ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/334093755

Design, Implementation and Validation of a Novel Real-Time Automatic Piano Transcription System

Thesis · February 2017

DOI: 10.13140/RG.2.2.31541.52960

CITATIONS
0

READS
201

2 authors:

Adria Galin
Autonomous University of Barcelona
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

David Castells-Rufas
Universitat Autònoma de Barcelona & Barcelona Supercomputing Center
105 PUBLICATIONS **440** CITATIONS
SEE PROFILE

All content following this page was uploaded by Adria Galin on 28 June 2019.

The user has requested enhancement of the downloaded file.

Universitat Autònoma de Barcelona

Master's Thesis

Design, Implementation and Validation of a Novel Real-Time Automatic Piano Transcription System

Author: Adria Galin

Supervisor:

Dr. David Castells-Rufas
Microelectronics and Electronic Systems Department

Escola Tècnica Superior d'Enginyeria (ETSE)
Universitat Autònoma de Barcelona (UAB)

February 2017

Declaration of Authorship

I, Adria Galin, declare that this thesis titled, 'Design, Implementation and Validation of a Novel Real-Time Automatic Piano Transcription System' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Historia vero testis temporum, lux veritatis, vita memoriae, magistra vitae, nuntia vetustatis, qua voce alia nisi oratoris immortalitati commendatur?"

Cicero, De Oratore, II, 36.

Universitat Autònoma de Barcelona

Escola Tècnica Superior d'Enginyeria (ETSE)

Abstract

Thesis: Design, Implementation and Validation of a Novel Real-Time Automatic Piano Transcription System

By: Adria Galin

To date, the accurate Music Transcription has been considered to be a rather abstract concept. Our goal was to explore this field of study, which until today has been not attended to. In order to be able to transcribe music automatically, a new type of algorithm has been created. The present dissertation represents the continuation on the Authors research field introduced in Galin A. et al (2015). This algorithm is composed of the Beat Tracking, Note Acquisition and Post Processing processes; during which development, multiple simulations have been run in order to perfect them. With these simultaneously running algorithms, we have reached our goal and were enabled to transcribe music accurately.

Universitat Autònoma de Barcelona

Escola Tècnica Superior d'Enginyeria (ETSE)

Abstract

Thesis: Design, Implementation and Validation of a Novel Real-Time Automatic Piano Transcription System

By: Adria Galin

Hasta la fecha la Transcripción de la Música Automática se ha considerado un concepto bastante abstracto. Nuestro objetivo ha sidoo explorar este campo de estudio. Con el fin de poder transcribir música automáticamente, se ha creado un nuevo tipo de algoritmo. Este documento establece la continuación de la linea de investigacion presentada en Galin, A. et al (2015). Este algoritmo se compone de los procesos: Beat Tracking, Note Acquisition y Post Processing; durante los cuales se han desarrollado múltiples simulaciones para perfeccionarlos. Con estos bloques que funcionan simultáneamente, hemos alcanzado nuestro objetivo y hemos podido transcribir la musica con precisión.

Universitat Autònoma de Barcelona

Escola Tècnica Superior d'Enginyeria (ETSE)

Abstract

Thesis: Design, Implementation and Validation of a Novel Real-Time Automatic Piano Transcription System

By: Adria Galin

Fins la data la transcripció de la música automàtica ha estat considerat com un concepte bastant abstracte. El nostre objectiu ha estat el d'explorar aquest camp d'estudi. Per tal de ser capaç de transcriure música de forma automàtica, un nou tipus d'algorisme ha estat creat pels autors. Aquest document estableix la continuació de la linia d'investigació presentada en Galin, A et al (2015). Aquest algoritme està compost pels seguents blocs: Beat Tracking, Note Acquisition i Post Processing; durant el desenvolupament d'aquests darrers, múltiples simulacions s'han dut a terme per tal de perfeccionar-se. Amb aquests presents algoritmes que s'executen simultaniament, hem assolit el nostre objectiu i ens hem capacitat per a transcriure la musica amb precisió.

Acknowledgements

Firstly, I would like to express my deepest gratitude to all the people involved that have in one way or another contributed to my thesis writing as well as my academic career path.

Namely, I would like to thank my thesis supervisor, Dr. David Castells Rufas at the Department of Microelectronics and Electronic Systems for his guidance and support during this time. He has not only helped me improve my written work, but has also helped me redefine my ideas so that they would become clearer and more comprehensive. Also, I owe him my gratitude for the much needed counseling on professional as well as personal level. Further, I would like to thank all of the coordinators of the Bachelor's and Master's Degree in Telecommunication Systems Engineering - Professors Dr. José A. López Salcedo and Dr. José López Vicario, for all of their assistance not only throughout the thesis-writing process, but also for consistent help during the academic year, for which I am more than grateful.

Further, I owe my gratitude to Dr. Antoni Morell Perez and Emmanuel Zenou -from Universitat Autònoma de Barcelonat department of Telecommunications and Systems Engineering and Institute Superieur de l'Aeronautique et de l'Espace at the Department of Computer Vision and Signal Processing respectively- for not only providing me with excellent academic background knowledge in the area of Instantaneous Spectrum to build upon, but foremost, for always cheering me on and supporting my decision in pursuing my further academic career.

Also, I would like to take this opportunity to express thank all of my professors involved in the program that have shaped my academic career by sharing their wisdom with me so that I will be ready for my future career in Telecommunication Engineer. And lastly, I would like to thank my family and friends for encouragement, moral support and attention; which enabled me to finish my studies until a successful end.

And with this record, my sense of gratitude to all, who directly or indirectly, have lent their hand in this venture. I am aware that in this short letter, I have not been able to mention everyone that has positively contributed not only to my thesis writing, but also to my life, but just know that I am truly grateful to all of you.

Contents

Abbreviations

ADSR	**A**ttack **D**ecay **S**ustain **R**elease
AM	**A**mplitude **M**odulation
BIN	**B**inarized Process
BPM	**B**eat **P**er **M**inute
BPS	**B**eat **P**er **S**econd
BS	**B**inarized **S**ignal
ERB	**E**quivalent **R**ectangular **B**andwidth
ESF	**E**ffective **S**ampling **F**requency
FM	**F**requency **M**odulation
fo	**F**undamental frequency
fs	**S**ampling Frequency
FT	**F**ourier **T**ransform
GT	**G**round **T**ruth
IF	**I**nstantaneous**F**requency
IS	**I**nstantaneous**S**pectrum
K	**S**ensibility **F**actor
MAPS	**M**idi **A**ligned **P**iano **S**ounds
MIDI	**M**usical **I**nstruments **D**igital **I**nterface
MP	**M**usical **P**itch
PF	**P**eack **F**inder
PMV	**P**eack **M**edian **V**alue
S	**S**cenario
SPEC	**S**pectogram
STFT	**S**hort **T**ime **F**ourier **T**ransform
TDS	**T**ime **D**etection **S**ensibility

TE **T**eager **E**energy

TO **T**eager **O**perator

Chapter 1

Introduction

1.1 Motivation

Music has to a great extent influenced generations in multiple ways. It has been one of the primary methods of how people were able to express themselves, including their thoughts and emotions. To this day, one has never been able to capture music to its full extent and observe all that it has to offer, provided that the particular individual was not blessed with the gift of an absolute note recognition; which is more than rare. For this very reason, we have decided to start exploring this field of study in order to capture music in its purest form. In other words, we have set out to challenge the limitations of human nature and substitute them by the means of modern detection algorithms.

Firstly, the Automatic Music Transcription aims to obtain a symbolic representation of the music content from an observation of one or more audio signals. For the music representation, the transcript should contain a specific key information such as the height of the played notes, rhythm, tempo and time signature. This information obtained is essential and can be then used to derive other, more abstract musical information; including tone, agreements and the structure of a song, etc,. Shall one be interested in the tone of the music signals, that is, the portion of the signal corresponding to the instruments playing notes - singing voice, piano, etc. - the transcription shall detect the beginnings and ends of notes also known as activations (onset / offset), in addition to the height of each of the notes that makes up for the analyzed signal. Such information is sufficient to obtain a precise symbolic representation of any musical event.

The objective, as stated above, is hence, to create such algorithm that is able to detect all of the essential parameters of the input signal and be able to present it as a musical score. These parameters -based on an observation and/or multiple observations of this input signal- will consequently serve as a firm basis to transcribe the context heard adequately to meet the desired objectives.

1.2 Outline of the Dissertation

A brief overview of the thesis structure will be the presented, excluding the introduction itself.

Next, in the second chapter, some of the important parameters which will help with the understanding of the basic principles, which our field of study is based upon, will be presented.

In the course of the third chapter, the Beat Tracking algorithm will be addressed together with its implementation. The Beat Tracking is one of the essential components of the Automatic Music Transcription, which in particular will provide for a tracking rhythm component of the music transcription.

The fourth chapter shall attend to the concern of the Note Acquisition as the second algorithm of the Automatic Music Transcription. Its objective is to capture a specific range of sound of the input signal, which in return shall be identified and transcribed.

Therefore in the fifth chapter the Post Processing block will be presented. This block is necessary in order to increase th e performance of the AMT system.

In the sixth an overview of all of the above will be summarized and briefly presented.

During the course of the seventh chapter a comparison within the state of the art proposal will be addressed and a innovative common way of measuring the performance of a give AMT system will be described.

1.3 The Greater Picture

The objective of this thesis is to be able to transcribe any type of music source automatically. This process begins with an initial Input signal, which is being heard; which will be subsequently analyzed and transcribed. This process consists of two separate analysis sections - Beat Tracking and Note Acquisition. Beat tracking is concerned with obtaining the rhythm of the given input signal. Similarly, Note Acquisition process' goal is to capture and detect the pitch of the different notes together with allowing to identify multiple notes captured at the same time during their starting point of onset. Finally the Post Processing block aims to increase the performance of the AMT system by means of adding decision making. With all the previous three blocks an Automatic Music Transcription takes place.

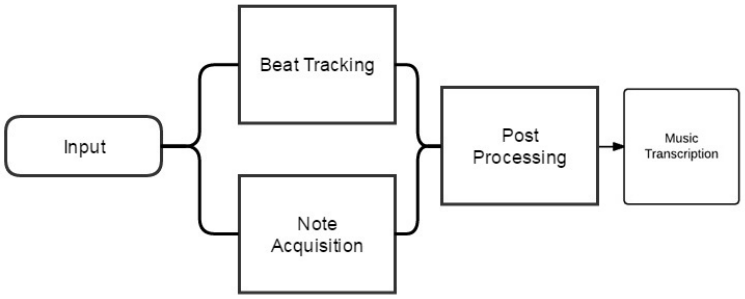


Figure 1.1: Block Diagram of the Music Transcription.

In the graph above, Figure 1.1, the block diagram of the different elements of the process may be observed. In the following chapters, each of the individual blocks of the graphs will be attended to and explained separately. Furthermore, it is important to bear in mind that in order to transcribe correctly, the blocks - Beat Tracking, Note Acquisition and Post Processing, must be working simultaneously. Additionally, all of these processes function as separate, individual units, which are not dependent on each other and therefore, the functionality of one does not affect the other. However, all are necessary for the music transcription.

Chapter 2

Basis and Musical Parameters

This section will allow us to present the processing elements of an audio signal and the musical concepts that shall be regularly referenced to in the following chapters. We have also found it useful to present briefly some of the tools and models that are not directly connected to the field of our study in retrospect. These elements are present, for they are necessary basis for understanding of the thesis itself. Further, we wanted to briefly highlight some of the context of the research conducted, in order to point out some of its advantages and explore its further possible implications.

2.1 Fundamental Frequency and Height

The perception of sound emitted by a harmonic or quasi-harmonic source - for instance the tonal instrument - generally corresponds to the fundamental frequency, also called frequency F0. The frequency F0 is associated with the first part of the model theory of the harmonic source, which will be discussed later on; identified by a peak in the amplitude spectrum of the respective signal. Further, other partial integer multiples of the frequency F0 are represented in the quasi-harmonic source spectrum.

By definition, a frequency is measured in hertz (Hz) or radians per second Rad/s - angular speed- ; although there are other measurement scales[1]. In particular, Equivalent

[1]*The following list of methods will not be directly mentioned nor necessary for our particular field of study. Therefore, it shall be considered as additional information to enrich the reader. For further information, please refer to: Stevens et al. (1940)

Rectangular Bandwidth scales (ERB), Mel and Bark3 were defined in relation to a psychoacoustic model on the human collection.

Musical Pitch or Pitch is a quantity that can associate any given concrete note with its corresponding frequency. The height is calculated for a reference frequency by the following formula:

$$Pitch(f) = P_{ref} + O \cdot log_2 \frac{f}{F_{ref}} \tag{2.1}$$

where Pref and Fref are the reference pitch and frequency respectively. The constant O sets the number of subdivisions of octave4. In Western music based on equal temperament, we define 12 different musical notes per octave or O=12. In the MIDI standard used worldwide by all musicians, sets these values to Pref = 69 , Fref = 440Hz.

The problem of estimating the height for the automatic transcription of music is associated with the estimation of the corresponding fundamental frequency. In some of the applications, the objective is to recognize names of the notes played. Hence, in the case of using a 12-note scale corresponding to the western music structure, the dimensions of the problem are significantly reduced, because there are only 12 possibilities. This will then provide for one of the solutions. That is to identify the frequency F0 or a multiple of its nF0 form for n in Z. In music applications, a prior knowledge of the tuning fork used (for example: A4 = 440 Hz) isolates regions of the spectrum, in order to find the fundamental frequencies. Thus, adjacent frequencies detected are reduced to the corresponding musical scale.

2.2 Uncertainty Principle: Heisenberg

In this segment, the trade-offs which are being in made in our scope of analysis between temporal resolution and frequency resolution will be briefly justified. Further, some background information about the principles on which are justification arguments are based upon will be presented.

Heisenberg Principle also known as Uncertainty Principle states that it is impossible to determine simultaneously both the position and velocity of a particle with any great

degree of accuracy or certainty. The more precisely one is known, the less precisely the other can be known. It is a statement about the nature of the system itself as described by the equations of quantum mechanics.

Any use of the words "position" and "velocity" with an accuracy exceeding that given by the uncertainty equation is just as meaningless as the use of words whose sense is not defined. (Physical Principles of the Quantum Theory, 1948)

We can link the Heisenberg Principle in Fourier Analysis. In this case we will not speak in terms of position and momentum (velocity), now the relationship will be between temporal signal and its Fourier Transform (FT).

The uncertainty principle will become especially important in chapter 4, concerned with the Note Acquisition process.

2.3 Mono-component and Multi-component Signals

Instantaneous Frequency (IF) is an important characteristic for those signals whose their spectral components changes in time, so Instantaneous Frequency is a parameter that change in time. The concept of a Mono-component signal can be interpreted as the frequency of a sinusoidal signal that corresponds or adjusts in the local point under analysis. Consequently the IF only has meaning in mono-component signals, where only exists one frequency or a finite range of frequencies, which changes as a function of time. For multi-component signals, the notion of IF for each time instant looses meaning, and it is necessary to do an analysis of the variation of each component individually.

From a Mechanic point of view frequency can be defined as the number of oscillations per unit of time. Vibration can be any type of movement from the point p0 to p1, and an oscillation is a complete movement, so the sequence p0 - p1 - p0. Using this process as a model, then we can define frequency for any type of arbitrary vibration.

Ville with the previous work of Van der Pol defined IF as -See also equation 5.1 and 5.2 -:

$$f_i(t) = \frac{1}{2\pi} \cdot \frac{d}{dt}[\arg z(t)] \tag{2.2}$$

The importance of IF is that it allows us to get a measure of the localization in frequency of the concentration of energy of the signal as a temporal function. This propriety explains the importance of the Instantaneous Frequency. For multi-component signals we will use the next model:

$$x(t) = \sum_{k=1}^{N} x_k(t) \tag{2.3}$$

Where xk(t) corresponds to each one of the N mono-component non-stationary signals, with envelopments variants in time ak(t) and also IF variant in time fik(k), analitical signal asociated with xk(t) can be extressed as:

$$z_k(t) = a_k(t) \cdot e^{k\varphi_k(t)} \tag{2.4}$$

Where:

$$\varphi_k(t) = \int_{-\infty}^{t} f_{ik}(\tau) d\tau \tag{2.5}$$

As we may observe in the equation above, all of the signals containing multi-components will be treated as an extension of signals composed of mono-components. However, assuming this extension, the equation 3.2 for mono-component signals does not apply to the equation containing the information about the multi-component signals.

2.4 Temporal and Transient Envelope

In this section, we wish to explore the relationship and the possible impacts of the presence of the note onto our temporal input signal.

In a more of a general setting, transient processes are associated with brief events resulting in significant changes in the nature of the studied signal.

However, in the particular case of music, these processes occur mainly during and/or between the starts and ends of notes (onsets and offsets) and have a particular envelope can be diagrammed in the next figure:

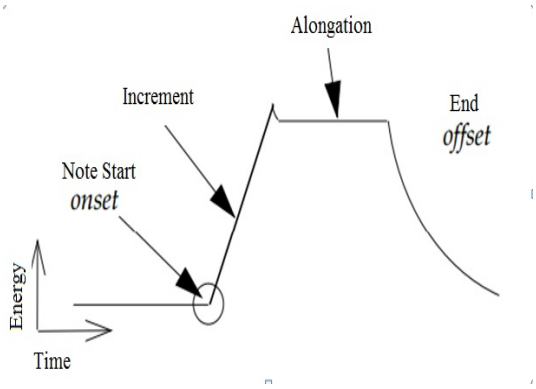


Figure 2.1: Onset and Offset of a determinate Note

The top note is located at the at the beginning of the activation of each note and consequently results in an increase in the energy level of the signal.

The note can then be maintained either by a resonance, which naturally follows each sound, or by a deliberate action of the musician with the intention of prolonging the note's resonance. This results in a weakening characteristic period from the end of note.

This characteristic envelope can be described by the model Attack Decay Sustain Release (ADSR) whose implementations are still used to this day in the field of musical synthesis. It has to be noted that the modulation of the previous envelope will coincide with the frequency of the played note.

2.5 Rhythm and Pulse Recognition

Rhythm and pulse are musical information that help to structure the musical events in time. Thus, for automatic transcription of music signals, this information may be useful for an a prior for example to increase the detection accuracy of the beginnings and ends of notes. The pulse is defined as an accentuation that occurs periodically in time. Thus, the regularity of the pulse sets the calculated tempo in beats per second. The rate is then given by number of beats per a given time-unit. Its value depends on the musical content (phrases organization), the will of the composer and the musical writing conventions. It is therefore quite possible to write musical themes identical but with different metric pulsations. This information can be deduced from a symbolic representation of music.

Rhythm recognition of an acoustic signal transforms the input samples into a list of individual acoustic events. Then exist the assignation of this values of notes duration

of this events. It has to be noted that grouping notes in units is not an easy work, because human interpretation is not perfectly precise, also because the music notation is ambiguous. It means that identical rhythms or really similar rhythms can be written as different ways. Like tone-detection, rhythm recognition has to ignore the insignificant variation due to find the essential rhythm.

In the next section we are going to analyze the Beat Tracking, so the way to get and recognize the beat of each signal data.

Chapter 3

Beat Tracking

The main propose of this section is to be able to obtain the BPM, also known as Beats Per Minute, of a determine sound signal. The proposed system should be able to estimate the tempo as well as locate and/or identify the events[1], which will help us to recognize the true existence of an event's occurrence. For more details concerning the Beat Tracking please refer to Galin, A. (2015).

3.1 State of Art

Beat Tracking has been studied in many fields of science from a variety of approaches. For instance, its application has been widely used in the field of medicine - to estimate the cardiac pulse from an echocardiogram (ECG). Hence, the acoustic domain of Beat Tracking has also a relevant significance for onset detection of the given input signal provided.

Further, there are many ways of obtaining and identifying the resent events (beats) of a signal. In this section, the most important publications from different influential authors will be summarized and compared.

To begin, M. Goto (2001) has been able to estimate the events (beats) by finding the local Maximum Power. In order to be able to do so, he initiated his research with vectors, where each one was working within a different frequency bands. So, the author

[1] An event coincides with a single pulsation of the input signal.

applied the following bankfilter ranges: 0–125 Hz, 125–250 Hz, 250–500 Hz, 0.5–1 kHz, 1– 2 kHz, 2–4kHz, and 4–11 kHz. He then worked with each vector band separately. The onset was to be detected only in the case of achieving an increased power level of each and every vector of the seven bands, which form the bankfilter.

In addition, F. Wu et al.'s (2011) research confirmed the existence of a relationship between the signal's power and its relative position within the unit compass. In other words, the authors acknowledged that the first beat of each compass will be the predominant one in terms of its power. Simultaneously, the third beat will be slightly less dominant than the first beat, yet still carrying quiet the impact. And finally, the second and the fourth beat will be much less dominant. Moreover, the authors also utilized the methodology mentioned above by Goto (2001) of increasing energy. To do so, F. Wu et al. (2011) used the derivate function in order to compute these increases. In the case of a decrease detection, a value of 0 is assigned, in order to avoid having negative values. It is important to bear in mind that these authors assumed a model of a BPM changing over time; further implying it not necessary being constant throughout time.

At the same time J. Laroche (2003) in his article assumed the different power levels of various beats in a compass. This author also introduced the calculation of fluctuation of energy, related to the increase of the power level presented by the previous authors. His estimation used the Fast Fourier Transform (FFT). In this particular case, J. Laroche did not utilize bankfilters, but rather, he applied a window into the signal in order to eliminate the high frequencies presented. The author obtained the energy flow, as captured in the following formula:

$$\widehat{E}(i) = \sum G(|X(f, t_i)|) - G(|X(f, t_{i-1})|) \tag{3.1}$$

$$E(i) = \widehat{E}(i) if > 0; 0 = Otherwise \tag{3.2}$$

J. Laroche (2003) also affirmed that the event detected at one point of time does not need to be constant, since the rhythm of a song itself is not necessary consistent throughout the entire song. Furthermore, he admited that some music styles make it harder to identify the pulsation of the rhythm (beats). With respect to the work of Wu et al

(2011), Laroche further developed the idea by identifying the relationship between the power of the four beats of the compass to different music style - rock, pop, latin, etc (refer to the following graph .....).

Table 1. Average relative amplitudes of downbeat, first quarter-beat, half-beat, and third quarter-beat for various genres, in the energy flux signal $E(R, t)$.

Genre	0	1/4	1/2	3/4
Techno	1.0	0.1	0.6	0.4
Pop	1.0	0.1	0.8	0.2
R&b	1.0	0.2	0.5	0.2
Rock	1.0	0.2	0.7	0.2
Reggae	1.0	0.3	1.1	0.5
Latin	1.0	1.0	0.9	1.5

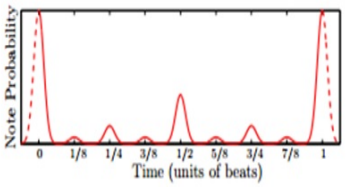
Table1: Relationship between the power of the four beats of the compass to different music styles. J. Laroche (2003)

J. Laroche also developed an algorithm, which adapts the BPM to the changes of song's rhythm during its duration. However, such beat changes are not very common in the modern music as stated by the author.

On the contrary, based on the results of J. Laroche, which stated that beat changes are not common in modern music, V. Panchwadkar et al (2013) assumed the BPM to be constant over time. For the purposes of their research, they used a Kalman Filter. In their case study, they also worked with the bankfilter, much like that of Goto (2001). Unlike Gotos bankfilter, they have chosen to work with only three filter ranges which are the following: 0Hz-200, 200Hz - 5000Hz, 5000Hz and above. For each events estimation, an Onset Detector was used for all of the three bands. Further, the power variation was measured, allowing for an estimation of the periodicity of the high peaks. Finally, Kalmans filter applied a weight to each of the three output ranges and estimated the tempo.

As observed in the work of Goto (2001),an algorithm was created, such that was able to obtain the instantaneous BPM, in order to detect the tempo of a song in all of its instances, considering the possibility of a tempo change (however uncommon may they be according to J. Laroche). However, he later further elaborated on the idea of a possibility of a non-changing tempo and considered the tempo to be constant throughout time.

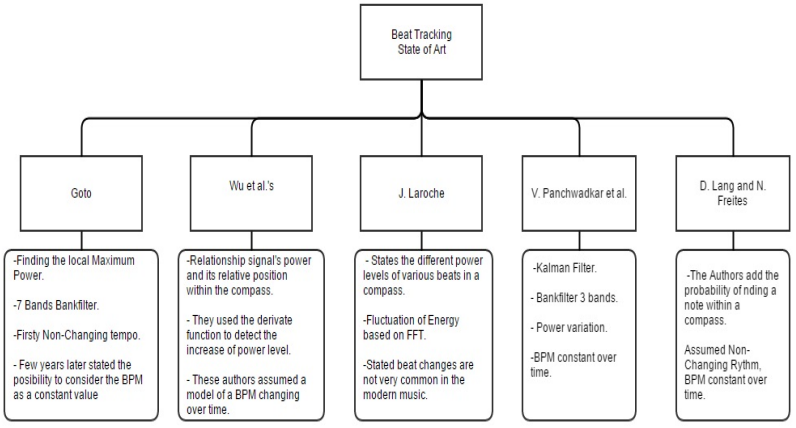
Finally, D. Lang and N. Freites also considered a constant BPM in the course of the song. In addition to J. Laroches associated relationship between the power of the four beats of the compass to different music style, Land and Freites (2012) add the probability of finding a note within a compass. These probabilities may be found in the following figure:



Gaphic1: Probability of finding a note within a compass. Land and Freites (2012)

As observed above, these results were coherent to those of J. Laroche, which stated that the beat with the most power was, in fact, the first beat (0 and 1). Then the third beat (1/2) being of a lower relevance, followed by the second (1/4) and fourth (3/4) which were found to be of least importance. Furthermore, the notation of the authors was normalized in the unit of compass. For this very reason, all of the values of the four beats were included in the range between 0 and 1.

The State of Art of the Beat Tracking process presented above can be summarized as follows:



Gaphic2: Beat Tracking State of Art

The purpose of this section was to shed some more light on the variety of theories and approaches to the problem of Beat Tracking. This section, hence, included the most influential authors in this field, such as Masataka Goto and Jean Laroche, which both

have published many articles in this particular field. Nevertheless, all of the authors - included as well as the non-included ones- have contributed to the development of the Beat Tracking field of study[2].

3.2 Proposed Solution

To be able to comprehend the functioning of the philosophy of the system as a whole, the understanding of rhythmic recognition process is called for.

The rhythm recognition process may be divided into the following steps:

1. Read the Input Signal and Compute Its Energy.
2. Pseudo-Energy and Beat Recognition.
3. Verification of the Results by Checking Other Selection Levels.

Hence, each of the above mentioned steps shall be analyzed in order to get a better grasp of understanding of the dependently-related proposed steps which will confirm the Beat Tracking process.

3.3 Read the Input Signal and Compute Its Energy

Beat Tracking is the first part of our algorithm necessary to the Automatic Music Transcription process. However, prior to proceeding to the Beat Tracking process itself, a few steps must be taken beforehand. Hence, prior to computing and identifying the beat of the input signal, we must first calculate the energy possessed by the input signal based on R. Peters(2003). The input signal will be the active source of data which we are interested to transcribe.

All of the above described processes of data capture are executed by the Teager Operator, which functioning is described in a greater detail in the section below. We have been based on Teager Energy instead of the Squared Energy by the fact of the Short-Term Energy Estimation stated by A. Potamios et al. (2009), for more information please refer to this publication.

[2]For further information refer to articles in the bibliography.

Once this data is analyzed and processed, we may then proceed to the capture of the input signal pulsation, into the Beat Tracking process itself.

3.3.1 Teager Operator

The Teager Operator (TO) is an operator used to estimate the energy that is required to generate an oscillation in amplitude as well as in frequency. So, it has to be noted that more energy is required to generate higher frequency signals (E. Kvedalen 2003). Harmonics simple movement says that the energy required to generate an oscillation is proportional to the power of its amplitude and its frequency. (E. Kvedalen 2003, M.J. Lipsey 2002 and R. Peters 2003)

Teager Energy (TE) is defined mathematically by the equation 4.1 displayed below. Further, it is important to take note that this equation expresses the energy of the real-valued analog signal x(t).

$$\Phi_c[x(t)] = \dot{x}^2(t) - x(t) \cdot \ddot{x}(t) \tag{3.3}$$

This operator and its discrete counterpart were introduced by Teager and systematically studied by Kaiser and others. For a pure sinusoid x(t)=Acos(wt), where A, w are constant reals, application of the operator results: (E. Kvedalen 2003 and R. Peters 2003)

$$\Phi_c[x(t)] = A^2 \cdot {\omega_o}^2 \tag{3.4}$$

The quantity is known as the Teager Energy of the signal; it is proportional to the energy required to generate the displacement x(t) in a mass-spring harmonic oscillator.

More generally, let's consider a real signal x(t) with joint amplitude and frequency modulation. If the AM and FM modulation indices a(t) and phi(t) are not too large, then the application of the operator results:

$$\Phi_c[x(t)] \approx a^2(t) \cdot \dot{\varphi}^2(t) \tag{3.5}$$

To add on, we will now use the previous definition of Teager Operator as it is described in the equation 4.1. However, this it it will be applied in discrete time, such that the

derivate in discrete time is applied. This equation is defined as follows:

$$\dot{x} = x[n+1] - x[n] \tag{3.6}$$

Therefore, once the first derivate is derived, the second derivate may be taken, which is captured in the following equation:

$$\ddot{x} = x[n] - x[n-1] - x[n-1] + x[n-2] = \tag{3.7}$$

$$x[n] - 2x[n-1] + x[n-2]; \tag{3.8}$$

Further, by applying the previous definition, we now obtain the equation in discrete time:

$$\dot{x}^2 = x[n] + x[n+1] - 2x[n]x[n-1]; \tag{3.9}$$

While at the same time:

$$x[n] \cdot \ddot{x}[n] = x^2[n] - x[n-1] \cdot x[n] - x[n-1] \cdot x[n] - x[n-2] \cdot x[n] \tag{3.10}$$

Substracting the both results following the first equation, 4.1, finally we obtain the discrete Teager Operator which is defined by:

$$\Phi_d = x^2[n-1] - x[n-2] \cdot x[n] \tag{3.11}$$

To sum up, the energy operator corresponds to the full energy of a system. The Schrödinger equation describes the space- and time-dependence of slow changing (non-relativistic) wave function of quantum systems. The solution of this equation for bound system is discrete (a set of permitted states, each characterized by an energy level) which results in the concept of quanta.

Computing our Teager Energy in discrete time obtained in 4.11, we can see:

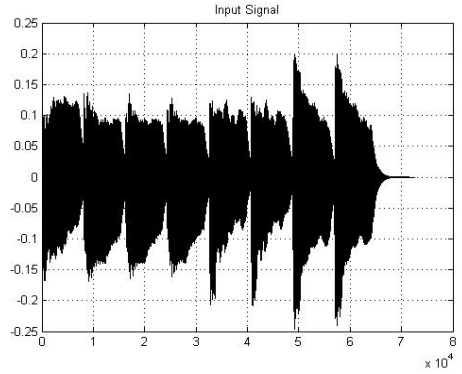


Figure 3.1: Input Signal

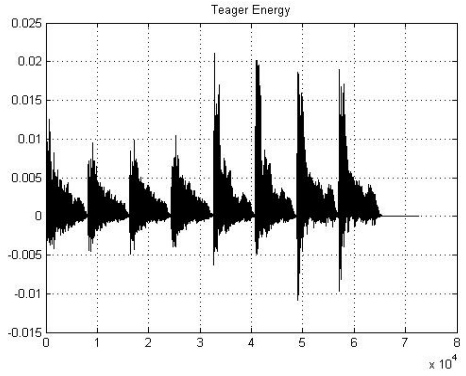


Figure 3.2: Discrete Teager Energy computed using 4.9

In the previous plots we may observe the differences between the original signal and its Energy, using the Teager Operator described before. The previous input signal corresponds to a piano playing a C scale.

3.4 Pseudo-Energy and Beat Recognition

This is the fundamental part of the Beat Tracking Algorithm. This part aims to identify each event[3] and calculate its periodicity to obtain an precise and accurate BPM value.

We will work using the energy auto-correlation signal, which proposes to identify its maximums over all its frequencies. The discrete auto-correlation of a given signal at the lag l is defined by:

$$R_{yy}(l) = \sum y(n)\bar{y}(n-l) \tag{3.12}$$

[3]Supra note 1

where:

$R_{yy}(l)$ corresponds to the calculated autocorrelation of y at the l lack.

y(n) corresponds to the Teager Energy of the input signal.

Thereafter, by applying the previous definition and plotting the autocorrelation of the original signal and the autocorrelation of its energy, we obtain the following:

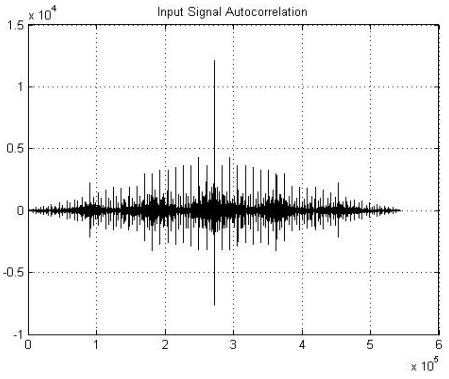


Figure 3.3: Autocorrelation using Input Signal

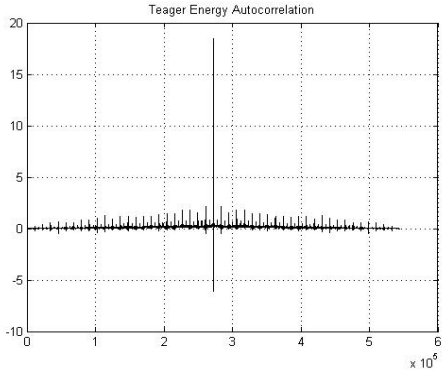


Figure 3.3: Autocorrelation using the Energy Signal computed by 4.9

As observed in the graphs above, it is apparent that working with the Teager Operator enables us to compress the original signal, making its changes in amplitude more abrupt. It will helps us to get an accurate value of BPM.

Moreover, the previous plots demonstrated the occurrence of autocorrelation of an offset due to the transformation itself by the Teager Operator. For this very reason the adaption of the previous signal using some commands is required, which will then allow for a that after will allow for a work with accuracy.

Zooming the Autocorrelated Energy Signal, it is noted the difference between the beats described by J. Laroche (2003) and F. Wu et al.'s (2011):

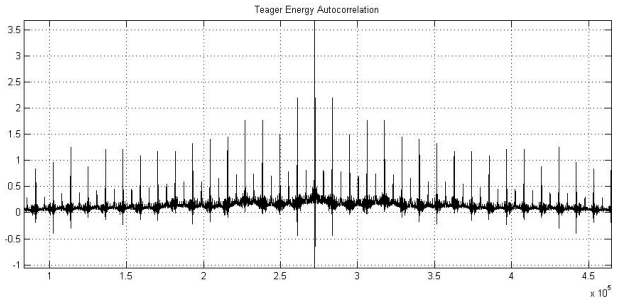


Figure 3.4: Zooming the figure 4.3

In order to correct for the previously received offset, the application of a derivate function to the autocorrelation function is called for. Also it shall be noted that the derivate function will provide a high pass filter, making the differential changes more abrupt. Furthermore, this application will also improve the computation by exaggerating its beats components.

The following is the discrete derivate function that shall be applied into the computation, which is described as:

$$\dot{x}[n] = x[n] - x[n-1] \tag{3.13}$$

By using an absolute value after applying the derivate function, we get the following signal:

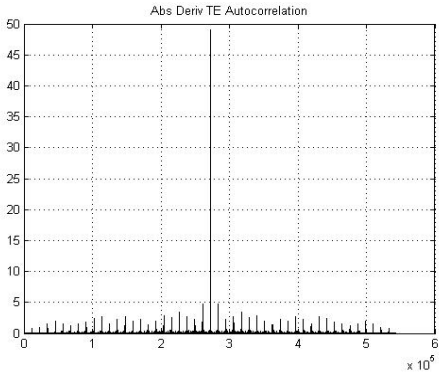


Figure 3.5: Derivate Energy Autocorrelation Function.

As may be observed above, the desired train pulses is obtained, where the objective is now to identify the distance between the individual peaks to compute the estimation of BPM. It has to be noted that there were too many ways to implement the blocking DC

signal, we used the derivate function because later it will be used in other parts of the Beat Tracking process and this way this element will be effective in terms of Hardware implementation.

After blocking the DC signal a Threshold is called for. However complicated, the sensibility value needs to be fixed, which will then depend on the following factors: the Energy of the signal, the length of the Signal as well as a K factor of Sensibility factor, which will determine how the sensibility is restricted. Hence, the sensibility of our algorithm can be described as following:

$$S = Sensibility \tag{3.14}$$

$$S = \frac{K}{2} \cdot \int \frac{\partial \Phi_c}{t \cdot \partial t} \tag{3.15}$$

In discrete domain:

$$S = \frac{K}{2 \cdot n} \cdot \sum \frac{\partial}{\partial t} \cdot \Phi_d \tag{3.16}$$

where:

K is the Sensibility Factor.

n is the number of samples of the analyzed signal.

Finally, the discrete Teager Energy calculated in 4.9 corresponds to Φ_d

First of all we will set the Sensibility Factor to K=14; which justification shall be clarified later on in the chapter. Once the K factor is set we will continue with the Binarization Process. This process aims to assign value of 1, if the absolute value of the autocorrelation of the TO signal is equal or higher that the Sensitive value S. However, if the signal amplitude is lower than the Sensitive value S, a value of 0 is assigned. After the Binarization Process, it is obtained:

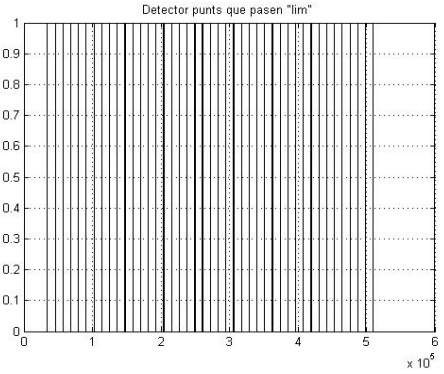


Figure 3.6: Binarized Signal.

Further, it should be noted, the obtained Binarized Signal also possesses properties of symmetry due to the proprieties of the Autocorrelation itself. This implies that in order to compute with the most efficiency, we should analyze only half size of the binarized signal without losing any component. This way, the new vector size is reduced by a factor of 2.

So it is obtained:

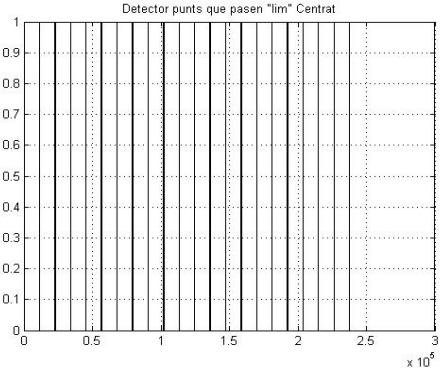


Figure 3.6: Binarized Signal. Reduced a length factor 2 from the Figure 4.6

Now the objective is to obtain the distance between the peaks, we shall bear in mind that an array obtained with a concatenated values of '1' followed by an array of '0'. This way, the comportment of the Binarized signal is modeled, also represented as Bs in the following equation:

$$B_s = [ones(k_1), zeros(q_1), ones(k_2), zeros(q_2), ..., ones(k_m), zeros(q_m)] \tag{3.17}$$

where k and q are variables that shall be estimated.

To solve the previous issue at hand, where we want to estimate the distance in number of samples between '1' and '1'. Further, the Binarized signal (Bs) is composed by the number of samples on the X axis and '1' or '0' on the Y axis, hence, the Binarized signal.

Using the discrete derivation described before in 4.4. Applying the derivate function to the Binarized Signal will convert the axis, where in Y axis, the number of samples between '1' and '1' is obtained by the proprieties of the derivation function itself. This way, on the y axis will be represented the distance between '1' and '1' in terms of samples.

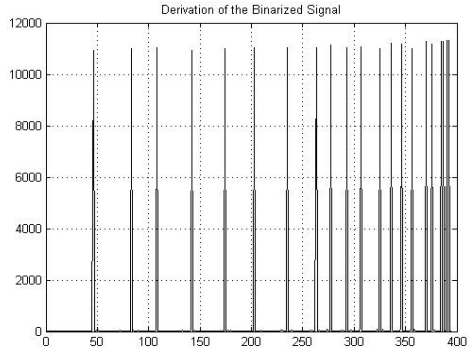


Figure 3.7: Derivation of the Binarized Signal.

Next, the PeakFinder function detects the peaks and obtains their amplitude. An intelligent algorithm has been implemented to discard the double amplitude values caused by the non-detection of some beats, as well as the detection of amplitudes with a high level of changing of the mean of all peaks.

At this time, the requirement of the Peak Mean Value is called for. PMV will consider the mean value of the peaks obtained in the previous figure. It has to be noted that the peak of the figure computes the difference between '1' on the function 4.15. PMV is defined as:

$$PMV_K = \frac{1}{N} \cdot \sum_{n=0}^{N} P_n \tag{3.18}$$

where:

P is the amplitude value of the n peak

N are the total detected peaks.

Finally, once the peaks are obtained and its mean computed; the procedure to the BPM shall be realized:

$$BPM_K = \frac{60 \cdot Fs}{5 \cdot PMV} \tag{3.19}$$

where:

5 is the factor due to the signal adaptation at the beginning of the process.

Fs is the Sampling Frequency.

PMV was calculated in 4.16.

As observed during this section, the BPM value will depend on the Sensibility Factor [K].

3.5 Verification of the Results by Checking Other Selection Levels

This section is concerned with the last part of the Beat Tracking, which consists of verification of the obtained value of BPM.

This verification is based on the computation of different receivers by changing their amplitude sensibility seen at 4.14. If the sensibility is big enough, we may not be able to detect some peaks and will; henceforth, miss the presence of some of the beats. On the other hand, when amplitude sensibility is too small, we will may falsely detect a non-existing event, caused by the relative noise of the signal energy responsible for this false detection.

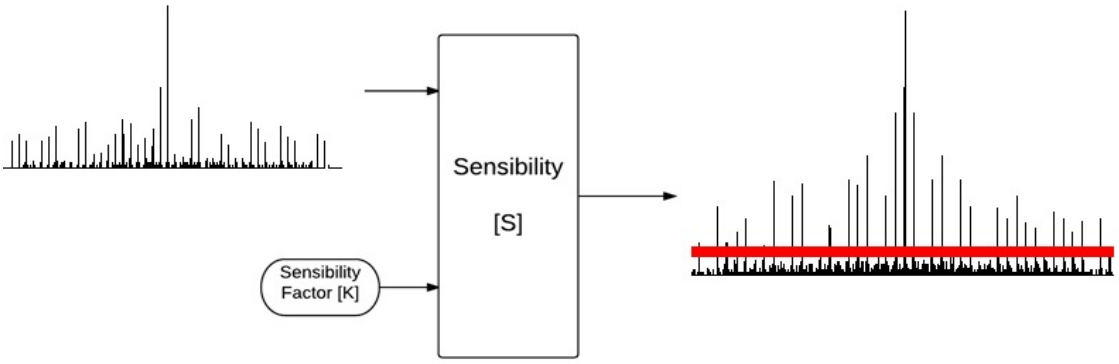


Figure 3.9: It is needed a Sensibility Factor [K] in order to compute the Sensibility [S] in red.

The verification process will consist of checking of all of the ranges of sensibility and computing the corresponding BPM value for each single sensibility S.

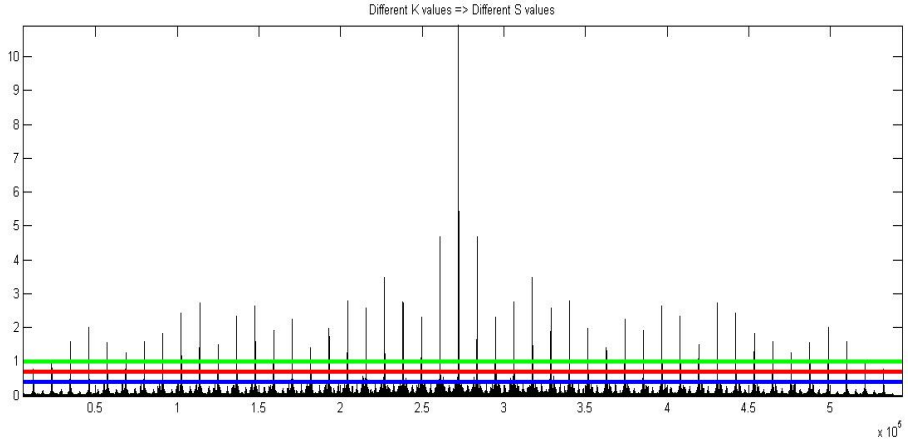


Figure 3.9: In the figure above, it is observed different values of K, where K and Sensibility S are related by 4.14. The blue Sensibility [S] corresponds to a value of K=8; the red S is provided by a K=14; and finally, the green sensibility is computed by a K=20.

Color	Blue	Red	Green
Sensibility Factor [K]	8	14	20

The verification process will consist on the calculation of the BPM by using different K values from:

$$K = 8 : 1 : 20; \tag{3.20}$$

As it is noted when K is less than 8, the Sensibility [S] it is small so we will get all the imperfections of the auto-correlation signal, and not an accurate value of BPM will be obtained. However if K is higher than 20, the probability of missing part of important beats will increase considerably; and also, not an accurate value of BPM will be obtained.

Respect the thirteen values of BPM obtained, first of all and inspection will be done. This inspection consist of an analysis of outliers; and then, once the outliers has been removed the median function will take place in order to get the final BPM value.

Sensitibity Factor [K]	8	9	10	11	12	13
BPM_K	112.7977	98.3826	97.1276	96.7813	98.9806	97.6654

[K]	14	15	16	17	18	19	20
BPM_K	97.5484	96.9941	96.8255	96.8255	96.8255	96.8255	96.8255

As noted, this process will conclude with the selection of the final BPM among all of the BPM values previously obtained for each sensibility value.

Using the median function to the previous array we finally get the following value:

$$BPM = 96.9941 \tag{3.21}$$

3.5.1 Error Balance

Is the previous result acceptable? Which is the error that can be acceptable? This question will be analyzed in this subsection.

In case the algorithm presents an error of difference between the real BPM and the computed BPM of 0.2, Beat Deplacement and Compass Deplacement definition takes place following the next equation:

$$T_{BeatDeplacement} = \frac{1}{\epsilon} \tag{3.22}$$

where:

T is the time to get a deplacement of a beat.

The error between the computed BPM and the real BPM is noted as ϵ.

In the previous case exposed, asuming a BPM error of +- 0.2 it means it will be a displacement of a beat each 5 minutes. And assuming that the distribution of the beats are based in the occidental distribution where a compass is formed by 4 beats, the deplacement of a compass will occur following the next equation:

$$T_{CompassDeplacement} = 4 \cdot T_{BeatDeplacement} \tag{3.23}$$

In addition, with an error of 0.2 BPM a Beat Deplacement will occur every 5 minutes and a Compass Deplacement every 20 minutes. It is quite rare to get an input signal longer than 20 minutes, so the algorithm will point always on the right compass. If the error is less than 0.2 the values of Deplacement will following the equations 4.22 and 4.23.

3.6 Chapter's Concepts Revision

To conclude, a brief overview of the chapter's main concepts will be analyzed in this section and put into a perspective; which will allow for a full understanding of the process' complexity at hand. Furthermore, such global analysis, such as this one, shall be a part of the following chapter as well. Resuming from all of the information from the previous sections of this chapter - input signal reading and energy computing, pseudo-energy and beat recognition, together with the results verification by checking other selection levels.- the following block diagram of the fully-detailed Beat tracking process is represented:

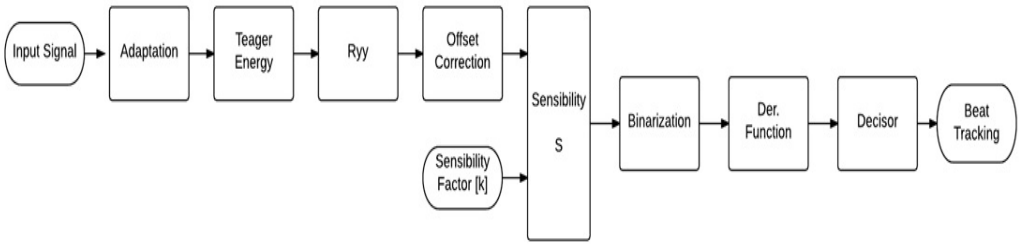


Figure 3.10: Beat Tracking Block Diagram

Firstly, as observed in the figure above, the process begins with an input digital signal; which in order to proceed with its analysis, should possess certain characteristics. Were these properties not possessed, the adaptation of the input signal will take place. This is followed by the block 'Teager Energy', which corresponds to the first part of the chapter. Next, the blocks including autocorrelation, derivate function, together with sensitivity computing were explained in the second section of this chapter. binarization, derivate function as well as decisor correspond to the final stage of the process of the Beat Tracking as addresses in the previous section.

In the last picture of this Chapter the Beat Tracking will be visualized at the same plot as Teager Energy of the input Signal, a high synchronization must be observed, the black lines represent the event detection from Beat Tracking, and the gray signal correspond to the Teager Energy of the Input Signal. As seen in the previous sub-section, assuming a determinated value of error balance it will takes place when time analysis is growing considerable. So, as seen considering an error value of 0.2 BPM it will produce a Beat Deplacement (BD) every 5 minutes Signal. And a Compas Deplacement (CD) every 20 minutes assuming a 4/4 rhythm structure. Following 4.20 and 4.21.

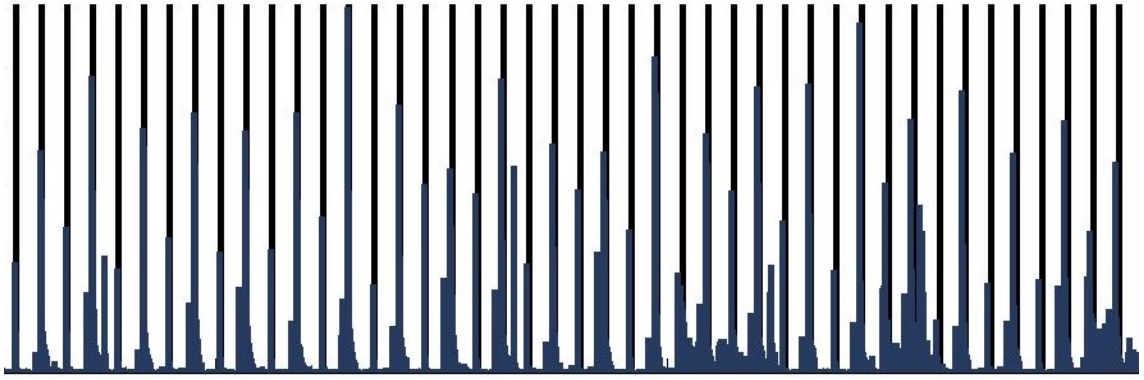


Figure 3.11: Beat Tracking vs Energy Signal

A deplacement will be produced when the event line - black line of the previous plot - has been compensated by one peak signal. It has to be noted that the black signal has been created by the BPM value, where it has been introduced the the Beat, accordingly to the BPM value obtained. On the first compass of the input signal, a synchronization has been released in order to align both signals. A compass time length will be four times the time beat assuming a 4/4 music structure, the most commonly western music structure.

As we have seen in this Chapter, the goal of Beat recognition has been assumed. In the following Chapter, the Note Acquisition will be analyzed.

Chapter 4

Note Acquisition

The Note Acquisition consists of detection of the presence of a determinate note or multiple notes as well as the estimation of its height for each instant of time of the input signal. Thus, this process is essential for the automatic transcription of music signals. For more details concerning the Note Acquisition block please refer t0 Galin, A. (2015).

4.1 State of Art

The concept of the note acquisition is closely related to the instantaneous frequency of a given signal. Due to this relationship, we have been able to connect some of the previously exposed concepts of the Musical parameters' chapter to some of the current topics and concerns of today's research in this field. Such concerns shall be addressed in this upcoming section.

As has been previously pointed out in the musical parameters' chapter, Van der Pool (1946) has defined the term of an instantaneous frequency as the temporal derivate of a phase of a given signal. Van der Pool (1946) then defined instantaneous frequency as follows:

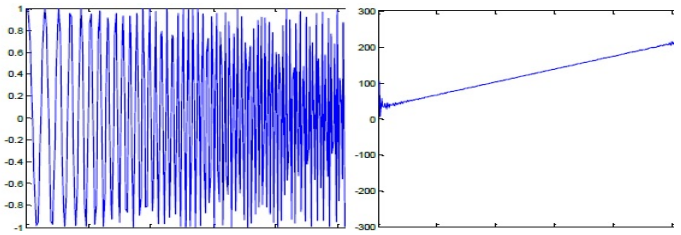
$$\omega_i(t) = \frac{d}{dt}[argz(t)] = \phi(t) \tag{4.1}$$

where:

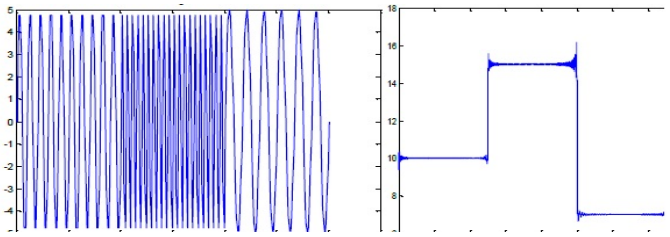
$$z(t) = s(t) + jH[s(t)] = a(t)e^{j\phi(t)} \tag{4.2}$$

Van der Pool (1946) has been able to obtain the above displayed results based on the publication of Gabor (1946) and Ville (1938).

For the purposes of our project, we have implemented the definition of Van der Pool in Matlab software, where we have utilized two examples. Firstly, using a single signal representing the constant change of the frequency during its duration as a chirp. Secondly, we chose a signal with frequency containing abrupt changes.



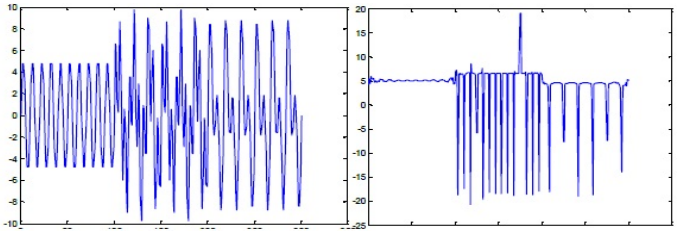
Gaphic4.1: Chirp Instantaneus Frequency Van der Pool (1946). Constant frequency change.



Gaphic4.2: Sin Instantaneus Frequency Van der Pool (1946). Abrubt change.

We can observe how the Van der Pool's (1946) definition is adequately adapted and is further capable of computing its instantaneous frequency in a very efficient manner in terms of computation[1].

However, this definition is valid only for the mono-component signals. In the case of multi-component signals, this definition is not able to provided the expected values. We may observe in the next example - Graphic 5.3 - the addition of two pure tones (sinus signals).



[1]For further information, please refer to Van der Pool's (1946)

Gaphic4.3: Sum of Sin signals. Instantaneus Frequency Van der Pool (1946).

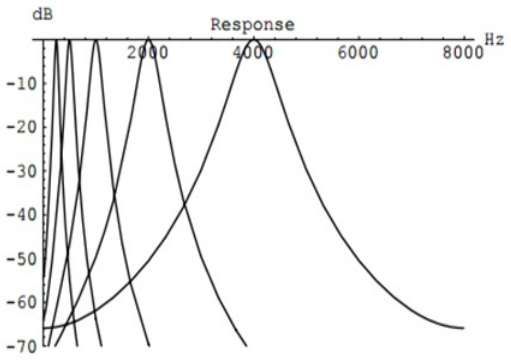
But as the above process of obtaining instantaneous frequency of such signal has proven to be flawed, other methods have been developed to approach this issue.

These methods contain the pseudo-instantaneous concept, meaning that our results will not be strictly instantaneous, but rather, will provide different frequency components. Such variables of the Van der Pool definition, may be grouped in two categories:

1) Gammatone filter

Firstly, Gammatone filter is one of the most commonly used models of the auditory system. It is described by an impulse response corresponding to a product of a gamma distribution and a pure tone. It is a linear filter based upon the audio impulsion response of R. Patterson (1992).

In the following graph, the differences of frequencies' resolution between low and high frequencies are being displayed. It may be observed that in low frequencies, great resolutions are obtained. On the other hand, high frequencies are not presented in as high of resolutions. All of these characteristics are simulate those of the human audio system[2].



Graph1: Gammatone Resolution. Extracted from Patterson(1992)

2) Short-time fourier transform

Short-time fourier transform (STFT) is another technique of analyzing instantaneous frequencies. In this field L. Cohen together with its many publications has became one of the most prominent authors (i.e. : L. Cohen (1995)).

[2]For further information, please refer to Patterson (1992)

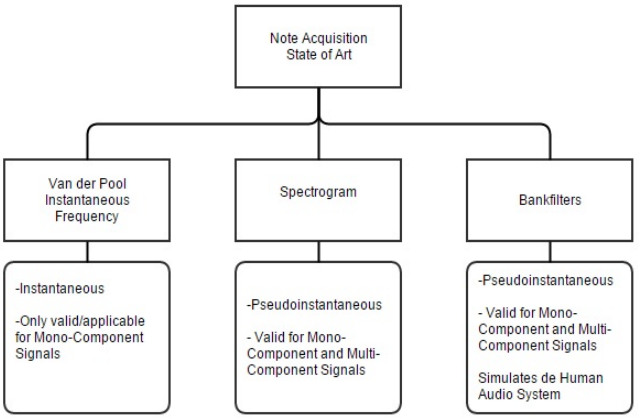
Apart from the basis of this concept laid down by the L. Cohen in his publications, new variations of this technology of analyzing instantaneous frequency have been developed over time. For instance, one of such technologies is Spectrogram, which has become of the most widely used tools in this field. (F. Katz 2013) The mathematic definition of this technique has been defined in the chapter of Music Parameters.

Further, this technique too, may be found in a variety of different forms and formats, such as concentrated Spectrogram (K. Czarnecki 2012) Another variant is the Instantaneous Frequency Estimation using aggregate Spectrogram. (2012).

In the same fashion, as previously addressed in the Beat Tracking state of art section (Laroche 2003), the author has been able to obtain the mean probability to find a note depending on its position in the compass. The authors mentioned above are considered to be some of the most influential authors in the field of instantaneous frequency computation in low frequencies as audio signals.

As observed above, the three most important typologies to obtain the instantaneous frequency of a given signal has been analyzed. Van der Pool (1949) proposed a way based on the derivation of the analytical signal with the limitation of a single tone detection. Due to this limitation new methodologies has been developed; Patterson (1992) worked using a bankfilter simulating the Auditive Human System, at the same time Cohen (1995) stated a huge publication with new methodologies all based on frequency domain resulting the spectrogram the most influential. The last two authors omit the instantaneous restriction to get all of the frequency multi-component conforming a given signal.

The State of Art of the Note Acquisition process presented above can be summarized as follows:



Gaphic2: Note Acquisition State of Art

4.2 Proposed Solution

Notes recognition of an acoustic signal transform the input samples into a single matrix of individual acoustic events. Moreover, each individual note is assigned a specific value, which is dependent on its height. May it be noted, the notes are grouped by different units, which determination of each particular grouping is not always deterministic to begin with. That is so due to the rather limited ability of humans' interpretation of music, which lacks precision. Further, the music notation itself is considered to be ambiguous.

In this chapter, the Note Acquisition is analyzed in order to explain the methodology of how each of the notes of the input signal are being recognized. Moreover, in addition to the recognition of height of each note, its location and duration must be assigned.

The structure of the chapter of Note Acquisition will be the following:

1. Input Signal Adaptation
2. Instantaneous Spectrum
3. Estimation of Instantaneous Notes

4.3 Input Signal Adaptation

The adaptation of the input signal is necessary for the proper functioning of the algorithm. Its function in the process consists of normalization of the sampling frequency of the input signal notes as 'fs', which is normalized to the value of 44100 Hz; which is considered optimal. The reason for this is that choosing any value lower than the above, would cause for a lack of precision in the computation. Whereas, apart from the value of 44100 Hz upwards, the higher the value selected, the higher the size of the matrices with the data contained; and hence, the lower the efficiency to analyze such data.

After imposing a fs value of:

$$F_s = 44100Hz \tag{4.3}$$

A decimate function will take place in order to reduce the computation of the process; as will be noted in the next subsections, however resolution of high frequencies will be lost. Applying a decimate factor of 5, we were able to reduce significantly the length of the input signal by a factor 5. So finally, an effective frequency sampling will be introduced, after the adaption to the signal to 44100Hz and then its decimation as following:

$$F_{S_{effective}} = \frac{44100}{5} = 8820Hz \tag{4.4}$$

It has to be noted that decimate factor can be changed, or even removed; depending on the designer specifications.

4.4 Instantaneous Spectrum

After the process of normalization is completed; we may continue with our work, for the fs of the input signal is now equalized to 44100 Hz. Afterwards, the analysis of the input signal is realized. In order to capture the Note Acquisition, the conversion from the time domain into the frequency domain is required.

One of the requisites for the Note Acquisition is the conversion using an instantaneous algorithm that is able compute the input signal in time domain into the frequency

domain. This may be accomplished by many various methods available[3]; however, the program chosen by us is the Spectrogram. This is due to its high quality performance, which functions uniformly and suits all of the signals' characteristics available.

Furthermore, the Spectrogram implemented shall not have a window with an associated resolution smaller than the difference of the semi-tone distance, in order to be able to detect it with a high enough precision. Similarly, were the window to be greater, the associated resolution would too, increase; however, the duration of the calculation would pro-long to the point, where the instantaneous factor would vanish completely. Henceforth, one shall find the proper trade-off between the instantaneous aspect and the frequency resolution.

The Instantaneous Spectrum algorithm used as mentioned above is the Spectrogram, defined by:

$$T_{SPEC}(t, f) = |\int x(\tau) \cdot g(\tau - t) \cdot e^{-j2\pi f\tau} \partial\tau|^2 \tag{4.5}$$

where:

T is the associated spectrogram of the input signal at the time window g

x is the input signal.

g is the temporal window filter. Hamming filter has been used.

In the next subsections different kinds of Tone Detection cases will be analyzed. The purpose of such analysis lies with exhibiting the variety of different properties each case possesses as well as analyzing its characteristics.

4.4.1 One Tone Detection

Firstly, prior to analyzing any more complex music piece as the input signal, we have chosen to capture the tones of a music scale. For this, an arbitrary scale was chosen, played by the piano. To begin with the analysis, the detection of a single tone will take place. For this very reason, a piano playing a C scale was chosen to be the Input Signal.

This analysis is for the demonstrated in the Spectogram below.

[3]*For all of the descriptions of the methodologies available, please refer to the following publication: Galin (2015)

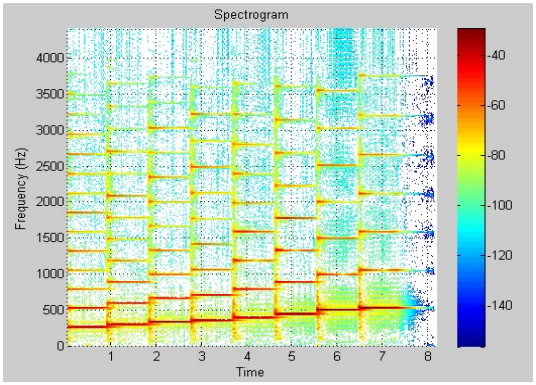


Figure 5.1: Pseudo-Instant of the previous signal.

Where during a pseudo-instant as noted above spectrogram is computed with 514 samples. Analyzing some determinate pseudo-instant:

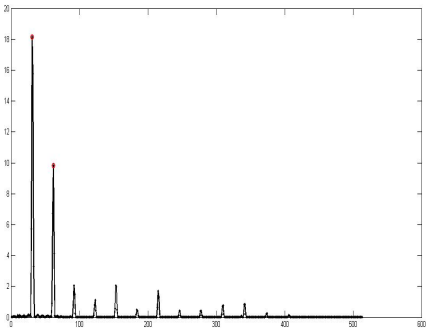


Figure 4.4: Spectrogram of a piano playing a C Scale. It has to be noted the presence of Harmonics tones of Fo as seen in the Chapter 3. In this case exist only one note playing and its Harmonics. The second Harmonic must be removed later to identify the existing notes.

4.4.2 Two Tones Detection

Building on the fundamentals laid down by the previous section, once that was achieved with accuracy, the Bach's famouse piece, the "Prelude", was analyzed in order to capture its notes in this part taking place. For this very reason the Bach Prelude will be analyzed as the Input Signal.

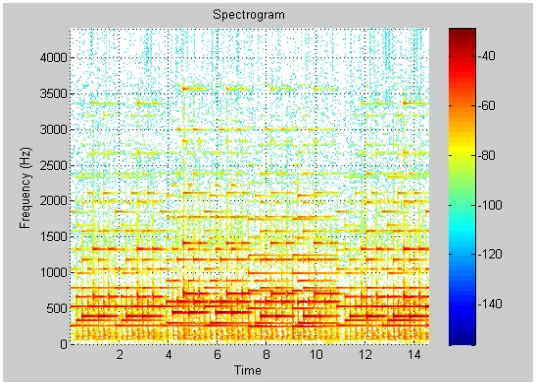


Figure 4.5: Spectrogram of the Bach Prelude

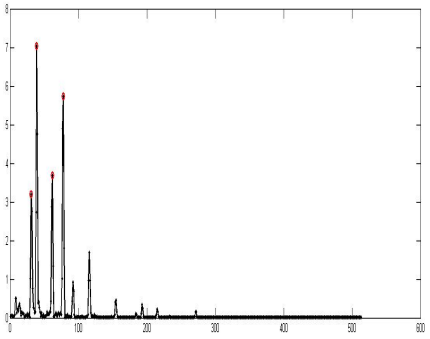


Figure 5.4: Pseudo-Instant of Bach Prelude.

As we may observe in the first figure of this section, this graph corresponds to the Spectrogram of the Bach's Prelude. The second graph is the Pseudo-Instant representation of the previous graph. In this case, one shall take note of the existence of two tones being played simultaneously at the same time with its Harmonics.

4.4.3 Multiple Tones Detection

Finally, once these two pieces of input signals were analyzed correctly, in process of which our algorithm we had a chance to improve, the results provided evidence which lead us to believe that the algorithm was capable of analyzing tones within a more complex musical structure. Such structure may be found in the more modern music - i.e. songs of the Beatles which include different instruments as well as voice signals.

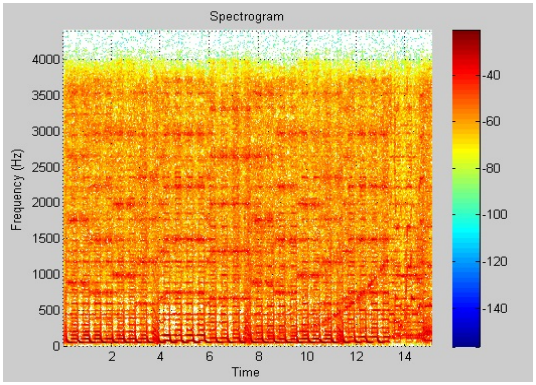


Figure 4.6: Pseudo-Instant of the Beatles - 'Twist and Shout' signal.

As we may observe in the first figure of this section, this graph corresponds to the Spectrogram of the Beatles' Twist and shout. The present of multiple notes has become in this case more eminent than in the previosu simulations. The second graph is the Pseudo-Instant representation of the previous graph.

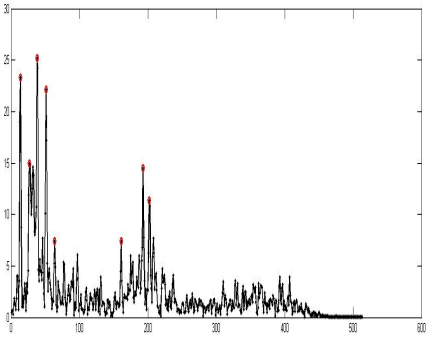


Figure 4.7: Pseudo-Instant of the previous signal. In this case exist different tones at the same time with its Harmonics.

In this case it has to be noted the existence of many tones at the same time with its Harmonics present. In this case there are exited low frequencies as well as high frequencies.

After running the algorithm in these three separate sessions, discrepancies among the different characteristics of each were observed. Henceforth, the variations of characteristics of the input signals are of a vital importance to our algorithm, which condition the different results attained.

4.4.4 Specifications

In this subsection a review of some of the Note Acquisition Characteristics will be analyzed in order to quantify the performance of the process. First of all, as pointed out in 5.2, the 8820 Hz will be the Effective Sampling Frequency (ESF) which corresponds to the signal rate firstly after its adaptation to 44100 Hz and consecutively decimated by a factor 5 as seen in the first part of this Chapter.

Concerning the filter used in order to compute the Spectrogram it has been used a Temporal Hamming Window Filter, which its frequency response presents high enough selectivity[4]; in terms of the Hamming filter it is composed by length l samples, so Hamming Filter and its temporal window length will be defined as:

$$g_{Hamming} = 0.54 - 0.46 \cdot cos\left(\frac{2\pi n}{l_{Hamming} - 1}\right) \tag{4.6}$$

$$T_{Hamming} = \frac{l_{Hamming}}{EFS} \tag{4.7}$$

Where 4.7 is the definition of the Hamming Filter and 4.5 corresponds to its time window.

It has been chosen a length l of 883 samples in order to get a Hamming Window of 0.1 seconds following 5.5. It has to be noted the existing symmetry between both sides of Hamming filter, where will present 441 samples side-to-side and where the center sample will be the time reference marker.

In the next figure the temporal window can be observed.

[4]*Comparing different kind of filters with its spectrum, please refer to the following publication: Galin (2015)

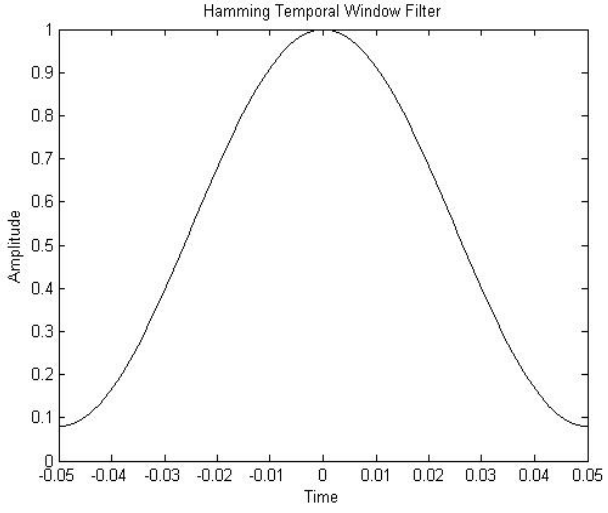


Figure 4.8: Hamming, Temporal Window Filter

Once the filter has been designed, the application of the Spectrogram will take place following 5.3; as seen in the previous sections, the spectrogram will be conformed by 514 samples which will provide us the sufficient values to recognize with accuracy and precision in terms of detecting the semi-tone difference, following the Heisenberg Principle and Musical Parameters introduced in Chapter 3.

The Spectrogram computation will be periodically performed accordingly to the input signal presence. At this point the Time-Detection Sensibility (TDS) must be introduced. This parameter will allow to obtain the time resolution of the Onset Note, described in Chapter 3. It means how fast can be our algorithm at least to detect the Activation Note. TDS will be defined as:

$$TDS = \frac{\vartheta}{EFS} \tag{4.8}$$

where:

The numerator represents the number of spectrogram computations per unit of input samples. This value must be an integer represented as ϑ.

EFS is the Effective Sampling Frequency from 5.2.

We have arbitrary chosen a value for the numerator and set it to be 9. With this value, the Time-Detection Sensibility of 0.001 seconds is obtained. This implies that our algorithm will compute the Spectrogram using a time window of 883 samples every 9 samples of the numerator. This also further signifies that the process will be capable of detecting each Activation Note, with at least 0.001 seconds of delay, or in other words, we may not be able to detect notes with a duration shorter than that of 1ms.

However arbitrary the nature of the value of 9 may be, it does have its reasoning behind it as for why it has been chosen from the variety of options. Henceforth, its explanation will be addressed, as without it, its justification may come across as too presumptions. Firstly, we do not recommend to set a TDS to be a low value. The reasoning behind this is that the low value would entail for the hardware to work at a high computational rate in order to to compute the spectrum. One shall take a note of the fact that with this configuration can present a TDS value of 1/8820 at its maximum capacity.

Furthermore, low value implies that the spectrogram will be computed every sample of the input signal, calling for a High Hardware System to be needed. Further, yet at the same time, the result may be the same as if we have done significantly fewer calculations with a higher value, as the calculations observed in the case of a lower value would become repetitive, causing for a loss of efficiency.

This is so due to the Notes Activation value being quite small. More, Hamming Filter would have to be working whilst containing too many samples.

In the next figure, it is observed the functionally basis of this process, essential part of the Note Acquisition. With the following Configuration Settings as:

EFS = 8820 Hz

$\mathrm{l}_{spect} = 514 samples$

$\mathrm{l}_{ham} = 883 samples$

$\mathrm{t}_{ham} = 0.1s$

mu = 9

TDS = 1ms

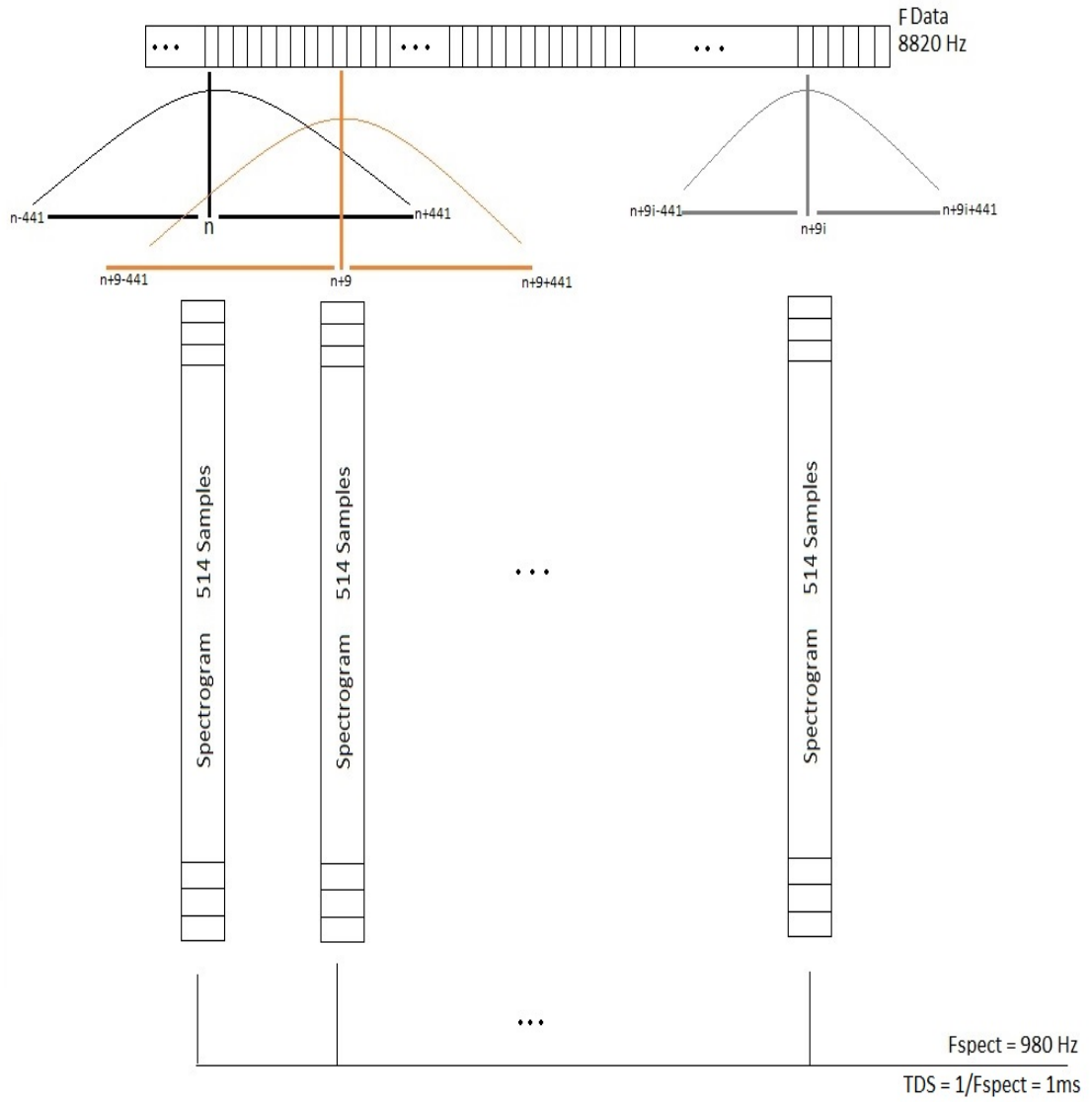


Figure 4.9: Functional

As observed in the figure above, the schematic of this section is visualized. It has to be noted that length of Hamming Window to avoid blind spots of the Input signal must not be less than ϑ.

The marker as seen during this subsection indicates the higher value of the time window filter and corresponds to the central point.

4.5 Instantaneous Notes Estimation

Once the input signal is adapted in terms of fs as well as converted into the frequency domain, one is able to proceed to the Note Acquisition process itself by the means of analysis of the previous instantaneous Spectrogram obtained. May it be noted that the use of the Spectrogram also considers the instantaneous factor of calculation itself as an abstract concept, for in its purest form it is non-existent. Hence, where referred to as 'instantaneous', the pseudo- instantaneous concept is implied.

Afterwards, each of the entries recorded and properly computed is then analyzed accordingly. In order to do so; however, we must attend to each of the entries individually, where each of them consists of 514 samples, as it has been pointed out in the previous section of the chapter. Once accessed and the screening of Spectrogram completed, the peak detector will be applied. The main objective of the peak detector is hence, to compute the position of each of the peaks of for all of the pseudo-instant.

Thereafter, the resulting matrix for each pseudo-instant shall contain the exact position of all of the peaks. In this particular case, our algorithm has been set to detect up to 5 peaks per pseudo-instant. As represented in the second chapter, the association between f0 of the height of the quasi-harmonic source and the position of the peaks has been deterministic. One must assure oneself that any of the frequencies detected does not correspond to a false multiple of f0 from the nature of a quasi-harmonic source.

Once the peaks are obtained as well as non-false-double-frequencies have been detected and removed where necessary; we may proceed to the conversion position of the peak to the corresponding Note. For this very reason, the codification of every single Note is called for. To begin with, the codification of each Note is composed out of two corresponding parameters - firstly, the obtained note codified on a scale 1-12 corresponding to A-G; including the semi-tones; and the secondly, the scale of 1-9 in height. For instance, the algorithm for codifying C of scale of 5 coincides with the codification value of 25.

Latin Notation	English Notation	Codification
DO	C	1x
DO #	C#	2x
RE	D	3x
RE #	D#	4x
MI	E	5x
FA	F	6x
FA #	F#	7x
SOL	G	8x
SOL #	G#	9x
LA	A	10x
LA #	A#	11x
SI	B	12x

In the figure 4.10, the codification procedure may be observed where x correspond to the scale value. This parameter must be an integer.

Furthermore, each scale will have its very own frequency, with which it will be associated. Upon such categorization of scales, a table of all the different frequencies is constructed and displayed. Moreover, each note of music will have its own frequency attached. Resulting will be a relationship between each scale with each note as a multiplication of frequencies. In the next table the association of Frequency and Notes corresponding to the Scale 4 must be observed.

English Notation	C4	Cs4	D4	Ds4	E4	F4	Fs4	G4	Gs4	A4	As4	B4
Frequency [Hz]	261	277	293	311	329	349	369	392	415	440	466	493

After completing and following all of the above mentioned steps, we are able to compute all of the notes that correspond to each pseudo-instant. With all of the pseudo-instant notes' data captured, we are able to construct the Note Acquisition process.

4.6 Chapter's Concepts Revision

In the following section, a brief summary of this chapter, as a part of our analysis, will be made. Further, it is with the intention to put into retrospect all of the essential aspects of the Note Acquisition algorithm in order to provide a more of an aggregate image of the process itself.

After a thrall analysis of the Note Acquisition process' components - adaptation, instantaneous Spectrum as well as instantaneous note estimation- the summary of the process in retrospect is added; in order to point out the exact flow of processes in the computation.

The following is the block diagram of the fully-detailed Note Acquisition process:

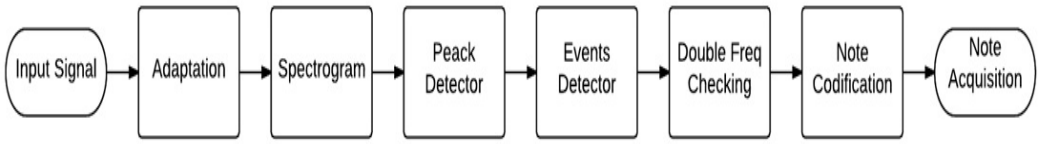


Figure 4.11: Note Acquisition Block Diagram

Firstly, as observed in the figure above, the process begins with an input digital signal; which in order to proceed with its analysis, should possess certain characteristics. Were these properties not possessed, the adaptation of the input signal will take place.

This is followed by the block 'Spectrogram', which corresponds to the second part of the chapter. After calculating the Spectrogram, the relative Peaks need to be obtained, in order to detect the sheer existence of a musical event. Once the occurrence of the event is confirmed, taking into consideration that each peak is assigned its value in form of a note, ex-ante to the detection process; these peaks of the input signal detected are now noted adequately.

After obtaining the associated notes of the detected peaks, we must now make sure the multiple frequencies were not present in our analysis. This is essential, for if the disharmony were, in fact, present, it may pollute our results by misdetection of notes.

Finally, the previous blocks including Peak Detector, matrix events, double frequency checking together with note codification correspond to the last stage of the process of the Note Acquisition as addresses in the previous section.

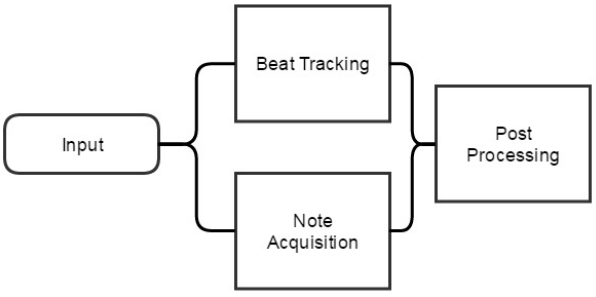
Chapter 5

Post Processing

In our solution, once the Beat Tracking and Note Acquisition processes takes place, there are required some processing techniques in order to increase the performance of the ATM system.

This post processing present three separate parts: the Decision Making, the MIDI Conversion and the Graphical Representation.

All this sub-parts are going to be discussed in the present section. The block diagram of the process is being displayed in the following figure:



Gaphic5.1: Post Processing Schematic

In terms of a clock analysis, one can say that the Beat Tracking and Note Acquisition blocks are working continuously by means of a short-time iterations - since this two blocs are working regularly for every input of data-, the Post Processing block provides the low frequency response in the order of time magnitude to be compared with the duration of the notes, where it is treated the data from the previous blocks.

5.1 Decision Making

The Decision Making part of the Post Processing block is of essential in order to determine the presence of a given note, and in fact, here is where the system confirms the presence of a given event.

This Post-Processing block is composed by several sub-blocks which are depicted in the following chart.

Gaphic5.2: Post Processing Block Schematic

Now a brief overview of each sub-block will be performed and its purpose will be highlighted.

First of all the Burst Selection block aims to join the detected events from the Beat Tracking and Note Acquisition. This detected events are unified in this block creating notes.

The Harmonics Detection performs again - it is also present in the Note Acquisition block, in terms of Double Frequency Search. - an accurate search of harmonic issues but in this case, this the harmonic detection is performed over an overview of the note, in this case the decision about the presence of an harmonic is not only performed in a single temporal analysis but in a high level extraction during the overall temporal presence of the note itself.

Therefore the Fall avoidance is a sub-block which takes into account the possibility of loosing the presence of a given note during some period of time due to a given reason -for example the saturation of the receiver, the microphone, by a sudden excitation of multi-frequencies, this block was designed thinking about real case AMT system, where one can find the presence of noise and other sources of signal at the input of the receiver -microphone-. This block aims to detect this effect, and reconstruct the detection even if the note has been lost.

Next the Onset Detection block takes place in order to identify and separate two consecutive notes of the same high. This block is very valuable when multiple instruments

play at the same time and in a given period of time they play the same note. A timbre analysis is not performed in this block. Also this block provides the unambiguity when sustain effect is present.

Then finalizing this Decision Making we design the Length Selection block, which makes sure that the selected notes are in accordance and no inconsistencies are found. At the moment we are not dealing in this block with an input signal harmonic analysis, we check that the notes are consistent, coherent and that is used as the final validation stage.

Once the detected notes have been validated by the systems - it has to be noted that this task must be in an efficient way, we cannot lose notes, otherwise the indicators will decrease the performance of the AMT system as seen in Chapter 7.- we can proceed to the treatment of such information, in the following subsection we are going to deal with this topic.

5.2 MIDI Conversion

The idea of this section is once the detected notes have been validated, to represent in a standard way the output of the ATM System. There exist several standard methods for doing so, we have been motivated for using MIDI standard due to its interoperability[1] . The idea is to build the ATM system in an interoperable platform, this way the use of an standard is required, therefore the way of codifying each detected note must be in accordance with such standard. For this very reason we decided to convert to MIDI format.

The MIDI specification can be represented like the following equation:

$$MIDI(i,j) = [T(j), Ch(j), N(j), V(j), S(j), E(j)]; \tag{5.1}$$

where:

T Corresponds to Track Number of the j Note.

Ch Corresponds to Channel Number of the j Note.

[1] In Chapter 4 there has been presented and designed a Note codification scheme

N Corresponds to j Note Number codified from 1 to 150.

V Corresponds to the volume the j Note

S Corresponds to the start of onset-time in seconds from the Note j.

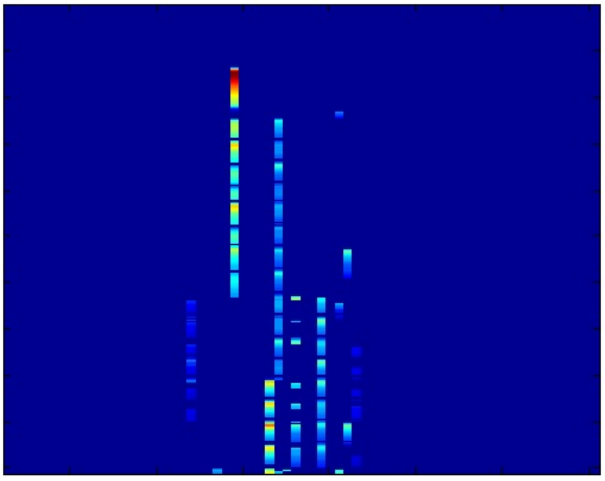
E Corresponds to the end-time of the Note j in seconds.

Therefore the objective of this block is to convert the validated data from the Decision Making block to a MIDI file.

5.2.1 Piano Roll

In this subsection we are going to introduce the Piano Roll. The Piano Roll is a way of representing a given acoustic signal in terms of high and duration.

Once the Beat Tracking and Note Acquisition processes takes place, and the validation has been performed one is able to represent graphically the obtained results. The Piano Roll is one common way to represent it. The Y-axis of the Piano Roll represent the temporal line while the X-axis is related to the Note. The axis can be ex-changed. There exist a close relationship about the MIDI format and the piano roll. An example of a piano roll is shown below:



Gaphic5.3: Piano Roll Example extracted from MAPS Database ref. MAPS MUS-chpn-p4 AkPnBcht

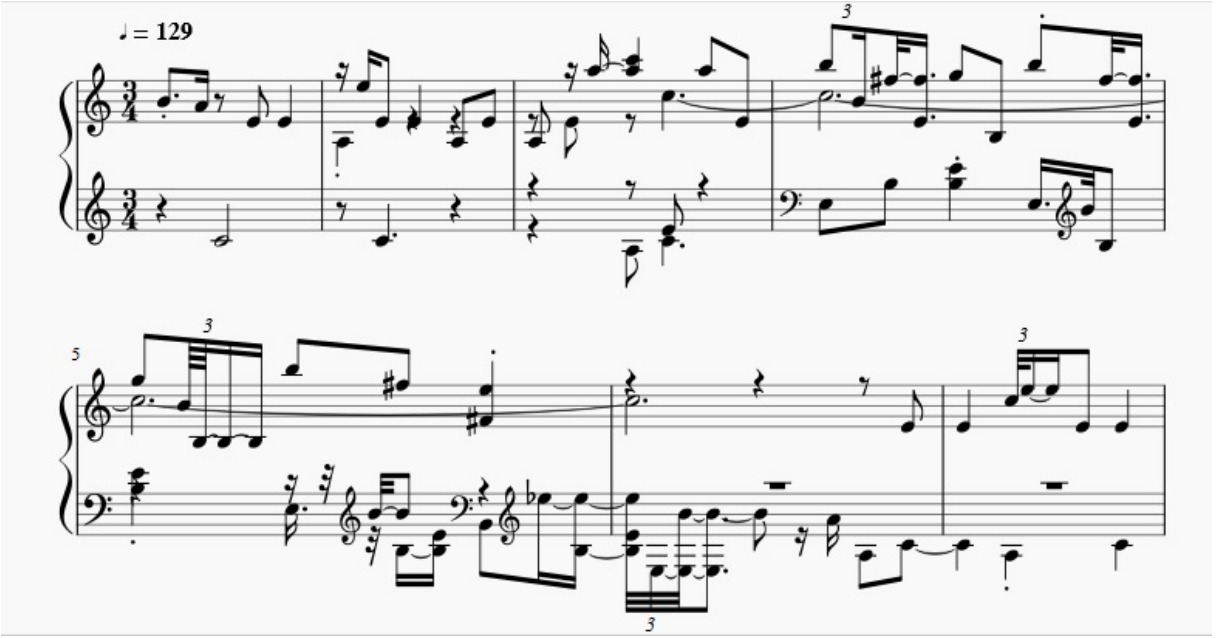
We can find several ways of representing a Piano Roll. For example in the Graphic 5.1 above the temporal intensity is being considered. It means that is being represented the strength of each note during its duration.

Another way of representing the Piano Roll is the binarized model, where it is only displayed whether one note is present or not. In fact, this type of information is the basis in order to provide the musical representation.

5.3 Graphical Representation

The graphical representation is of essential in order to obtain the output of the AMT system. The graphical representation process aims to represent the obtained results from the previous blocs in a way which can be understandable, usually by means of a pentagram.

Once the MIDI file has been generated, it is only a matter of exporting this file and open it in a platform compatible with MIDI format. The idea as mentioned in this Chapter was to use a standard method in order to have easily solved the Graphical Representation block.



Gaphic5.3: Piano Score obtained after the Transcription Process.

Chapter 6

Process Overview

In this chapter we will make a brief overview of the functioning of the different blocks analyzed during the previous chapters: Beat Tracking, Note Acquisition, Post Processing together in order to obtain an accurate music transcription.

In order to be able to transcribe music, many different stages and different processes first needed to take place, in order to proceed to the Automatic Music Transcription. The three separate algorithms had to be created - Beat tracking, Note Acquisition and Post Processing, for the process of transcription to function correctly.

6.1 Beat Tracking

Beat tracking is the designated algorithm, with the sole purpose of determining the tempo (BPM) and the occurrence/ non-occurrence of events in the signal analyzed. Therefore, the system is now able to recognize different rhythms, which is accomplished by different processes, which are the following:

A. Read the Input Signal and Compute Its Energy.

B. Pseudo-Energy and Beat Recognition.

C. Verification of the Results by Checking Other Selection Levels.

And with each of these processes of the Beat Tracking being in check, half of the Automatic Music Transcription is done.

6.2 Note Acquisition

Note Acquisition is the other algorithm, which compliments the Bet Tracking, determining note or multiple notes as well as the estimation of its height for each instant of time of the input signal. All of the data captured by the notes' recognition process of the input samples is then transformed into a single matrix of individual acoustic events with an assigned a special value, dependent on its height and grouped by different units. Moreover, in addition to the recognition of height of each note, each note's location and duration is assigned. Note Acquisition process has been divided in three steps:

A. Input Signal Adaptation

B. Instantaneous Spectrum

C. Estimation of Instantaneous Notes

6.3 Post Processing

The Post Processing process is the brain of the ATM system, where according to the inputs from the Beat Tracking and Note Acquisition the system is able to identify or delete a given event. In addition to this decision making, the Post Processing stage also converts the data into a MIDI file in order to be displayed in a graphical way. Post Processing has been divided in the following subsets:

A. Decision Making

B. MIDI Conversion

C. Graphical Representation

6.4 Assemble

Once we have been able to perfect each of the above mentioned parameters, we were able to proceed with transcription. Finally, the full algorithm block diagram will be as following:

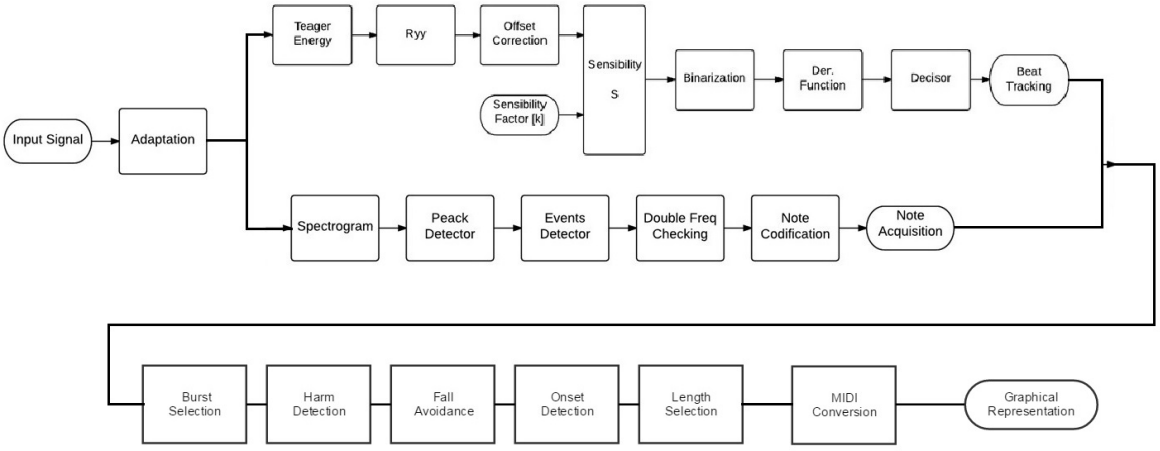


Figure 6.1: Entire Automatic Music Transcription process'

In the graph displayed above, the entire Automatic Music Transcription process' functioning, where all process' - Beat Tracking, Note Acquisition and Post Processing - algorithms work simultaneously.

After all of the above mentioned calculations are complete, a result is obtained. However, in order to further verify our result and compare it to the original input signal, the calibration and calculation over a given Ground Truth is of essential in order to test the performance of the solution. Provided that our results coincides with the original input signal, we may now conclude that our algorithm has worked correctly and has been able to transcribe music automatically.

Chapter 7

State of Art Comparison

In the present chapter we are going to analyze how well is our ATM solution behaving related to the State of the Art proposals.

7.1 Evaluation Methods

Evaluating an AMT system means an evaluation of how well it performs; and furthermore, the need of collecting evidence is present by the fact that it may be applied by an end user. The first AMT systems - Moorer, J. A (1977) and Piszczalski, M et al (1977) - were evaluated by visual comparison of ground-truth, and they automatically obtained scores from the ATM system. This practice remained for some time during the following years, but with the growth of scientific efforts towards AMT, it became necessary to use objective performance measures that could not only be automatically calculated over large databases, but also be immediately applied for the performance comparison of different transcribers.

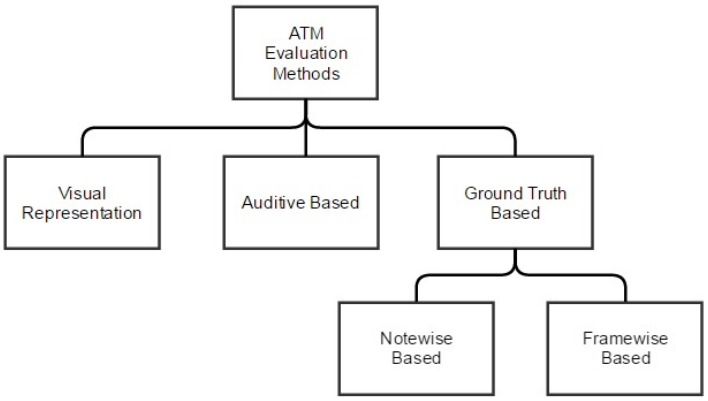


Figure 7.1: ATM Evaluation Methods.

By analyzing the State of the Art, there exist several ways of measuring how well or how bad is the ATM model, but most of the times this measuring tool is referred to a Ground-Truth. The Ground-Truth (GT) provides an absolute reference of the input signal, The behaving of the ATM system must be compared to the proper GT according the input signal.

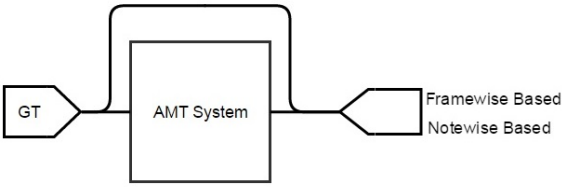


Figure 7.2: Schematic of Evaluation Methods GT Based.

Extracted from the literature there exist two different ways to compare the ATM system with the Ground Truth, the Framewise based and the Notewise based models.

The Framewise based model aims to compare every bin of time obtained from the ATM system - mainly defined by the Spectrogram - extract such detected notes and compare with the GT. The magnitude of the Framewise based model is adimentional and is given by percentage, where the computation takes into account the ratio between the properly transcribed Frames - extracted from the GT - compared to the total present frames. One can say that the Framewise based model is a relationship between the detected correctly notes over the length of the present notes -extracted from GT-.

On the other hand, the Notewise based model considers the detection of the notes itself. A Notewise based model will consider the onset, the high and the duration of the notes, if all this three parts are within the tolerance then the note has been properly transcribed[1]. In this approach it need to define the following parameters:

$$Recall = \frac{\#ofcorrectlytranscribednotes}{\#ofnotesinGround - Truth} \tag{7.1}$$

$$Precision = \frac{\#ofcorrectlytranscribednotes}{\#ofnotesinAutomaticTranscription} \tag{7.2}$$

$$F - measure = 2\frac{RecallPrecision}{Recall + Precision} \tag{7.3}$$

The Notewise Based model will be given by the value of F-measure. It has to be noted that by having a highly sensitive ATM model, the number of notes will increase, so the number of false positive will also increase and F-measure will be reduced. The F-measure indicates the compromise between the system.

It has to be noted that the Notewise Based model must consider a given temporal tolerance in order to consider whether a given note has been properly transcribed or not. Several authors use different tolerance in order to obtain F measure under the Notewise Based model.

7.2 Common Dataset

The use of a common dataset is of essential in order to compare different ATM systems proposals. In fact some of ATM systems are highly sensible to the topology and proprieties of the input signal, behaving well only in a given subsets of inputs - for instance some ATM solution may not be performing well when several notes takes place simultaneously. Others may not perform well when a height range of frequencies -height and low notes - are present in a given input signal. In a response to that, there exist the need of testing the ATM Solution in several environments.

[1]Some authors only considers the onset and the high in order to consider weather or not a note has been properly transcribed.

For the presented very reason we have been working with the MAPS dataset, the MAPS - standing for MIDI Aligned Piano Sounds - provides high-quality recordings of a Yamaha Disklavier, that is, an automatic piano. It contains 40 GBytes of recordings with corresponding Ground-Truths (GT), and may be freely downloaded. It has to be noted that the MAPS dataset has been commonly used in several publications including: Dessien, A et al (2010), Nam, J et al (2011) and O'Hanlon, K et al (2012).

The most common music category that is dealt with in AMT research is the piano solo in line with the selected MAPS dataset. The first systems for that task were designed in the 1990s- by Martin, KD (1996), Privosnik, M et al (1998) and Keren, R et al (1998).

Later, more sophisticated techniques were proposed - Abdallah, SA. et al (2004), Barbancho, I. et al (2004), Bertin, N. et al (2009), Boogart, C. et al (2009). Raphael, C. (2002), Guibin, Z. et al (2007).

One may think about the use of a piano only instrument; the piano is a polyphonic instrument, that is more than one note can be played at the same time. Also, there is no direct contact between the musician and the vibrating string that produces the sound - by pressing a key, the musician triggers a mechanical interface-, hence a predictable behavior may be expected from the spectral envelope related to each note. Moreover, since it is assumed that there is only one instrument in the acoustic signal, all detected notes should be assigned to it, making timbre analysis unnecessary.

The MAPS dataset provides several environments where one can find several inputs in order to test the ATM system in different environments which will be discussed later on in the present section.

Besides MAPS there exist other few datasets which are currently available, highlighted the Opolko, F et al (1987), Goto, M et al (2003). They are usually made up of isolated tones from various musical instruments and/or musical recordings. Then, when necessary isolated notes may be added in order to generate chords.

The mentioned above databases provide a large quantity of sounds and were generally obtained after considerable efforts.

Concerning MAPS, it provides recordings with CD quality (16-bit, 44kHz sapled stereo audio) and the related aligned MIDI files and Ground Truth (GT). The overall size of the database is about 40GB, i.e 65 hours of audio recordings.

This large amount of sounds and reliable ground truth has been generated through an automatic generation process, the use of a Disklavier (MIDI piano) and of high quality sysnthesis software based on libraries of samples permitted a satisfying trade of between the quality of the sounds and the time consumption needed to produce such a quantity of annotate sounds.

In order to favor generalization to many audio scenes, several grand pianos and upright pianos have been played in various recording conditions, including various rooms and close/ambient has been taken into account.

The content of MAPS dataset is divided in four sets, which are detailed below:

ISOL set: isolated notes and monophonic excerpts. Thus aims at testing single-pitch estimation algorithms or at training multipich algorithms when isolated tones are required.

RAND set: Chords with random pitch notes. It provides chords composed of randomly-chosen notes. It was designed in order to evaluate the algorithms in an objective way, without any a priori musical knowledge which is commonly performed in the literature on multipitch estimation.

UCHO set: Usual chords from Western Music. Thus, these chords are useful to assess the performances with an a priory knowledge and are made with notes that are harmonically related.

MUS set: pieces of piano music. These high quality files have been carefully hand-written in order to obtain a kind of musical interpretation as a MIDI file.The note location, duration and loudness have thus been adjusted by hand by the creator of the database. About 238 pieces of classical and traditional music were actually available when MAPS dataset was created.

7.3 Figures

In the present section we are going to analyze the performance of our AMT system compare with the state of the art. In the following table the Notewise model is expressed in percentage which corresponds to the output of the F-measure, and the Framewise

model is also expressed in percentage. The technique being used for each model as well as the publication and the database used is also included in the following table.

Notewise	Framewise	Technique	Publication	Database
-	85	SVM with Memory	Constantini, G et al (2009)	Poliner-Elis
-	79	DBN	Nam, J et al (2011)	Poliner-Elis
78.2	76.3	Sparse NMF decomposition	O'Hanlon, K. et al (2012)	MAPS Database
-	74.4	DBM	Nam, J et al (2011)	MAPS Database
71.6	**64.1**	**Present Document**	**Galin, A. et al (2017)**	**MAPS Database**
71.5	65.5	NMF with beta-divergence	Dessein, A. et all (2010)	MAPS Database
-	70	SVM	Poliner, GE. et al (2007)	Poliner-Elis
-	63.6	MLP network	Marolt, M. (2004)	MAPS Database
-	46	HMM and specialist signal processing	Ryynamen, M. et al (2005)	Poliner-Elis
-	39	MLP network	Marolt, M (2004)	Poliner-Elis
Notewise	Framewise	Technique	Publication	Database

Unreported results are identified with a dash. It has to be noted that this results show the performance of a given AMT solution in a given input data according to that we must take this values as a maximum, the performance is expected to decrease when the input is not optimized for the given model.

Concerning the above results, the notewise model, as mentioned before in this chapter needs a tolerance parameter, in order to identify whether a note has been well transcribed or not. This tolerance in the literature usually is of the order of 20ms. Concerning the framewise model, since it is not always taking into account the notes which are not good transcribed, its results could not necessary mean a good automatic transcription.

It has to be mentioned that the presented method in this dissertation still has a lot of gap for improvement by means of tunning some of the sub-parts in order to optimize the system according to the input signal. We have ensure the good performance of the AMT system for a several scenarios which provides the conviction of performing a good automatic transcription in a several environments.

7.4 Comparison Proposal

The idea behind this section is to propose a common method of inter-comparison between the different AMT solution providers.

Up to date, as noticed during this document, several authors have calculated their AMT system by means of an F-measure or framewise model which was extracted from a subset of input signals: for example during the first thirty seconds of a given MAPS database song, in a subset of environments by means of maximum number of notes played at the same time, or by means of given number/type of instruments.

The MAPS dataset provides a huge amount of different piano solo music environments. The idea behind this section is to state the need of testing under the same conditions several AMT systems.

Since every author is testing its solution according their AMT solution in this section we propose a method of evaluating by considering a multi-variable process. Being i the number of different Scenarios (S) and being n the number of notes of each scenario. We propose the following condition:

$$Notes\left\langle S_i\right\rangle = Notes\left\langle S_{i+1}\right\rangle = n \tag{7.4}$$

So, according the previous equation, the total number of evaluated notes will be:

$$TotalNotes = \sum_{i=0}^{i=S-1} Notes\left\langle S_i\right\rangle = Sn \tag{7.5}$$

Once defined the note length of the S scenarios, one can compute the F measure:

$$F - measure = 2\frac{RecallPrecision}{Recall + Precision} \tag{7.6}$$

So, according the previous definitions:

$$F - measure = 2\frac{\frac{\#ofcorrectlytranscribednotes}{\#ofnotesinGround-Truth} \cdot \frac{\#ofcorrectlytranscribednotes}{\#ofnotesinAutomaticTranscription}}{\frac{\#ofcorrectlytranscribednotes}{\#ofnotesinGround-Truth} + \frac{\#ofcorrectlytranscribednotes}{\#ofnotesinAutomaticTranscription}} \tag{7.7}$$

being Sn the number of notes in Ground-Truth. So:

$$F - measure = 2\frac{\frac{\#ofcorrectlytranscribednotes}{Sn} \cdot \frac{\#ofcorrectlytranscribednotes}{\#ofnotesinAutomaticTranscription}}{\frac{\#ofcorrectlytranscribednotes}{Sn} + \frac{\#ofcorrectlytranscribednotes}{\#ofnotesinAutomaticTranscription}} \tag{7.8}$$

In other words, we have stated the need of testing every ATM system under the same conditions, but as well under a range of different Scenarios like mono-component, multi-component, different notes at the same time, etc. All of them being equal in weight, so in number of notes in a Notewise based model.

Then the number of correctly transcribed notes and the number of AMT transcribed notes will be dependent on the solution provider. The important point behind that is that everyone are going to test its AMT under the same scenarios, the same number of notes per scenario, then the comparison between solutions is absolute. It has to be noted that since the number of notes are equal in the different Scenarios and not the temporal length of each S, then the model must be Notewise Based.

Chapter 8

Conclusion

The main objective of this thesis was to create a way for a music transcription to be realized without the necessity of a human mediator. Hence, our goal was to create an algorithm, which would be able to analyze and consequently transcribe music accurately. This transcription of the input audio signal heard would be the visual representation of this input signal, including the detained information about its main parameters, such as height of the played notes, rhythm, tempo, time signature and other, less conventional, such as tone, agreements and the structure of a song. Furthermore, the beginnings and ends of notes – the activations, as well as the height of each of the note of the signal, was to be also transcribed.

Once the above analysis has been approved, the development of the computational aspects themselves has started. This process has consisted of three different, separate aspects - Beat Tracking, Note Acquisition and Post Processing. All of which are mutually dependent and are essential for the music transcription process. The way these three function simultaneously has been better explained in the previous chapter.

Further, in the course of this thesis we have also looked at other methods of approaching part of the problemology involved in our project - i.e. capture of the beat. The methods included were those, of the most renowned authors in their designated fields. We have also took an interest in the evolution of each particular problem together with its changing resolution over time. Hence, some issues may be resolved differently with the ever-advancing technology.

In many instances we had to take decisions about the complexity of our solutions when facing a given problem. We have always favored those, which have been considered the most practical. Moreover, the efficiency was not the only determinant for choosing that particular option. The solutions applicability was also being taken into consideration.

During our work, we have encountered many issues and problems along the way, which; however, have turned out to be beneficial in the end. For, every flaw that has occurred, has helped; for it has pointed out the eminent possibility for a further improvement. And this applies to the current version project as it is. This project may always be improved upon and many more aspects of detection and music transcription may be added. However, we hope to believe that the algorithm in its current state as it is being presented, will serve as a firm basis for further development, continuation and improvement of our work.

8.1 Metrics

In order to conclude the work exposed in the present document a brief list of the capabilities of the proposed AMT system is depicted:

- As mentioned in Chapter 6, the proposed AMT solution presents an average of 71.6 percentage of Notewise base accuracy and a 64.1 percent concerning Framewise based model. It has to be noted that this values were obtained by means of different pieces of MAPS dataset, so we are not referring to maximal values. There still exist a big gap of improvement in terms of accuracy.
- Regarding the maximum notes that can be transcribed at the same time, up to date within the present configuration the ATM system does support a transcription of up to ten notes simultaneously. This maximal number of notes can be increased by several means.
- The ATM system does not consider any a-priory information. This makes the system able to work with an unlimited range of input signals.
- The sampling rate of the microphone is set to 44.1kHz. The effective ATM system processing clock is of 8820Hz.

- The system provides 980 spectrograms per second. Concerning the length of the spectrogram it is fixed to 514 samples. A Hamming filter and other techniques are used in order to increase the frequency accuracy.
- The ATM system can considerably increase its performance by means of adding more microphone chains. The system does support multiple inputs.
- The BPM of the final transcription will be between 60 and 240.
- The Range of the Transcribed Notes are from C-1 which corresponds to 8.177Hz to G7 which corresponds to 3324Hz.
- The minimum note duration of the transcription will be of 0.1s, so shorter notes of 0.1s are not going to be detected nor transcribed.
- It is supported the transcription in a noisy environment, where the acoustic saturation is present. This block is still under development.
- Up to now the proposed ATM system does not include timbre identification.

8.2 Future Work

In this section we are going to mention the next steps to be addressed in the future, as the implementation in a stand alone dedicated hardware the proposed AMT solution.

Also a next line to be addressed is a topic related with the timbre identification which will enhance considerably the scope of the proposed model. Another point which is currently being addressed is the capability of the algorithm on working in a real environment, taking into account the presence of a given noise source into the input signal, by doing the system strong against noisy environment also will increase the applications behind the system itself.

Another important work plan to be considered is the fact of detecting more notes at the same time - up to now this value is set to ten - for a sort of applications it is required to increase the maximum number of detected notes at a given time period.

Concerning the inter-comparison between other AMT systems, as stated in this document, it is required a common way of quantize the performance of a given AMT Solution.

The proposal presented in Chapter 6 needs to be defined and applied by the different Authors in order to provide absolute results.

Bibliography

[**1**] AAMIR, M. ZAMIN, A. ALI, M (2007). "*Instantaneous Frequency Estimation Using Aggregate Spectrogram*" Volume 4681, 2007, pp 484-493. Accessed 08.04.2015. Available from:
http://link.springer.com/chapter/

[**2**] ABDALLAH, SA. PLUMBLEY, MD (2004) "*Polyphonic music transcription by non-negative sparse coding of power spectra.*".Proceedings 5th international conference music information retrieval (ISMIR'04), Barcelona, Spain. Accessed 26.09.2016.

[**3**] AUGER, F. FLANDRIN, P. (1995). "*Improving the readibility of time-frequency and time-scale representations by the reassignment method*". IEEE Transactions on Signal Processing. Accessed 11.12.2014

[**4**] BARBANCHO, I. BARBANCHO A. JURADO, A. TARDON, L. (2004) "*Transcription of piano recordings.*".Appl Acoust 65(12):1261–1287. doi:10.1016/j.apacoust.2004.05.007. Accessed 28.09.2016. Available from:
http://www.sciencedirect.com/science/article/B6V1S-4D7CDP7-2/

[**5**] BERANEK, L. (1954). "*Acoustics*". McGraw Hill. Accessed 04.02.2015

[**6**] BERTIN, N. BADEAU, R. VINCENT, E (2009) "*Fast Bayesian nmf algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription.* ".IEEE workshop on applications of signal processing to audio and acoustics, WASPAA '09, pp 29–32. doi:10.1109/ASPAA.2009.5346531. Accessed 28.09.2016.

[**7**] BOOGART, C. LIENHART, R. (2009) "*Note onset detection for the transcription of polyphonic piano music.*"Proceedings of the IEEE international conference on multimedia and expo, ICME 2009, pp 446–449. doi:10.1109/ICME.2009.5202530 Accessed 28.09.2016.

[**8**] COHEN,L (1993). "*Instantaneous "anything"*". In IEEE Int. Conf. Acoust, Speech and Signal Proc. Accessed 10.10.2014

[**9**] COHEN, L (1994)." *The frequency and scale content of biological signals*". In Records of Twenty-Eighth Asilomar Conf. on Signals, Systems and Computers. Accessed 14.10.2014

[**10**] CONSTANTINI, G. TODISCO, M. PERFETTI, R.(2009) "*On the use of memory for detecting musical notes in polyphonic piano music.*" In: Proceedings of the European conference on circuit theory and design, ECCTD 2009, pp 806–809. doi:10.1109/ECCTD.2009.5275106, Accessed 10.01.2017.

[**11**] CZARNECKI, K (2012). "*Concentrated Spectrogram of audio acoustic signals - acomparative study*". Accessed 29.12.2014. Available from: https://hal.archives-ouvertes.fr/hal-00810604/document

[**12**] DESSEIN, A. CONT, A. LEMAITRE, G (2010) "*Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence.*". Proceedings of the 11th international Society for Music Information Retrieval conference (ISMIR 2010), Utrecht, Netherlands. Accessed 23.10.2016.

[**13**] DEVILLE, Y. (2011) "*Signaux Temporels et Spatiotemporels - Analyse des Signaux, Théorie de l'Information, traitement d'Antenne, Séparation Aveugle de Sources*". Accessed 17.01.2015

[**14**] DOSSAL,C. PEYRÉ,G. FADILI,J.(2009). "*A numerical exploration of compressed sampling recovery*". In Proc. SPARS'09. Accessed 02.11.2014.

[**15**] EMIYA, V. BADEAU, R. DAVID, B (2010) "*Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle.*". IEEE Trans Audio Speech Lang Process 18(6):1643–1654 Accessed 02.09.2016.

[**16**] FERNANDES, T. GARCIA-ARNAL, J. ATTUX, R. LOPES, A (3) "*Survey on automatic transcription of music*". The Brazilian Computer Society 2013. J Braz Comput Soc (2013) 19:589–604 DOI 10.1007/s13173-013-0118-6. Accessed 25.08.2016.

[**17**] GALIN, A. (2015). "*Instantaneous Frequency Analysis*". Institute Supérieur de l'Aeronautique et de l'Espace.

[**18**] GALIN, A (2015) "*Automatic Music Transcription*". Bachelor Thesis at Universitat Autonoma de Barcelona. Under Supervision of: Dr. Prof. Antoni Morell.

[**19**] GOTO, M. (2001). "*An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds*" Journal of New Music Research, Vol. 30, No. 2, pp. 159–171. Accessed 22.05.2015. Available from:
https://staff.aist.go.jp/m.goto/PAPER/JNMR2001goto.pdf

[**20**] GOTO, M. HASHIGUCHI, H. NISHIMURA, T. OKA, R(2003) "*RWC music database: Music genre database and musical instruments sound database*" ISMIR. Baltimore, MD, USA, Accessed 28.09.2016.

[**21**] GUIBIN, Z. SHENG, L.(2007) "*Automatic transcription method for polyphonic music based on adaptive comb filter and neural network.*"Proceedings of the international conference on mechatronics and automation, ICMA 2007, pp 2592–2597. doi:10.1109/ICMA.2007.4303965. Accessed 28.09.2016.

[**22**] HEISENBERG, W. (1948). "*The Physical Principles of the Quantum Theory*". Kindle Edition. pp 114-135. Accessed 15.10.2014.

[**23**] KATZ, F (2013). "*Phonetics and Spectrograms: Putting Sounds on Paper*". Accessed 26.02.2015. Available from:
http://www.dummies.com/how-to/content/phonetics-and-spectrograms-putting-\discretionary{-}{}{}sounds-on-paper.html

[**24**] KEREN, R. ZEEVI, YY. CHAZAN, D. (1998) "*Multiresolution timefrequency analysis of polyphonic music.*". Proceedings of the IEEE-SP international symposium on time-frequency and timescale analysis, pp 565–568, Pittsburgh, PA, USA Accessed 26.09.2016.

[**25**] KOTZ, S. JOHNSON, N. (1992). "*Breakthroughs in Statistics. Springer Series in Statistics*". Accessed 09.01.2015.

[**26**] KVEDALEN, E. (2003). "*Signal processing using the Teager Energy Operator and other nonlinear operators*". University of Oslo, Department of Informatics. Accessed 13.10.2014. Available from:
http://folk.uio.no/eivindkv/ek-thesis-2003-05-12-final-2.pdf

[**27**] LANG, D. DE FREITAS, N (2012). "*Beat Tracking the Graphical Model Way*". Department of Computer Science University of British Columbia, Vancouver, BC. Accessed 13.04.2015. Available from:
http://papers.nips.cc/paper/2745-beat-tracking-the-graphical-model-way.pdf

[**28**] LAROCHE, L. (2003). "*Efficient Tempo and Beat Tracking in Audio Recordings*". J. Audio Eng. Soc., Vol. 51, No. 4. pp. 226-233. Accessed 15.06.2015. Available from:
http://www.ee.columbia.edu/~dpwe/papers/Laro03-beattrack.pdf

[**29**] LIPSEY, M. J. (2002). "*On the Teager-Kaiser Energy Operator 'Low Frequency Error'.*". University of Oklahoma. Accessed 19.10.2014. Available from:
http://hotnsour.ou.edu/joebob/PdfPubs/MWSCAS2002Matt.pdf

[**30**] NAM, J. NGIAM, J. LEE, H. SLANEY, M (2011) "*A classification-based polyphonic piano transcription approach using learned feature representations.*". Proceedings of the 12th international society for music information retrieval conference (ISMIR 2011), 24–28 Oct 2011, Miami, FL, USA Accessed 23.10.2016.

[**31**] MAROLT, M. (2004) "*A connectionist approach to automatic transcription of polyphonic piano music.*" IEEE Trans Multimed 6(3):439–449. doi:10.1109/TMM.2004.827507. Accessed 15.01.2017.

[**32**] MARTIN, J.R. (2009). "*Management Accounting: Concepts,Techniques Controversial Issues*". Chapter 11 Conventional Linear Cost-Volume-Profit Analysis. Accessed 29.04.2015. Available from:
http://maaw.info/MAAWTextbookMain.htm

[**33**] MARTIN, K. D (1996) "*A blackboard system for automatic transcription of simple polyphonic music.*". Technical report. Accessed 02.10.2016.

[**34**] MOORER, J. A (1977) "*On the transcription of musical sound by computer.*". Comput Music J 1(4):32–38. Accessed 12.12.2016.

[**35**] O'HANLON, K. NGANO, H. PLUMBEY, M (2012) "*Structured sparsity for automatic music transcription.*". Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 441–444. doi:10.1109/ICASSP.2012.6287911. Accessed 25.10.2016.

[36] OPOLKO, F. WAPNICK, J(1987) "*Megill university master samples*" Accessed 28.09.2016.

[37] PANCHWADKAR, V. PANDE, S. VELANKAR, M. (2013) "*SurveyPaperon Music Beat Tracking*". International Journal of Research in Computer and Communication Technology, Vol 2, Issue 10. pp 953-958. Accessed 09.06.2015. Available from: http://www.researchgate.net/publication/264212002_Survey_paper_on_music_beat_tracking

[38] PATTERSON, R.D. MOORE, B.C.J. (1986) "*Auditory Filters and Exitation Patterns as Representations of Frequency Resolution*". Frequency Selectivity in Hearing. Academic Press Ltd. pp 123-177. Accessed 30.05.2015

[39] PATTERSON, R.D. (1992) "*An Efficient Implementation of Gammatone Filters*". Accessed 14.05.2015. Available from: staffwww.dcs.shef.ac.uk/people/N.Ma/resources/gammatone

[40] PETERS, R. (2003). "*Signal Peak-Tracker based on the Teager-Kaiser Energy (TKE) Operator*". Physics Department Mercer University, Macon, GA. Accessed 17.10.2014. Available from: http://arxiv.org/ftp/arxiv/papers/1010/1010.5166.pdf

[41] PISZCZALSKI, M. GALLER, B.A (1977) "*Automatic music transcription.*". Comput Music J 4(1):24–31 Accessed 12.12.2016.

[42] POLINER, G.E. ELIS, D. (2007) "*Improving generalization for classification-based polyphonic piano transcription*" In: Proceedings of the 2007 IEEE workshop on applications of signal processing to audio and acoustics, pp 86–89. doi:10.1109/ASPAA.2007. Accessed 11.01.2017.

[43] POTAMIOS, A. DIMITRIADIS, D. MARAGOS, P (2009) "*A Comparison of the Squared Energy and Teager-Kaiser Operators for Short-Term Energy Estimation in Additive Noise*". pp 2569-2581. Accessed: 03/02/2015. Available from: http://cvsp.cs.ntua.gr/publications/jpubl+bchap/DimitriadisPotamianosMaragos_ComparisonSquaredAmpl-TKOper-EnergyEstimation_ieeetSP2008.pdf

[**44**] PRIVOSNIK, M. MAROLT, M. (1998) "*A system for automatic transcriptionv of music based on multiple agents architecture.*". Proceedings of MELECON'98, pp 169–172 (Tel Aviv 1998) Accessed 28.09.2016.

[**45**] RAPHAEL, C. (2002) "*Automatic transcription of piano music.*"Proceedings of the 3rd international conference on music information retrieval: ISMIR 2002, pp 15–19, Paris, France Accessed 28.09.2016.

[**46**] RYYNAMEN, M. KLAPURI, A. (2005) "*Polyphonic music transcription using note event modeling.*" In: Proceedings of the IEEE workshop on applications of signal processing to audio and acoustics, pp 319–322. doi:10.1109/ASPAA.2005.1540233. Accessed 15.01.2017.

[**47**] VAN DER POOL, B. (1948) "*Estimating and Interpreting The Instantaneous Frequency of a Signal*". PROCEEDINGS OF THE IEEE, VOL. 80, NO. 4. Accessed 25.10.2014

[**48**] VINCENT, E. BERLIN, N. BADEAU, R (2008) "*Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription.*"Proceedings of the IEEE international conference on acoustics, speech and signal processing, ICASSP 2008, pp 109–112. doi:10.1109/ICASSP.2008.4517558 Accessed 28.09.2016.

[**49**] WU, F. LEE, C. CHANG, K. WANG, W. (2011) "*A two-fold dynamic programming approach to beat tracking for audio music with time-varying tempo*". 12th International Society for Music Information Retrieval Conference. pp 191-196. Accessed 20.06.2015. Available from:
http://ismir2011.ismir.net/papers/PS2-4.pdf

If not stated otherwise, the graphs in this thesis were of my own creation and are henceforth part of my intellectual property, which I choose to enforce legally. Were any these rights vialated, appropriate consequences would follow.

In all of the other graphs which were not mine and hence were presented as such, the author and the source were given a proper credit.

View publication stats