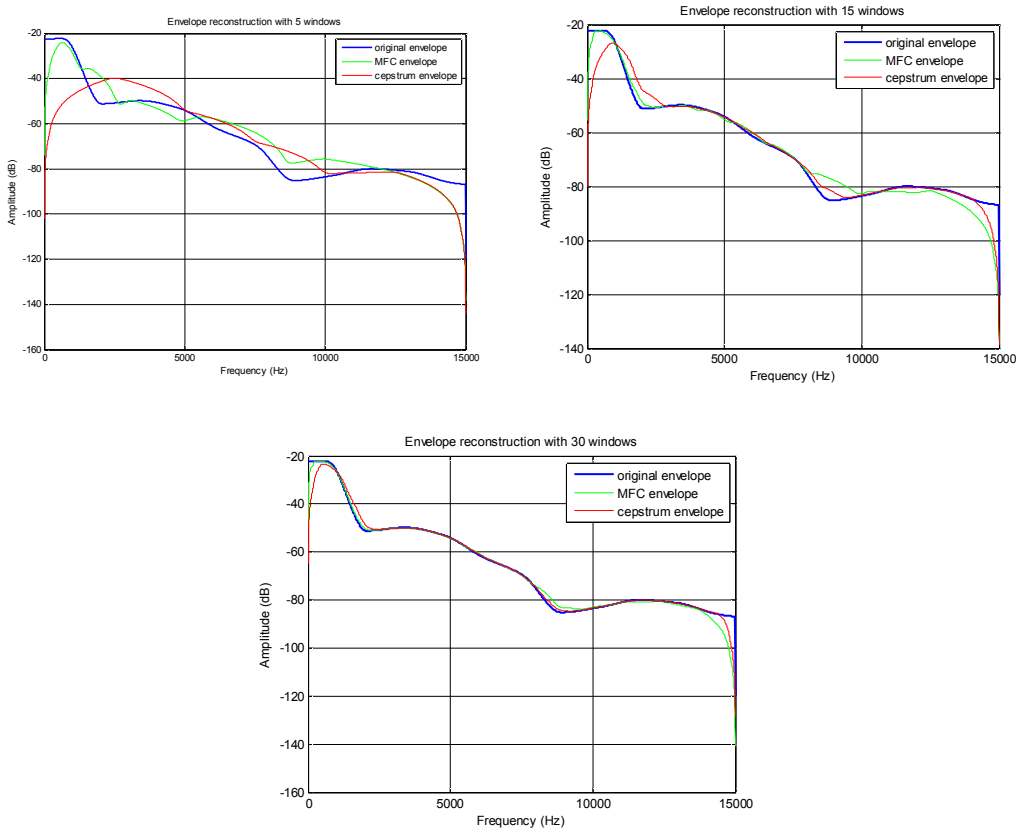

Figure 37: Comparison between the original envelope, the cepstral representation and the MFC representation for different number of window.

The

Figure 38 highlights the fact that in every case the MFC representation has a correlation coefficient higher than the cepstrum representation. This is due to the reparation of the audio content which is mainly located between 300 and 5000Hz, frequencies that are accurately represented by MFCC. The results of the MSE (Mean Square Error) do not match exactly the conclusion since the MFC representation MSE is higher than the cepstrum MSE. The cause is probably that MFCCs are more imprecise at the cutoff frequency than the linear cepstrum representation. This lack of accuracy at the end of the spectrum is not crucial since there is only few harmonic content in the high frequency. Therefore, looking at those results we decided to use 15 MFCCs in order to compress one envelope, thus the representation of the odd, even and residual envelopes is composed by a total of 45 variables for each frame.

	5 windows		15 windows		30 windows	
	MSE	Corr coef	MSE	Corr coef	MSE	Corr coef
MFC representation	5.0512	0.9344	2.5213	0.9717	1.2449	0.9847
Cepstrum representation	7.1166	0.8070	2.3766	0.9587	1.0780	0.9846

Figure 38: Comparison of different methods of representation for the envelope, with different resolution

4.3. RE-SYNTHESIS

The re-synthesis process (*Figure 39*) aim at recovering sound from the previously extracted parameters: the pitch[3] and the spectral envelopes. The pitch gives the position of the fundamental, which permits the reconstruction of the harmonic series, corresponding to the multiples of the fundamental. After having reconstructed the harmonics envelopes, we use them to modulate the amplitude of the harmonic series previously generated. The residual envelope is then filled by white noise in order to approximate the residual of the original sound. As the result of that approximation does not match exactly the amplitude of the original residual, it is necessary to adjust its gain according to the harmonic part. The last step of this additive synthesis [24] is to apply the IFFT (Inverse discrete Fourier transform) over the reconstructed spectrums (harmonic and residual) and to add the two resulting time signals together.

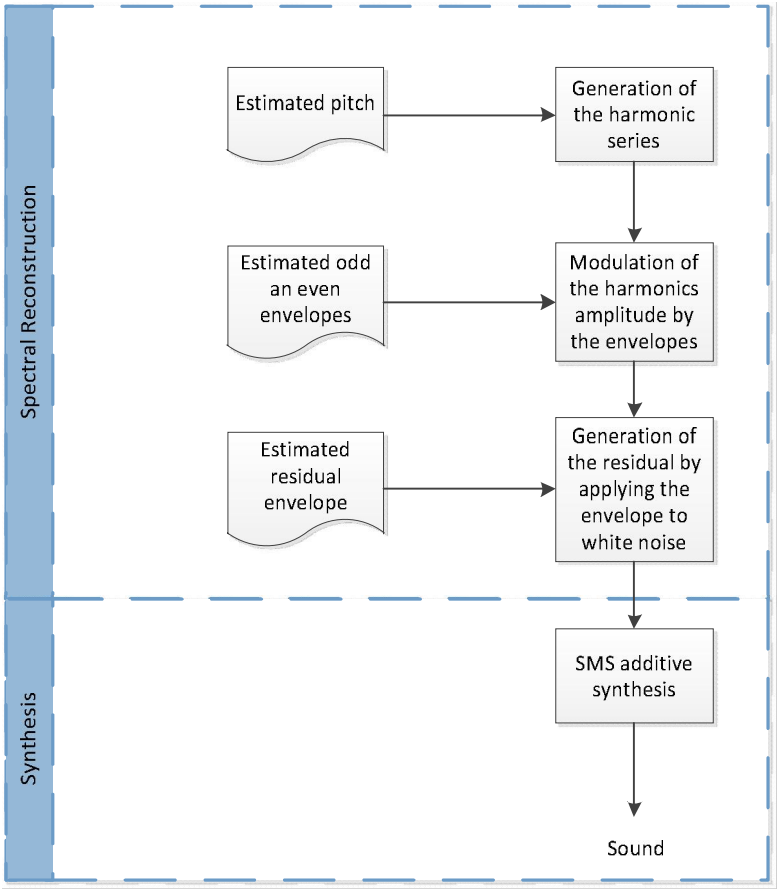

Figure 39: Re-synthesis process

[3] Here we call pitch the frequency of the first harmonic which is not necessarily the higher harmonic.

5. MODELING OF GESTURE-SOUND RELATIONSHIP

The present chapter starts by giving an overview on the method used for sound synthesis and indirect acquisition. The second section is a brief introduction on artificial neural network technology and is followed by a presentation of different model architectures tried throughout the prediction experiments. Later on we will have a discussion on provided results of predicted timber and then the chapter will end up with a presentation of the script used for sound synthesis.

5.1. OVERVIEW OF THE METHODS USED FOR INDIRECT ACQUISITION AND SOUND SYNTHESIS.

This section is an introduction to the methods used to carry out the sound synthesis and the indirect acquisition. This first description of the systems will give a general view of the process and will allow a better understanding of the work explained in this chapter.

5.1.1. SOUND SYNTHESIS METHOD

The main purpose of this thesis is to design a recorder sound synthesis method by using the existing correlation shown by Flecher[3] between timber and blowing pressure.

The flowchart [*Figure 40*] shows the principals processes involve in the synthesis methodology. The previously recorded sound and the corresponding blowing pressure are firstly windowed in order to work in a frame by frame fashion. Then an analysis of the timbre in the frequency domain is carried out over the sound of the recorder with the aim to extract significant spectral parameters describing the spectral variations of the timbre. Those frame by frame timbre descriptors and blowing pressure are at that point used to build a multimodal database. With this database we now able to train artificial neural networks (ANNs) which are capable of creating statistical rules to link the blowing pressure to the timbre descriptors. Once the statistical model built, it can estimate the shape of the spectrum by inputting the blowing pressure. The generated spectral estimation is then reversed from frequency to time domain thanks additive synthesis.

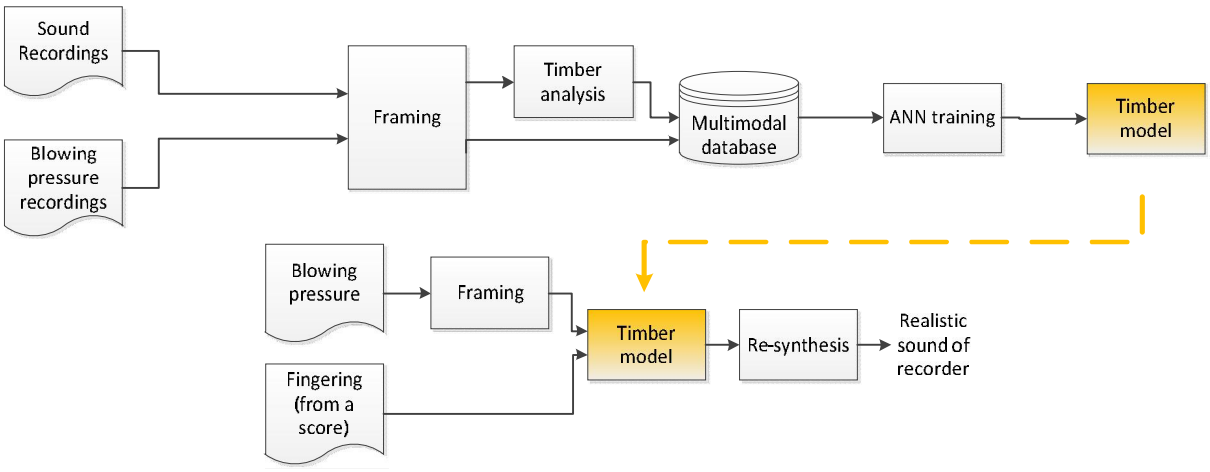

Figure 40: Synthesis method, from recordings to synthetic sound

5.1.2. INDIRECT ACQUISITION OF BLOWING PRESSURE

As it is possible to estimate the spectral parameters from the blowing pressure, by keeping the same method and by modifying the training process of the ANNs, the model can provide an estimation of the blowing pressure applied to the recorder to generate sound (*Figure 41*).

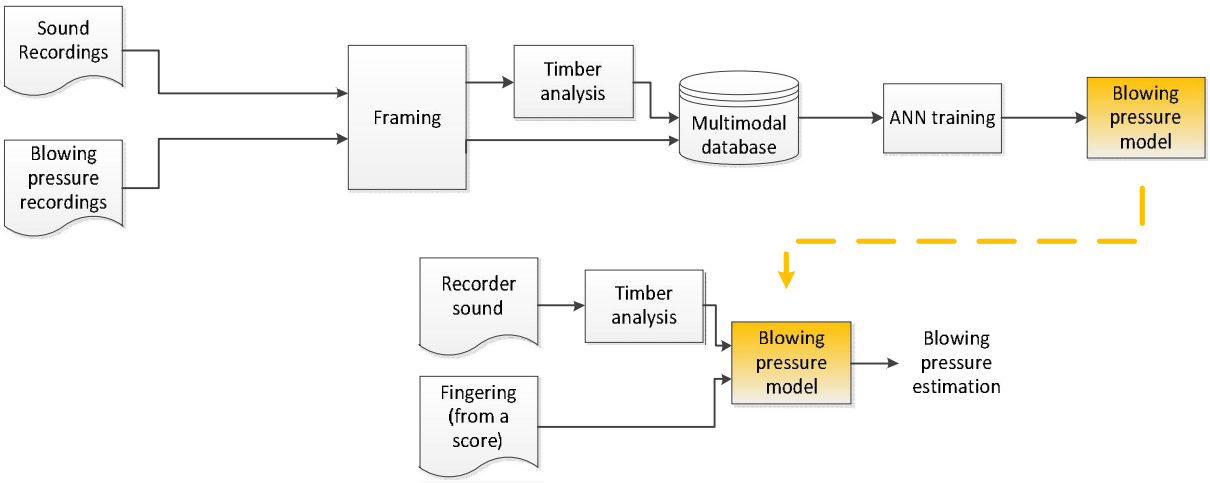

Figure 41: Method for indirect acquisition of blowing pressure

5.2. ARTIFICIAL NEURAL NETWORK TRAINING

Starting by giving a brief introduction on the basis of ANN (Artificial Neural Network), this section describes the algorithms used to train the timbre and the gesture models, explains the different possible options for selecting the input parameters and explicates the choice of the model architectures by providing results.

5.2.1. INTRODUCTION

An ANNs are non-linear statistical data modeling tools inspired by biological nervous systems. They can be used to model complex relationships between inputs and outputs or to find patterns in diverse types of data. They consist of interconnected processing elements (neurons) working together to solve specific issues. The learning process acts in fact on inter neural connections which are adjusted accordingly to the information flowing through the network.

Neural networks works following a cycle of mainly 3 states: comparison, adjustment and training. The cycle is repeated until a particular input leads to a specific target output. The *Figure 42* illustrates such a situation, where the network is adjusted, based on the comparison of the actual output and the desired target, up to the network output matches the target.

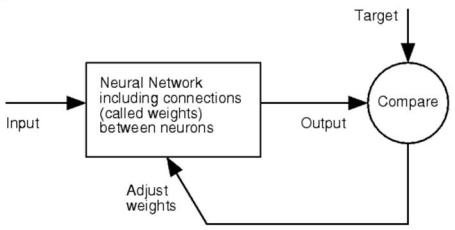

Figure 42: Working process of artificial neural network

ANNs are composed by neurons (*Figure 43*). They have a single input *p* connected with an associated weight w and a transfer function *f*. The input data is transmitted through the single connection that multiplies its strength by its corresponding weight *w*. The weighted input *wp* goes then to the transfer function *f*, which produces the output *a*. In general ANNs are also fitted with an extra connection called bias. Several transfer functions as linear, symmetric saturating linear, log-sigmoid and hard-limit are useful to process the internal data regarding the shape of final output wanted. The formula of the neuron output is: $a = f(Wp + b)$.

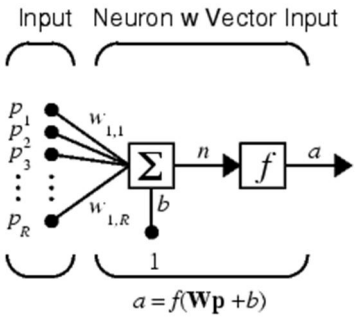

Figure 43: Representation of one neuron

A network of neurons can have several layers. Each layer has a weight matrix W, a bias vector b, and an output vector a. The network shown in *Figure 44* has P inputs, S neurons in the first layer, S2 neurons in the second layer, etc. It is common for different layers to have different numbers of neurons. A constant input 1 is fed to the bias for each neuron. The network architecture can have any shape but the most common configurations are feed-forward networks where all connections go forward without any loops or feedback. Networks are usually trained with a different set of data that is not used for the evaluation. The most typical training algorithm is back propagation. The principal steps of back propagation algorithms are:

- Forward propagation of the input through the neural network.
- Back propagation of the output in order to generate the deltas of all output.
- Multiply output delta and input to get the gradient of the weight.
- Bring the weight in the opposite direction by subtracting its ratio from the weight.

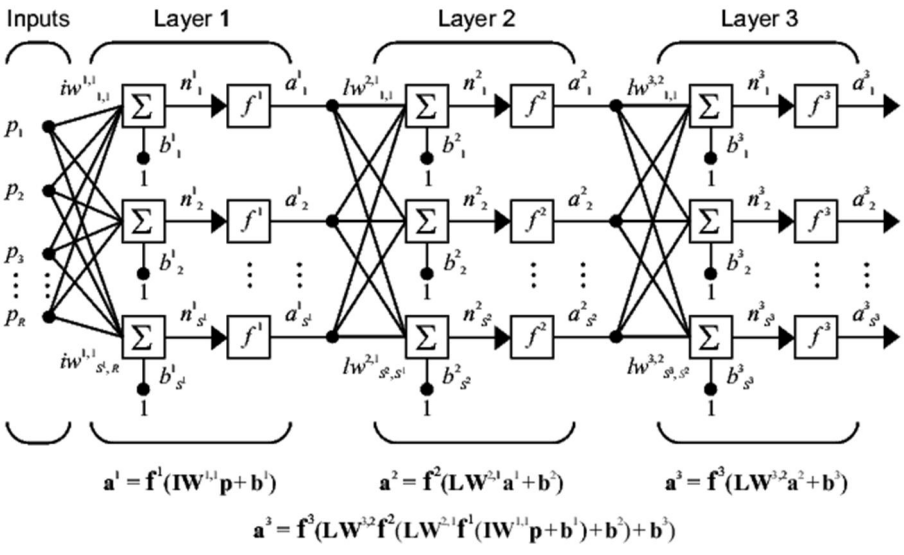

Figure 44: Representation of a 3 layers feed-forward network

Feed-forward networks are considered as static networks since the output is directly computed from the input. Another type of networks is dynamic networks which contain

feedbacks or delays and where the output depends not only on the current input to the network, but also on previous inputs or outputs of the network.

5.2.2. ANN ARCHITECTURE

This section describes the two types of ANN architectures employed to train the timbre model and the model for indirect acquisition of gesture.

5.2.2.1. Timbre to pressure and pressure to timbre estimation.

The architecture used to train two models, one able to pressure estimation (from timbre) and a second able of timbre estimation (from pressure), is a two layers cascade-forward with back propagation ANN. The MATLAB function *newcf* creates cascade-forward networks (*Figure 45*) which are similar to traditional feed-forward networks, but add a weight connection from each layer to the successive layers. For example, the two-layer network has connections not only from layer 1 to layers 2, but also from the input to layer 2. Giving as arguments *trainscg* and *learngdm* to the function, we set the back propagation training function to update weight and bias values according to the scaled conjugate gradient method, and the back propagation weight and bias learning according to the gradient descent method. The transfer functions of the different layers are hyperbolic tangent sigmoid for the first layer and purely linear for the second layer.

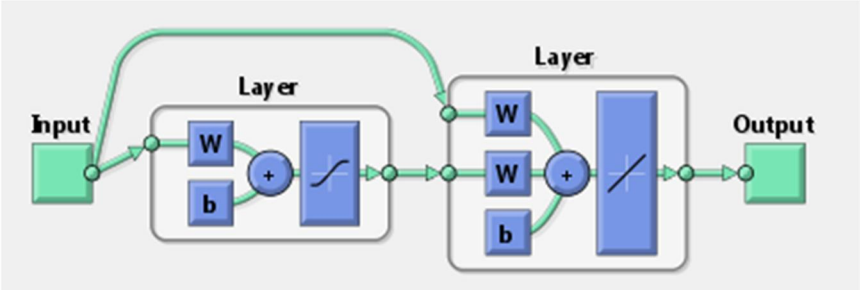

Figure 45: two layers cascade-forward with back propagation ANN

5.2.2.2. Pitch establishment estimation.

The architecture used to train a model able to pitch establishment estimation is a three-layers feed-forward with back propagation ANN. The MATLAB function *newff* creates a traditional feed-forward network (*Figure 46*). The back propagation is set with the scaled conjugate gradient method as training function and the gradient descent method as back propagation weight and bias learning process. The transfer functions of the two first layers are hyperbolic tangent sigmoid and symmetric saturating linear for the output layer. The choice of

the symmetric saturating linear transfer function has been done regarding the target output which is a Boolean signal.

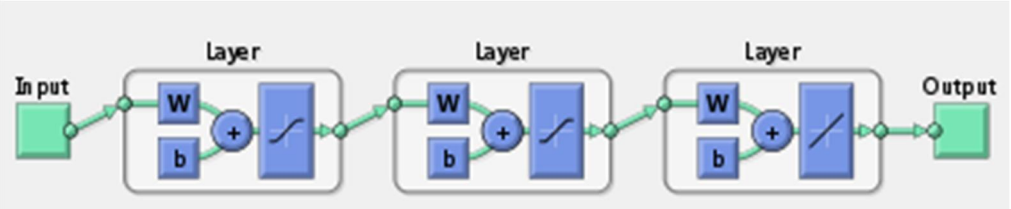

Figure 46: three-layers feed-forward with back propagation ANN

Both structures are using the mean squared normalized error function (MSNE) to estimate the prediction performance. The MSNE is a function measuring the network performance according to the mean of squared normalized errors. Normalized errors are calculated as the difference between targets and outputs after they are each normalized to [-1, 1].

5.2.3. INPUT PARAMETER SELECTION

In this section we evaluate and compare network setups with different input parameters for ANN training. All the following parameters are computed and stored in the multimodal database but only few of them are improving the ANN performance. In a first time the parameters are presented and later on, a table (*Figure 48*) provides their correlation coefficient with each other is studied.

Selection of Parameters:

- Pitch: it is indeed a fundamental parameter used to reconstruct the spectrum, therefore since we need do predict it, we need to include it in the training dataset.

- Blowing pressure: the blowing pressure is also a central parameter, for timbre prediction (as an input) and for gesture prediction (as a target), consequently it has to be included in the data set.

- RMS energy: The RMS energy of the audio signal could be good material for both models to predict the residual but since we want to use the less parameters as possible and that the residual is not the more critical part of the audio signal we decided to abandon it for the prediction.

- MFCCs: MFC coefficients are essentials for timbre prediction since they are the target, but not that significant for blowing pressure prediction. The computation of the correlation coefficient (*Figure 48*) shows that there is an interesting relation chip between the first residual MFCC and the blowing pressure but nothing significant regarding the odd and even harmonics MFCCs.

- First derivative of blowing pressure: Temporal parameters are very important, especially because a delay due to the wave speed propagation is exciting between the blowing pressure and the resulting sound. The derivative of the blowing pressure gives information on the pressure of the previous and the next frame but in our framework these temporal parameters do not improve the performance of the training.

- First derivative of pitch: In the case of the first derivative of the pitch we can observe the same inefficacity than with the first derivative of the blowing pressure. In order to include temporal information to the models it could be better to experiment a dynamic ANN.

The following table (*Figure 48*) has been computed using the correlation coefficient calculation (*Figure 47*) between each parameter. The best coefficients are highlighted in bold.

Assuming that X and Y are both compared variables and E is the expected value operator, the correlation coefficient rx,y is computed as follows:

$$r_{x,y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

Figure 47: Correlation coefficient formula

	pitch	**pressure**	**RMS**	**MFCC odd 1st**	**MFCC even1st**	**MFCC res 1st**
pitch		0.5566	0.7542	-0.9394	0.9373	0.8918
pressure	0.5566		0.5566	-0.3811	-0.3119	0.6966
RMS	0.7542	0.6038		-0.5245	-0.5360	**0.9268**
MFCC odd 1st	**-0.9394**	-0.3811	-0.5245		**0.9896**	-0.7064
MFCC even1st	0.9373	-0.3119	-0.5360	**0.9896**		-0.7038
MFCC res 1st	0.8918	**0.6966**	**0.9268**	-0.7064	-0.7038	

Figure 48: Correlation coefficients of spectral descriptors between themselves

5.2.4. MODELS TRAINING FOR SOUND SYNTHESIS

This subsection explains how the models are trained using different data set (full data set, or silence free data set) and different structure of ANN. The "full data set" is composed by all the tracks which have been recorded at 90BPM for one fingering (crescendo and repetitions), so that it remains the data from the exercises performed at 120BPM for the models validation.

During the tests a second kind of training data set has been built with the aim to improve the accuracy of the learning process. This second training data set called "silence free data set" gathers the data from the same tracks than the "full data set" but discarding the frames where no harmonics are detected.

Throughout the trials it has been noticed that normalizing the data set before training the models can decrease the prediction error. Therefore the data set is normalized using the MatLab function "mapminmax" which map the minimum and the maximum value of the set to -1 and 1 and then the data set is used to train the models. Since the models are trained with normalized data they must be fed with normalized values. After the prediction, the results must be converted back to the original scale.

With the help of the MatLab function "MSNE", the mean of squared normalized errors of each model is computed. This normalization insures that networks with multiple outputs will be trained so that the accuracy of each output is treated as equally important. Without normalization outputs with larger values would be treated as more important.

As the computing cost of the training of the models is growing up with the augmentation of input and output parameters, the models are trained for one fingering. Once a method providing a good accuracy of prediction is found, other models are trained with that method for the remaining fingerings. Each model is then evaluated on the prediction of timbre from a track of blowing pressure, the errors are compared, and the predicted pitch and envelopes are also plotted versus the original descriptors values.

5.2.4.1. Timbre Model 1

This first model is trained using as input the blowing pressure and envelopes and pitch as targets (*Figure 49*). In this case the training is done on a two layers cascade-forward with back propagation ANN, using the full data set and returns a MSNE of 0.0455.

Once the training is done, the model is fed with a recording of blowing pressure and the result of the prediction is then synthesized by mean of additive synthesis (*Figure 50*).

Figure 49: Training for model 1 able of envelopes and pitch prediction

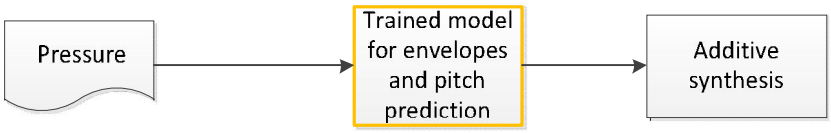

Figure 50: Sound synthesis from the predicted spectral descriptors

The *Figure 51* shows a plot comparing the real pitch (blue) with the synthesized pitch (red). We can notice here the large error between both pitches which cause during the synthesis a frequency error about 20Hz. The *Figure 52* shows a plot of the predicted envelopes compared to the real ones. It is clearly seeable that we get here a non-negligible error as well. The addition of the pitch error, the envelope error affects also the synthesis and produces a sound with the wrong pitch and non-realistic timbre.

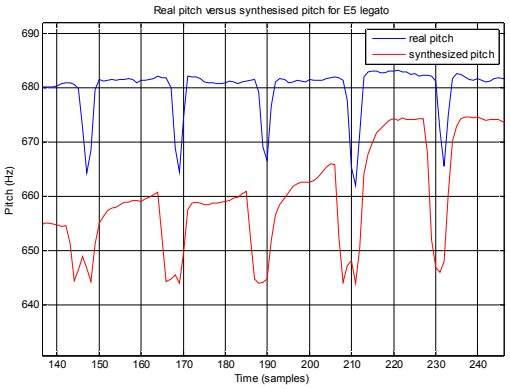

Figure 51: Real pitch versus synthesized pitch by model 1

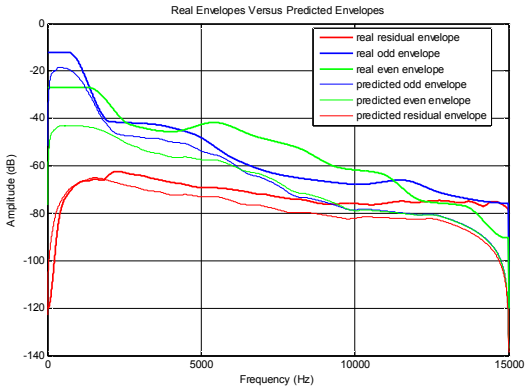

Figure 52: Real envelopes versus synthesized envelopes by model 1

5.2.4.2. Timbre Model 2

The model 2 is built based on another philosophy; assuming that it could be easier to predict fewer parameters, each parameter is predicted by one independent model. Consequently there are 4 artificial neural networks of two layers cascade-forward with back propagation which are trained inputting the blowing pressure from the full data set and use as target one of the envelopes or the pitch (*Figure 53*). After the training, the provided MSNEs are respectively 0.0387 for the model predicting the odd envelope, 0.0331for the model predicting the even envelope, 0.0527 for the model predicting the residual envelope and 0.2787 for the model predicting the pitch.

In order to predict the timbre each trained model receives as input the blowing pressure and restitutes one spectral parameter. The predicted envelopes and pitch are then gathered and transmitted to the additive synthesis (*Figure 54*).

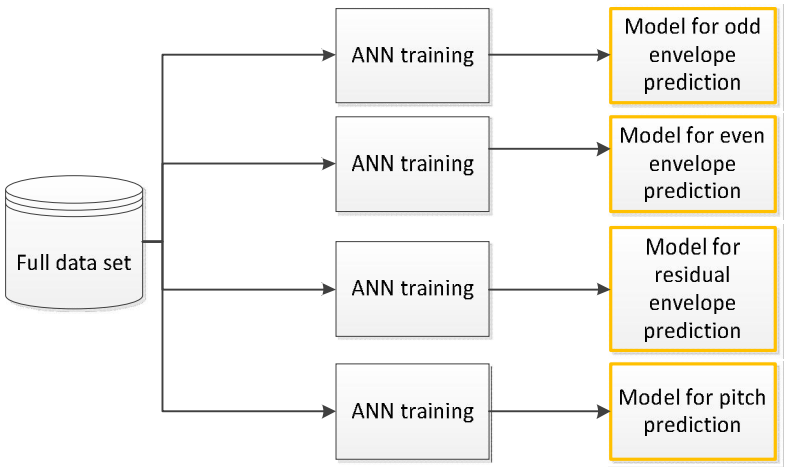

Figure 53: Training for model 2 able of envelopes and pitch prediction

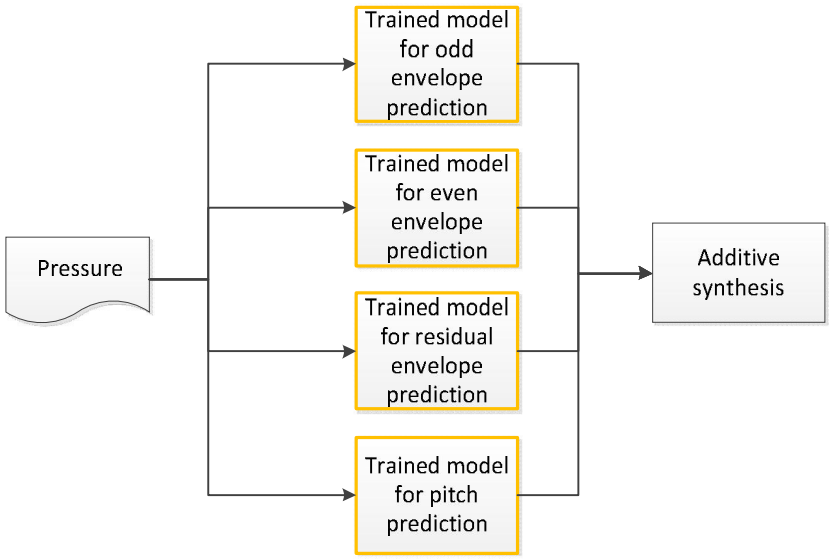

Figure 54: Sound synthesis from the predicted spectral descriptors

The prediction of the timbre by means of this second model provide also a large pitch error about 20Hz (*Figure 55*). Likewise the *Figure 56* shows that the envelopes are not well predicted. In general, the results obtained with this method are worse than for the first model. Thus we can conclude that using one model to predict each descriptor does not improve the performances.

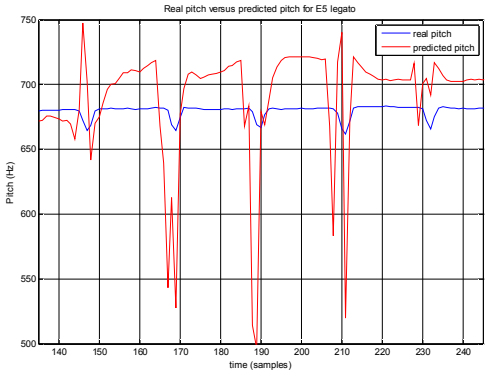

Figure 55: Real pitch versus synthesized pitch by model 2

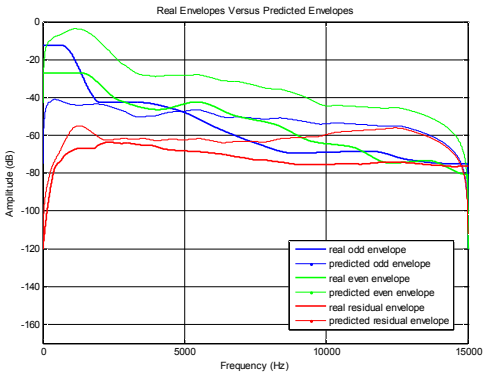

Figure 56: Real envelopes versus synthesized envelopes by model 2

5.2.4.3. Timbre Model 3

Aiming at improving the first model which got better results than the second one, we can train the first model with a "silence free data set". Since the models will only predict well the pitch activity, it is need de design a second network (a three-layers feed-forward with back propagation) which will recognize regarding the blowing pressure if the instrument is generating sound. In other word, a second artificial neural network predicts the pitch establishment which turns on or off the model predicting the timbre. Therefore two networks are trained, one taking as input the blowing pressure and as targets the spectral parameters, and a second model taking as input the blowing pressure and as target a Boolean representation of the pitch (0 if the pitch is inferior at 50 per cent of the expected pitch and 1 if it is superior). The MSNE of the model predicting the envelopes and the pitch is 0.0515 and the MSNE of the model predicting the pitch establishment is 0.9523.

Figure 57: Training for model 3 able of envelopes and pitch prediction

Figure 58: Training for model 3 able of pitch establishment prediction

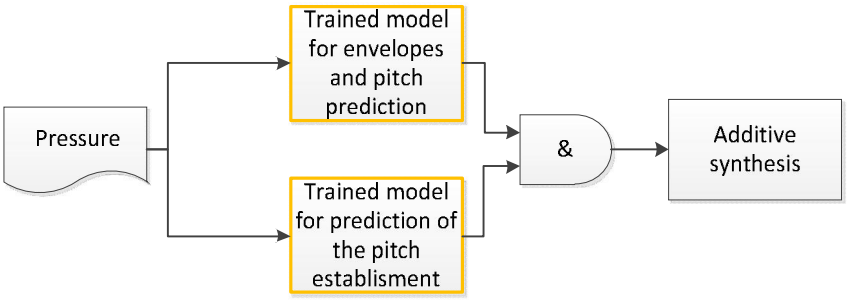

Figure 59: Sound synthesis from the predicted spectral descriptors and switched by the predicted pitch establishment

The *Figure 60* shows that feeding the model with only sequences where the pitch is active improve the accuracy of the pitch prediction. In the present case the error is reduced and become in average about 6Hz. In the *Figure 61* we can also see that the envelopes prediction is also enhanced with a better similarity between the predicted envelopes and the original envelopes. Being much better, the results obtained with this method allow generating by means of additive synthesis realistic sounds of recorder.

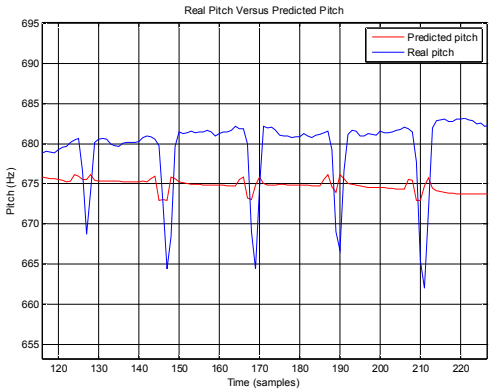

Figure 60: Real pitch versus synthesized pitch by model 3

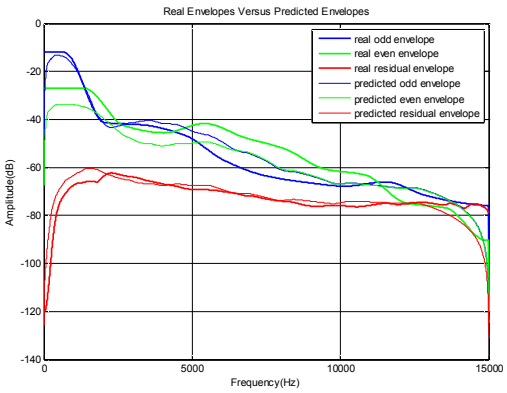

Figure 61: Real envelopes versus synthesized envelopes by model 3

5.2.5. MODEL FOR BLOWING PRESSURE ESTIMATION

In order to predict the blowing pressure from the spectral analysis of the timbre of the recorder (indirect acquisition), several methods have been tried but no significant improvements have been noticed. The selected method is composed by an ANN of two layers cascade-forward with back propagation trained with the spectral parameters as input and the pressure as target (*Figure 62*). With this configuration the MSNE resulting from the training is about 0.0238. At that point the prediction of the blowing pressure can be done by inputting the spectral parameters in the trained model (*Figure 63*).

Figure 62: Training for model able of pressure prediction

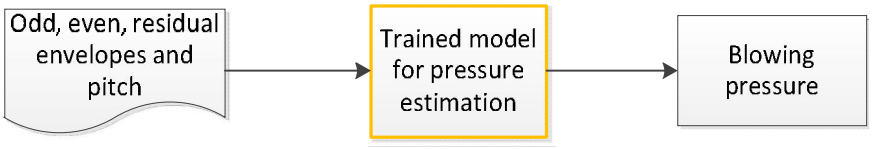

Figure 63: Prediction of the blowing pressure from the spectral descriptors

The result of the indirect acquisition can be observed in the *Figure 64* where it can be seen that the prediction (blue curve) is very noisy and not perfectly matching the real pressure (red curve). In order to improve the result it possible to smooth the predicted values by using the "smooth" function of MatLab with 7 as coefficient (green curve). Even if the indirect acquisition does not provide perfect results, the patterns corresponding to the articulations are detectable.

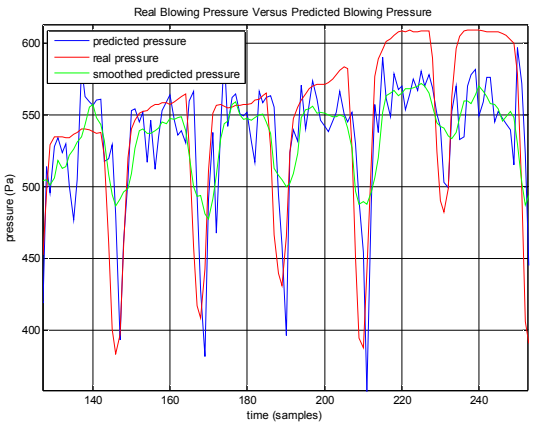

Figure 64: Real pressure versus predicted pressure + smoothed predicted pressure

5.3. SYNTHESIZER SCRIPT

Aiming at synthesizing a sequence of notes, a script able to switch between the different timbre models (one for each fingering) has been built. This script takes as input a track of blowing pressure and its corresponding score. In the present case the musical phrase which provides the blowing pressure is one from the recorded scales. Basically the scrip acts as follows:

- Loading the selected track of blowing pressure.
- Loading the fingerings corresponding to the track using the XML version of the score.
- Computing the onsets of each note/fingering and match them with the track of blowing pressure.
- For each note (between two onsets) loading the corresponding timbre model, predicting the spectral envelope and pitch from the blowing pressure.
- Reconstructing the spectrum.
- Concatenating all frames of the reconstructed spectrum.
- Synthesizing the spectrum by means of additive synthesis (SMS).

This preliminary algorithm allows to synthesize sound with only few parameters. Even if the resulting track sounds is quite realistic, the *Figure 65* shows that the transients are not well respected and that the notes duration is always longer than in the original track. The introduction of a model predicting the RMS energy of the notes could improve the general shape of the transient and make them more natural. On the *Figure 66* we can see that on the transients, the pitch is also badly predicted. This is mainly due to the way of training the models (discarding the silences so the transients) but it can be relativized by the fact that in the transient part the RMS energy is low; consequently the pitch of the sound is not hearable. In compensation we can observe that the pitch is mostly well predicted during the continuous part of the notes.

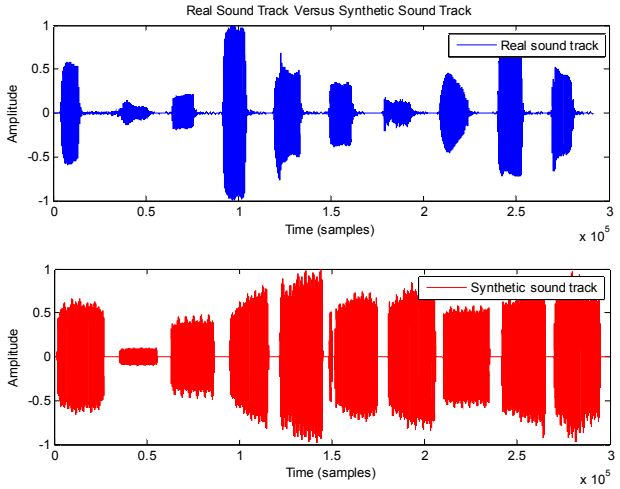

Figure 65: Comparison of the real sound track with the synthesized sound track

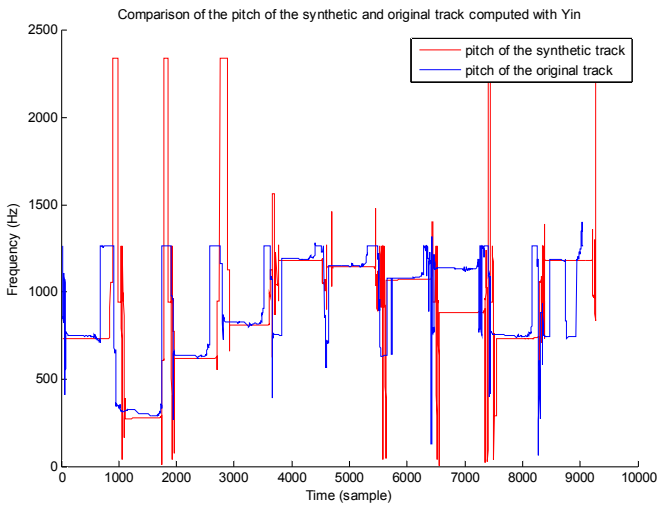

Figure 66: Comparison between the real pitch and the pitch of the synthesized sound track

6. CONCLUSION

Through the presented study of the relationship between the blowing pressure and timbre characteristics, it has been shown that it is possible to synthesize the timbre of the recorder instrumental gestures, and to estimate the blowing pressure by looking at spectral descriptors.

We developed and adapted a spectral representation of the timbre of the recorder by means of tree spectral envelopes representing odd harmonics, even harmonics and residual. This adapted representation allows to decrease the amount of data needed to reconstruct the spectrum (45 MFCCs and 1 pitch value, thus 46 parameters for each frame) and then permits to enhance the prediction of the timbre models which are based on artificial neural network.

As a preliminary test we could synthesize a short sequence of notes and notice that even if the ANN prediction can provide good estimations, a better synthesizing method is needed. Undeniably since the transients are not taken into account by our methods there are some discontinuity problems between notes (transition between modes of vibration, note attacks and note releases). It is also noticeable that the temporal dynamics of the instrument are not well respected. Actually the sound should come delayed regarding the blowing pressure but the type of ANNs used for the training do not take into account the temporal dimension of the relationship concerning sound and pressure. A last point is that the recorded database is not large enough to ensure a watertight training of the models. Having a bigger database would certainly have provided better results but would have without doubt required a superior computational power.

Nevertheless the outcome is promising, fairly good results were obtained and more of the problems already have an existing solution. Furthermore having developing this timbre modeling framework, we can contemplate to apply it to flute-like instruments or even to other instruments of the wind family.

6.1. FUTURE WORK

6.1.1. IMPROVING THE PREDICTION QUALITY

In this work we highlighted the fact that the quality of the sound strongly depends on the performances of the predictive models. Improving their performances could bring a significant gain of sound quality. With the aim to enhance the accuracy of the models it could be very interesting to train those using recurrent neural networks which are more complicated to implement and require more computational power, but are expected to perform better. In fact, by means of an included memory they can model a dynamic system, where the output not only depends on the actual input vector but also on previous ones and on the states of the system.
One type of recurrent networks that best suits this needs is the NARX network. Nevertheless it could also be useful to try other types of machine learning technics as decision tree learning, support vector machines or clustering.

In the section 4.2 we can observe a difference in the quality of the envelope representations by decreasing the number of bands used by the filters. It is not sure that this loss of accuracy significantly modifies the perception of the synthesized sound. Accordingly we could study the number of point necessary to represent the spectrum and to define a threshold behind which the human hear cannot perceive the difference.

6.1.2. BUILD A COMPLETE SET OF TIMBRE MODELS

Due to a lack of time, it has not been possible to compute the whole set of timbre models (one for each fingering), consequently it was not possible to synthesize any piece of music. By carrying out this task a full piece of music could be synthesized from the score and the corresponding blowing pressure. We could also imagine training a model gathering every single fingering, and would be inputted by the blowing pressure plus the wanted fingering. In section 5.3 we obviously notice that the synthetic sound suffers from a lack of accuracy on the transients predictions. One solution could be to build a specific model only trained for the prediction of transient regions.

6.1.3. DEVELOPMENT OF THE DATABASE

In the presented work the database is kept simple and just gathers the most obvious playing technics as the dynamics or the articulation. A possibility to enhance that database would be to create a musical score where the musician will use other playing technics as vibrato, tremolo or other ornaments.

6.1.4. REAL-TIME POSSIBILITIES

Another important point is the development of real-time application. It is conceivable to build a MaxMsp patche able of real-time synthesis by means of statistical timbre models previously trained. We can also imagine a patch aiming at real-time indirect acquisition which

will allow to observe the blowing pressure applied by the player during the performance without any internal sensors.

6.1.5. COMPLETE SYNTHESIZER

The last but not the least point will be to realize a synthesizer of gestures and by connecting it to the timbre synthesizer which has been developed in this thesis, being able to synthesize any piece of music by inputting the corresponding score (no need of the associated blowing pressure) and by choosing the type of articulation, the dynamic or even the mood of the virtual player.

BIBLIOGRAPHY

[1]. *FLETCHER NH. ACOUSTICAL CORRELATES OF FLUTE PERFORMANCE TECHNIQUE. AMERICA. 1975.*

[2]. *FLETCHER NH. NONLINEAR THEORY OF MUSICAL WIND INSTRUMENTS. APPLIED ACOUSTICS. 1990.*

[3]. *FLETCHER NH, ROSSING TD. THE PHYSICS OF MUSICAL INSTRUMENTS 1998.*

[4]. *MARTIN J. THE ACOUSTICS OF THE RECORDER. MOECK 1994.*

[5]. *NICOLAS MONTGERMONT, BENOIT FABRE, PATRICIO DE LA CUADRA. FLUTE CONTROL PARAMETERS: FUNDAMENTAL TECHNIQUES OVERVIEW. INTERNATIONAL SYMPOSIUM ON MUSIC ACOUSTICS, 2007.*

[6]. *TZANETAKIS G, KAPUR A, TINDALE AR. LEARNING INDIRECT ACQUISITION OF INSTRUMENTAL GESTURES USING DIRECT SENSORS. COMPUTER.*

[7]. *PHILIPPE BOLTON, RECORDER MAKER. HTTP://WWW.FLUTE-A-BEC.COM.*

[8]. *THE UNIVERSITY NEW SOUTH WALES, SYDNEY, AUSTRALIA.* http://www.phys.unsw.edu.au.

[9]. *INTERVIEW OF THE TEACHER OF RECORDER JUAN VIVES.*

[10]. *MONTGERMONT N, FABRE B, CUADRA PD. GESTURE SYNTHESIS: BASIC CONTROL OF A FLUTE PHYSICAL MODE. 2008*

[11]. ESTEBAN MAESTRE GOMES. MODELING INSTRUMENTAL GESTURES: AN ANALYSIS/SYNTHESIS FRAMEWORK FOR VIOLIN BOWING. TESI DOCTORAL, UNIVERSITAT POMPEU FABRA, 2009.

[12]. DEMOUCRON MATTHIAS, CAUSSÉ RENÉ. SOUND SYNTHESIS OF BOWED STRING INSTRUMENTS USING A GESTURE BASED CONTROL OF A PHYSICAL MODEL. IRCAM, 2007.

[13]. WANDERLEY, M. M. NON-OBVIOUS PERFORMER GESTURES IN INSTRUMENTAL MUSIC. MOST, 37-48. (1999).

[14]. CADOZ,C. INSTRUMENTAL GESTURE AND MUSICAL COMPOSITION. IN INTERNATIONAL COMPUTER MUSIC CONFERENCE.1988: INTERNATIONAL COMPUTER MUSIC ASSOCIATION.

[15]. TRAUBE, C., DEPALLE, P., & WANDERLEY, M. (2003). INDIRECT ACQUISITION OF INSTRUMENTAL GESTURE BASED ON SIGNAL, PHYSICAL AND PERCEPTUAL INFORMATION. AREA, 42-47.

[16]. CHEN, E. W., WALES, N. S., & NSW, S. VOCAL TRACT INTERACTIONS IN RECORDER PERFORMANCE INTERNATIONAL CONGRESS ON ACOUSTICS MADRID (2007).

[17]. WALES, N. S. THE RELATION BETWEEN THE VOCAL TRACT AND RECORDER SOUND QUALITY. LANGUAGE, (JANUARY). (1998).

[18]. SCAVONE, G., SILVA, A., & WEST, S. S. FREQUENCY CONTENT OF BREATH PRESSURE AND IMPLICATIONS FOR USE IN CONTROL. SENSORS (PETERBOROUGH, NH), 93-96. (2005).

[19]. ALFONSO ANTONIO PÉREZ CARRILLO. ENHANCING SPECTRAL SYNTHESIS TECHNIQUES WITH PERFORMANCE GESTURES USING THE VIOLIN AS A CASE STUDY. TESI DOCTORAL.,UNIVERSITAT POMPEU FABRA, 2009.

[20]. SCHWARZ, D. CURRENT RESEARCH IN CONCATENATIVE SOUND SYNTHESIS. COMPUTER, (ICMC). (2005).

[21]. ALFONSO PÉREZ CARRILLO, JORDI BONADA. MODELING THE INFLUENCE OF PERFORMANCE CONTROLS ON VIOLIN TIMBRE. MUSIC TECHNOLOGY GROUP, UNIVERSITAT POMPEU FABRA.

[22]. N.H. FLETCHER AND LORNA M. DOUGLAS. HARMONIC GENERATION IN ORGAN PIPES, RECORDERS, AND FLUTES. DEPARTMENT OF PHYSICS, UNIVERSITY OF NEW ENGLAND, ARMIDALE NEW SOUTH WALES 2351, AUSTRALIA .1980

[23]. CHEW, E, ZIMMERMANN, R., SAWCHUK, A. A, PAPADOPOULOS, C, KYRIAKAKIS, C, TANOUE, C, ET AL. (N.D.). UNIVERSITY OF SOUTHERN CALIFORNIA VITERBI SCHOOL OF ENGINEERING. A SECOND REPORT ON THE USER EXPERIMENTS IN THE DISTRIBUTED IMMERSIVE PERFORMANCE PROJECT. 2005.

[24]. X SERRA, J SMITH III - COMPUTER MUSIC JOURNAL, 1990 – JSTOR SPECTRAL MODELING SYNTHESIS: A SOUND ANALYSIS/SYNTHESIS SYSTEM BASED ON A DETERMINISTIC PLUS STOCHASTIC DECOMPOSITION *Computer Music Journal*, *14*(4), 12-24.

[25]. SCHONER, B., COOPER, C., DOUGLAS, C., & GERSHENFELD, N. (1999). DATA-DRIVEN MODELING OF ACOUSTICAL INSTRUMENTS. JOURNAL FOR NEWMUSICRESEARCH, 28:28–42.

[26]. A. CHAIGNE AND J. KERGOMARD. ACOUSTIQUE DES INSTRUMENTS DE MUSIQUE. BELIN, 2009

[27]. FRANÇOIS BLANC, PRODUCTION DE SON PAR COUPLAGE ECOULEMENT/RESONATEUR ACOUSTIQUE : ETUDE DES PARAMETRES DE FACTURE DE FLUTES PAR EXPERIMENTATIONS ET SIMULATIONS NUMERIQUES D'ECOULEMENTS. THESE DE DOCTORAT DE L'UNIVERSITE PIERRE ET MARIE CURIE, 2009

ANNEXES

A. Pictures of the recorder prototype

Figure 67: View of the mouthpiece, the bloc and one of the sensors (photo from Josep Tubau)

Figure 68: View of the mouthpiece with the bloc mounted (photo from Josep Tubau)

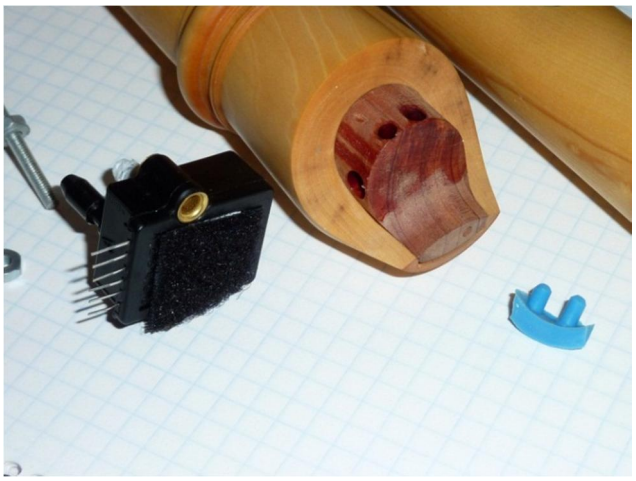

Figure 69: View of one sensor plus the mouthpiece (photo from Josep Tubau)

B. COMPUTER-AIDED DESIGN OF THE PROTOTYPE OF RECORDER

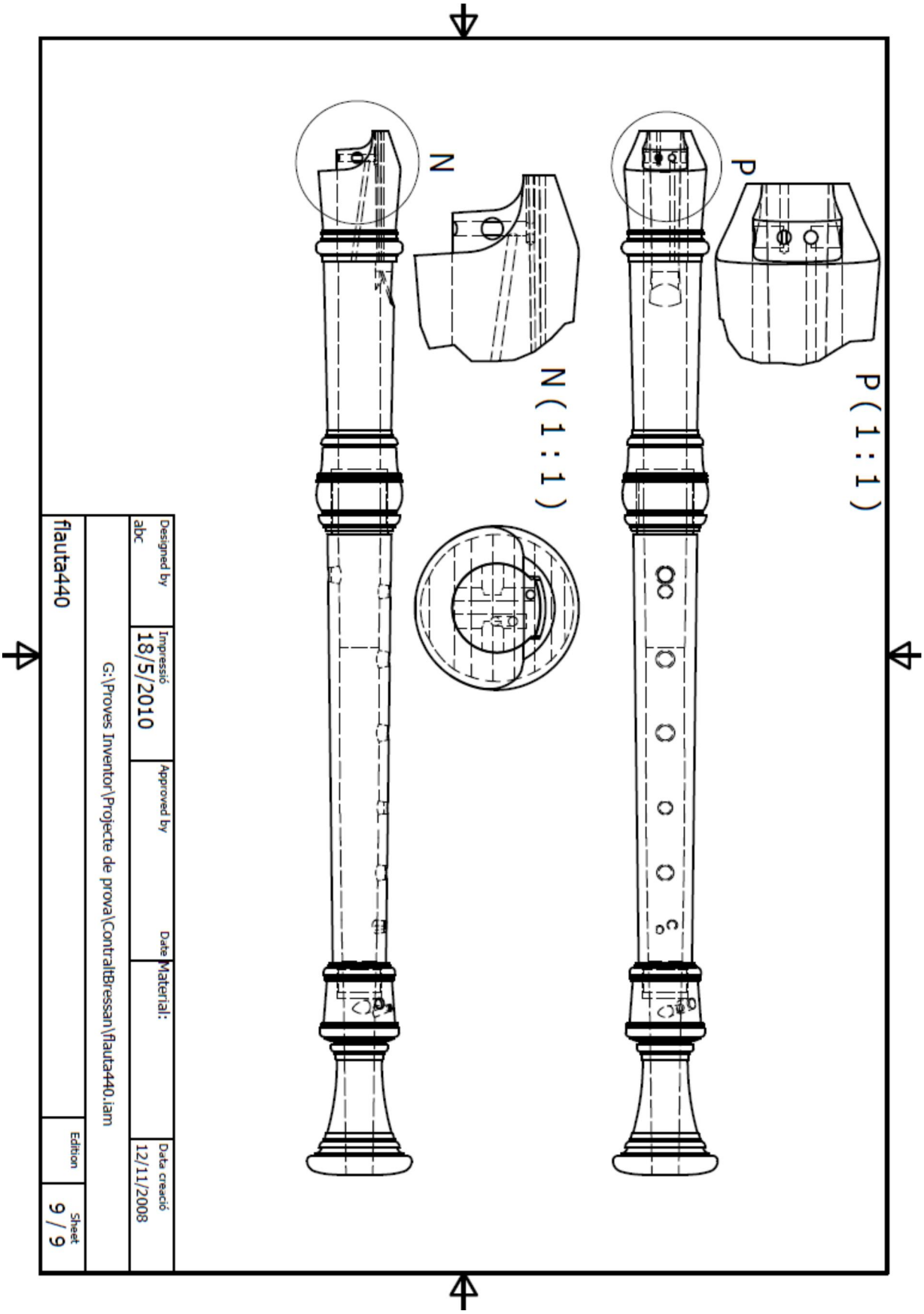

Figure 70: CAD of the recorder. Mouthpiece plus body (schematic from Josep Tubau)

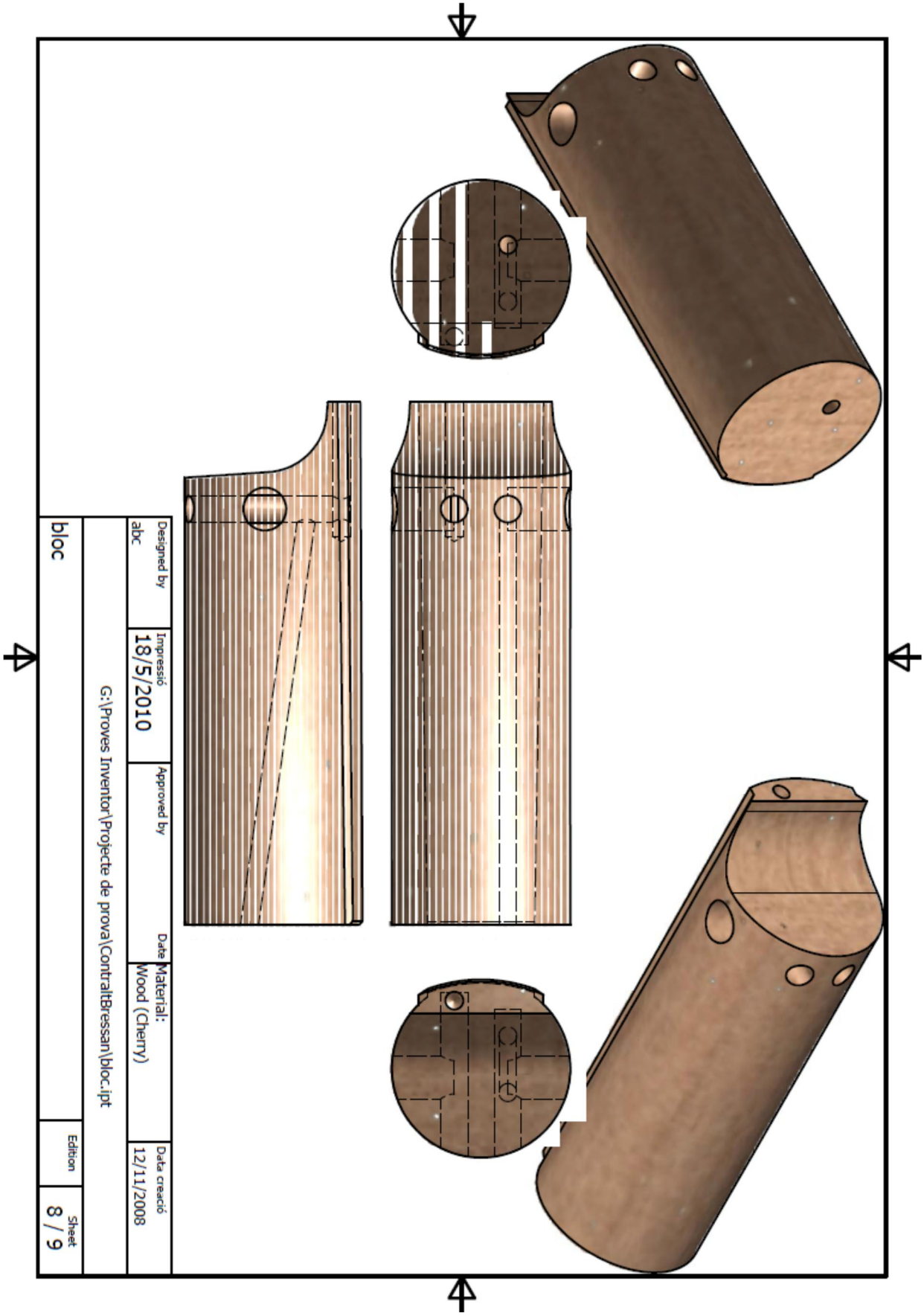

Figure 71: 3D view of the bloc of the recorder (schematic from Josep Tubau)

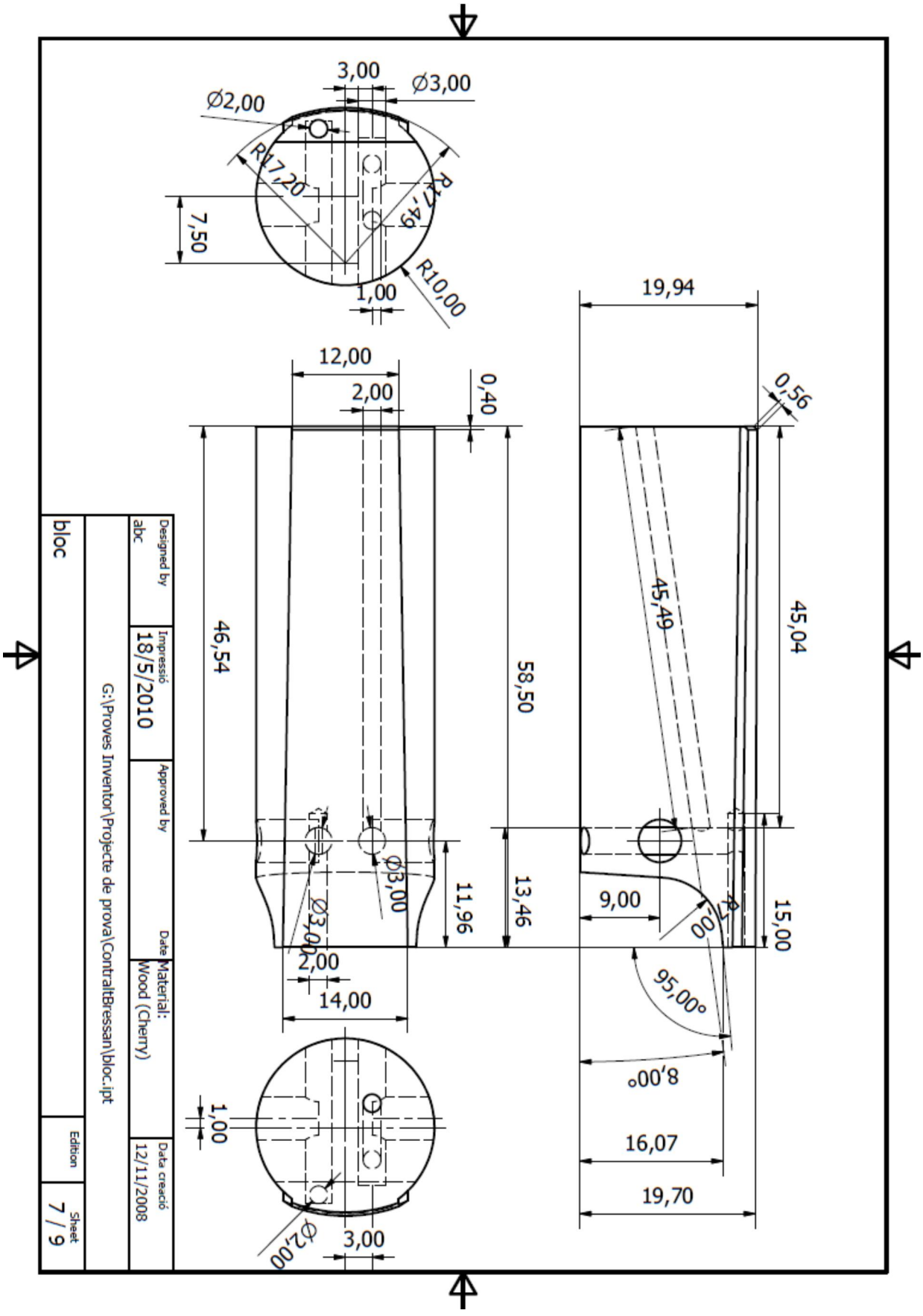

Figure 72: CAD of the bloc of the recorder (schematic from Josep Tubau)

C. MUSICAL PIECES

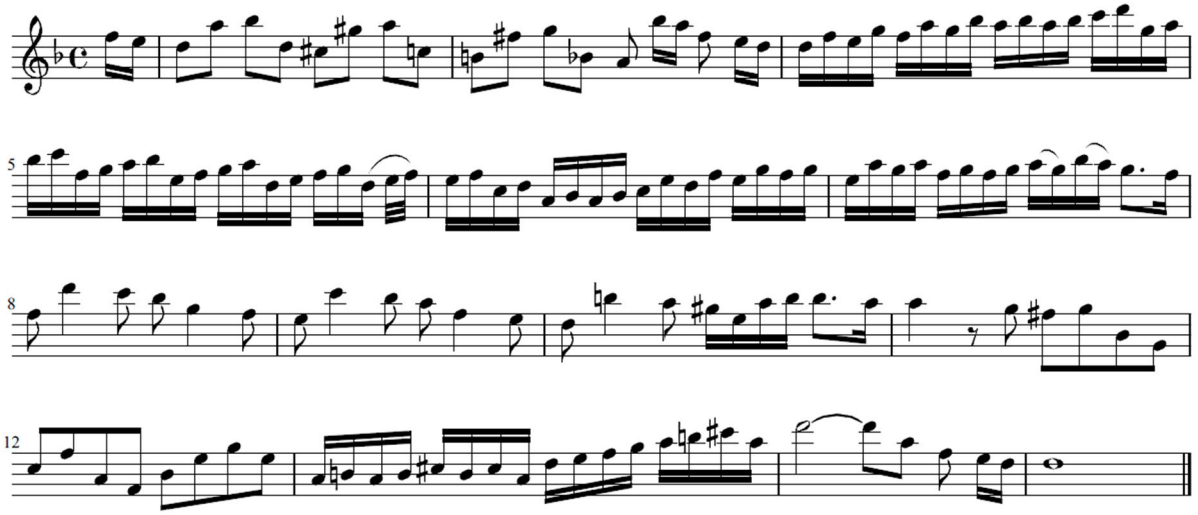

Figure 73: Piece 1, Prelude by Henry Purcell

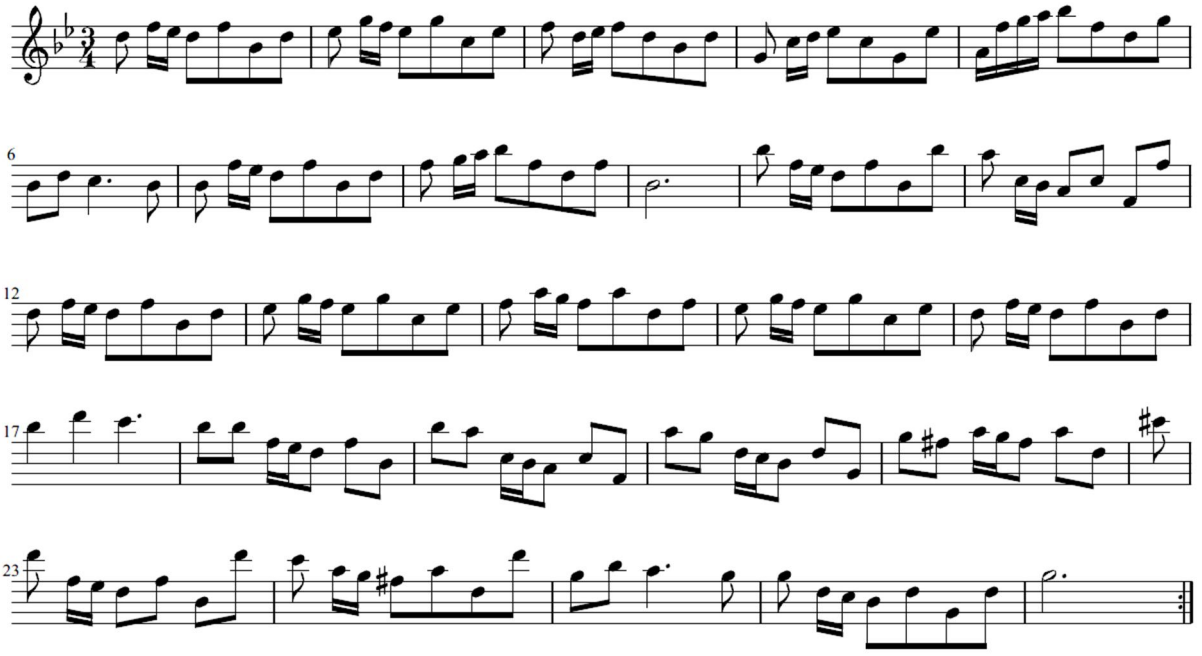

Figure 74: piece 2, Prelude by Torelli

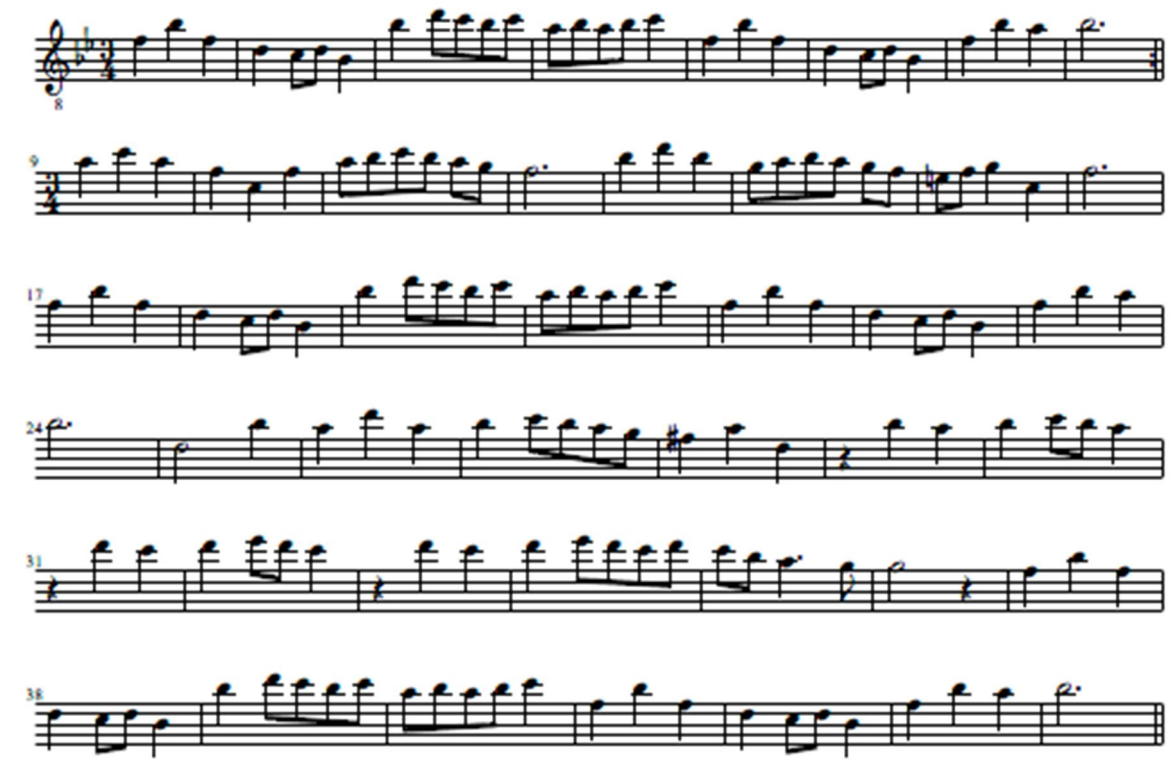

Figure 75: pieces 3, Prelude by Pepusch

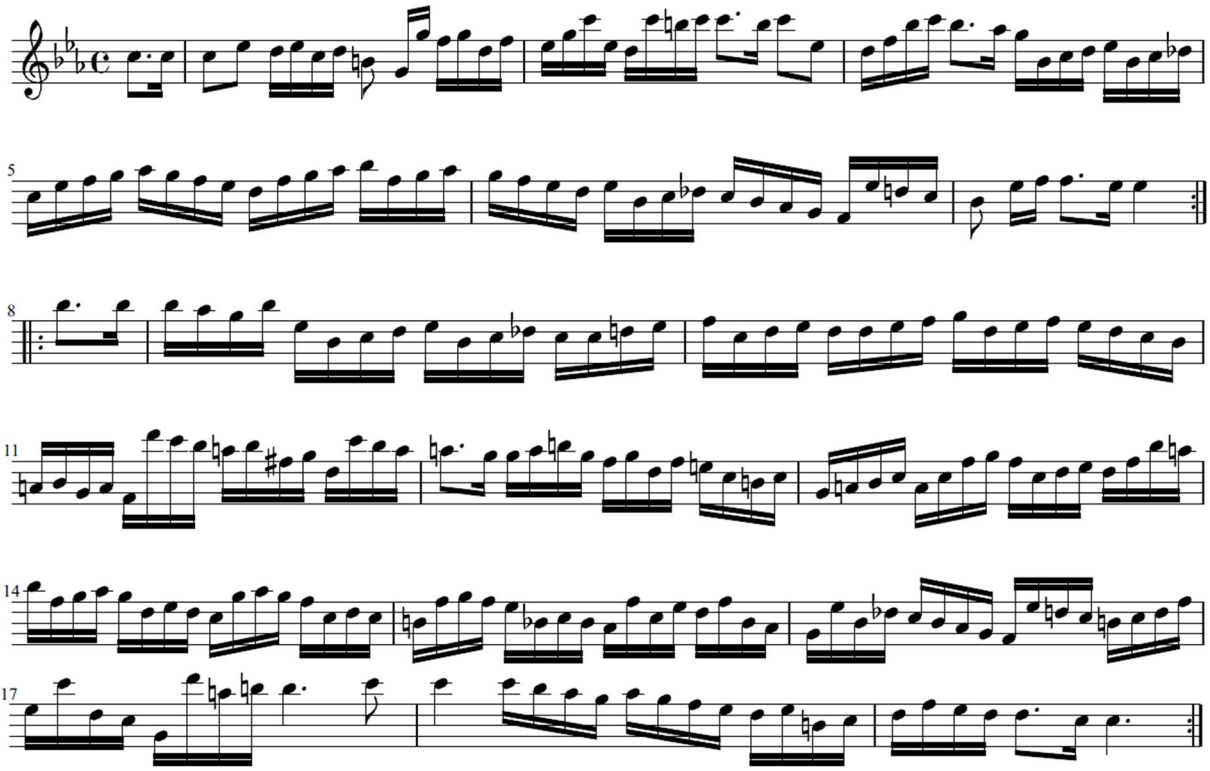

Figure 76: Piece 4, Prelude by Nicola

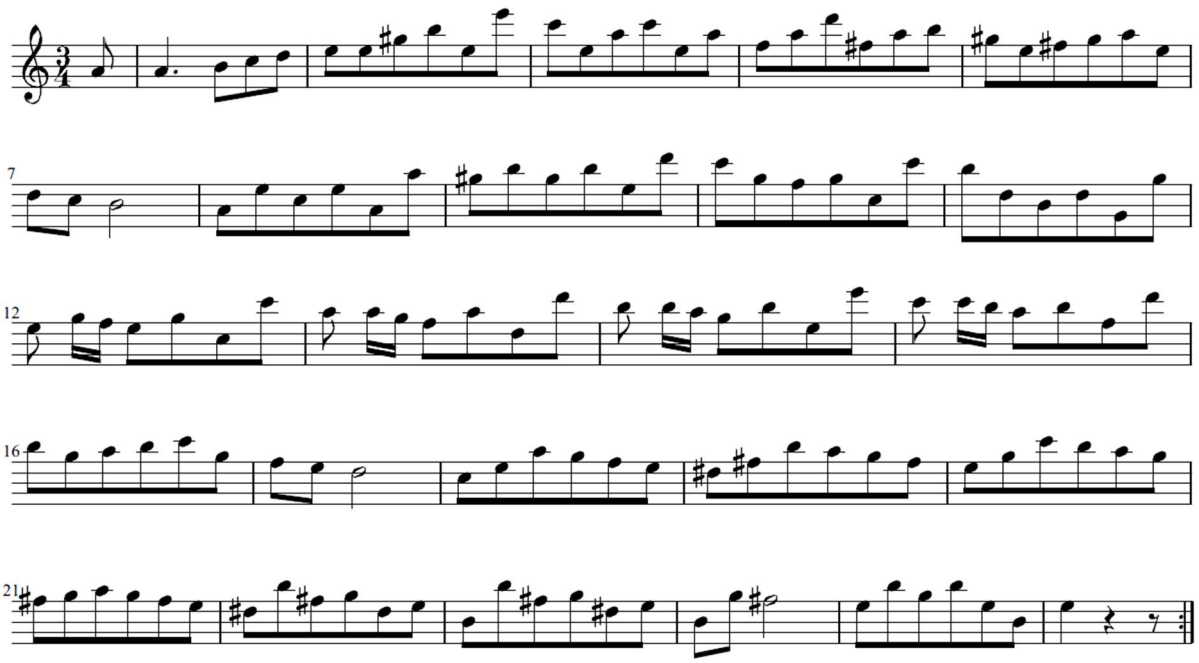

Figure 77: Prelude by Ziani

D. Table of harmonic / non-harmonic second mode of vibration

Fingerings	SMS analysis	Second modes
G6	ok	harmonic
G5	ok	harmonic
G4	ok	harmonic
G#6	ok	harmonic
G#4	ok	harmonic
F6	**wrong**	**not harmonic**
F5	ok	harmonic
F4	ok	harmonic
F#6	**wrong**	**not harmonic**
F#5	ok	harmonic
F#4	ok	harmonic
E6	**wrong**	**not harmonic**
E5	ok	harmonic
D6	**wrong**	**not harmonic**
D5	**wrong**	**not harmonic**
D#5	ok	harmonic
D#6	**wrong**	**not harmonic**
C6	ok	harmonic
C#6	**wrong**	**not harmonic**
C#5	**wrong**	**not harmonic**
C5	ok	harmonic
B5	**wrong**	**not harmonic**
B4	**wrong**	**not harmonic**
A6	ok	harmonic
A5	**wrong**	**not harmonic**
A4	ok	harmonic
A#5	**wrong**	**not harmonic**
A#4	ok	harmonic

Figure 78: Table of the harmonic/ non-harmonic second modes according to the fingerings: For each fingering we observe the spectral evolution of the recorder timbre when the blowing pressure is increased and we determine whether or not this evolution bring new harmonic series. In the of new harmonic series the evolution is called non-harmonic

E. Table of the mean square normalized error of each timbre model built

Fingering	F4	A5	D5	E5	F5	G5	D6	E6	F6
Performance of ANN (MSNE)	0.0376	0.0364	0.0406	0.0515	0.0443	0.0425	0.0379	0.0433	0.0399

Figure 79: Mean square normalized error of each timbre model