AMPLITUDE MODULATED SINUSOIDAL MODELING FOR AUDIO ONSET DETECTION

F.J. Rodriguez-Serrano, P. Vera-Candeas, P. Cabañas Molero, J.J. Carabias-Orti, N. Ruiz Reyes

Telecommunication Engineering Department, University of Jaen
Alfonso X El Sabio 28, 23700, Linares, Jaen, Spain
phone: + (34)953648581, fax: + (34)953648508, email: fjrodrig@ujaen.es

ABSTRACT

Onset detection is a key application in music processing. Beat detection algorithms and some music transcribers usually perform onset detection as the starting point of their processing. In music transcription of polyphonic signals, onset detection is very helpful because it aids to place note-event starting times. In this paper, a new technique to implement an onset detection system is proposed. In sinusoidal modelling, the energy burst of non-stationary tones are detected by means of linear prediction in the frequency domain. In frequency, the tone peak and its nearby samples does not match with the window transform when the tone is not stationary at the current frame. This property can bedetected whith linear prediction in the frequency domain. When perceptually significant tones are detected as unstable in a time frame, the system alerts about an onset at this frame. The proposed onset detection system is evaluated over two sound databases obtaining encouraging results.

Index Terms - Music onset detection, linear prediction,sinusoidal modeling, Hidden Markov Model, perceptual modeling.

1. INTRODUCTION

Onset detection provides very usefull information to music signal processing applications. Onset detection is needed as a previous task for beat detection, in other words, the onset information is post processed to obtain beat times of piece [1]. In applications devoted to align automatically the score (the MIDI file) and the recording, an onset detector can improve the results [2]. Onset detection is usually combined with a pitch estimator to obtain an automatic alignment.

Onset detection is very helpful for some music transcribers [3] in order to determine the starting times of note-events. An onset has tipically unstable frequency regions beside it, so a pitch analysis near the onset tends to fail. Otherwise, when onset frames are known, the system labels these frames as unestable and focus on analyzing stable regions.

Onset detection systems are often based on searching for frequency changes in the spectrogram or looking for significant energy increases in the time domain signal. Then, statistical models are employed in order to obtain the estimated starting times of musical events [3]. The main problem of using the time domain signal is that a new frequency can appear while the dominant ones are active. Because of this, a frequency-based analysis is generally performed to avoid onset time loses [4].

In this paper, we propose an onset detection system based on sinusoidal modeling. For each frame, first peak locations are detected. Then, linear prediction in frequency for only a few samples close to each peak is implemented. The main idea is to detect energy increases in consecutive frames for each sinusoid detected by the modeling. Frames are analyzed by the sinusoidal model with strong overlapping between them to merge the information of consecutive frames for each peak location. In this way, a combined time and frequency domain analysis is performed to label some sinusoids as unstable in a range of frames. When there are a sufficient set of unstable sinusoids in a range frames, an onset is detected by our system. This decision is made taking into account the psychoacoustic significance [5] of unstable peaks.

The paper is organized as follows: section 2 explains the theoretical principles used by the system, section 3 is devoted to detail the processing steps implemented by the system, and finally, in section 4 and 5 results and conclusions are addressed.

2. FUNDAMENTALS

Sinusoidal analysis is almost universally implemented by dividing time signal in frames and applying a time window before computing the Fourier transform. This analysis leads to mismatches when sinusoids are not stationary. One solution to this problem is switching to short frames when not stationary signals are detected [6]. Another solution is to analyze the time amplitude modulation by means of linear prediction in frequency domain [7][8].

The information about time envelope is really available in the frequency domain. When a sinusoid is not stationary, its peak (in frequency) appears but does not fit with the actual form of the main lobe for the time window. In fact, the actual form of the sinusoid main lobe informs us about the amplitude modulation in time of the sinusoid. In this paper, the detection of energy burst for sinusoids is obtained from frequency domain information.

The use of frequency domain information to obtain time envelope of a signal is common on audio coding applications [9][10]. In [10], noise coding is adapted to have the same time envelope as the signal in order to avoid pre-echo effects due to coding. In Figure 1, the estimated time envelope in a frame by a 4 order linear predictor in frequency is shown.

Linear prediction in the frequency domain requires the estimation of the autocorrelation function in frequency. For a signal $x(n)$, the DFT at the t-th frame is denoted here as $X_t(k)$. The estimation of the correlation function can be computed as,

$$\hat{r}_{XX}(l) = \frac{1}{F-l-1} \sum_{k=0}^{F-l-1} X_t(k) X_t^*(k+l) \qquad (1)$$

where F is the number of samples in frequency.

 ISSN 2076-1465

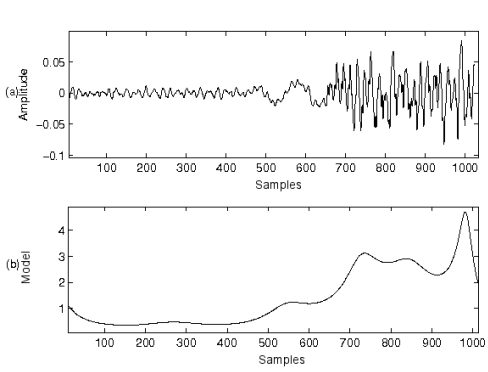

Figure 1: *4-pole AR-model from a signal. (a)Frame signal, (b)4-order estimated time envelope*

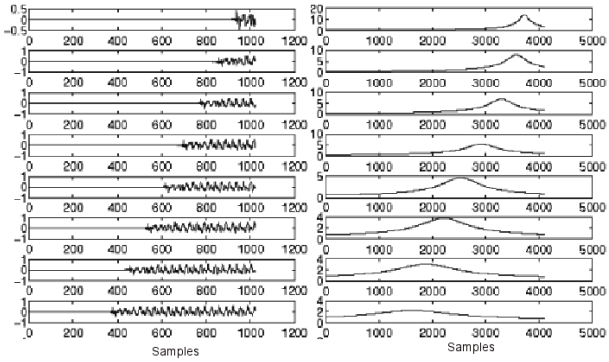

Figure 2: *Frame to frame evolution of the AR-model Left: Time signal Rigth: 1-pole time envelope*

However, this tool can not be applied directly to onset detection because only onsets that provoke an increment in the envelope in time of the signal would be detected. In music signals, it is common that some onsets do not affect to the global signal envelope. In these cases, the onsets can be detected by searching for new energy in some frequency regions [4][11]. We propose here to use linear prediction in the frequency domain but restricted to the neighborhood of each detected peak. In this way, time envelope of each sinusoid is analyzed in order to estimate onsets.

The main problem of this approach is that time resolution is very poor. Supposing that stationary and non-stationary sinusoids can appear at the same time frame, the global autocorrelation fuction should not be estimated from all frequency samples to obtain the time envelope of different sinusoids. Instead, the autocorrelation function has to be estimated in the neighborhood of each peak location in frequency. With this restriction, only a few samples can be utilized to estimate the local autocorrelation function and low order models can be estimated. The estimated time envelope for only one order linear predictor is shown for a non-stationary signal.

The localized autocorrelation function in frequency around the peak location k_p can be estimated as

$$\hat{r}_{XX}^{k_p}(l) = \frac{1}{2L+1} \sum_{k=k_p-L}^{k_p+L} X_t(k) X_t^*(k+l) \qquad (2)$$

where L is the shift in frequency from the peak to used for the estimation.

In order to improve the time resolution of the system, overlapping between frames in time must be performed. Following this principle, the tracking of a burst of energy in time can be carried out. Figure 2 shows the estimated time envelope for a tone with one order linear predictor in frequency. Autocorrelation of frequency samples is estimated with only 5 samples around the sinusoidal peak ($L = 2$).

The information to be used in order to detect the presence of a onset for each sinusoid is the phase of the coefficient estimated in the one order linear predictor. The phase of the pole indicates approximately the time center of the onset energy with respect to the current frame. This information can be used as an indicator of the energy burst position along consecutive frames and is here utilized as the main clue for onset detection. In Figure 3, the phase evolution in the z-plane of the models obtained in Figure 2 are shown. Using this analysis a good discrimination between onset and offset detection can be obtained.

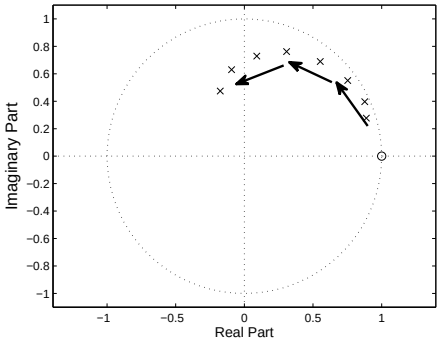

Figure 3: *Pole phase evolution along frame to frame evolution*

For only one order linear prediction, the coefficient of the predictor filter is computed from the localized autocorrelation function as follows,

$$c_t^{k_p} = -\frac{\hat{r}_{XX}^{k_p}(1)}{\hat{r}_{XX}^{k_p}(0)} \qquad (3)$$

As can be seen in equation 3, the phase information is carried out only for the localized autocorrelation function at $l = 1$ ($r_{XX}^{k_p}(1)$).

3. SYSTEM DESCRIPTION

This section describes the proposed onset detection system as a whole. First, the block diagram of the system is presented. Then, each processing block is detailed explaining the information that flows between blocks.

3.1 System structure

The system can be modeled by a block structure like the one shown in Figure 4. The input is the music signal to be processed. This signal is directly processed by the sinusoidal model and the perceptual model. The perceptual model is used by two processing blocks of the system. Here, we have used the perceptual model proposed in [5].

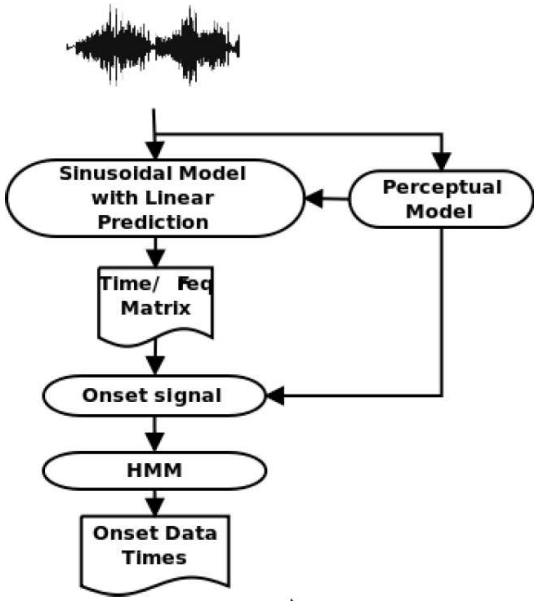

Figure 4: *Onset Detection System Block Diagram*

3.2 Sinusoidal model with linear prediction

Sinusoidal modeling assumes a signal model that can be expressed as,

$$x(n) \approx \sum_{q=0}^{Q} \alpha_q(n)\cos(2\pi f_q n + \varphi_q(n)) \tag{4}$$

where $x(n)$ represents the original audio signal, $\alpha_q(n)$ are amplitudes, f_q frequencies and $\varphi_q(n)$ phases for each tone. Each tone has a modulated amplitude, this AM consideration helps the modeling of transients. In [9] it is demonstrated that AM is a good improvement for this particular kinds of signals. The accurate estimation of phases $\varphi_q(n)$ for each tone is an important issue if the signal is going to be resynthesized.

Here, sinusoidal modeling is implemented by means of perceptual matching pursuits [12]. This approach extracts all perceptually significant sinusoids from a signal frame following the perceptual model proposed in [5]. It must be stressed that amplitude, frequency and phase of each tone are here constant along the current frame.

Once all sinusoids in a frame have been detected, the localized autocorrelation function in frequency is estimated at each peak location following equation (2). Then, the coefficient of one order linear predictor is computed by equation (3). This information is therefore associated to each modeled tone.

From this information, the onset location for each tone can be estimated using the coefficient phase $\varphi(c_t^{k_p})$. This phase is directly related with time envelope. When there is no onset, the phase is close to zero, while as shown in Figure 3, the phase follows the onset location along frames. This phase tends to point out to the middle of the energy burst in each frame. Taking this property into account, we can estimate the sample where the energy burst is placed respect the begining of the frame as $N - \frac{2N\varphi(c)}{2\pi}$ where N is the frame length. However, this estimation of the sample where the onset appears is not very accurate using only a one order linear predictor. Instead, we compute the distance in frames hops from the current one to the frame where the onset appears.

$$fr = \frac{(\frac{N}{2}) - (\frac{\varphi(c_t^{k_p})N}{\pi})}{O} \tag{5}$$

where fr is the distance in frames hops and O the number of overlapping samples between frames. These onset distances are computed for each modeled tone. Also, this information can be converted to the absolute frame in which the onset is produced.

3.3 Time/frequency matrixes

First, sinusoids are ordered in frequency following a logarithmic scale. In order to reduce the information, only the most perceptually important tone at the frequency range of each MIDI note is maintained, the remaining tones at this range are rejected. Following this approach, the sinusoidal model produces a matrix where the time are frames and the frequency has a sample for each MIDI note. Using this discretized model, a stationary sinusoid is generally modeled as a vector in a MIDI note for some consecutive frames. We obtain the following matrixes: complex amplitudes, perceptual significances, frequencies and onset locations in frames.

The time/frequency matrix for onset locations in frames which contains onset information is processed and cleaned. First, a grouping process is implemented to wipe all redundant information from the matrix and only cells which gives neccesary information about onset will be mantained. The cleaning condition is the following: only those consecutive cells in time with a distance in frames less than one frame (see equation 5) are maintained.

3.4 Onset signal

Now we have all the information to extract onset times, we compose an onset signal from the information of onset location and perceptual significance matrixes. This signal is conformed by adding at each frame all the perceptual significances of those tones that point to the current frame according to the onset location matrix. To sum up, the perceptual significance of all non-stationary sinusoids whose energy increases belong to the current frame are added to estimate this *onset signal*.

When a note is played, for example in a piano onset, all harmonic partials belonging to the note are added at the same location in the onset signal. This property makes robust the onset detection for harmonic signals.

3.5 HMM

This onset signal is processed by a Hidden Markov Model to set activation times[13]. This block is needed because not all peaks at the onset signal really represent an onset. Some of them are caused by spurious sequences of poles detected at the analysis stage. They usually have low perceptual importances. Only significant peaks at the onset signal activates the HMM block and marks an onset region. Onset signal is adapted to represent probabilities that goes into the HMM. it has two states, on (onset is active) and off(stable region). For each detected onset period the frame that corresponds to the maximum of the onset signal is set as the onset time. Figure 5 shows an example of the onset signal and HMM activation time for an excerpt of piano.

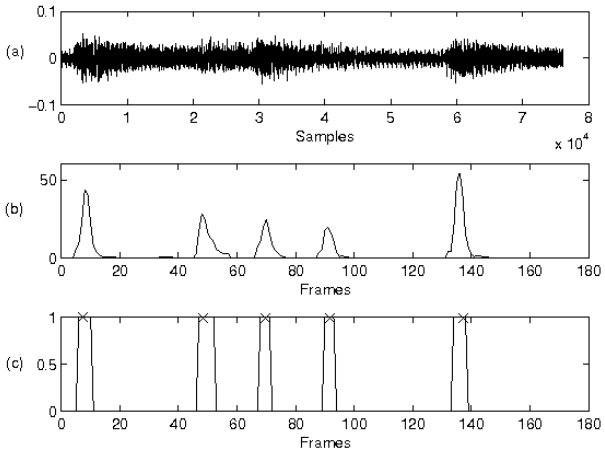

Figure 5: *HMM over energy signal and onset determination. (a) Audio signal, (b) Onset signal, (c) HMM determination. The onset time location is represented by a 'x' marker.*

4. EXPERIMENTAL RESULTS

Our system has been tested versus the audio transcription software, Sonic, for piano signals from MAPS(MIDI Aligned Piano Sounds) database[14],[15], and the annotated onset database proposed by Juan Pablo Bello[16]. It has a variety of music including drums, string instruments, electric instruments, singing voice, piano and a mix of them. This is to assess the proposed onset detector performs wll with a variety of signalsword that this is not only a piano onset detector, it is applicable to other music signals.

A data set of piano sounds where selected from MAPS data base, these files where input for both system, Sonic and the proposed one, in order to compare their results. Sonic [17][18] was selected because of its level of accuracy and its relevance at the bibliography. This software is more than an onset detector, it is a transcription system which gives a MIDI file as an output. We have extracted the onset information from the corresponding MIDI data estimation.

MAPS database contains a lot of types of piano sounds, some of them are isolated or random notes, which are not interesting for testing our system, and other ones are real played compositions. For each test file a MIDI and a text files are given, both of them contain the onset information. In MAPS database con be found sounds taken at several recording places. For our testing procedure we have taken a pair of them, which represents the best and the worst situation: a concert hall (Table 1) and a church (Table 2). Insted of syntesized piano excerpts [19], real piano composition from MAPS database is here used to evaluate the onset detection system.All files at Bello's database where processed, this database was selected in order to have results with other types of music signals,not only piano music.

Three measures have been implemented; *Precision* (P), expression (6), which takes false positive detections into account, *Recall* (R) expression (7), which takes false negatives fails into account and the *F-measure* (F) which is a summary of them.

Concert Hall Recordings						
	Sonic			Proposed		
FILE	P	R	F	P	R	F
alb_se3	0.99	0.72	0.83	0.99	0.94	0.97
bach_846	0.92	0.82	0.87	0.93	1	0.97
bach_847	0.94	1	0.97	0.99	0.99	0.99
bk_xmas5	0.86	0.67	0.75	0.85	0.94	0.89
chp_op31	0.91	0.87	0.89	1	0.83	0.90
chpn_op25	0.96	0.78	0.86	0.95	0.97	0.96
chpn_op66	0.85	0.71	0.77	0.85	0.81	0.83
MEAN	0.92	0.79	0.85	0.94	0.92	0.93

Table 1: Experimental results: Concert Hall

Church Recordings						
	Sonic			Proposed		
FILE	P	R	F	P	R	F
alb_se3	0.95	0.71	0.81	0.98	0.83	0.90
alb_se4	0.92	0.71	0.80	0.93	0.83	0.87
alb_se7	0.98	0.81	0.89	0.96	0.85	0.90
appass_1	0.75	0.44	0.55	0.97	0.69	0.80
bach_850	0.78	0.90	0.84	0.90	0.99	0.94
bk_xmas1	0.96	0.71	0.82	1	0.88	0.94
bor_ps6	0.97	0.60	0.74	0.98	0.88	0.92
MEAN	0.90	0.68	0.77	0.96	0.84	0.90

Table 2: Experimental results: Church

$$P = \frac{cd}{cd + fp} \tag{6}$$

$$R = \frac{cd}{cd + fn} \tag{7}$$

$$F = \frac{2PR}{P + R} \tag{8}$$

Where *cd* are correct detected onsets, *fp* are false positive samples and *fn* are false negative ones.

At the first evaluation, results for the proposed method are always greater than Sonic ones. In average (see Tables 1

Bello Database							
FILE	P	R	F	FILE	P	R	F
arab60s	0,96	0,94	0,95	Jaillet66	0,70	0,73	0,72
dido	0,83	0,80	0,82	Jaillet67	0,83	0,94	0,88
fiona	0,81	0,55	0,65	Jaillet70	1	0,93	0,97
Jaillet15	1	1	1	Jaillet73	0,71	0,83	0,77
Jaillet17	1	0,78	0,88	Jaillet74	1	1	1
Jaillet21	0,61	0,59	0,6	Jaillet75	1	0,50	0,67
Jaillet27	1	0,90	0,93	jaxx	0,48	0,53	0,51
Jaillet29	1	0,71	0,83	metheny	0,92	0,73	0,81
Jaillet34	1	0,93	0,96	PianoDB	0,4	1	0,57
Jaillet64	0,78	0,44	0,56	tabla	0,91	0,83	0,87
Jaillet65	0,91	0,71	0,80	violin	0,85	0,81	0,83
wilco	0,75	0,74	0,74	MEAN	0,83	0,76	0,78

Table 3: Experimental results: Bello Database

and 2) it is 10% better than Sonic. F-score is the compared measure because it takes precision and recall measurements and it is the one used at MIREX for the onset detection competition. Our score is penalized by false negatives samples showing a high precision rate. The same occurs with Sonic, but in this case the difference between *Precision* and *Recall* is higher.

The second evaluation (see Table 3) shows that the system has good results for other types of music. Results has a lower rate than the first ones because piano signal has very suitable characteristics for this task. However, this evaluation could be compared versus others algorithms at MIREX 2010 in the "Audio Onset Detection" task where it is participating but results are not avaiable at the time of writing this paper. This is a sign that the system is robust and gives good results for a variety of music types.

5. CONCLUSIONS AND FUTURE WORK

As it is shown in results, our system is really precise and robust with piano signals. Then it can be a helpful block for music signal processing system. Our system gives as an output stable periods of the signal for any analysis. In this way, a lot of problems because of transients are avoided. We have show experiments with piano signals and all other music types, the system has good results with harmonic instruments, drums and speech too. Piano signal results are better as expected, but all result are in good range of accurancy to consider this system as a previous stage to help other kind of signal processing.

In future works, we will substitute linear prediction by the time/frequency reassignment methods[20]. Poor resolution in time is obtained by using only one order linear predictor, we think that this resolution can be improved with reassignment techniques.

ACKNOWLEDGEMETS

We want to give our acknowledgement to Valentin Emiya for allowing us to use his piano sounds database [14] and Juan Pablo Bello for share with us his annotated onset database [16] to obtain the results of this paper.

This work was supported by FEDER, the Spanish Ministry of Science and Innovation under Project TEC2006-13883-C04-03, and the Andalusian Business, Science and Innovation Council under project P07-TIC-02713.

REFERENCES

[1] S.Dixon and G Widmer, "Evaluation of the Audio Beat Tracking System BeatRoot", *Journal of New Music Research*, 36, 1, pp 39-50.

[2] Y. Meron and K. Hirose, "Automatic alignment of a musical score to performed music," *Acoustical Science and Technology*,Vol. 22 , No. 3 pp.189-198,2001.

[3] M. Marolt, A. Kavcic, M. Privosnik, "On detecting note onsets in piano music," *Proceedings of MELECON 2002*,Cairo, Egypt, 2002.

[4] M. Gainza,B. Lawlor,"Onset Detection Using Comb Filters' *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*pp. 263-266, 2005.

[5] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens,"A new psycho-acoustical masking model for audio coding applications" *in in Proc. ICASSP*, 2002, vol. 2, pp. 18051808.

[6] Purnhagen, H.; Meine N, "MPEG4: Harmonic and Individual Lines plus Noise (HILN)," *ISCAS*,2000.

[7] M. G. Christensen and S. H. Jensen, "Computationally Eficient Amplitude Modulated Sinusoidal Audio Coding using Frequency-Domain Linear Prediction," *IEEE Int. Conf. Acoust., Speech, Signal Processing*.(2006).

[8] W.C. Lee, C.C.J. Kuo"Musical onset detection based on adaptive linear prediction" *In Proc. Int. Computer Music Conference,ICME 2006*, Toronto, Canada, 2006, pp. 957960.

[9] M. G. Christensen, A. Jakobsson, S. V. Andersen, and S. H. Jensen, "Linear AM Decomposition for Sinusoidal Audio Coding,"*in Proc. IEEE Int. Conf. Acoust.,Speech, Signal Processing*, vol. 3, pp. 165-168, 2005.

[10] J. Herre,"Temporal Noise Shaping, Quantization, and Coding Methods in Perceptual Audio Coding: A Tutorial Introduction," *AES 17th Int. Conf. High Quality Audio Coding*,Florence, Italy, September 1999.

[11] S. Hainsworth and S. Macleod and Malcolm"Onset Detection in Musical Audio Signals" *IEEE International Conference on Multimedia and Expo*, 2003.

[12] P.Vera-Candeas,N.Ruiz-Reyes,J.C.Cuevas-Martinez, M. Rosa-Zurera and F. Lopez-Ferreras,"Sinusoidal modeling using perceptual matching pursuits in the Bark scale for parametric audio coding" *IEEE Proc.-Vis. Image Signal Process.*,Vol. 153, No. 4, pp. 431-435, August 2006.

[13] W. Chai and B. Vercoe, "Folk Music Classification Using Hidden Markov Models," *Proc. of the International Conference on Artificial Intelligence, ICAI01*,2001.

[14] V. Emiya, *Transcription automatique de la musique de piano*, These de doctorat, Telecom ParisTech, 2008.

[15] V. Emiya, R. Badeau and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEE Transaction on Audio, Speech and Language processing*.(to be published).

[16] Bello, J.P.; Daudet, L.; Abdallah, S.; Duxbury, C.; Davies, M.; Sandler, M.B., "A Tutorial on Onset Detection in Music Signals", *Speech and Audio Processing, IEEE Transactions* , vol.13, no.5, pp. 1035- 1047, Sept. 2005.

[17] M. Marolt,"SONIC : transcription of polyphonic piano music with neural networks" *Proceedings of Workshop on Current Research Directions in Computer Music*, Barcelona, Spain, November 15-17, 2001.

[18] Sonic Software http://lgm.fri.uni-lj.si/matic/SONIC/sonic.zip

[19] C. G. v. d. Boogaart, R. Lienhart"Note onset detection for the transcription of polyphonic piano music" *Technical Report, Institute of Computer Science, University of Augsburg*, Ausburg, May 2009

[20] F. Auger and P. Flandrin,"Improving the readibility of time-frequency and time-scale representations by the reassignment method" *IEEE Transactions on Signal Processing*,vol. 43, pp. 1068-1089, May 1995.