Proceedings of the 2nd International Conference on Information Technology for Application (ICITA 2004)

Finding music beats and tempo by using an image processing technique

Gabriel Pablo Nava and Hidehiko Tanaka

Abstract-- **In this paper we propose a method to find the music beats and tempo in sampled audio signals. Previous systems that have been proposed 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. Our approach, based on a basic image processing technique and the frequency centroid to detect onsets and a Bayesian probability network to infer the beats and tempo, showed to be able to find the beats and tempo in signals of different kinds of music. The idea is to accentuate the sound transients in the spectrogram by employing an image edge enhancement technique before computing the frequency centroid, then a relative differential will make evident the transients, finally the possible representatives of beats will be evaluated at a probabilistic hypotheses network where the actual tempo is inferred.**

I. 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 tracking systems is 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, tracking the rhythmic structure of music provides many applications with the overall music event map of a music composition and enables them to participate more interactively with the user.

Earliest approaches consisted of beat detection systems that take as input high-level symbolic representations of music, like MIDI sequences, [1][2][3]. However, these systems found limited applications were only MIDI signals were available. Some other works where developed to take as input acoustic signals but they perform the signal analysis off-line, this is, using prerecorded samples (non real-time) [4][5][6]. [4] proposed a method in which the signal is divided into frequency bands and the amplitudes envelopes of each band modulate a noise signal corresponding to that band, then the modulated signals are combined and the resulting signal, according to [4], contains the same rhythmic information as the original audio signal. More recent systems were developed to work with audio signals and in real-time, for example [8] presented a real-time tempo tracking that uses time domain analysis of the amplitude envelope extracted from acoustic signals and then forming several clusters to infer the basic tempo of the composition. The system proposed in [8] can process music samples from any gender and with any time signature but the user has to specify some parameters (that probably not every user is familiar with) in order to obtain a reasonable percentage of correct beat detection; this system assumes that the onsets of the beats are somehow explicit by considering hard attacks of the onset times. Another approach to beat detection and rhythm tracking was presented in [7]. This system is able to process acoustic signals in real-time, infer the beat structure at higher levels of hierarchical organization and recognize some drum patterns. It is a system that takes into account complex music knowledge and perform a huge signal processing in order to detect the beats from the raw signals, and because of this reason the computational cost is enormous (it was implemented on a distributed memory parallel computer, in connection with other terminals through an Ethernet and using a special music control protocol [7]).

The system proposed here can detect the beats and track the tempo by performing basic frequency analysis and employing a basic image processing technique that is used to enhance and detect edges in gray-scaled images. Our system is not limited to correctly detect the beats of a particular music style, and tempo tracking can range from 30 beats per minute (bpm), in contrast with [7] that performs well for pop music with a metrical signature of 4/4, and assuming a roughly constant tempo between 61 – 120 bpm. And there is no need to specify any parameter by user. Like people without any music theory knowledge are able to tap and detect the music beats, our system makes no assumptions of any musical knowledge. The signal processing is based on fast operations that make the computational cost quite lower than that of [7]. Then the method proposed is intended to work in a single personal computer.

II. System Overview

In Figure 1, the overall system for beat and tempo tracking is shown. It consists of three main sections: preprocessing (spectrogram enhancement), beat extraction, and hypotheses

Gabriel Pablo Nava and Hidehiko Tanaka are with the Department of Information and Communication Engineering, Graduate School of Information Science and Tehcnology, The University of Tokyo, Tokyo 113-8656, Japan (telephone: 03-5841-7413, e-mail: {pablo/tanaka}@mtl.t.u-tokyo.ac.jp).

ICITA2004 ISBN 0-646-42313-4

network. The input signal is first antialiasing-filtered in order to reduce noise produced by temporal modulations in harmonic mixtures, and down sampled. Then the signal is split into three frequency bands making it possible to track beats of different frequency ranges. In this stage, the spectrogram of the signal is constructed and enhanced. The beat extraction stage finds the cues that are more likely to represent beats. The possible beats are integrated to form hypotheses that will be evaluated by probabilistic network.

A. Preprocessing.

One of the principal problems in tracking beats and tempo of music rhythm is detecting original beats from the raw signal due to the fact that usually the beat cues are not explicit [7][8]. Even more, the hypotheses evaluation network relies greatly in the beat information generated by the beat extraction stage. Therefore, an accurate algorithm is needed in order to discover those cues that have high probabilities to represent true beats.

In this approach, the beats are detected by tracking the changes in the spectral power distribution from a spectrogram that has been enhanced by using a Laplacian convolutional kernel which is a simple image processing technique employed to enhance the edges of an image. Then by finding the repetitive cues from the signal of a frequency centroid function, candidates of beats are passed to the hypothesis generation stage.

From the starting point of the preprocessing stage, the signal is filtered into three frequency bands, as follows:

- Lowpass band (25 – 250 Hz)
- Midpass band (250 – 2500 Hz)
- Highpass band (2.5 – 10 kHz)

Since the beat detection is based on the frequency centroid of the incoming signal, we have found that dividing the signal into three bands is enough in order to detect the beats present at different frequencies ranges. For example, drum beats will become more evident in the Lowpass band, while beats of piano, guitar, and other instruments will be detected at the Midpass and Highpass bands. After dividing the signal into sub-bands, the subsequent signal processing is the same for the three bands.

In figure 2 the preprocessing and beat extraction blocks are shown in detail. After filtering, the spectrogram of a 2 sec. segment of the input signal is constructed by sliding a Hanning window of size 1024 samples with an overlapping of 57%, and with the sampling frequency of 22050 Hz, we achieve a resolution in time of 20 ms. Then the FFT is computed and arranged into frames to form the spectrogram. In these conditions we obtain a three dimensional matrix with dimensions *f, t* and $S(f_i,t_k)$, where:

$f = f_1, f_2, f_3, \ldots, f_M$ = frequency axes; M = number of frequency bins.

$t = t_1, t_2, t_3, \ldots, t_N$ = discrete time axes; N = total number of frames within a 2 sec segment of signal.

$S(f_i, t_k)$ = spectral power in the bin f_i at moment t_k; $i = 1,2,3,\ldots, M$; and $k = 1,2,3, \ldots, N$.

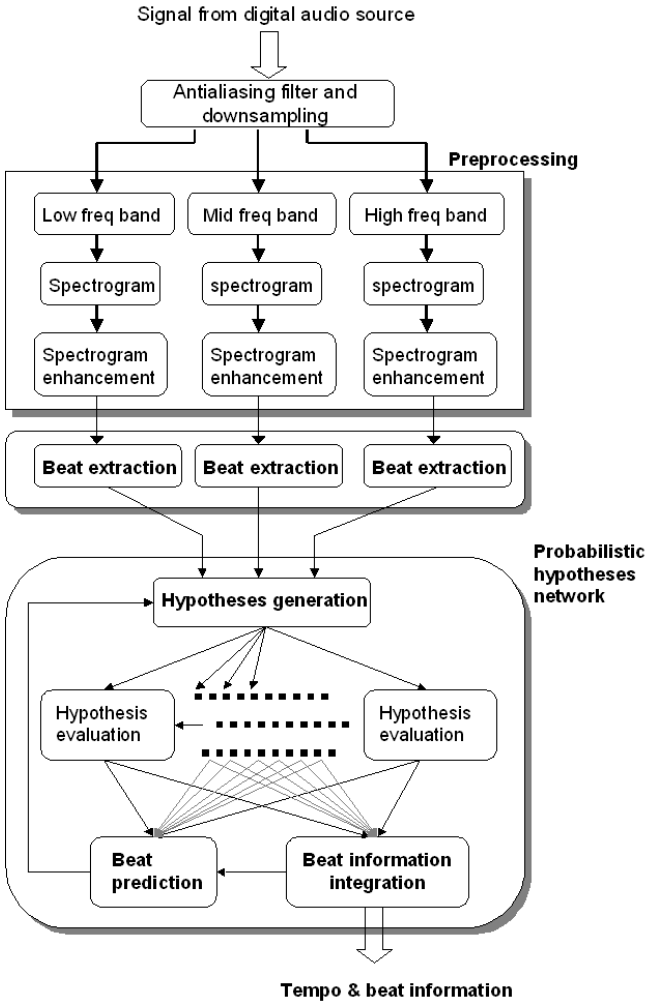

Fig. 1. Framework of the beat and tempo tracking system.

In order to accentuate the changes in energy distribution in the current 2 sec. song segment spectrogram, a Laplacian filtered version of the spectrogram scaled in decibels is created, then both spectrograms are normalized, and scaling the Laplacian spectrogram in decibels is added to the original to get a spectrogram with enhanced transients of energy distribution.

The technique used in this step is a basic image processing algorithm utilized to intensify and detect the edges of gray-scaled image. Thus, the enhanced version of the original spectrogram is given by:

$$S_{m,L} = \mathrm{mod}\left(S'_{m,ij}\right) + \mathrm{mod}\left(L'_{m,ij}\right)$$

where:

$$S'_{m,ij}\left(S_m\right) = \alpha_1 \frac{S_{m,ik}}{\max\left(S_m\right)}; \quad and \quad L'_{m,ij}\left(L_m\right) = \alpha_2 \frac{L_{m,ik}}{\max\left(L_m\right)}$$

S_m is the corresponding *m-th* sub-band spectrogram matrix scaled in dB's obtained after dividing the signal into frequency bands. L_m is the Laplacian filtered version of S_m. S' and L' are the corresponding normalized matrixes of S_m and L_m.

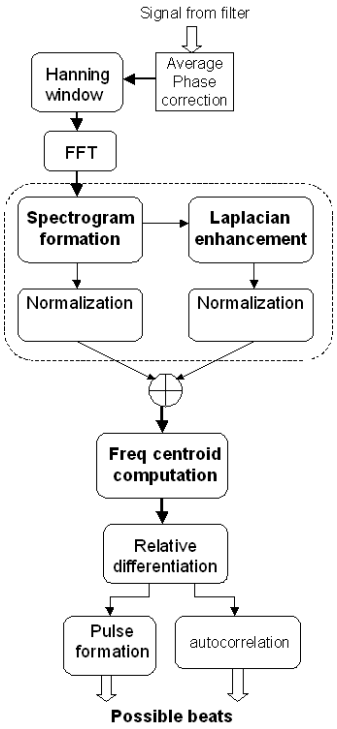

Fig. 2. Beat detection process in detail.

The Laplacian matrix L_m is computed by convolving an omni directional LoG (Laplacian of Gaussian) kernel of seven points that will give high values at the point where there are changes in the energy distribution while smoothing noisy transients. These changes are then tracked by a frequency centroid function.

B. Beat extraction.

The frequency centroid is a measure that indicates the balance of spectral power at the specified instant of time. Thus, the effect of the Laplacian enhancement over the spectrogram is to make the transients of energy distribution more detectable by frequency centroid function, therefore, even beats with low energy will produce notable unbalances that can be detected by the centroid function. This last parameter is a very well known function that is computed as follows:

$$C(t_k) = \frac{\sum_{i=1}^{N} \left| Pw(S_{L\,i,k}) \right|^2 f_{i,k}}{\sum_{i=1}^{N} \left| Pw(S_{L\,i,k}) \right|^2}$$

where Pw = power contained in the enhanced spectrogram.

In Figure 3 an example of the enhanced spectrogram (2nd graph from the top) of a segment taken from a pop song is shown. The centroid signal is displayed in the graph below the enhanced spectrogram.

In order to take into account the transients produced by weak beats, we apply a relative difference that was previously used by [5].

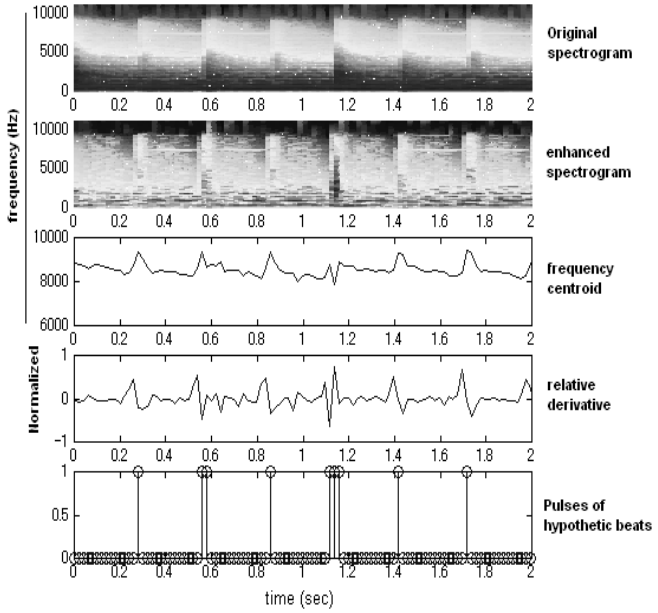

Fig. 3. Signals of the preprocessing and beat extraction. These signals correspond to a 2 sec segment of a pop song ("Fields of gold" of Sting). The train of pulses (lower graph) represents the possible beats of the audio segment processed.

This relative differential takes the derivative of the logarithm of the signal. In the graph of relative differentiation in Figure 3, we can appreciate the signal produced by this operation. Strong and weak peaks are present as well. Finally the signal produced by the relative differential is autocorrelated within the segment of 2 sec to reveal partial periodicities. This will help in posterior steps to construct the beat hypotheses and to infer the basic tempo.

Combining the statistical characteristics of the centroid signal and zero –crossing detection, a threshold is applied to form a train of pulses that represents possible beats candidates in order to create reliable hypothesis in the next stage.

C. Hypotheses network

The beat prediction and tempo inference are based in the probabilistic hypotheses generated from the possible representation of beats (train of pulses) and from the information provided by the temporal periodicities revealed by the autocorrelation. In figure 4, the network of a single hypothesis node is illustrated. As observed in figure 4, the first *a priori* knowledge that serves for subsequent hypothesis inferences is given by the partial periodicities of the autocorrelation, the intervals between the beats detected and the current hypothetic beats detected in each frequency band. Then the beats hypotheses are evaluated in levels in the more general network illustrated in figure 5. At the first level, let us assume that we wish to find the belief (*BE*) of a true beat induced at node *A*, this is, the first evaluation of the current beat hypothesis that will be propagated. Then letting D_h be the hypothesis of a partial periodicity detected, and D_H be the hypothesis that is propagated to the rest of the nodes. Under this

condition we can write:

$$BE(A) = P(A \mid D_H, D_h)$$

where A is the a vector of the probability hypotheses generated by the hypotheses generator $A=(a_1, a_2, a_3, \ldots, a_m)$.

By using Bayes theorem and introducing a normalizing factor, the hypothesis that will be propagated to the nodes of the general network is:

$$BE(A) = \mu(H_{prev}, t)P(D_h \mid A)P(A \mid D_H)$$

where $\mu(H_{prev}, t)$ is the normalization factor and dependent of the previous hypotheses tested and of time.

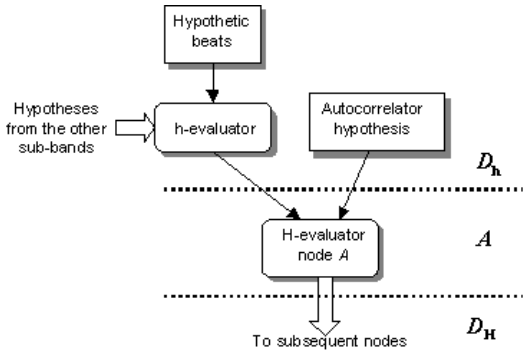

Fig. 4. Probability network of a single hypothesis node.

A similar reasoning is applied to evaluate the hypotheses coming from the different branches of the network and from the three frequency sub-bands. As seen in figure 5, the tested hypotheses are combined to aid the prediction of the next beats. This process continues as time progress and new nodes are created due to the new information coming from the beat extraction stage. Consequently, the historical probability knowledge influences the decisions of further evaluations of hypothetic music beats, as can be observed in figure 5.

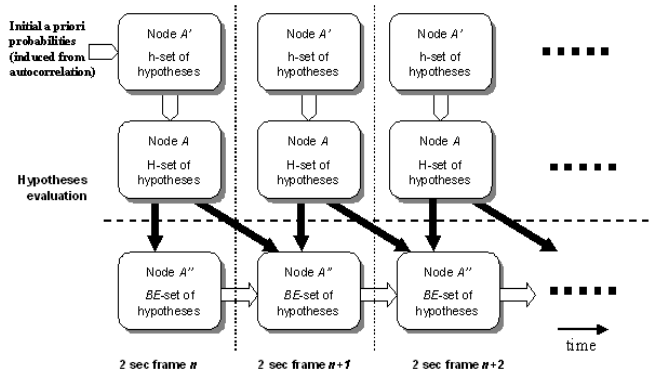

Fig. 5. Beats that were detected in more that one frequency band have high probabilities to represent true beats and also serve as reference to find more true beats.

The validated hypotheses of true beats are confirmed or discarded by their re-evaluation in the next processing frames. Those hypothesis that are not confirmed while the music performance progress in time, tend to disappear by the action of the normalizing factor $\mu(H_{prev}, t)$, while those that are confirmed are more likely to be taken into account to predict the next beats and infer the basic tempo.

When the process of a 2 sec music segment is completed, the system takes the following 2 sec segment and repeats the process, but this time there is the a priori knowledge from the beat information inferred previously.

DISCUSSION

The system has been tested until now with a total of 86 songs of different styles taken from [9]. Another set of training data was taken from commercial audio CD's. The music styles that have been tested range from pop to classic, including jazz, rock, instrumental, country, and dance music. At this point the system has been able to detect the basic beats that define the tempo of the song and most of the actual onsets become detectable since the level of the relative differential signal. However, songs that include high levels of voiced sounds can get the system confused due to the high concentration of spectral energy at the voice frequency range.

CONCLUSION

We have proposed a system to find the beats and tempo of music signals. Our approach, in contrast with previous once, is based on the analysis of changes in the spectral power distribution and by using a basic image processing technique the transients that correspond to the true beat sounds become more likely to be detectable at a relatively low computational cost. The system has shown until now robustness to signals that do not contain explicit onsets in the amplitude envelope of the sound signal. However, the training music database will be expanded to cover more music styles (like folk music, reggae, African, etc), and with various tempo ranges and complexities. The convolutional LoG kernel used in this work showed to improve the beat detection to some extent. Using more sophisticated image processing tools suitable to the algorithm requirements can help to retrieve the beat information from music signals with more accuracy.

REFERENCES

[1] Dannemberg and Mont-Reynaud, "Following an improvisation in real-time", International Conference in Computer Music, 1987.
[2] Desain, P. & Honing, H. "Quantization of musical time: A connectionist approach". Computer Music Journal. 13(3):56-66, 1989.
[3] Allen, P & Dennenberg, R. "Tracking musical beats in real time". Proceedings of the International Computer Music Conference, International Computer Music Association, 1990.
[4] Scheirer, E. "Tempo and beat analysis of acoustic musical signals". Journal of the Acoustical Society of America, 103(1), 1998.
[5] Klapuri A., "Sound onset detection by applying psychoacoustic knowledge", IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1999.
[6] Large, E. and Kolen, J. "Resonance and the perception of musical meter". Connection Science. 6:177-208. 1994.
[7] Goto, M. and Muraoka, Y. "Real-time beat tracking for drumless audio signals". Speech Communication, 27(3-4):331-335. 1999.
[8] Dixon S., Goebl W. and Widmer G., "Real-time tracking and visualization of musical expression", Second International Conference ICMAI 2002 Proceedings, 2002
[9] Audio CD's of the Real World Computer (RWC) Music database , 2001