
**Music beat and tempo tracking with
Laplacian and Bayesian networks**

**ラプラシアン及びベイジアンネットワークを用いた
楽曲におけるビートとテンポの追跡**

Gabriel Pablo Nava

Master of Science Thesis

January 30, 2004

Supervisor: Professor Hidehiko Tanaka

Information and Communication Engineering Department

Graduate School of Information Science and Technology

The University of Tokyo

Abstract

Nowadays, there are many computer music algorithms applied to a wide number of applications in the market. Furthermore, intelligent music systems that interact with the user are becoming more popular. Systems that are able to “listen to the music” are finding a growing acceptance as they aid the automation of tasks that result difficult to accomplish for humans without special training, as for example music transcription and rhythm recognition.

In this thesis, the model proposed has the aim of finding the music beats and tempo in sampled audio signals, and keeping track of them as the music of the input signal progresses in time. Previous systems that are based on the analysis of the amplitude envelope of the audio signal, but they found difficulties when dealing with signals that do not display explicit hard note onsets that can be reflected in the amplitude envelope. The proposed approach is based on a basic image processing technique and frequency centroid to detect onsets, and a Bayesian probability network to infer the beats and tempo. Tests with audio segments of different kinds of music styles showed the robustness of the model to detect the beats and infer the basic tempo in signals with complex mixture of sounds. The idea behind the model is, to accentuate the sound transients in the spectrogram by employing an image edge enhancement technique before computing the frequency centroid. A relative differential will make the transients evident, then. Finally, the possible representatives of beats will be evaluated at a probabilistic hypotheses network where the actual tempo is inferred.

Preface

The work presented in this thesis was developed at Tanaka/Sakai Laboratory of the Department of Information and Communication Engineering from the Graduate School of Information Science and Technology, The University of Tokyo.

I want to express my most sincere thanks to my professors Hidehiko Tanaka and Yuichi Sakai for their constant assistance and guidance during the investigations of my Master's program. The interactive instruction of my supervisor, professor Tanaka, was such a key support giving professional advices and providing unconditional materials and resources for the progress of my work. I want to mention also the staff of Tanaka/Sakai Laboratory for their kind help to manage the formalities of this work, and my colleagues of the Laboratory for their aid to improve this research. I am especially grateful to Dr. Ichiro Ide for his professional and curricular guidance to this work while motivating my research.

Finally, I have not found the words to express my love and my most infinite thanks to my parents, Gabriel Pablo and Maria Elena Nava, for their love and invaluable endless support, and to my sister Alma Sonia and my brother Mario Alberto, for their warm moral encouragement.

Tokyo, January 30, 2004

Gabriel Pablo Nava

Table of contents

Abstract	ii
Preface	iii
1) Introduction	1
2) Theoretical background	4
2.1 The rhythmic structure of music.	4
2.2 The beat and tempo tracking problem.	7
2.3 Signal processing for beat and tempo tracking.	8
2.3.1 Frequency spectrogram of a signal.	8
2.3.2 Frequency centroid.	9
2.3.3 Image edge enhancement.	10
2.4 Bayes probability theory and Bayesian probability networks.	11
3) Previous researches.	14
3.1 Multi-agent models.	16
3.2 Multiple oscillator models.	17
3.3 Probabilistic models.	18
3.4 Rule based models.	19
3.5 Perceptual models.	20
4) A Bayesian probability model for music	
beat and tempo tracking with Laplacian filter.	22
4.1 System framework.	22
4.2 Signal processing.	25
4.2.1 Sub-band analysis filters.	27
4.2.2 Spectrogram processing and enhancement.	30
4.3 Beat extraction.	36
4.3.1 Frequency centroid of the enhanced spectrogram.	37

4.3.2	Relative differentiation applied to the frequency centroid signal.	38
4.3.3	Beat labeling (Pulse formation).	39
4.4	Bayesian network model for beat and tempo tracking.	42
4.4.1	Hypotheses generation.	44
4.4.2	Hypotheses evaluation network.	48
4.4.3	Beat and tempo information integration.	51
4.4.4	Propagation of tempo knowledge.	51
5)	Model performance.	53
5.1	Method of evaluation.	53
5.2	Results.	57
5.3	Discussions.	57
6)	Conclusions.	60
	References.	63
	Appendix A.	69

Chapter 1 Introduction

In the last few decades the number of applications of computer music algorithms has increased in a considerable way that interests of researchers in computer music have made this field split into several specialized areas. Rhythm recognition systems are one of the areas that is very well known for the problem of finding the notes present in the music data and inferring the tempo and beat structure hierarchy at high levels of organization. In many music applications such as automatic music transcription, sound source location and identification, video and audio synchronization, etc., a beat and tempo tracking subsystem is incorporated in order to identify the timing sequence of the music cues in the signal that is being processed. And in general, recognizing the rhythmic structure of music, provides many applications with the overall time event information of a music composition and enables them to participate more interactively with the user.

On the other hand, the problem of inferring the beat and tempo as the basic structure of rhythm, is a topic that has attracted the application of several computational models from different fields. As in the model that will be introduced in this thesis, areas such as audio signal processing, image processing and probability theory are combined to propose a solution to aim the task of tracking beats and tempo in audio signals.

Since finding the beats or tapping with one foot at the unison of music,

results to be a primitive task for humans, the topic of Beat and Tempo recognition has been somewhat underestimated, and it would seem that developing a computer algorithm that is able to recognize the time sequencing characteristic of music does not require complex operations. However we should take into account that the auditory system of humans has been evolving since our existence, to a degree that listening to music and retrieving musical information can be performed without becoming aware of these complex processes. Human ear has mastered such tasks and reached a level of complexity that until now many researchers have been trying to model. Thus, developing a computer system that can recognize patterns in music signals is still a good challenge for investigators involved in this field.

In this thesis I will propose a model that attempts to find the music beats in audio signals, and with this information, infer the implicit tempo. After the system succeeds in these operations, it will then keep track of both, the new incoming beats and the possible variations in tempo that can be introduced by the music performer.

This problem arose originally from the necessity to implement a subsystem whose task was specifically that of extracting the rhythmic characteristic of a music composition performed with MIDI instruments, with the objective of aiding automatic transcription of melodies in a more general system proposed in [KT93]. However, the model of this thesis has been extended so that it can process not MIDI but real world audio signals from music of different genres. In [KT93] and even before, several approaches have been published with partial satisfactory results, which will be reviewed in Chapter 3.

The model implemented here is based purely in signal processing techniques and probability theory. No prior music knowledge was incorporated. Therefore, I will not attempt to prove or formulate any psychoacoustic principle rather than the idea that, in order to retrieve and infer higher levels of music knowledge there is a need of accurate techniques to discover the cues implicit in the audio signal on which the

inference of higher level music knowledge relies. Thus, in this thesis I will introduce one possible solution to satisfy this need.

The thesis is then organized as follows: Chapter 2 presents the theoretical background that is closely connected to the principles and theory used in the model. In Chapter 3, a review of the previous researches most directly related to this work, is introduced. The actual system proposed for beat and tempo tracking is presented in Chapter 4, which is followed by the methodological evaluation of the model performance presented in Chapter 5, together with the discussion of the results and the general work of this thesis. Finally, the conclusions to this work are stated in Chapter 6.

Chapter 2 Theoretical background

The field of music pattern recognition is considerably wide nowadays. Studies and researches in its many subareas have been even more motivated by the new discoveries and developments made in related interdisciplinary areas. Rhythm pattern recognition is one of the music studies that has received support from computational algorithms proposed in areas such as digital signal processing, artificial intelligence, musical psychoacoustics, probability theory and statistical signal processing, etc. There are many tools and techniques to aid the task of recognition of rhythmic structures in music signals, which have been used or developed in the attempts to model music rhythm perception. Thus, in this section, I will make a review of those computer tools that are connected and fit in the approach proposed in this thesis for beat and tempo tracking. Some concepts of music theory and the statement of the specific problem of beat and tempo tracking will be discussed as well.

2.1 The rhythmic structure of music

The auditory system of humans is capable to analyze sounds of any nature from several sources vibrating simultaneously, giving as a result a composed sound with different characteristics in its spectral composition. Hence, the existence of an infinite variety of sounds that surrounds our acoustic environment. But from this ocean of sounds, humans have particular interest on music signals. Music is distinguished from other audio signals in that it is composed of melody,

harmony and rhythm. Rhythm and harmony are considered by some authors [CK99] to be complementary since the same piece of music can be analyzed interchangeably from the point of view of harmony or rhythm. However, this last feature of music is still an ambiguous concept since it is a feature purely perceptual for the listener. In some literatures, a basic definition of rhythm is as follows:

“rhythm: periodical accent and duration of notes in music.”

Oxford Dictionary and Thesaurus, Oxford University Press, 1996.

I will not enter into the discussion around the formal definition of this concept, but I will focus on the components that work together in order to give the music its metrical time structure.

In general, the sense of rhythm is caused whenever there is a relation between timing and duration of music notes, and when some notes are accentuated in sound loudness, pitch or timbre. However, this is valid for the case of music. More broadly speaking, when humans listen to articulated sounds (these are, sounds that are combined or changed harmonically), we tend to identify the time limited acoustic events that persist and keep a rough constant relationship as time progresses. These acoustic pulses are what human brain perceives as *beats*. A beat is a part of the metric structure of music and is considered as the most important property of rhythm. The beat is well defined by the tapping of a foot when people are listening to music. Thus, beats are the immediate perception of rhythm and they define what is called in music the *tempo* of the performance. The tempo is then, a measure that indicates the speed at which the sounding music composition is played. Tempo is usually given in *beats per minute* (bpm).

Now, since the beat can be perceived at different levels of timing, the metric of music can be visualized as a hierarchical organization, in which several components take a place in this hierarchy and combine to form the rhythm. In Figure 2.1 an example of this hierarchical organization of the beats is illustrated. It can be seen from Figure 2.1

that due to the existence of grouping levels, the basic tempo can be inferred once we have found one of the metric levels of the music. This principle will be used in the implementation of my model as will be discussed later.

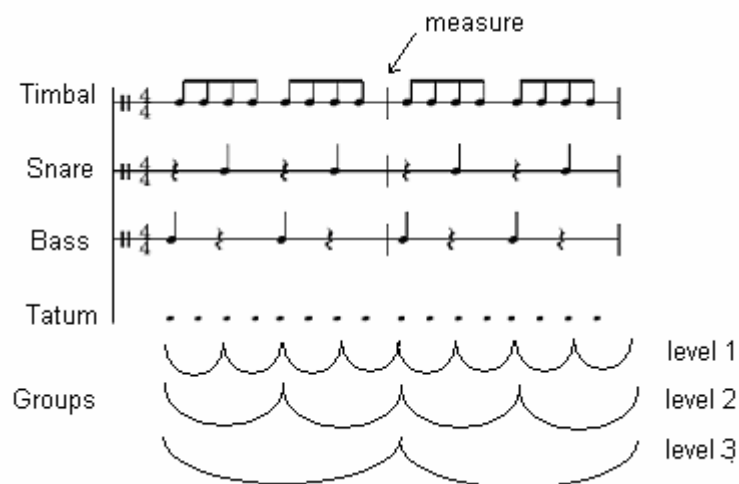


Figure 2.1. Example of the hierarchical organization of the beats. Tatum is the minimum pulsation perceptible in the performance.

I would like to make a distinction here, between the beats and *onsets*. As seen before, the first is defined by the series of pulsations that are separated by equally spaced intervals of time (assuming constant tempo).

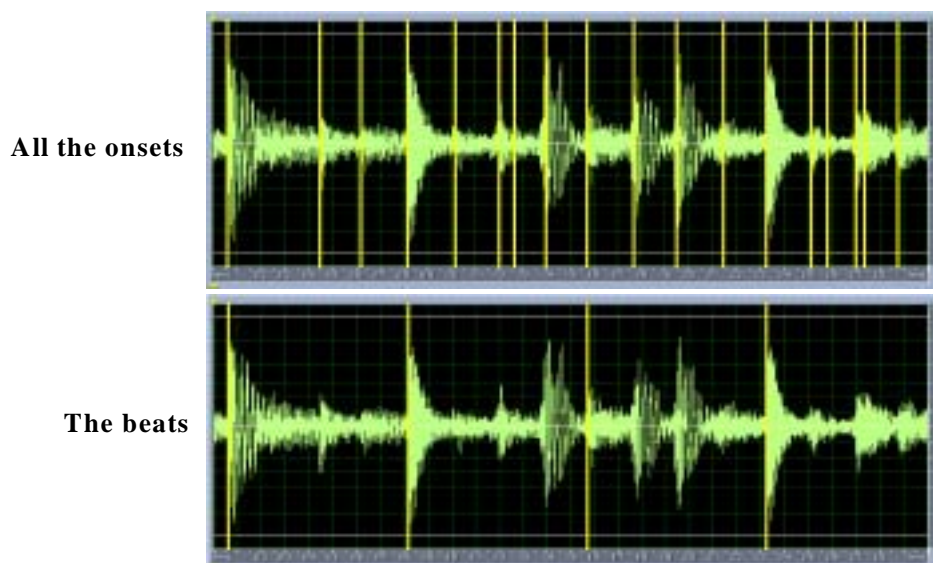


Figure 2.2. Examples of the same music segment showing the onsets

and the beats.

While onsets are present whenever a musical note is played. This can be better appreciated on Figure 2.2.

2.2 The beat and tempo tracking problem

We have seen in the paragraphs above the relationships between the musical events of an audio signal. We have discussed the concepts that organizes the timing metric of the basic cues that define the tempo. Now I will make the definition of the problem of beat and tempo tracking that has motivated the implementation of the model that will be presented in further chapters.

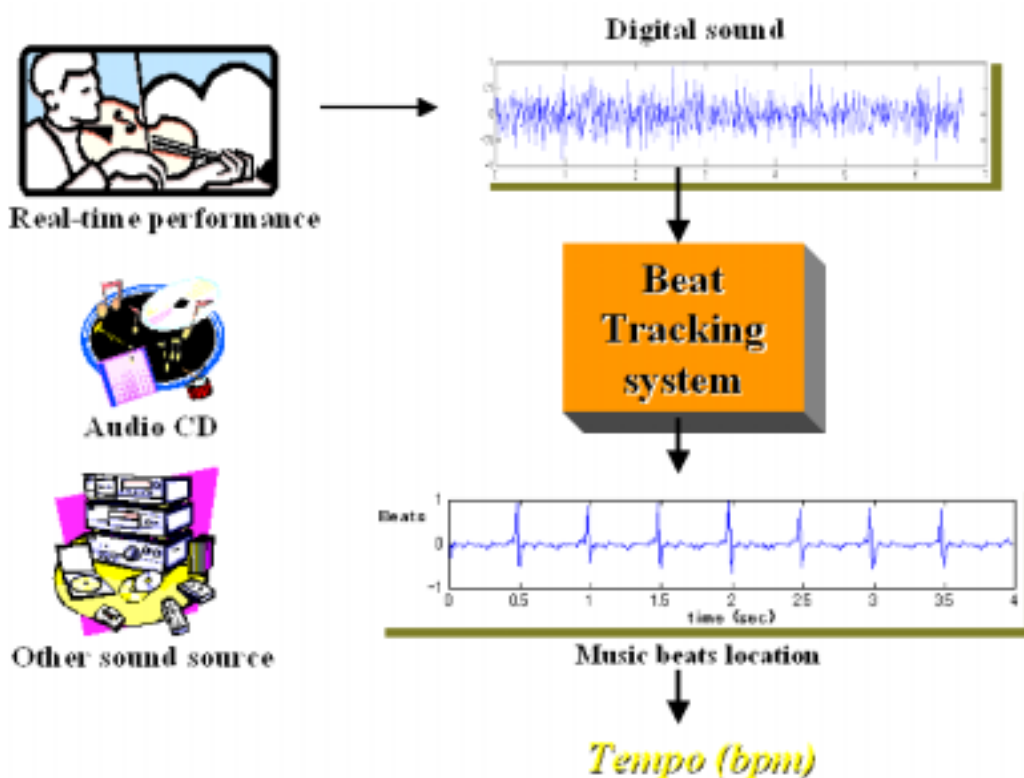


Figure 2.3. Automatic beat and tempo extraction from complex acoustic signals.

The basic idea, which is summarized in Figure 2.3, is to develop a

functional system that will attempt to recognize the timing sequences in the input signal that it is fed with. As output, the system will provide us with the measure of the time metric of the audio signal and will keep track of possible changes in the relationships of the beats in time. It means that this model will not necessarily recognize every single onset of the music, nor try to infer hierarchical organizations of higher level, but it will rather be aware of the beats that mark the actual tempo and its variations as the music performance goes on. Thus, the problem is limited to develop of an algorithm to find, first the most salient beats by observing the input data, and second, the tempo in bpm so that the system can follow the changes in the timing of both variables: beat and tempo. Another particularity of this approach is that it will be capable to analyze and infer the tempo of not only signals under a musical context but also of any audio signal that present periodic acoustic patterns. No assumptions of musical knowledge will be considered in the implementation of this model, therefore, the problem is visualized from the point of view of signal processing and probability theory.

Thus, I proceed to make a review of the tools that are key in the design of my beat and tempo tracking system.

2.3 Signal processing for beat and tempo tracking

In many literatures, basic and advanced algorithms for signal processing have been published. For our purpose, we are interested in those that are applicable to the analysis of audio signals, and more particularly to music signals. As will be explained in Chapter 4, one of the signal analysis tools that is a basis for the beat detection task, is the *frequency spectrogram*. Therefore I will begin this theoretical review with this topic.

2.3.1 Frequency spectrogram of a signal.

The Fourier Transform is a well known technique to decompose any

given signal into its spectral components. Furthermore, the Short Time Fourier Transform (STFT) has the capability to show the spectrum of a signal as a function of time. In other words, the spectral vector of coefficients is computed for short intervals of time taken from the signal where these intervals can overlap each other, in a way that the complete signal is analyzed. This is performed by sliding a window of a finite number of digital samples, that usually is much smaller than the length of the original signal. The size of this window, the overlapping factor and the sampling frequency as well, determine the resolution of time of the resulting spectral analysis. Thus what we obtain is a 2-D arrange of coefficients that is commonly called the *spectrogram* of the signal. This kind of analysis is also called in some literatures as *time-frequency* domain. I will refer to it simply as the *spectrogram*.

The spectrogram of a discrete signal $\mathbf{x}[\mathbf{n}]$ is computed using the window $w[m]$ of size R with the expression:

$$X[k, nh] = \sum_{m=0}^{R-1} x[nh - m]w[m]e^{-2\pi km/N} \quad (1)$$

where $k = 0, 1, 2, \dots, N - 1$ are the N frequency bins for which the Fourier coefficients are being calculated. h is the hop of the sliding window, giving thus, nh the discrete time index of the spectrogram. More details in time-frequency analysis are available in [Mit01] and [PM96].

2.3.2 Frequency centroid.

In many applications that perform signal analysis, it is required to extract features from the spectral information given by the spectrogram. One of these features is the *frequency centroid*. This parameter is a measure that indicates the position of the center of balance of energy contained in all the frequency bins of the spectrum. The frequency centroid of a power spectrum $X[f]$ is given by:

$$C = \frac{\sum_{i=1}^N |X(f_i)|^2 f_i}{\sum_{i=1}^N |X(f_i)|^2} \quad (2)$$

where $|X(f)|$ is the power at frequency bin f_i . I will comment more on the frequency centroid in Chapter 4.

2.3.3 Image edge enhancement.

Image processing is another area that has been actively developing algorithms for signal processing applied to images. A basic operation that is performed in images is the detection of edges by enhancement of changes in contrast, as can be appreciated in the example of Figure 2.4.

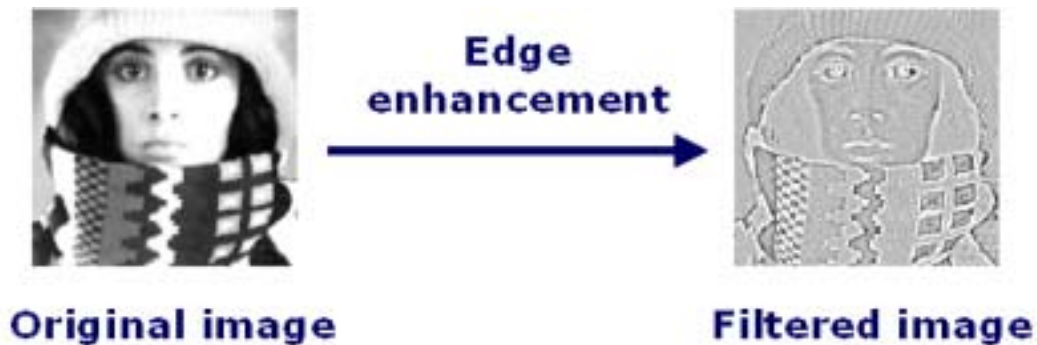


Figure 2.4. Example of edge enhancement of images.

Edge enhancement basically consists in the detection of abrupt changes in the intensity of the image and amplifying them by a factor. This operation is normally performed by convolving the matrix containing the intensities of the image with a much smaller matrix that is known as *kernel* or *mask*. A number of masks have been developed, which give different responses with the type or direction of the edges. For the purposes of this thesis, I will mention in the next paragraphs the edge enhanced method that is based on Laplacian filtering.

The Laplacian is a 2-D isotropic measure of the 2nd spatial derivative of an image. The Laplacian of an image highlights regions or rapid intensity change regardless the direction of change. Thus, the Laplacian $L(x,y)$ of an image with intensity pixel values $I(x,y)$ is given by:

$$L(x, y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \quad (3)$$

As mention before, this operation is usually computed by convolutional methods. In section 4.2.2, more details of the Laplacian edge enhancement algorithm used for my model will be introduced as well as its application to audio signal spectrogram enhancement.

2.4 Bayes probability theory and Bayesian probability networks

Bayes theory has become a powerful tool widely used in all areas where there is a need to perform inference of data based on *a priori* knowledge of the possible situations that occur in a certain event. Classical statistical models do not permit the introduction of prior knowledge into the calculations. This proceeds from the rigorousness of the method to prevent the introduction of extraneous data that might influence the results. However there are situations in which the use of prior knowledge is a powerful contribution to the inference process.

The Bayes theorem is mathematically expressed as:

$$p(H | E, c) = \frac{p(H | c) * p(E | H, c)}{p(E | c)} \quad (4)$$

where the belief of hypotheses H can be updated once we have observed the evidence E under the context c . The terms of equation (4) are interpreted as follows:

- $p(H | E, c)$ = the *posteriori probability*, or the probability of H after having observed E under c .

- $p(H|c)$ = prior probability of H given c .
- $p(E|H, c)$ = the *likelihood* that determines the probability of the evidence assuming the hypotheses H under the context c .
- $p(E|c)$ is independent from H and can be considered as a normalizing factor.

There are countless real world situations in which the probability of a certain event is conditioned by the probability of the previous one. The concept of conditional probability is useful and is the basis for implementing more complex relations of events such as Bayesian belief networks.

The formulation of a Bayesian belief network is then as follows. Given a subset X of variables x_i where x_i belongs to U , if one can observe the state of every variable in X , then this observation is called an instance of X and is denoted as:

$$X = p(x_i | x_1 \dots x_{i-1}, c) = p(x_i | \Pi_i, c) k_X \quad (5)$$

for the observations $x_i = k_X$. All the set of instances of U determine the joint space of U . $p(X=k_X|Y=k_Y, c)$ denotes a generalized probability density that satisfies Equation (5) given $Y = k_Y$ for a given state information c . $p(X|Y, c)$ denotes then, the generalized probability density function for X , given all possible observations of Y .

Then, a Bayesian network for domain U represents a joint generalized probability density over U . These representations consist of local conditional probability density functions combined with a group of conditional independence assertions that allow the construction of a global probability density function. To this point, the chain rule of probability can be applied to find these values, thus:

$$p(x_1 \dots x_k, c) = \prod_{i=1}^k p(x_i | x_1 \dots x_{i-1}, c) \quad (6)$$

An imposition of Bayesian network theory is that each variable must be

a set of variable that renders x_i and $\{x_1, \dots, x_{i-1}\}$ conditionally independent. Therefore:

$$p(x_i | x_1 \dots x_{i-1}, c) = p(x_i | \Pi_i, c) \quad (7)$$

In this way, a Bayesian Network encodes the assertions of conditional independence in the Equation (6). Then, a Bayesian network can be visualized as a directed acyclic graph in which each variable U corresponds to a node, and the parents of that node corresponding to x_i are the nodes corresponding to the variables p_i .

When presenting the Bayesian network for the beat and tempo tracking model, I will not step into the details on the formulation of the domain of the sets of variables, but instead I will explain the possible situations considered to set the conditional relationship of each node. More details in how to construct a Bayesian network can be found in [Heck96].

Chapter 3 Previous researches

In this chapter, I will make a review of the most relevant approaches done in the area of beat and tempo tracking. In recent years, the number of researches on this field has been in explosion being reflected in a good number of literatures published. Those approaches that have introduced revolutions in the concepts of music beat and tempo recognition will be considered here. I will also try to propose a classification in order to clearly see their functionality and different tendencies that the researchers have followed towards the implementation of an automatic system for beat and tempo tracking.

In [Dix01] a formal comparison of tempo trackers is discussed and a first classification of beat and tempo tracking systems is presented. Due to the amount of new approaches that have appeared recently, I will try to update the classification made in [Dix01] and other works.

A first and most general classification that I will mention is based on the purpose of the design of the system: models that try to imitate the human perception of music, and the models intended to satisfy specific tasks. A second classification divides the systems into those that perform beat and tempo tracking in real-time and those that process prerecorded signals. Some authors have called these systems as *on-line* or *off-line* processing models respectively. One more classification can be done according to the nature of the input data that they process: models that work on acoustic audio signals and models for higher level of symbolic data (such as MIDI sequences). In Figure 3.1, this classification is summarized.

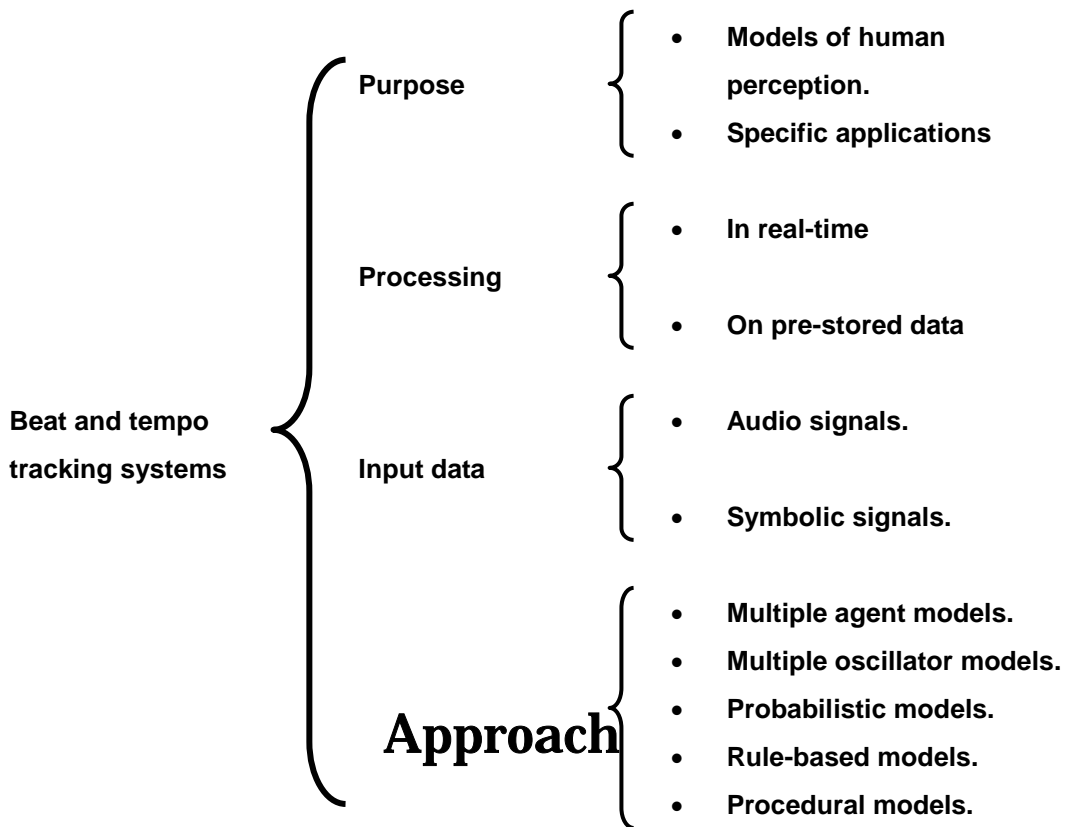


Figure 3.1. Classification of current beat and tempo tracking systems.

In the Figure above, the category of the lower part is in some sense ambiguous since the techniques used in some approaches are extensions or combinations of other works that have been retaken for further investigations, such is the case of the rule-based models and the multi-agent models.

Since the first three categories of Figure 3.1 are more general, the particular categories that fall in the subcategory “Approach” will be discussed here in order to take a glance in how other works have conceptualized the problem of beat and tempo tracking and proposed solutions. Thus, I will mention here the most relevant and recent models that have shown notable results in this matter.

3.1 Multi-agent models

Multi-agent models are systems that base their inference of tempo on the formulation of multiple hypotheses that represent different possible estimates of features such as pulse period, phase, etc. A salient value is computed for each hypothesis by iterations. While performing the tracking operation, the number of hypotheses can be increased or reduced by pruning or splitting, according to their validation as the system finds new cues. A single hypothesis is called an *agent*. Multi-agent models need to be initialized with a number of hypotheses in order to keep track of the new agents that are generated.

Allen and Dannenberg [AD90], formulated a multi-agent model that relies on a set of heuristic rules that penalize the situation of the occurrence of beats, and then uses a beam search engine to find the beat metric that represents the beat transcription. This model works on MIDI signals, and the algorithm requires the user to enter the initial down beat in order to start the beam search.

Rosenthal [Ros92] presented another system to analyze symbolic meter on polyphonic music. The system attempts to find the beat from the beginning of the music piece by computing an inter onset interval histogram that is then transformed into harmonic parameters, and then the resulting coefficients are convolved with a Gaussian function which is then weighted with an a priori tempo distribution. The period with maximum value is selected to compute the tempo at the final step.

Goto and Muraoka [GM98] published an advanced system that performs onset detection on audio signals by spectral analysis of the input signal. In multiple frequency bands, the detection of onsets is performed and then agents are assigned to these onsets coming from the bands. The multiple-agents from each sub-band compute histograms of inter onset intervals (IOI) to determine the beat period. Beats of bass and snare onsets of drums are identified and used to infer additional information about the metric of the audio signal.

Dixon, Goelb and Widmer [DGW02] recently introduced a system that operates on audio signals. Onset detection is performed by high-pass filtering the incoming signal. Then a full wave rectifier and a moving average filter are applied, and finally by peak picking, the IOI are measured and clustered into histogram representation where each IOI represents a “class”. The beat positions are determined by iterating over the onsets, and the hypotheses of beat periods are used to perform beat tracking.

3.2 Multiple oscillator models

In these kinds of approaches, multiple copies of a basic oscillator are used to respond and phase-lock to different meter hypotheses. Each oscillator only responds at a characteristic frequency range. The oscillators are excited with a train of pulses. If an oscillator matches with the period of the characteristic frequency, then it is phase-locked. Thus, a bank of oscillators is used to respond at different frequencies and the one with the highest resonance is taken as the recognized pulse.

Large and Kolen [LK94] described a model based on non-linear oscillator units that when stimulated with a train of pulses that match their frequency range, respond by synchronizing another pulsation. An arrange of six oscillators are fed with the same pulse stimuli and those oscillators that match will respond with different metrical levels while others will fail to respond at all.

Scheirer [Sch98]. In this model, the beat tracking is performed with independent oscillators arranged at each single frequency sub-band. The final beat tracking results from the combination of the energies of the sub-band oscillators. This model was the first to introduce comb filters as oscillator units with the idea that a comb filter oscillator will resonate at integer multiples of its characteristic frequency. Therefore, different tempi are tracked by their corresponding oscillator at multiple

sublevels.

Eck [Eck01] presented a model based on oscillator units called Fitzhugh-Nagumo relaxation oscillators. These oscillators were originally developed to model dynamics of neural action potentials. A network of 20 oscillators is interconnected through a specific coupling function. This model results to be the most complex from the multi-oscillator based models reviewed here.

3.3 Probabilistic models

Probabilistic models consider the onset times and other music phenomena as random processes by nature. Thus, the beat cues are observations that are corrupted by uncertainty. Therefore these models try to overcome the task of beat detection by taking into account this uncertainty and applying probabilistic methods to determine the original beat timing.

Cemgil, Kappen, Desain and Honning [CKDH01], developed a system that estimates the beat trajectory by applying a Karman filter to a local periodicity matrix that they call a Tempogram. The tempogram of periodicity data represents energy as a function of period in local timeframe. Then the Kalman filter estimates the optimum parameters of a linear dynamic system where the beat position is a hidden variable.

Cemgil and Kappen [CK03], proposed the structure of a model on which switching variables represent discrete note locations and the continuous hidden variables describe the tempo. The authors introduce Monte Carlo methods for the integration and optimization of the filtering, and maximum a posteriori state estimation tasks which represent tempo tracking and rhythm quantization operations. This method works over MIDI signals.

3.4 Rule based models

The design of these models was developed in conjunction with a series of experiments on rhythm patterns in order to model the perception of rhythm. Then, the basis of the model relies on a number of heuristic rules that guide the system when attempting to perform tracking.

Parncutt [Par94] considers two important features in his work compared to the previous models: moderate tempo and accents phenomena. A phenomenal accent is measured as the sum of terms defining the pitch accent, loudness accent, durational accent and the relations between these. This model relates directly the perceived beat with the IOI, duration accents and moderate tempo. The estimation of metrical accents and expressive timing is performed in addition to the estimation of the beat.

Temperley and Sleator [TS99] implemented a model that is driven by rules specifying situations of the beat occurrence. For example, they consider that beats should be aligned with onsets, that beats should be spaced regularly and that strong beats should align to onsets of longer events. The evaluation of these rules gives a score value that is employed to infer the metric of the tempo by using the Viterbi algorithm.

Laroche [Lar01] proposed a tempo tracking system in which he made the assumptions of constant tempo in the input acoustic signal and that the metric of the beat is subdivided into four tatums. The onset detection technique used in his model is similar to that used in [Sch98], however the number of frequency sub-bands employed is not specified. The model uses a four component Gaussian mixture to estimate the likelihood of onset locations when performing beat tracking.

Laroche [Lar03] proposed another system that works off-line and assuming that the music has a relatively pronounced beat. The algorithm for beat tracking is based on the detection of rapid changes in

energy within a short-term frequency representation, reducing with this, effects of masking of those spectral components that define the actual beat. Then by performing a least-square optimization, the best tempo and downbeat location are identified.

3.5 Procedural models

These models are characterized specifically by the particular procedure they follow to perform beat and tempo tracking. However, the methodology to approach the problem varies among them. Perhaps, one characteristic that they share is the application of signal processing techniques for beat detection and tracking.

Smith [Smi99] proposed a model in which a train of pulses constructed from a stream of symbolic onset times, is decomposed with a Morlet continuous wavelet transform into representations of time-frequency. Then with this representation, a series of steps process these representations in time-frequency domain in order to detect the beats.

Foot and Uchihashi [FU01] have proposed the idea of using what they called *self-similarity matrix*. Although the autocorrelation that they use to measure the beat period is a standard technique, their main contribution was the construction a 2-D similarity matrix by measuring the distance between two points of the audio signal. They show that the diagonal line patterns represent the different metrics of rhythm in the input signal.

Tzanetakis, Essl and Cook [TEC01] implemented a model in which they use four analysis frequency bands that consist of octave wavelets, and then they integrate the signals from the four bands after having applied rectification, low-pass filtering, decimation and normalization, and finally the beats are found from the autocorrelation of this excitation signal.

Sethares and Staley [SS01] introduced the periodicity transform in their model. The input signal is transformed into frequency domain and split into 23 frequency bands from which the rms envelopes are calculated. The envelopes are transformed to periodicity domain by using the periodicity transformed and the greatest value is chosen to compute the period of the beats.

Paulus and Klapuri [PK03] presented in their percussive audio signals transcriptor a novel method for labeling of onsets. The system first performs coarse onset detection. Then features are extracted from the onsets and arranged in vectors that are clustered with a label attached to every cluster describing the rhythmic information of each event. The labeling process is based on the metrical positions of the sound events that have been measured.

In addition to the systems mentioned above, there have been other parallel software commercial implementations for solutions to beat and tempo tracking in different applications. Some examples are: 1) **Sonic Foundry Acid Pro 3 software** [SF01], a packet that performs analysis of audio signals. The software, while attempting to find the positions of the beats, asks the user for additional information in order to verify the decisions taken by the algorithm. 2) **Protools 5.1.1** [Pro02], which is a powerful software for the edition of digital audio signals, incorporates a “beat detective” that reveals the position of the beats in the waveform in edition. Finally I will mention 3) **InTime Tempo Tracking system** [InT02], a software developed specifically for real-time tempo tracking of MIDI signals.

Chapter 4

Music beat and tempo tracking with Laplacian and Bayesian networks

In this chapter, a model for beat and tempo tracking is described. After the overview of the system, each subsection will be discussed in detail for a better comprehension. First the signal pre-processing is exposed together with the theoretical design of the filters employed. In the section of Spectrogram enhancement by Laplacian filtering, a short review of the image edge enhancement technique with Laplacian will be retaken in order to make clear its application to spectrogram enhancement. The extraction of beat and tempo hypotheses is then explained, and finally the Bayesian network for the evaluation of these hypotheses is presented.

4.1 System framework

The proposed system in this thesis is composed of three main sections: Preprocessing, Beat extraction, and Bayesian probability network. In Figure 4.1 the overall model for beat and tempo tracking is shown. At the input, the signal is assumed to be in digital form sampled at 44.1 kHz and 16 bits of resolution. This signal can be taken directly from an audio CD so that this brings the advantage to apply this model to track beats and tempo of any commercial audio CD. In any case, the signal source can be from any device as long as the digital signal is in the format mentioned above. Segments of 2 sec. in length of the input signal are taken and processed progressively to keep track of the tempo. In the case of stereo music, the signals are converted to mono aural format. All segments are low-pass filtered, and down-sampled to 22 kHz. Since the most important information of tempo and beats is contained basically in

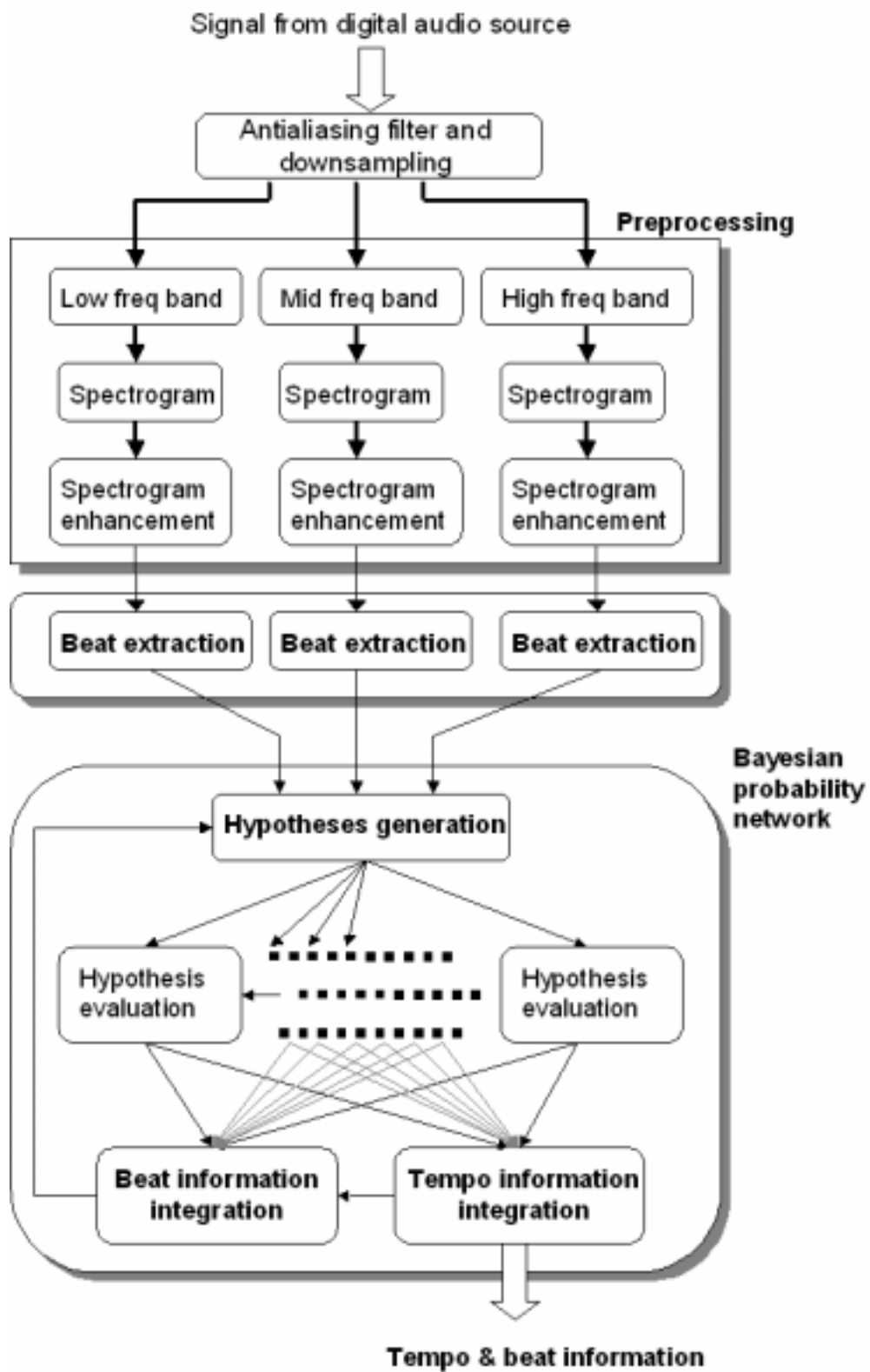


Figure 4.1. Framework of the beat and tempo tracking system.

the lower frequencies, we can down-sample the signal to reduce the overall computational cost.

As will be explained later, the extraction of beats is based on the computation of a frequency centroid function applied to the enhanced spectrogram of the original signal. The frequency centroid is a measure that gives indications of where the centroid of the spectral power is located. Thus, if the frequency centroid is computed for all of the N frames that conform the spectrogram along the time axis, then the frequency centroid will now indicate how the spectral power distribution has changed along time. In contrast, computing the frequency centroid at every moment in the complete frequency band (which is from 0 to 10 kHz after down-sampling), will lead to misses of some beats that are masked by stronger beats in different frequency ranges. In order to reduce this effect, after down-sampling, the signal is split into three frequency bands, as indicated in Figure 4.2, making it possible to track beats at different frequency ranges.

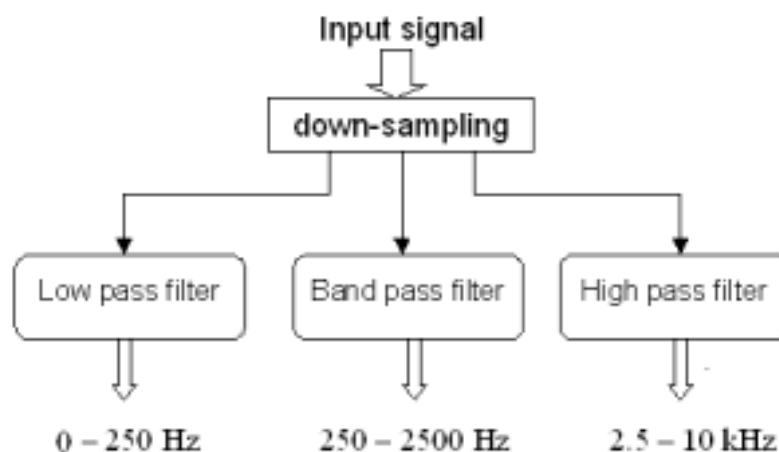


Figure 4.2. Sub-band separation of the input signal.

After filtering the signal, the subsequent signal processing is the same for the three sub-bands except for the values of some coefficients that will be introduced in the corresponding sections. Once the signal is divided, the spectrogram enhancement is performed by convolving a Laplacian matrix kernel with the power spectrogram of the sub-band

signals. The frequency centroids of the enhanced spectrograms are then computed and the resulting signals are passed to a differentiator which gives as an output the derivative of the signal in relation to its magnitude. At this level the indications of possible true beats are now detectable by the hypothetical beat extraction stage in which a train of binary pulses are constructed in order to serve as hypotheses of beats that will be eventually evaluated by the probability network. The Bayesian network model implemented here tries to infer the tempo by observing the hypotheses coming from the three frequency bands and explaining away those probabilities that will lead to the most probable true tempo and true beats. In Figure 4.3, the functional blocks of the Preprocessing and Beat extraction for each sub-band is shown in detail.

4.2 Signal Preprocessing

One of the principal problems in detecting and tracking beats and tempo of music signals is finding the cues that represent the real beats due to the fact that usually these cues are not explicit in the raw signal. Even more, although the probabilistic network tries to find the true beats from the hypotheses, the inference of the tempo relies greatly in the beat information generated by the beat extraction stage. Therefore, a reliable algorithm is needed in order to discover those cues that have high probabilities to represent true beats. To this point, the Preprocessing stage plays an important role, in the sense that makes the appropriate transformations to the signal, with the purpose of highlighting the locations of the probable true beats. Thus, the Preprocessing functions as a first filter of false beats that could have a heavy influence in the probability network, resulting on calculations of erroneous tempi with high values of probability to be the true tempo that will be propagated when keeping track of the tempo as the input music signal progresses in time.

In the following sections, the algorithm for these primary transformations is presented.

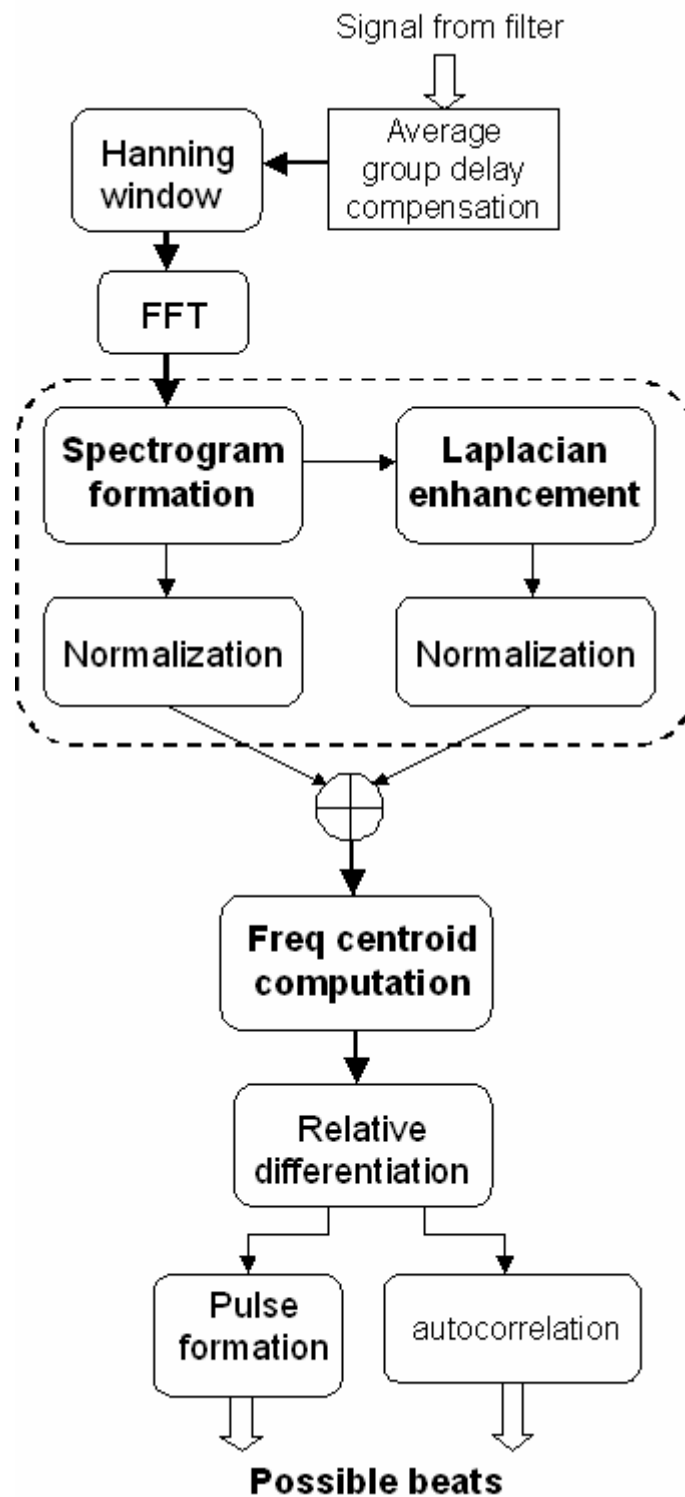


Figure 4.3. Preprocessing and beat extraction in detail.

4.2.1 Sub-band analysis filters

The number of analysis filters used in previous works varies considerable and they have been proposed according to different considerations in the design of the model. For example, the system proposed in [Kla99] has used 21 analysis filters, in [Shc98] six filters are used, five in [DSD02], while Smith has used as much as 28 filters in [Smi96]. In the model proposed in this thesis, a total of three analysis filters was found to be enough in order to discover the beats in the audio signal. The frequency bands of the filters are 0 – 250 Hz, 250 – 2500 Hz and 2.5 kHz – 10 kHz, as was shown in Figure 4.2 before. Experiments with more than four frequency analysis filters were also performed showing small improvements compared to the experiments with only three analysis bands, but with the difference that by using more frequency bands, the computation cost increases and the probability of detecting spurious beats is higher as well.

Since beats can be detected in any of the three frequency bands, their location in time after filtering has to be consistent in all the bands. Therefore, linear phase and stable magnitude response in the operating frequency band of each filter were important parameters in their design. However, these parameters were also trade in favor of simplicity and efficient computation of the filters, while trying to minimize the non-linear phase distortion in most of their operating band. Butterworth IIR type filters showed to perform well for this purpose with relatively low computational cost.

We start by considering a normalized low-pass prototype filter in which we use normalized frequencies. In Figure 4.4, a prototype low-pass filter characteristic is shown together with the parameters to be specified in order to start the filter design. For the design of the Band-pass and High-pass filters, the transformations of the operating frequencies are required.

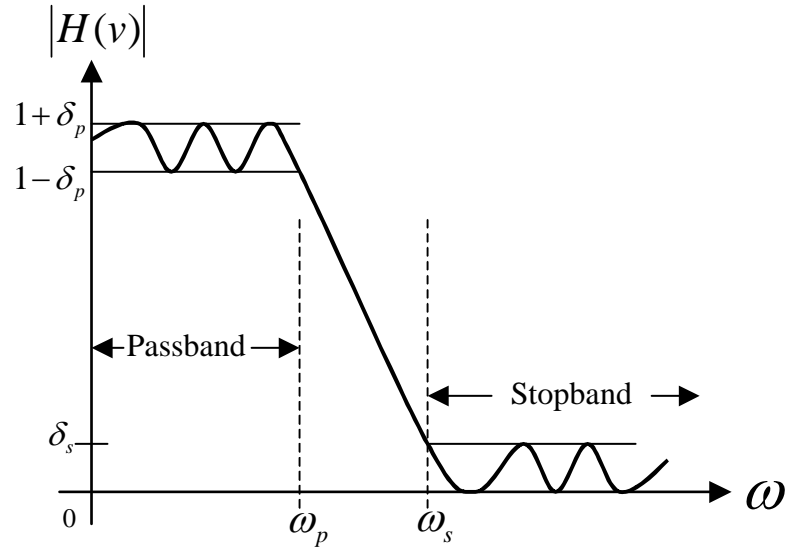


Figure 4.4. Parameter specifications for a prototype low-pass digital filter.

In the Figure above, ω_p and ω_s are the *passband edge frequency* and the *stopband edge frequency* respectively, and the peak ripple values δ_p and δ_s , are the limits of tolerances at the corresponding passband and stopband edges.

It is usually encountered that the specifications for digital filters are given in terms of the losses in dB's. In this case, the *peak passband ripple* α_p and the *minimum stopband attenuation* α_s are the loss specifications of the digital filter ([Mit01]) which are given by:

$$\begin{aligned} \alpha_p &= -20 \log_{10}(1 - \delta_p) \quad dB \\ \alpha_s &= -20 \log_{10}(\delta_s) \quad dB \end{aligned} \quad (8)$$

In the other hand, the filter design techniques are mostly developed in terms of the normalized frequencies, therefore, the edge frequencies should be normalized in order to apply a specific filter design algorithm. Thus, let f_N be the sampling frequency (in Hz), f_p and f_s the passband and stopband respectively (also in Hz), then the normalized edge

frequencies in radians are given by:

$$\omega_p = \frac{2\pi f_p}{f_N}$$

$$\omega_s = \frac{2\pi f_s}{f_N} \quad (9)$$

With the above specifications of edge frequencies, passband ripple and stopband attenuation, we are now in the conditions to calculate the order of the filter and then its transfer function. For this purpose, the appropriate design techniques are used depending on the type approximation desired for the filter, Butterworth, Chebyshev, elliptic, etc. As mention before, the approximation employed in this work is Butterwoth. The order and coefficients of the filters are then obtained with the aid of the MATLAB tools for filter design using the specifications computed with the formulas above. In Table 4a, the parameters used with the MATLAB files and the order of the resulting filters are presented. Figure 4.5 shows the *magnitude vs. frequency* response of the obtained filters.

Filter	Operating Band (Hz)	Normalized freq. edges (of the prototype low pass filter)	Ripple values (dB)	Order of the filter
Low pass	0 – 250	$f_p = 0.0215$	$\mapsto_p = 3$ $\mapsto_s = 30$	3
<i>Band pass</i>	250 – 2500	$f_{p1} = 0.0226$ $f_{p2} = 0.2274$	$\mapsto_p = 3$ $\mapsto_s = 40$	5
<i>High pass</i>	2.5 k – 10 k	$f_p = 0.2202$	$\mapsto_p = 3$ $\mapsto_s = 40$	5

Table 4a. Parameters of the analysis filters.

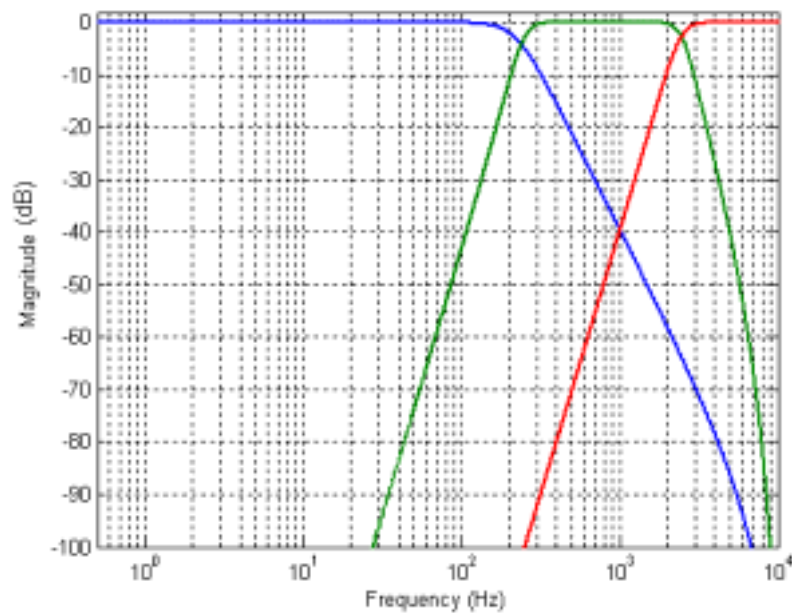


Figure 4.5. Magnitude response of the analysis filters.

4.2.2 Spectrogram processing and enhancement

There are a number of techniques to decompose a signal into a series of coefficients that represent the original time domain signal, Fourier transformed, Cosine transformed and Wavelet transformed are some examples. Some models of beat tracking perform beat detection in time domain, as in [DGW02], [Sch98] and [Kla99], while others extract parameters of the signal in frequency domain ([GM95], [HK02]); [DSD02] combine parameters from both, time and frequency domains, to find the beats, and in [TEC02] discrete wavelet transform is used.

A well known technique for signal analysis in frequency domain is the Short Time Fourier Transform (STFT). In the present work, the STFT is used to obtain the spectrogram of a 2 sec. music segment. The magnitude values of this spectrogram form the matrix that will be convolved with the Laplacian filter kernel in order to produce another matrix with prominent peaks at the locations of abrupt changes in energy. Having this filtered spectrogram, it is normalized and added to the original magnitude spectrogram of the audio signal to give the *edge*

enhanced spectrogram version. In this last spectrogram, the transients of the beats have been highlighted by effect of the convolution with the Laplacian kernel. We can mention at this point that, since the filters used to analyze the input signal into three frequency bands introduce phase delay in the filtered signal, before computing the corresponding spectrogram, an average group delay compensation step is added to correct the phase alteration of the filters. In Figure 4.6, the procedure for the spectrogram enhancement is illustrated in blocks.

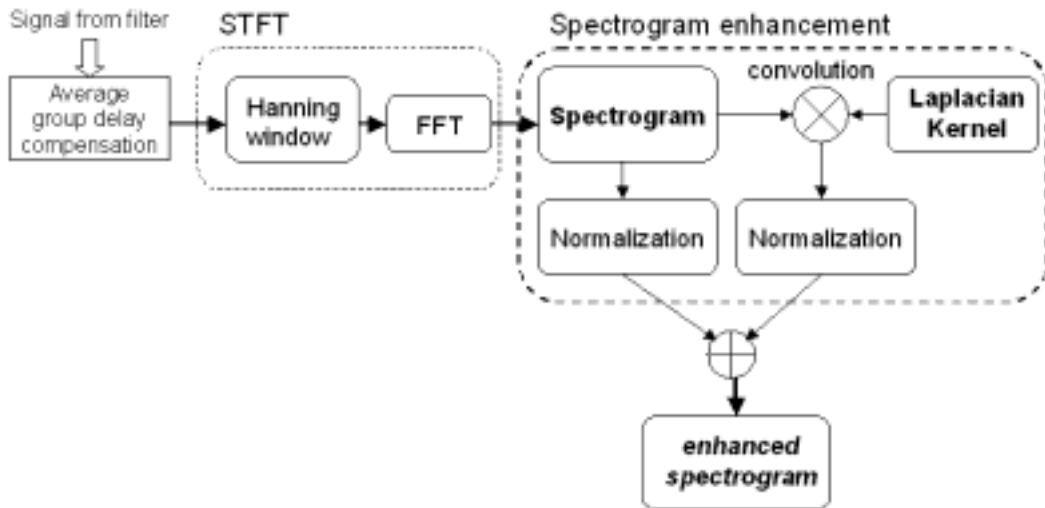


Figure 4.6. Spectrogram enhancement process.

As can be seen in the Figure above, the spectrogram is computed by sliding a Hanning window with a size of 1024 samples, and an overlapping of 57%, thus, with a sampling frequency of 22 kHz, a resolution of 20 ms is achieved. Then, the FFT is computed and arranged in consecutive frames so as to form the spectrogram.

Lets consider the time domain signal of the 2 sec. segment of music, expressed as $s[nh]$. Its spectrogram is computed as follows:

$$S[k, nh] = \sum_{m=0}^{R-1} s[nh - m]w[m]e^{-2\pi km / N} \quad (10)$$

where $k = 0, 1, 2, \dots, N-1$ are the N total frequency bins, and $w[m]$ is the sliding Hanning window of size R . Hence, the resulting spectrogram that will be enhanced, is a three dimensional matrix with dimensions f , t , and $\mathcal{S}(f_i, t_k)$, where:

- $f = f_1, f_2, f_3, \dots, f_M$: frequency axis; M : number of frequency bins.
- $t = t_1, t_2, t_3, \dots, t_k$: discrete time index of the N frames within the 2 sec. segment signal.
- $\mathcal{S}(f_i, t_k)$: power spectrum in bin f_i at moment t_k ; $i = 1, 2, 3, \dots, M$; and $k = 1, 2, 3, \dots, N$.

The magnitude of the power spectrum $\mathcal{S}(f_i, t_k)$ is taken to perform the convolution with the Laplacian kernel.

The idea here, is to treat the magnitude spectrogram as a pseudo image and use the Laplacian kernel to detect, or better said, to enhance the burst in energy that are present in the transient interval when an onset of a musical instrument appears. Due to the characteristics of music signals, these bursts of energy in the note onsets resemble changes of contrast at the edges contained in an image. This comparison can be appreciated in Figure 4.7. Edge enhancement with Laplacian, is a basic technique in image processing field used to detect edges of gray-scaled images. Although the Laplacian filtering kernel is a simple algorithm for edge enhancement, it was selected as the first approach for an application to music beat detection owing to its simple operations that it involves and its omni-directional characteristic to detect edges.

In general, edge detection techniques in image processing usually involve the convolution of two matrices, the image and the edge detector kernel. For the case of edge detection by Laplacian, the Laplacian kernel $L(f)$ can be expressed in its discrete form as follows:

$$L(f(x, y)) = f(x+1, y) + f(x-1, y) + f(x, y+1) + f(x, y-1) - 4f(x, y) \quad (11)$$

where $f(x, y)$ is the value of the image at pixels x and y .

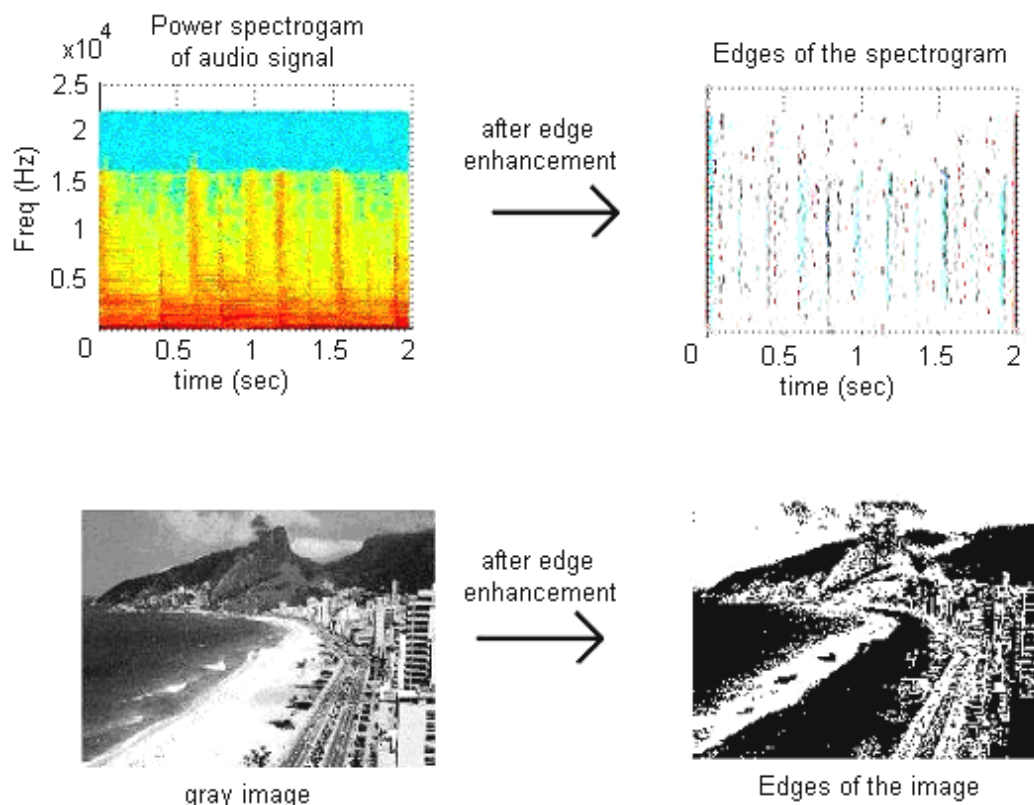


Figure 4.7. Comparison of edge enhancement of the spectrogram of an audio signal (upper) and a normal gray image (lower).

However, the convolution kernels are usually precalculated matrices quite smaller than the image. Some examples of Laplacian kernels are the following matrices:

0	1	0
1	-4	1
0	1	0

1	1	1
1	-8	1
1	1	1

-1	2	-1
2	-4	2
-1	2	-1

On the other hand, since Laplacian kernels are an approximation of the second derivative of the image, they are very sensitive to noise. Because of this reason, the actual kernel that it is used to enhance the

spectrogram, is a Gaussian smoothed version of the Laplacian. This is, a Gaussian function has been convolved in advance with the Laplacian kernel giving a convolution matrix that will reduce the high frequency noise prior to the differentiation effect of the Laplacian. This kernel is known in image processing as *Laplacian of Gaussian* (LoG), and its 2-dimensional expression with Gaussian standard deviation σ and centered in zero is given by:

$$LoG(x, y) = -\frac{1}{\pi\sigma^4} \left[1 - \frac{x^2 + y^2}{2\sigma^2} \right] \exp \left(-\frac{x^2 + y^2}{2\sigma^2} \right) \quad (12)$$

Figure 4.8 shows the *LoG* kernel used to enhance the spectrograms of the audio signals. This 8-point kernel is pre-calculated with $\sigma = 0.5$, and normalized, so that the convolution operation can use it immediately to perform the filtering.

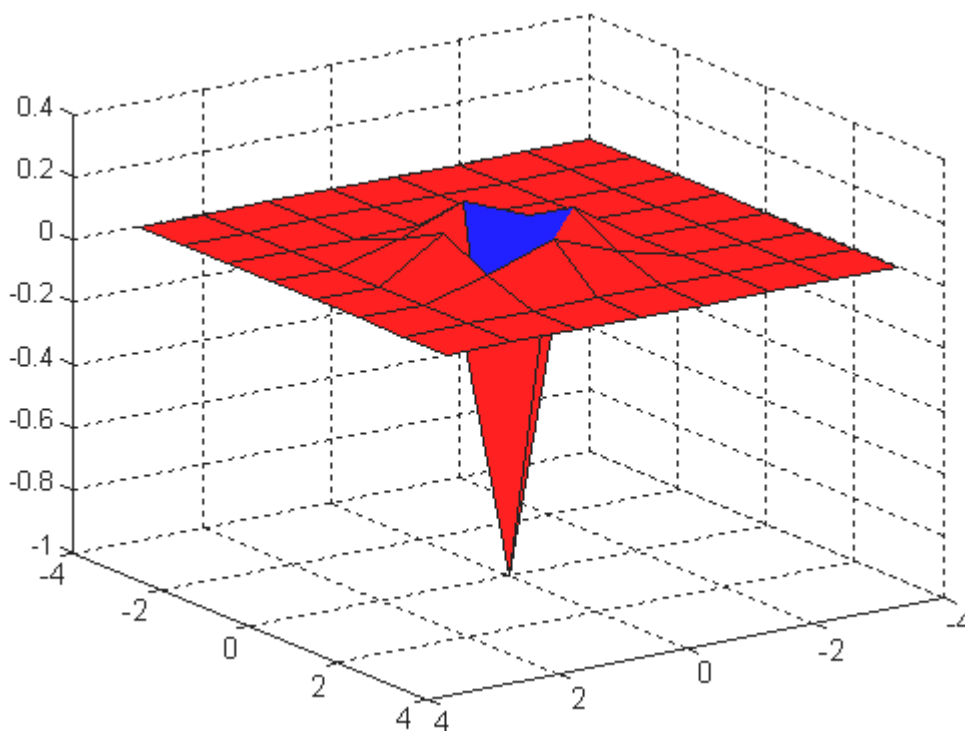


Figure 4.8. Laplacian of Gaussian kernel used for spectrogram enhancement.

Once the LoG kernel is computed, the two dimensional convolution with the magnitude of the power spectrogram matrix, is obtained as follows:

$$L_m(i, j) = \sum_{k_1=1}^m \sum_{k_2=1}^n |S(k_1, k_2)| K(i - k_1, j - k_2) \quad (13)$$

where:

- $K(i - k_1, j - k_2)$ is the kernel, m and n are the rows and the columns of the kernel, and in this case, $i = 1, 2, \dots, M - m$, and $j = 1, 2, \dots, N - n$.

The two-dimensional convolution operation returns a matrix of size $[M + m - 1, N + n - 1]$. Therefore, we take the central part of the convolution that has the same size as the original spectrogram so that both matrices can be normalized and summed.

The enhanced version of the original spectrogram is thus given by:

$$S_{m,L} = \text{mod}(S'_{m,ij}) + \alpha \text{mod}(L'_{m,ij}) \quad (14)$$

where:

$$S'_{m,ij}(S_m) = \frac{S_{m,ik}}{\max(S_m)}; \quad \text{and} \quad L'_{m,ij}(L_m) = \frac{L_{m,ik}}{\max(L_m)} \quad (15)(16)$$

S_m is the corresponding m -th sub-band spectrogram matrix scaled in dBs that was obtained after dividing the signal into frequency bands. L_m is the Laplacian filtered version of S_m . S' and L' are the respective normalized matrices of S_m and L_m . \mapsto is a scaling factor to control the level of L_m .

The enhanced spectrogram matrix has the same as the original spectrogram of the signal, except that now the edges of the abrupt changes of energy in the complete frequency band, have been highlighted, and this effect will be reflected in the subsequent processing where the frequency centroid function will be computed in order to start the labeling of the detected beats as hypotheses for possible true beats.

4.3 Beat extraction

Beat extraction aims to detect the raw beats and give as an output, a signal with labels at the positions where supposed beats were detected. These labels will serve as observations for generating probabilistic hypotheses of tempi, and as evidences for the Bayesian probability network to infer the actual tempo as well as to keep track of both, the beats and tempo of the audio signal that is in progress.

The output labeled signal is a binary train of pulses generated by a threshold stage that works based on the statistics of the differentiated frequency centroid signal. Figure 4.9 illustrates the procedure for the beat extraction task.

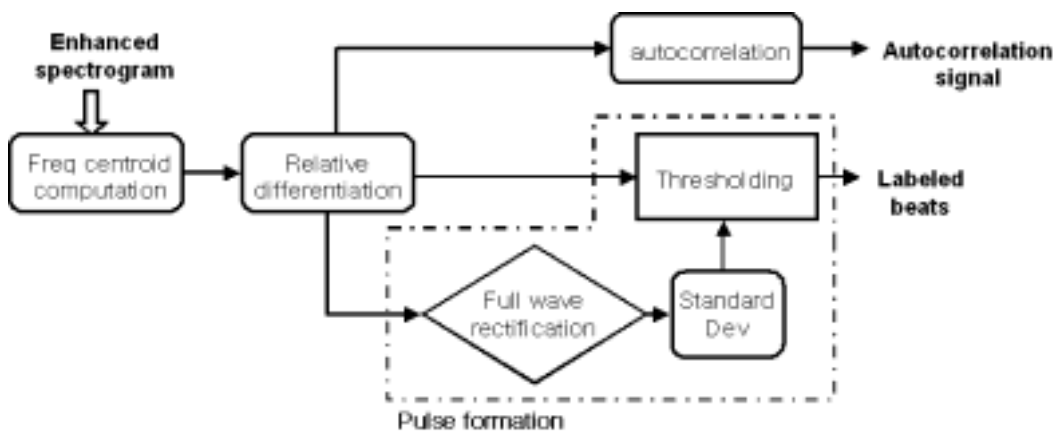


Figure 4.9. Algorithm for the beat extraction.

After enhancing the spectrogram, the frequency centroid is calculated for all the frames that compose the 2 sec. enhanced spectrogram. In order to detect the changes in energy distribution reflected on the frequency centroid function, we take its first derivative and normalize it to its own magnitude so that we small variations are taken into account as well. This *relative differentiation* has been used before in [Kla99] but applied to noticeable differences of sound intensities. The next step is to generate the binary train of pulses (labels of the beats) by applying a

threshold to the differentiated signal. The autocorrelation of this last signal is also estimated and passed to the Bayesian network so as to serve as prior knowledge when evaluating the tempo hypotheses. The measure of the tempo from the autocorrelation signal is not evaluated nor propagated to the next processes of the tempo inference but serves as a starting point of reference for the network.

4.3.1 Frequency centroid of the enhanced spectrogram

The *frequency centroid* is a measure that indicates where the center of balance of spectral power is located. In some literatures related to audio features extraction, the frequency centroid is referred as the *timbre* or *brightness* of the sound (for example in [Li00]). From here, *frequency centroid* or simply *centroid* will be used indistinctively.

Let $S_L(i, t_k)$ represent the matrix of the enhanced spectrogram. Then the frequency centroid of the 2 sec. segment spectrogram at every moment indicated by the time index t_k , is estimated as follows:

$$C(t_k) = \frac{\sum_{i=1}^N |S_L(i, t_k)|^2 f_i}{\sum_{i=1}^N |S_L(i, t_k)|^2} \quad (17)$$

where N is the total number of frequency bins.

As can be seen in the last equation, the frequency centroid is sensible to the energy present in the frequency bins f_i . Furthermore, most of the transients of the note onsets of music instruments have an interval of time in which noise of wideband frequency is present. This wideband noise will produce high peaks in the centroid signal. Thus, the effect of the Laplacian enhancement over the spectrogram was, to make the energy distribution changes of the transients more prominent, so that even beats with low energy will produce noticeable unbalances that can be detected from the centroid function.

4.3.2 Relative differentiation applied to the frequency centroid signal

Since the centroid function represents the changes in energy distribution, we are now interested in these differences at all moments of discrete time. Therefore, we take the first derivative of the centroid signal. But as mentioned before, some peaks might be higher than others (and in fact this occurs), making the implementation of a threshold stage more complicated and increasing the probability of false detections or misses of true beats. In order to reduce this effect, we normalize the derivatives of the centroid by its magnitude value. If $C(t_k)$ is the frequency centroid at time instants t_k , the relative differential is given as follows:

$$R(C(t_k)) = \frac{d[C(t_k)]/dt_k}{C(t_k)} \quad (18)$$

this is equivalent to take the derivative of the logarithm of the centroid:

$$R(C(t_k)) = \frac{d[\log(C(t_k))]}{dt_k} \quad (19a)$$

Then, substituting equation (17) in (19a), we have:

$$R(C(t_k)) = \frac{d}{dt_k} \left[\log \left(\frac{\sum_{i=1}^N |S_L(i, t_k)|^2 f_i}{\sum_{i=1}^N |S_L(i, t_k)|^2} \right) \right] \quad (19b)$$

arriving to:

$$R(C(t_k)) = \frac{d}{dt_k} \left[\log \left(\sum_{i=1}^N |S_L(i, t_k)|^2 f_i \right) - \log \left(\sum_{i=1}^N |S_L(i, t_k)|^2 \right) \right] \quad (20)$$

From this signal, the autocorrelation and the binary train of pulses are computed next.

4.3.3 Beat labeling (Pulses formation)

Before starting the labeling process, the relative differentiated signal is normalized to have mean zero. Since beats with relatively high energy at the lower frequencies will produce negative peaks in the normalized differentiated signal, both positive and negative peaks, have to be considered for the beat labeling process, and therefore, after normalization, the signal is full wave rectified. Then, its standard deviation is estimated. This parameter is one of the criteria for the threshold step, and is calculated as follows:

$$s(R_{nor}) = \sqrt{\frac{1}{N-1} \sum_{k=1}^N (R_{nor}(t_k))^2} \quad (21)$$

where the mean value has been set to zero after the normalization of the signal. R_{nor} is $R(C(t_k))$ normalized; N the total number of samples within the 2 sec. segment and t_k the discrete time index.

Hence, the threshold is applied under the following considerations: knowing that the signal has zero mean, we first detect zero-crossings; thus the first condition is that $R(t_k)R(t_{k-1}) < 0$; now, in order to differentiate the peaks produced by real note onsets from the small peaks produced by noise, we compare the rate of increment (or decrement) of the signal $R(t_k)$ with the value of the standard deviation of all the peaks; $|R(t_k) - R(t_{k-1})| \geq s(R_{nor})$. Combining these conditions, we have:

$$[R(t_k)R(t_{k-1})] |R(t_k) - R(t_{k-1})| \geq s(R_{norm}) \quad (22)$$

Since the first condition states that the product $R(t_k)R(t_{k-1})$ be less than zero when a zero-crossing is detected, Equation (22) becomes:

$$[R(t_k)R(t_{k-1})] |R(t_k) - R(t_{k-1})| \leq -s(R_{norm}) \quad (23)$$

or

$$R(t_k)R(t_{k-1}) |R(t_k) - R(t_{k-1})| + s(R_{norm}) \leq 0 \quad (24)$$

However, it was found that some onsets produce high peaks without crossing the zero level. For such a case, a third condition was considered for thresholding:

$$\left| R(t_k) - R(t_{k-1}) \right| \geq \beta s(R_{norm}) \quad (25)$$

where after experiments, this condition was determined to work well with $\downarrow = 2.9$.

Finally, the binary signal $P(t_k)$ is constructed as follows:

$$P(t_k) = \begin{cases} 1 & \left\{ \begin{array}{l} R(t_k)R(t_{k-1})\left| R(t_k) - R(t_{k-1}) \right| + s(R_{norm}) \leq 0; \text{ or} \\ \left| R(t_k) - R(t_{k-1}) \right| \geq \beta s(R_{norm}) \end{array} \right. \\ 0 & \text{other case} \end{cases} \quad (26)$$

In the last signal, a label of 1 represents the place where a probable onset is located, and evidently a zero indicates the absence of onsets.

Now we illustrate in Figure 4.10, an example of a 2 sec. segment spectrogram, together with the edge enhanced spectrogram and the signals at different stages of the Preprocessing and Beat extraction blocks. The sample audio segment corresponds to a pop music song ("Fields of gold" by Sting) analyzed at the higher frequency band.

As it can be observed in Figure 4.10, after the spectrogram enhancement process, the transients have been highlighted in relation to the relaxation interval of the music note. As expected, the enhanced transients produced higher peaks in the frequency centroid signal that are detectable after applying the relative differentiation. In the binary signal of Figure 4.10, the threshold has been applied giving as a result labels at the positions of possible true beats and at its neighborhoods, indicating that the actual beat is within that vicinity.

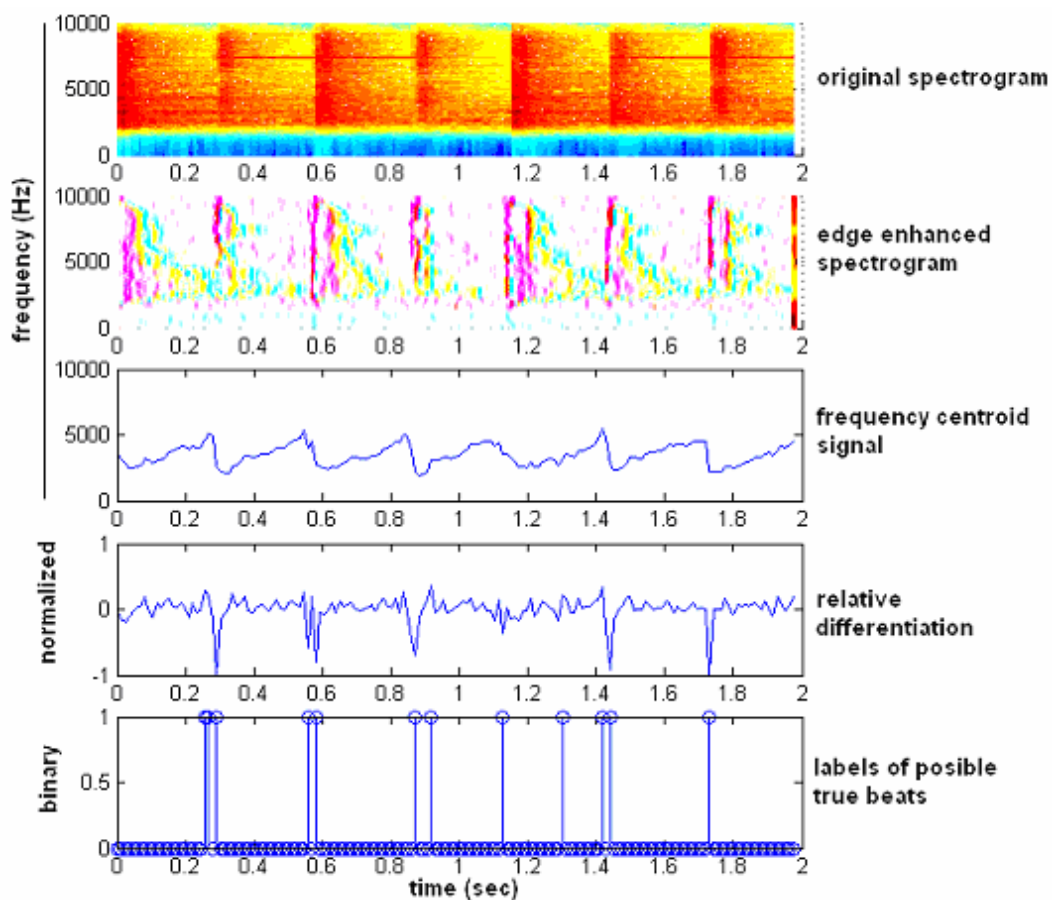


Figure 4.10. Example of signals in the Preprocessing and Beat extraction blocks at the higher frequency sub-band. Audio sample: 2 sec. segment of a pop music song (“Fields of gold” by Sting).

One could argue that at this point, a first approximation to the tempo can be easily estimated by measuring the first equally spaced pulses of this binary signal or by computing the autocorrelation and finding the first peak that, by concept, represents one periodicity of the autocorrelated signal. In fact, and as it was mentioned before, the autocorrelation of the signal $R(t_k)$ is calculated so as to function as a first reference for the Bayesian network when starting to track the tempo as the audio signal progresses in time. Nevertheless, the binary signal and the autocorrelation signal as well, still reveal spurious onsets that, in the case of measuring the tempo from the binary train of pulses, will lead to ambiguities and uncertainties about what the actual tempo is, and in the autocorrelation function, strong peaks at erroneous periods

of tempo will appear. Hence, this problem motivates the implementation of a probability network model that will have the main task of evaluating the hypotheses of true beats and tempo, and determine which is the most likely to represent the actual tempo and beats of the current audio signal.

4.4 A Bayesian network model for beat and tempo tracking

In this section, I will present a model based on Bayesian probability theory that aims the task of hypotheses evaluation, and by observing the beat labels, it will try first, to infer the actual tempo, and then to keep track of the beats and possible changes of tempo. The general network structure is recalled in figure 4.11

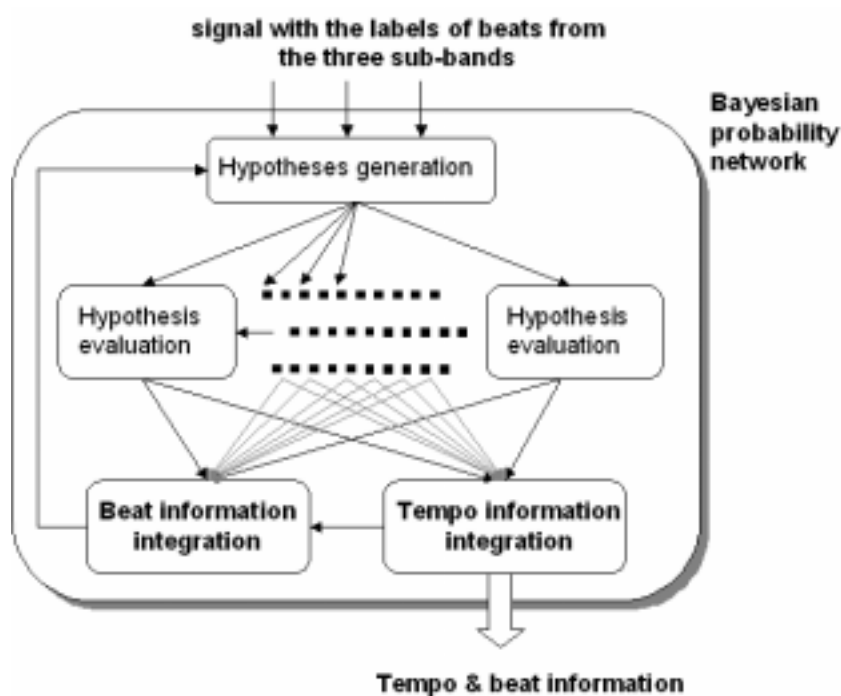


Figure 4.11. Probability network for beat and tempo tracking.

As can be seen in the figure above, the evaluation process starts with the generation of hypotheses. The first hypotheses generated are those of tempo. The system, tries to infer the tempo and then with the actual

tempo estimated, it will try to track the beats. However, as will be explain in detail later, the inference of tempo is performed depending on the hypotheses of beats. Therefore, the evaluation of beat hypotheses is followed by the evaluation of tempo hypotheses.

In Figure 4.11, the Hypotheses evaluation blocks are the core of this network for tempo inference. As a Bayesian network, the Hypotheses evaluation network is constituted by a number of nodes that represent a set of variables interconnected by their conditional probabilities. The variables in each node describe the situations of the probability of an actual tempo and beat in light of the evidences observed from the hypotheses. For example, the probability of true tempo is higher when the tempo hypothesis is confirmed by the presence of a beat label at the time location where this tempo hypothesis indicates a beat should be detected. More details of the nodes will be explained later.

Hence, the inference of tempo in this network is made purely based on the observations of the hypotheses generated, and the labels of possible true beats. In the other hand, we can mention to this respect that the network makes no assumptions of any musical knowledge prior to the evaluation. In contrast, other approaches perform beat and tempo tracking under some assumptions such as, the presence of patterns that follow certain music rules like specific time signature and chord change (as in [GM95]), and drum sounds in the audio signal (as in [HK02]), or that the user can enter some music parameters to the system (as in [DGW02]). The only assumption in the model of this work is that a pattern of repetitive acoustic events is present in the audio signal that is being processed. This follows the idea of the fact that humans without any musical knowledge can find the beats of the music they are listening to, and are able to tap to the same rhythm as that performance progresses. Even if they are listening to a single sound source in the absence of music context, the timing sequence of that sound can be recognized. Therefore, the model proposed here, is intended to track beats in sounds not only from musical instruments but also from any sound source that produces repetitive time limited

acoustic events, and their inter onset intervals are bigger than the time resolution of this system (20 ms).

The multiple evaluations of the hypotheses are finally integrated in the Beat information integration and Tempo information integration blocks respectively. In these blocks, a table with the results of the evaluation is constructed, and now the search for the hypothesis that has the highest probability to represent the true tempo is performed. As soon as the tempo is found, the beat tracking task starts. For this purpose, the Beat information integration block has available the table with the results of the evaluations of beat hypotheses. By searching in this table and taking into account the tempo already inferred, the beats are tracked from the train of pulses that contain the labels of the raw beats detected by the previous Beat extraction blocks. In other words, with the information of the detected tempo, the system is now able to predict the next beats with less uncertainty. However, the hypotheses evaluation process continues as the system is fed with a new segment of the next 2 sec. of the audio signal. Once a segment of music is input to the system, the signal processes and hypothesis evaluations are repeated.

In the next section, I will discuss in detail the structure of the Bayesian network and the process of the tempo inference.

4.4.1 Hypotheses generation

The binary signal from the Beat extraction stage is the first symbolic representation of the beats found in the audio signal and the basis for the Bayesian network to infer and track the tempo. Therefore, the hypotheses of tempo are generated from this signal. In order to determine the possible tempi that are implicit in the audio signal, the system looks at the inter onset intervals of the binary signal. The inter onset intervals are measured and these measures are then arranged in the form of a histogram. The first onset of the binary signal is taken as the reference onset and the interval to the next onset is measured; with

the same onset reference, the interval to the second onset is taken, then to the third, and so on until the interval to the last onset in the 2 sec. binary signal is measured. The second onset is now taken as reference and the same forward measurement is performed. This process of measurement is repeated until all the onsets have been taken as references, as shown in Figure 4.12. However, there is a restriction when executing these measures, if the interval between the reference and the first consecutive onset is less than three time indexes, the measure is discarded. The reason is that, it was found that onsets labels that are located less or equal to two time indexes, are likely to represent the same onset of the original audio signal, and the consideration of such small intervals lead the network to make false inferences of tempi. Figure 4.13 show the loop steps for the tempo hypotheses generation.

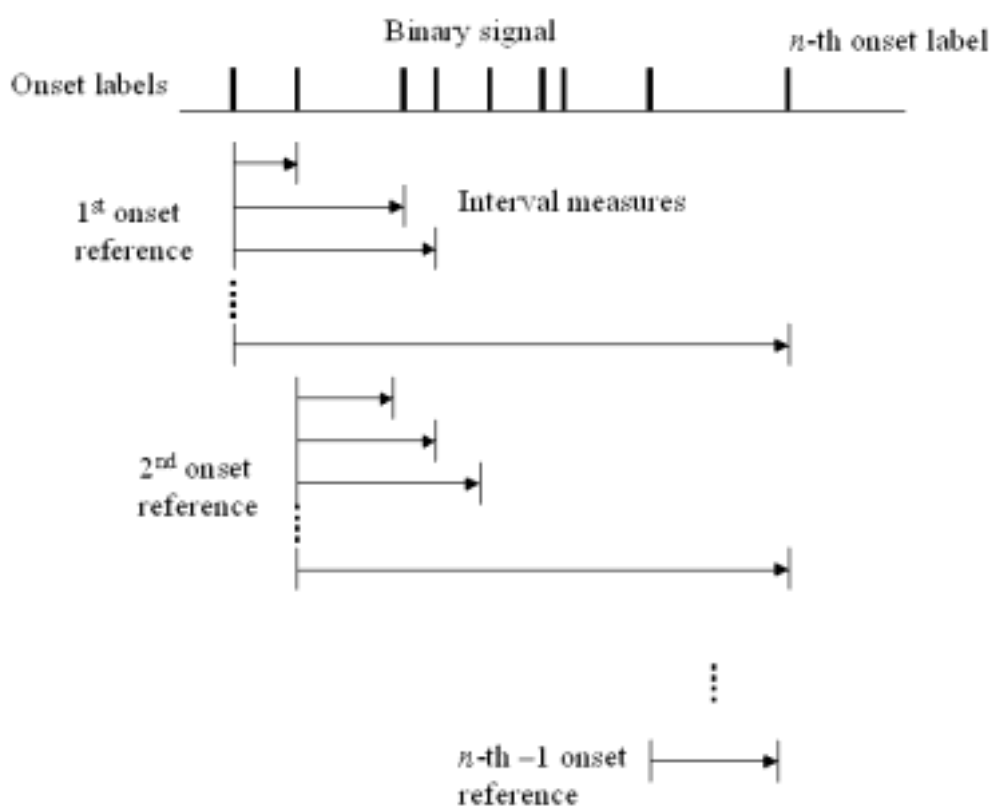


Figure 4.12. Inter onset intervals measurement.

```

for  $i = 1, 2, 3, \dots, n\text{-th-1}$  onset label
    take the  $i$ -th onset as reference
    for  $j = i\text{-th onset label} + 1$ 
        measure the interval between the
        reference onset and the  $j$ -th onset
        if the measure  $<$  three time indexes
            discard the current measure
        else store the measure
    end
end
histogram the measures stored

```

Figure 4.13. Algorithm for tempo hypotheses generation.

This algorithm of tempo hypotheses generation is applied to the three binary signals of the frequency sub-bands.

For the beat hypotheses generation, the process is simpler. The onset labels from the three frequency sub-bands are integrated, and matches of onset labels at the same location in the sub-bands are marked by the sum of the number of matching labels.

Figure 4.14 shows an example of histograms of tempo hypotheses at the three sub-bands, and Figure 4.15 depicts an example of beat hypotheses signal.

We should note in Figure 4.14 that the scale of tempo in beats per minute (bpm) ranges from 30 to infinite, however, the maximum tempo that the system can recognize is determined by its resolution in time. The range infinite was written just to be consistent with both scales, the one for IOI (in sec) and the one for tempo (in bpm).

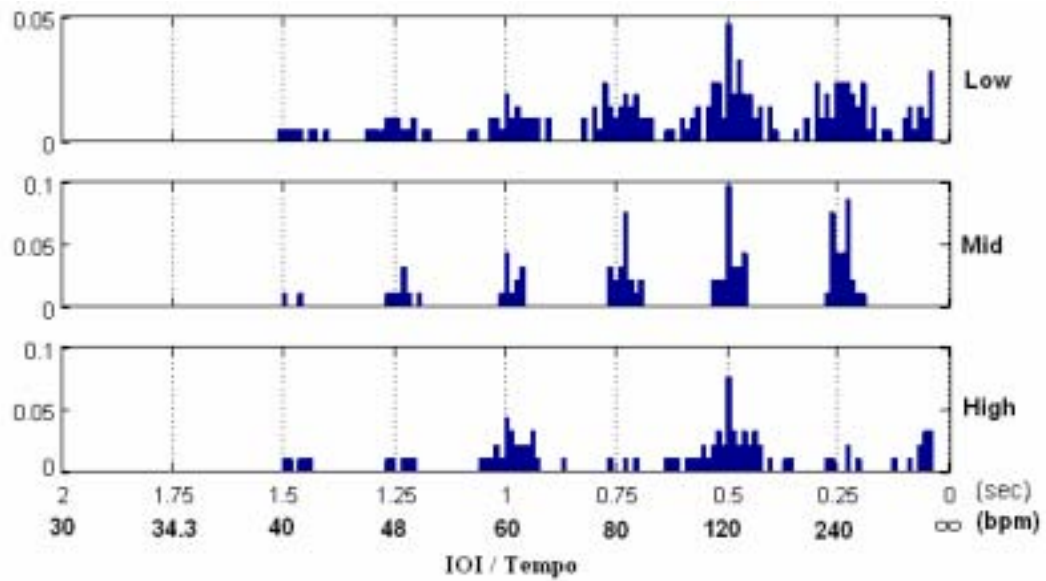


Figure 4.14. Example of histograms of tempo hypotheses at the Low, Mid (Bandpass) and High pass frequency sub-bands. The horizontal scale is given in seconds for the measures of Inter Onset Intervals (IOI), and in beats per minute (bpm) for the hypothetical tempi.

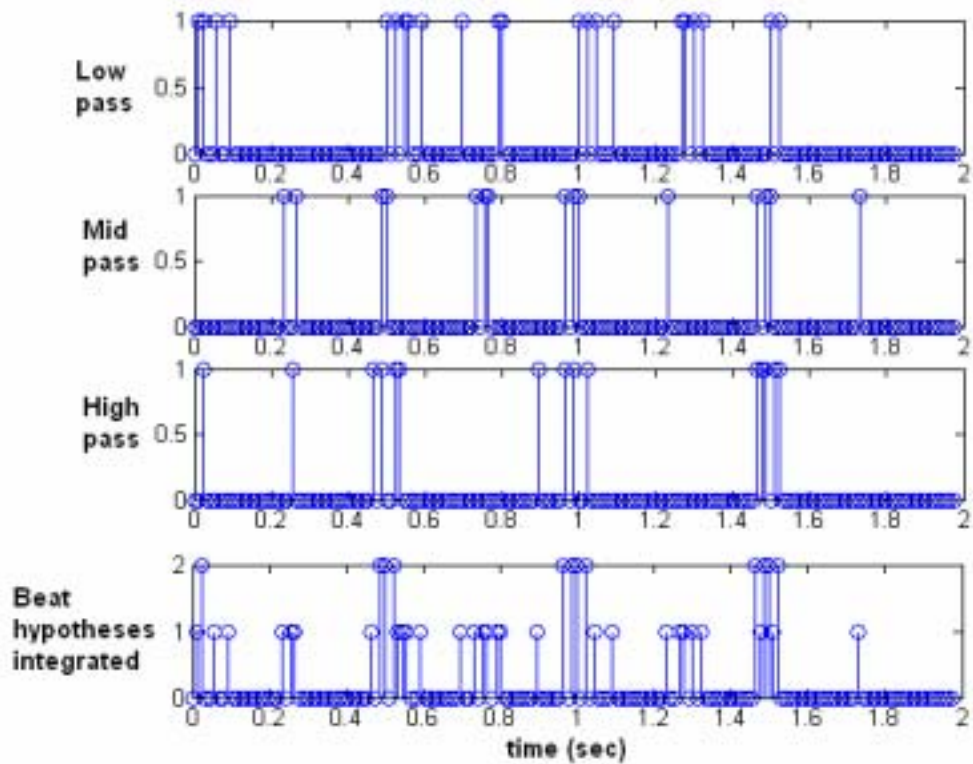


Figure 4.15. Example of beat labels at the sub-bands and the final beat hypotheses.

Another point to note in Figure 4.14 is the vertical scale of the histograms. The histograms have been normalized to the total number of measures in each sub-band so that the sum of the values in the histogram is 1.

4.4.2 Hypotheses evaluation network

The hypotheses evaluation network is composed by a number of nodes that encode the variables representing the different situations of the observed data that can explain the belief of the beat and tempo hypotheses in evaluation. Every hypothesis generated in the previous block, is evaluated according to the evidences that correspond to that particular hypotheses. This means that, the evaluation of the hypotheses within the same 2 sec. segment is assumed to be independent from one to another. However, the dependency of the subsequent new set of hypotheses is kept in order to maintain a consistency of tempo tracking as the audio signal progresses in time. As the new set of hypotheses (from the new 2 sec. segment of the audio signal) becomes available, the evaluation process is repeated but now the antecedent of the previous estimated tempo is considered. Thus, the evaluations are related in time by previous tempo and beats estimations made in their corresponding antecedent evaluation process.

Figure 4.16 shows the network nodes used to perform the evaluation of a single hypothesis.

Lets suppose that we want to test the belief of the True beat and True tempo hypotheses. As mention before, the first hypothesis tested is that of the True beat. This hypothesis depends on the fact that whether the beat in evaluation was detected in more than one frequency sub-band (labeled as Beat aligned in the network of Figure 4.16). If it is found that the same beat is present in the binary onset labels of not only one

sub-band but also in another one or in all of them, the probability of the beat hypothesis in evaluation to be an actual beat increases.

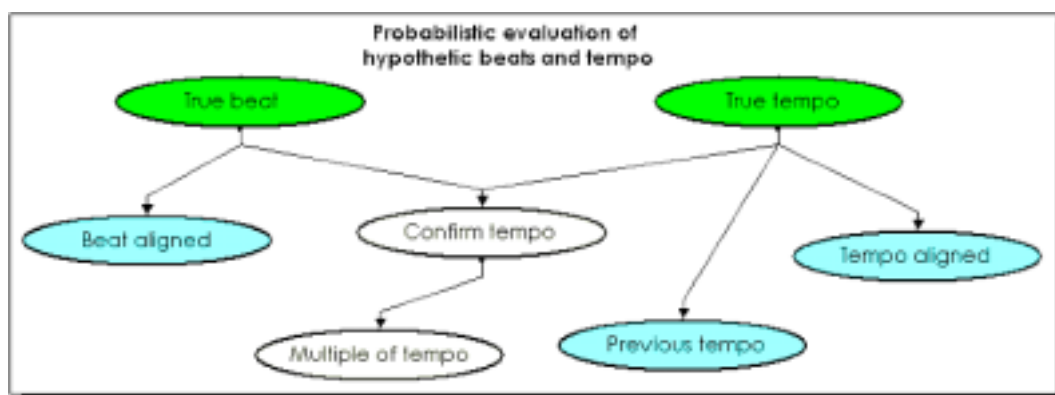


Figure 4.16. Bayesian network nodes for the hypotheses evaluation.

Nevertheless, the probability of the true beat does not rely only on the beat alignment, if the beat hypothesis in evaluation is also found to be located at the time where a beat is expected to be present, its probability to be a true beat is higher as well. In other words, the beat hypothesis confirms the actual tempo or an integer multiple of the true tempo. Thus, it is here where the information obtained from the autocorrelation is considered. Since at this moment, the current tempo hypothesis has not been tested yet, the immediate reference that the network takes is the tempo hypothesis given by the autocorrelation. As soon as the True tempo hypothesis in evaluation is tested, the probabilities of the network are updated with this new evidence.

In order to enter the evidence of tempo confirmation, the system searches for the beats that are supposed to be within the margins of the hypothetic tempo, taking as reference the hypothetic beat in evaluation. If the search does not match at least 80% of the current tempo predictions, the system looks now for matches to integer submultiples, starting with half of the tempo. Then, the search runs until the 4th integer submultiple. After this search, the evidences are entered to the corresponding nodes to update the probabilities. This process is illustrated in Figure 4.17.

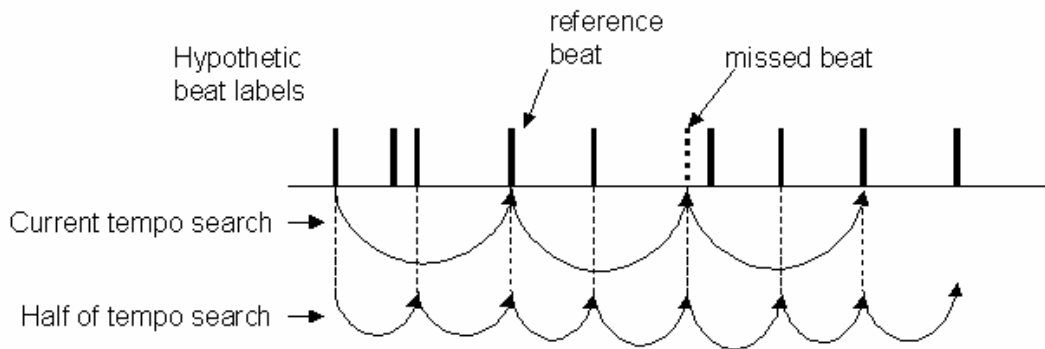


Figure 4.17. Search for tempo confirmation.

Once the Beat hypotheses have been tested for the current hypothetical tempo, now the evaluation of the True tempo hypothesis is performed.

Similar to the case of beat test, the tempo evaluation depends also on its confirmation by true beats at the current tempo level or at submultiples. However, in addition to this test, evidences of previous tempo estimations and same tempo hypotheses detected in more than one frequency sub-bands are also considered. The Previous tempo node in Figure 4.16, evaluates whether the tempo hypothesis in evaluation is similar to the true tempo inferred in tempo evaluations of previous sets of hypotheses. If the evaluation in process is the first evaluation that the system performs to the input audio signal, the evidence of previous tempo is not tested since evidently there is no previous tempo detected.

The tempo evaluation continues with the Tempo aligned test. This evidence corresponds to the fact of whether the same tempo hypothesis has been generated not only in one frequency band but also in the others. If this evidence shows to be true, consequently the probability of the True tempo is increased.

When the evidences are determined, the probabilities of the network are updated in order to estimate the final belief of the beat and tempo hypotheses in evaluation. This value is then stored.

In general, the evaluations at the Bayesian network are performed for all of the beat hypotheses with each of the tempo hypotheses. And the results of the evaluations are arranged in a table of belief values.

4.4.3 Beat and Tempo information integration

The next step is the integration of all the information given by the probability network. First the tempo hypothesis that is most likely to represent the actual tempo of the audio signal is searched in the table of beliefs. The tempo with the highest belief value is determined to be the actual tempo. With the final value of tempo estimated, the beats that belong to this tempo are integrated and confirmed with the original onset labels given by each of the frequency bands. This last operation aims to keep track of the beats present in the 2 sec. segment audio signal.

4.4.4 Propagation of the tempo knowledge

The process of evaluation discussed until now is that of the process in a 2 sec. segment of the input signal. The evaluation of new sets of hypotheses is repeated for the next incoming 2 sec. segment, and the operation continues in this way. Furthermore, the tempo values estimated in previous evaluations are propagated to future evaluations, resulting in a network that is related in time by the history of tempi estimated in past segments of analyzed signal, as can be appreciated in Figure 4.18. In the evaluation, the Previous tempo node of the network takes the tempo value that is being propagated as reference to compare with the current tempo hypothesis in order to determine the evidence that will be entered to the network so that the probabilities are immediately updated.

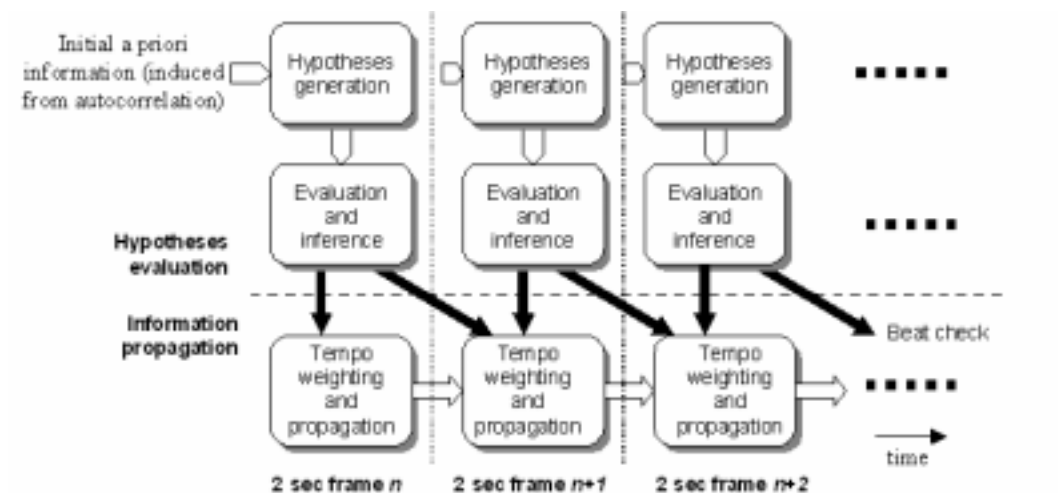


Figure 4.18. Relationship of the evaluation processes along time.

The system updates the tempo and beat tracking every time that it processes a 2 sec. segment of music. In other words, it takes a maximum of 2 sec. for the system to perform signal processing and probability computations in order to determine the tempo that is implicit in the input signal. This feature is comparable to the ability of humans who in the worst of the cases take more than 2 sec. of listening to the music before they can recognize the tempo and keep track of it. And since normally real music performances are longer than 2 sec., the system is supposed to be able to keep track of the beat and tempo in real-time mode.

Chapter 5 Model performance

As in all implementations of systems design to perform a task, the criterion to determine their proper functioning is through the evaluation of their performance. Hence, in the following paragraphs, I will present the method for the evaluation of the beat and tempo tracking system introduced in the previous chapter. Then, I will discuss the results and state some comments.

5.1 Method of evaluation

As it is discussed in the Appendix, during the training of the system, partial tests were performed with a short set of music samples. However in order to evaluate the overall performance of the system, the number of music samples was extended and prepared for the tests.

The database of music samples consisted of a set of 86 segments with a length of 30 sec. recorded in standard Windows PCM at a sampling frequency of 44.1 kHz, 16 bits of resolution, and stored in mono aural WAV format files. Some music samples were taken from [RWC01] which is a music database compilation developed for research purposes, and another set of samples were taken from commercial audio CD's. In the selection of the music samples, as much music genres as possible were tried to be considered so that the behavior of the system could be observed and evaluated. Note that genres such as African and Latin pop music which have not been considered in the performance tests of previous approaches, have been included in the database for the evaluation of my model.

In order to have a reference of comparison, the music samples have been prepared before tested in this model. In the preparation process, the beats have been labeled manually in the individual music sample segments by using a professional audio editor [CEP21], and with the aid of experienced musicians. In Figure 5.1, an example window of [CEP21] is illustrated when labeling the music samples.

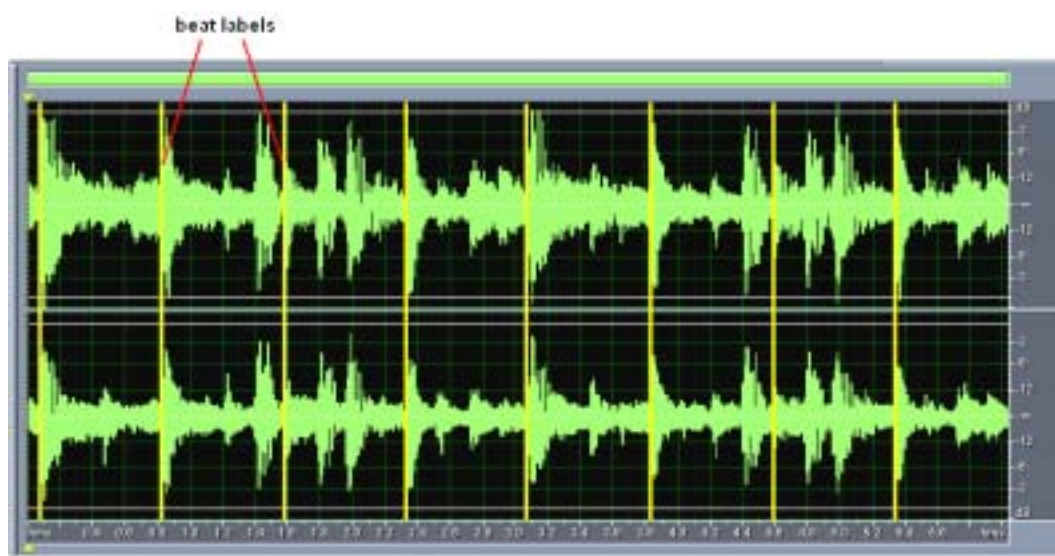


Figure 5.1. Example of beat labels in the preparation of the music samples.

In the Figure above, it can be observed that not all the onsets are labeled, but only those onsets that represent the beats defining the tempo have been labeled. To this aim, the participation of the experienced musicians was crucial when performing the discrimination among the onsets that are not in line with the metric of basic tempo.

With the labeled music samples, the tempo is pre-estimated by measuring the intervals between the beats and the average of the distances is taken to calculate the actual tempo of the sample.

The music samples were then fed to the system and the tests performed. Figure 5.2 depicts a preliminary window interface of the beat and tempo tracking system.

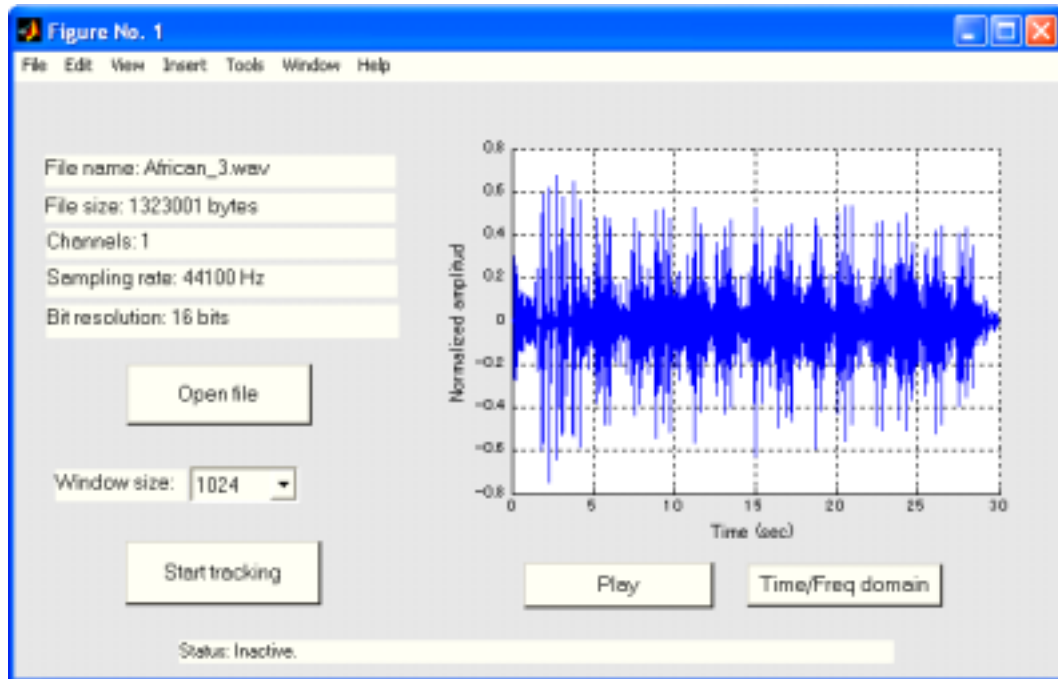


Figure 5.2. Sample window interface of the beat and tempo tracking system.

In the literatures on music beat and tempo tracking, different methods for the evaluation have been proposed. For example, in [GM97], the authors discuss the aspects to take into account when performing tests in beat tracking systems. They present advanced techniques for evaluation of their system and introduce a measure of the deviation of errors. In a more recent work, [CDK01] proposed another measurement for their beat tracker based on a Gaussian observation window. In [Sch00] the author performed simpler tests by comparing the results of his tempo tracker with the results of experiments performed with professional musicians. Since these methods vary according to the implemented system, the measure that I will use here is the standard rate of correct asserts against failures. This kind of measurement allows us to make general comparisons with other systems.

Thus, from the tests performed in the system, the beat tracking accuracy (*BTA*) was evaluated as follows:

$$BTA = \frac{1}{N} \sum_{i=1}^N \frac{Bcd(i) - Fb(i) - Mb(i)}{Tot(i)} \times 100\% \quad (27)$$

where:

- Bcd = correct beat detections.
- Fb = False beat detections.
- Mb = Missed beats.
- Tot = number of total beats manually labeled in the sample segment.
- N = Total number of 2 sec. segments that the system processed (in this case is equal to 15 since the length of each music sample is 30 sec.)

In this measure, 100% indicates a perfect accuracy in the beat tracking task.

For the tests of accuracy in tempo deviation (AT), the ratio was estimated as follows:

$$AT = \left| \frac{1}{N} \sum_{i=1}^N 1 - \frac{Tsys(i)}{Tm(i)} \right| \times 100\% \quad (28)$$

where:

- $Tsys$ = Tempo determined by that system within the i -th 2 sec. segment of music.
- Tm = Tempo found manually in the corresponding i -th 2 sec. segment of music.
- N = Total number of 2 sec. segments that the system processed (in this case is equal to 15 since the length of each music sample is 30 sec.)

In this case, a value that tends to 0 % indicates a perfect accuracy in the tempo tracking operation.

5.2 Results

With the compilation of 86 music samples, promising results were achieved from the tests of the system. Table 5.1 summarized these results. In this table, only the categories in which the music corpus was arranged are shown.

Music category	No. of samples tested	Average Beat tracking accuracy (BTA)	Average Tempo tracking deviation (AT)
African	3	83.4 %	2.6 %
BigBand	3	71.3 %	23. %
Bossa Nova	2	81.8 %	2.8 %
Classic	8	76.2 %	10.0 %
Country	3	90.0 %	4.3 %
Dance/Techno	12	98.2 %	1.2 %
Folk	3	89.5 %	8.1 %
Funk	2	88.7 %	5.2 %
Jazz	8	92.4 %	3.7 %
Latin	6	93.1 %	3.8 %
Pop	17	91.6 %	2.4%
Reggae	7	92.5 %	2.8%
Rock	12	86.4 %	8.5%
Total	86	Average: 87.3 %	Average: 6.0 %

Table 5.1. Average results from the tests of the beat and tempo tracking system.

5.3 Discussion

The tests performed in the evaluation of the systems reflect somehow the general performance of the model. However, due to the simplicity of the measurement method, it is difficult to determine the causes of the

errors and deviation while performing the tracking task. If we want to estimate with more accuracy the behavior of the system in different situations, we would have to consider other factors that lead the system to respond to certain kinds of music stimuli. For example, during the tests, it was found that signals that contain high levels of voiced sounds, tended to confuse the system. This effect can be observed in Table 5.1 in the categories of BigBand, African and Bossa Nova music. In BigBand music, the trumpet sounds are considerably as strong or if not stronger than the voiced sounds, leading to a poor level of smooth tracking. Similar is the case of Classic music, in which violins, flutes, trumpets, etc. with long duration tend to mask the instruments that mark the metric of the music composition. The explanation to these deviations can be attributed to the frequency centroid function. As I mentioned in previous Chapters, the frequency centroid is sensitive to changes in spectral energy distribution, specially at high frequencies where the change of energy produces a higher peak in the centroid signal than if the same change occurs at lower frequencies. Although the accuracy of tracking of the mentioned music genres (BigBand, African, Bissa Nova) is not as good as the rest of the genres, they are genres that have not been included in the databases of the evaluation of other beat and tempo tracking systems (Latin pop is another of this genres). In this sense, the system proposed here, has proved to work with a wider range of music genres, in contrast with other systems that typically consider at most five genres, or even less, they specialize in one genre, such as the system in [GM95] that its best performance is achieved when processing Pop music signals in a metric signature of 4/4.

In the rest of the genres, the system achieved beat tracking levels of 90% or higher, and tempo deviations less than 8.3 %, rate that was determined for this system to produce deviations in tempo perceptible by human ears.

The parameters that the system requires to the user to enter are another point to consider. Although the parameters and coefficients used in the functional blocks of this model were determined empirically,

they have shown to enable the system to achieve satisfactory results. Compared to other models that require some parameters to be specified by the user in order to perform well, as in [DGW02], parameters that might be unfamiliar to the user.

When implementing the model, the portability of the system was another aspect to mention. This algorithm was thought to be portable to any PC, without requiring huge power of hardware. The tests were performed in a Celeron equipment running at 800 MHz with 128 MB of RAM. Specifications that nowadays seem trivial compared to the capabilities of new computer equipments. In contrast to the model of [GM95] that runs in a distributed memory system using a specific music protocol developed by the authors. Thus, the simplicity of the model proposed in this thesis makes the system suitable to be incorporated in applications that require real-time beat and tempo tracking, however, as a drawback, at least 2 sec. of audio signal is needed in order to perform tempo inference. However, usually real world music signals are longer than 2 sec.

In general, the implementation and tests of the model have shown the viability for further improvements in this approach. The performance of the system can be improved by expanding the probability network so as to take into account more factors such as instantaneous tempo variations, implicit beats, changes in time signature and changes in chord. With these expansions the evaluation network would be provided with more evidences that could lead to inferences of true tempo with more reliability. Furthermore, the system would be more sensitive to changes in tempo and beat tracking.

Chapter 6 Conclusions

In this thesis, a new approach to the problem of music beat and tempo tracking in audio signals was presented. I have shown that the connection of the variety of theories on, audio signal spectral analysis, image processing and Bayesian probability networks, represent a viable solution for the aim of beat and tempo tracking task that has been increasingly studied in the last few years by researchers in the automatic music recognition field. The basic techniques used in this work are theories implemented in other works and have been widely used in other areas. Thus, my contribution resides, in the particularization and adaptation of these techniques to work together with the objective of processing the audio signals and discover the timing sequence of its musical events. Such an example of these techniques is the case of the image processing algorithm used here to enhance the transients on the note onsets. In none of the literatures published before, had been reported the use of specialized image processing tools applied to audio signal analysis in order to recognize musical feature patterns. On the other hand, the model of the Bayesian network proposed here is another particular contribution.

Recalling the framework of my approach, the model is formed by three main sections, the Preprocessing, Beat extraction and the Probability network:

- a) In the Preprocessing section of this model, the signal is decomposed into three frequency sub-bands. Signal decomposition into sub-bands has been extensively used in other works differing in the

number of analysis bands and their frequency distribution. The three sub-bands employed in the model proposed, showed to be sufficient signal decomposition, without meaning that this is the optimum number of analysis bands that can be used to get the maximum robustness of the frequency centroid function for the detection of onsets. Finding the optimum number of analysis sub-bands can be the solution to improve the overall robustness of the system to keep track of the beat and tempo even in audio signals that contain high levels of voiced sounds. The convolutional kernel used to enhance the spectrogram, is another subject for improvement of the model performance. The Laplacian kernel used here as a first approach to beat detection, can be substituted by a more sophisticated matrix that will lead to better results. A good number of refined techniques from the image processing field have been developed until now, thus, the selection and adaptation of the one suitable for the purpose of beat and tempo tracking is another motivation for future work.

- b) The frequency centroid of the Beat extraction stage is a signal feature that has shown to be a good indicative of the location of the onsets. However as mention before, its drawback is precisely its higher sensitiveness to changes in energy at high frequencies. A solution to overcome this problem might be reducing the frequency range in which it can move by increasing the number of analysis frequency bands as mentioned above. This, in combination with the implementation of a more advanced threshold technique would lead undoubtedly to better beat and tempo tracking figures.

- c) The assumption of no musical knowledge of any kind has been a point considered when designing the Bayesian probability network. In this sense, the network tries to resemble the tapping ability of humans without having received any musical training. The decisions and inference performed by the network is based purely on the observations of the data produced by the signal processes

previous to the network. We can make extensions to this model while keeping this philosophy, by enabling the network to consider more information observed from the audio signal features, such as instantaneous changes in tempo, changes in chord, presence of voice, implicit expressions of beating, etc. Although the model proposed here performed satisfactorily well to some extent, there is still place for improvement in the design of this network.

In general, the design of this system considered basic rules of music events and audio signals without looking into music rules of higher level. However, from the evaluation experiments, comparable performance to those tempo trackers that incorporate more complex musical knowledge has been shown. Although the integration of higher levels of musical knowledge seems to be the tendency towards the development of more advanced intelligent systems, I have demonstrated that the inference and further abstractions of symbolic music information relies firstly on the techniques to extract and observe relevant data from the audio signals. An intelligent music transcriptor, for example, in order to “*recognize*”, it has to “*listen*” to the audio stimuli first. Here is where systems such as the one proposed in this work, come to play an important role as a sub-part of a more complex system.

References

- [AD90] P. E. Allen and R. B. Dannenberg. Tracking musical beats in real time. In *Proc Int. Computer Music Conf. (ICMC)*, pages 140 – 143, 1990.
- [CDKH01] A. T. Cemgil, B. Kappen, P. Desain and H. Honing. On tempo tracking: Tempogram representation and kalman filter. *J. New Music Research*, 2001.
- [CK03] A. T. Cemgil and B. Kappen. Monte Carlo methods for Tempo tracking and Rhythm quantization. In *J. of Artificial Intelligence Research*. Vol. 18, 45 – 81, 2003.
- [CKDH01] A. T. Cemgil, B. Kappen, P. Desain and H. Honing. On tempo tracking: Tempogram representation and kalman filter. In *J. of New Music Research*, vol. 28, No. 4, pages 259 – 273, 2001.
- [DGW02] S. Dixon, W. Goebel and G. Widmer. Real time tracking and visualization of musical expression. In *Proc. 2nd Int. Conf. ICMAI 2002, Music and Artificial Intelligence*, Springer, pages 58 – 68, September 2002.
- [DSD02] C. Dubuxury, M. Sandler and M. Davies. A hybrid approach to musical note onset detection. In *Proc. of the 5th Int. Conf. on Digital Audio Effects (DAFx-02)*, pages 33-38, September 2002.
- [DW00] P. Desain and L. Windsor. *Rhythm: Perception and Production*. Swets & Zeitlinger Publishers, Netherlands, 2000.
- [Dix01] S. Dixon. An empirical comparison of tempo trackers. In *Proc. of the 8th Brazilian Symposium on Computer Music*, pages 832 – 840, August 2001.
- [Eck01] D. Eck. A network of relaxation oscillators that finds downbeats in

rhythms. Tech. report IDSIA-06-01, IDSIA, 2001.

- [FU01] J. Foote and S. Uchihashi. The beat spectrum: A new approach to rhythm analysis. In *Proc. IEEE Int. Conf. On Multimedia and Expo*, 2001.
- [GM95] M. Goto and Y. Muraoka. A real-time beat tracking system for audio signals. In *Proc. Int. Computer Music Conf. 1995*, pages 171 – 174, September 1995.
- [GM97] M. Goto and Y. Muraoka. Issues in evaluating beat tracking systems. In *IJCAI-97 Workshop on Issues in AI and Music – Evaluation and Assesment*, pages 9 – 16, 1997.
- [GM98] M. Goto and Y. Muraoka. Music understanding at a beat level: Real-time beat tracking for audio signals. *Computational Auditory Scene Analysis*, D.F Rosenthal an H. G. Okuno editors, pages 157 – 176, Lawrence Erlbaum Associates, USA, 1998.
- [HK02] T. Heittola and A. Klapuri. Locating segments with drums in music signals. In *Proc. of the 3rd Int. Conf. on Music Information Retrieval: ISMIR 2002*, IRCAM 2002, poster No. 271-272, October 2002.
- [HSSB96] m. Heath, S. Sarkar, T. Sanocki and K. Bowyer. Comparison of edge detectors: a methodology and initial studies. In *Proc. IEEE, Int. Conf. on Computer Vision and Pattern Recognition*, pages 143 – 148, 1996.
- [Har84] R. M. Haralick. Digital step edges from zero crossing of second directional derivatives. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 1, pages 58 – 68, Jan 1984.
- [Heck96] D. Heckerman. A tutorial on Learning with Bayesian Networks. Technical Report, Advanced Technology Division, Microsoft Co., November 1996.

- [InT02] Circular Logic. InTime Tempo tracking system software, <http://www.circular-logic.com>, 2002.
- [KAY93] S. M. Kay. Fundamentals of Statistical Signal Processing: estimation theory. Prentice-Hall signal processing series, Vol. 1, USA, 1993.
- [KT93] K. Kashino and H. Tanaka. A sound source separation system with the ability of automatic tone modeling. In *Proc. of the 1993 ICMC*, pages 248 – 255, 1993.
- [Kla99] A. Klappuri. Sound onset detection by applying psychoacoustic knowledge. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, volume 6, pages 3089-3092, 1999.
- [LK94] E.W. Large and J. F. Kolen. Resonance and the perception of musical meter. *Connection Science*, vol. 6, No. 2/3, pages 177-208, 1994.
- [Lar01] J. Laroche. Estimating tempo, swing and beat locations in audio recordings. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 135 – 138, October 2001.
- [Lar03] J. Laroche. Efficient tempo and beat tracking in audio recordings. *J. Audio Engineering Society*, Vol. 51, No. 4, page 226, April 2003.
- [Li00] Li Stan Z. Content-based audio classification and retrieval using the nearest feature line method. *IEEE Transactions on Speech and Audio Processing*, vol. 8, No. 5, pages 619 – 625, September, 2000.
- [MH80] D. Marr and E. Hildreth. Theory of edge detection. In *Proc. Royal Society of London*, Vol. B 207, pages 187 – 217, 1980.

- [MO90] F. R. Moore. Elements of Computer Music. Prentice-Hall Inc, USA, 1990.
- [MSBN01] Microsoft Bayesian Network, MSBNx Editor and toolkit V1.4.02. Microsoft Co. <http://research.microsoft.com/adapt/MSBNx>
- [Mit01] S. K. Mitra. Digital Signal Processing, A computer approach. McGraw-Hill series in Electrical and Computer Engineering, 2nd edition, McGraw-Hill, 2001.
- [Mur01] K. Murphy. The Bayes Net Toolbox for Matlab. Computing Science and Statistics, volume 33, 2001.
- [PK03] J. Paulus and A. Klappuri. Model-based event labeling in the transcription of percussive audio signals. In *Proc 6th Int. Conf. Digital Audio Effects (DAFX-03)*, 2003.
- [PM96] J. G. Proakis and D. G Manolakis. Digital Signal Processing, Principles, Algorithms, and Applications. Prentice Hall International Editions. 3rd Edition. USA, 1996
- [PNTA04] G. Pablo Nava and H. Tanaka, "Finding music beats and tempo by using an image processing technique", In *Proc. 2nd Int. Conf. on Information Technology and Applications ICITA 2004*, paper No. 410-5, January, 2004.
- [PNTAID04] G. Pablo Nava, H. Tanaka and I. Ide, "A convolutional-kernel based approach for note onset detection in piano-solo audio signals", to be appeared in *Proc. of the Int. Symposium on Musical Acoustics ISMA 2004*, March-April, 2004.
- [Par94] R. Parncutt. A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, Vol. 11, No. 4, pages 409 – 464, 1994.

- [Pro02] Digidesign. Pro tools 5.1.1 software, <http://www.digidesign.com>, 2002.
- [RPPD97] C. Roads, S. T. Pope, A. Piccialli and G. De Poli. Musical Signal Processing. Swets & Zeitlinger Publishers. Netherlands, 1997.
- [RWC01] RWC Music Database, 2003.03, Real World Computer Partnership, C MUSIC Corporation, Japan.
- [Ros92] D. F. Rosenthal. Machine Rhythm: Computer emulation of human rhythm perception. Ph.D. thesis, Massachusetts Institute of Tech., August 1992.
- [SB89] G. E. Sotak and K. L. Boyer. The Laplacian-of-Gaussian Kernel: A formal analysis and design procedure for fast, accurate convolution and full-frame output. *Computer vision, graphics and image processing*, Vol. 48, pages 147 – 189, 1989.
- [SF01] Sonic Foundry. Acid Pro 3 software, <http://www.sonicfoundry.com>, 2001.
- [SS01] W. A. Sethares and T. W. Staley. Meter and periodicity in musical performance. *J. New Music Reseach*, 2001.
- [Sch00] E. D. Scheirer. Music – Listening Systems, PhD Thesis, Massachusetts Institute of Technology, pages 94 – 99, June 2000.
- [Sch98] E. D. Scheirer. Tempo and beat analysis of acoustic signals. *J. Acoust. Soc. Am.*, vol. 103, No. 1, pages 588-601, January 1998.
- [Smi96] L. S. Smith. Onset-based sound segmentation. In D.S. Touretzky, M.C. Mozer, and M.E. Haselmo editors, *Advances in Neural Information Processing Systems*, vol. 8, pages 729-735. MIT Press, 1996.
- [Smi99] L. M. Smith. A multi-resolution time – frequency analysis and interpretation of musical rhythm. Ph. D. thesis, Univ. of Western Austrlia, July 1999.

- [TEC01] G. Tzanetakis, G Essl and Perry Cook. Automatic musical genre classification of audio signals. In *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, pages 205 – 210, October 2001.
- [TK99] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, USA, 1999.
- [TP86] V. Torre and T. A. Poggio. On edge detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 8, pages 147 – 163, March 1986.
- [TS99] D. Temperley and D. Sleator. Modeling meter and harmony: a preference-rule approach. *Computer Music J.*, Vol. 23, No. 1, pages 10 – 27, 1999.

Appendix

Tuning of parameters

In this appendix, the values of some parameters used in the model are presented. In section 4.2.2 Spectrogram enhancement of Chapter 4, the method for enhancement uses a parameter \mapsto in equation (14) that controls the level of enhancement of the spectrogram. In order to determine quantitative values of these parameters, some tests were performed with the system. A set of 30 music sample segments was prepared in order to be analyzed by the model. Among these test segments, samples of solo piano keys, rhythmic patterns of drums, and solo-guitar, were included. Then, the behavior of the system in terms of beat recognition was observed. The observations are resumed in Figure A1, A2 and A3, in which the data has been approximated in curves.

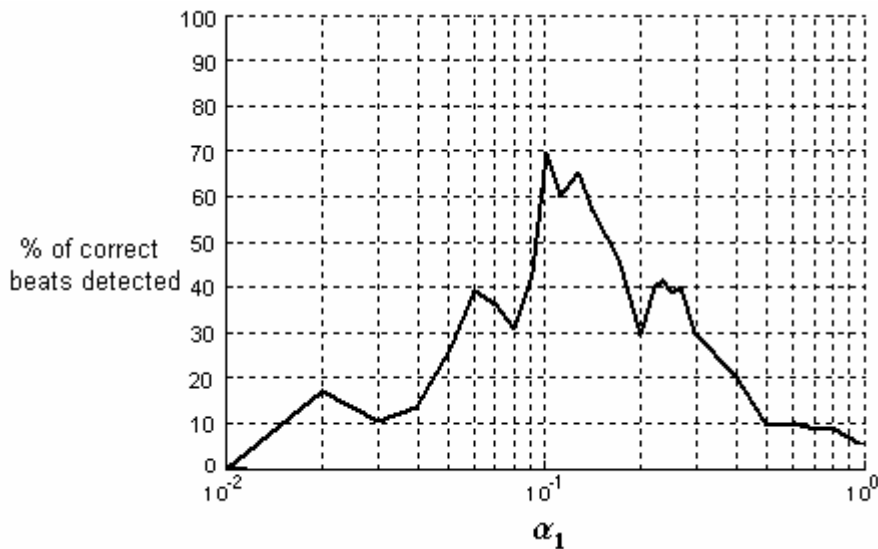


Figure A1. Values of parameter \mapsto_1 for enhancement of the low-pass band spectrogram.

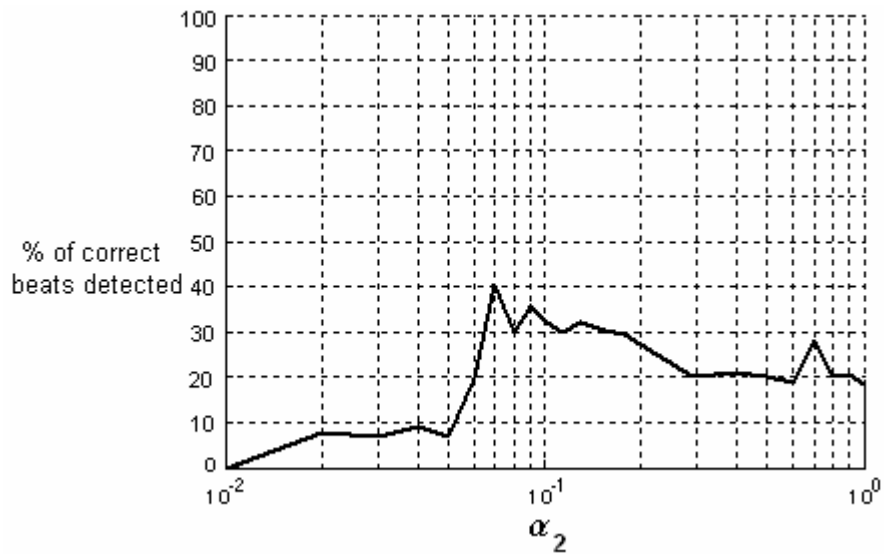


Figure A2. Values of parameter \mapsto_2 for enhancement of the Mid-pass band spectrogram.

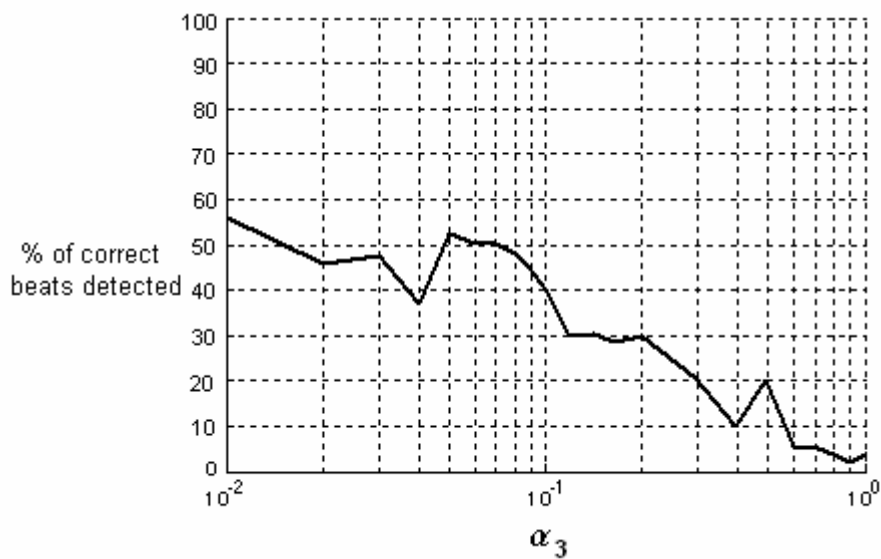


Figure A3. Values of parameter \mapsto_3 for enhancement of the High-pass band spectrogram.

By observing the graphs above, the best values for the parameters in equations (14) are: $\mapsto_1 = 0.1$, $\mapsto_2 = 0.07$ and $\mapsto_3 = 0.01$. These values are derived empirically as explained above, and strictly speaking, it does not mean that these values are the optimums. But as they serve for our purpose, we take them as immediate references.

For the case of the assessment of the Bayesian network, we consider all the possible situations that can occur while evaluating the beat hypotheses. For example, the nodes labeled as True beat and True Tempo in Figure 4.16 are variables of binary states, and by weighting the probabilities of their two possible outcomes, they are assessed with unbiased probabilities. A similar criterion is followed for the assessment of the rest of the nodes in Figure 4.6, but with the consideration of the number of states of each node. Figure A4 shows the probability network of Figure 4.6 with its respective assessments.

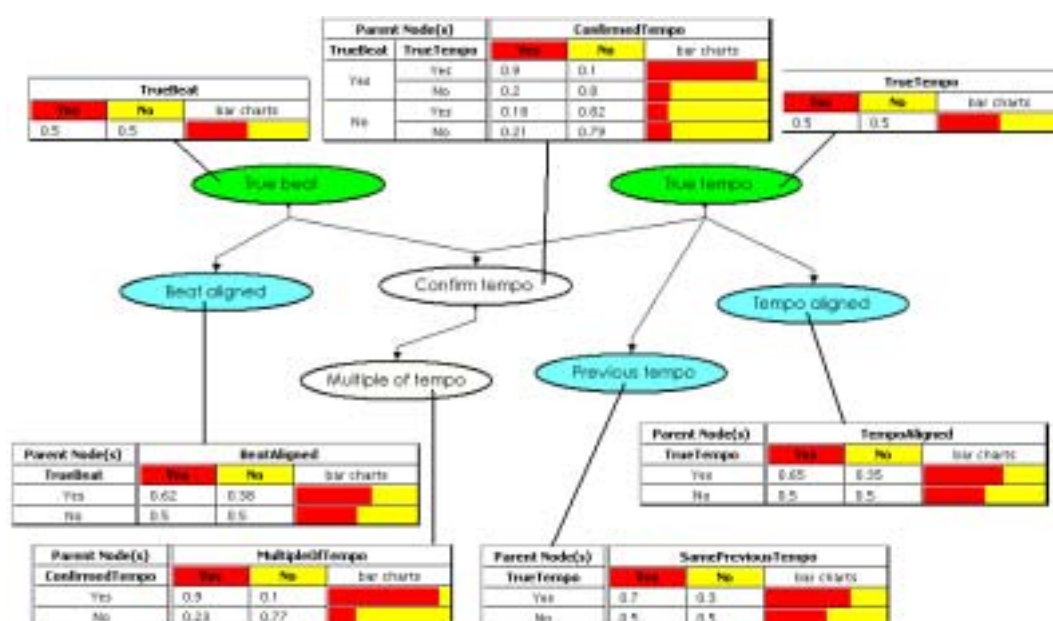


Figure A4. Assessment of the probability network for the evaluation of hypothetical beats and tempi.

The assessment of the network was done with the aid of [MSBN01], a software for the implementation and simulation of Bayesian network models. The actual implementation of the network was aided by the use of the Bayes Net Toolbox (BNT) for Matlab, developed by [Mur01], which is a toolbox available in the web for research purposes.