A convolutional-kernel based approach for note onset detection in piano-solo audio signals

Gabriel Pablo Nava[1], *Hidehiko Tanaka*[1], *Ichiro Ide*[2]

[1]Department of Information and Communication Eng., The University of Tokyo, Japan
[2]Software Research Division, National Institute of Informatics, Japan
[1]{pablo/tanaka}@mtl.t.u-tokyo.ac.jp [2]ide@nii.ac.jp

Abstract

This paper presents a new approach to piano onset detection based on a convolutional kernel modeled from the salient characteristics of the transients of piano onsets. Preliminary experiments performed with piano recordings, show that our method promises to achieve reasonable high level of onset detection even in complex signals such as piano jazz and piano improvisations with considerably fast tempi. Another advantage of this approach is the viability to be extended to detect onsets of other instruments with few modifications.

1. Introduction

Onset detection has been studied in computer music in order to aid tasks such as automatic music transcription, audio signals segmentation, tempo and beat tracking, etc. These systems usually include in their first processing stages an algorithm to discover in the raw signal, the reliable cues representing the real musical events that can be taken into account to infer higher levels of music information. Since the detection of onsets is usually the starting point in such systems, the subsequent processes rely greatly on the information revealed by the onset detector. Therefore, the implementation of an algorithm that can find and label the true musical events, becomes a problem that deserves attention.

In this work, we concentrated on the detection of note onsets in recordings of piano-solo signals. Researches on onset detection have been done in areas of beat and tempo tracking such as in [1][2][3]. The onset information that these systems retrieve is that of the beats defining the tempo and the rhythmic structure for the signal analyzed and not precisely the location of all the note onsets played in the original music performance. Thus, such systems are inappropriate for automatic music transcription for example, where the main task is to transcribe the actual notes from the audio signal. In recent approaches, [4] has presented a system that performs onset detection in the complex domain, in contrast with traditional onset detectors that consider only the energy of the signal. The system of [4] works well for onsets that concentrate in the lower frequency spectral components, and onsets that present in some degree hard attacks. Another approach for detecting onsets of piano notes was proposed in [5], which is based in a gammatone filterbank with output channels centered at piano tones, and a smoothing filter that works together with a peak-picking step. This work was improved in [6] by employing two smoothing filters with different time constant, in combination with a neural network. Although the smoothing filters give considerable cues for onset inference, the neural network is unable to handle onsets with proximities less than 50 ms due to the period of inactivity of the neurons [6]. In [3], a particularization was done for onset of piano signals and the approach was presented in [7], which integrates the basic technique of onset detection in time domain [3] with analysis in frequency domain and a genetic algorithm. Finally, [8] presented a probabilistic model in which the problem of onset detection is viewed as *surprising moments* that should be observed and evaluated by a probability model using independent component analysis.

In our approach, the idea is to analyze the signal in the frequency domain by using a convolutional kernel that will respond to the transient events whenever a piano note is present. Piano sounds have the particularity of presenting abrupt changes of energy in the fundamental frequency components as well as in the noisy components that conform the transient. This characteristic hard attack in the noise components is due to the fact that piano sounds are of percussive nature. By finding an optimum kernel that can respond to this characteristic, we can convolve it with the spectral data in order to make evident the transient events. In Figure 1a, there is an example of the characteristic spectrogram of a piano key sound showing the characteristics mentioned above. In the cases of more complex piano signals (as in Figure 1b) the overlapping spectral components make the accurate detection of onsets difficult, leading to spurious detection and misses of some onsets. To solve this problem, the signal is divided into sub-bands and a probabilistic network will evaluate the hypotheses of possible true onsets detected.

2. System overview

The over all system for onset detection is shown in Figure 2. The three main sections are: *preprocessing*,

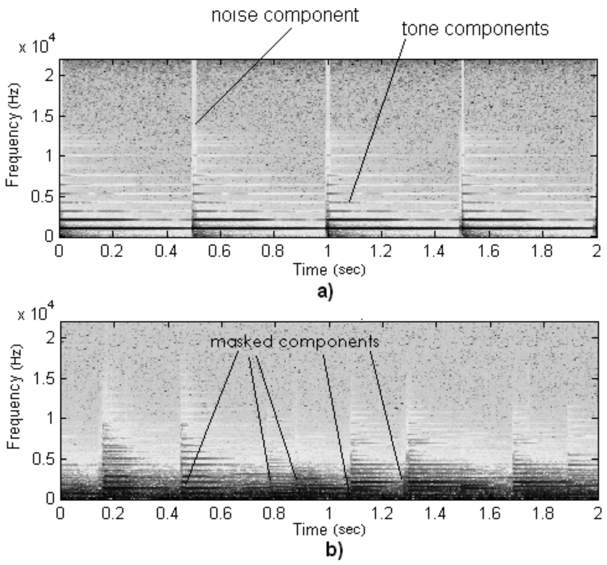

Figure 1.Eexample of spectrograms of piano sounds. a) A single piano note, b) A piece of piano jazz fragment.

onset hypotheses labeling and the *probabilistic network* for the evaluation. The input signal is antialiasing filtered and then down-sampled from 44.1 kHz to 22 kHz.

2.1. Preprocessing

In the cases of simple piano signals, as in Figure 1a, the appropriate kernel can be applied and the onsets can be located easily. However, in real situations of piano performances, many onsets cannot be detected due to the masking effect of some spectral components into others, as can be appreciated in Figure 1b where the magnitude of some harmonics are higher than other note tones. This motives us to split the signal into sub-bands. In this approach we use four sub-bands assigned as follows:

- Sub-band 1: (0 – 500 Hz)
- Sub-band 2: (500 – 2500 Hz)
- Sub-band 3: (2.5 – 5 kHz)
- Sub-band 4: (5 – 10 kHz)

After filtering, the sub signals are down sampled, then the spectrogram of each band is computed by sliding a Hanning window whose size and overlapping are different for each sub-band and are selected so as to achieve the best resolution of time in the four bands (the achieved resolution in this model is 11.6 ms). The resulting spectrogram is 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 axis; M = number of frequency bins.
- $t = t_1, t_2, t_3, \ldots, t_N$ = discrete time index of the N frames within a segment signal.
- $S(f_i, t_k)$ = spectral power in the bin f_i at moment t_k; $i = 1,2,3, \ldots, M$; and $k = 1,2,3, \ldots,N$.

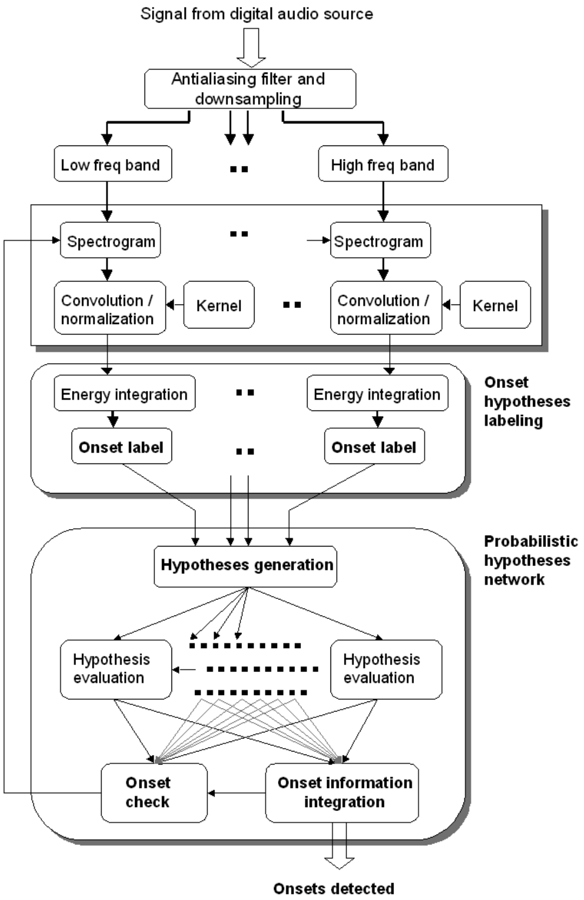
Signal from digital audio source
Antialiasing filter and downsampling
Low freq band
High freq band
Spectrogram
Spectrogram
Convolution / normalization
Kernel
Convolution / normalization
Kernel
Energy integration
Energy integration
Onset label
Onset label
Onset hypotheses labeling
Hypotheses generation
Probabilistic hypotheses network
Hypothesis evaluation
Hypothesis evaluation
Onset check
Onset information integration
Onsets detected

Figure 2. Scheme of the onset detection system.

The absolute values of the spectral power are taken to perform the convolution of the sub-band spectrogram with its corresponding kernel. The kernels are matrices, of different sizes for each sub-band, and their coefficients are computed from transients of piano notes. The samples of piano notes used to train and construct the kernels were taken from the piano recordings in [8]. For notes that are within the corresponding sub-bands of our system, parameters such as attack time, maximum peak, rising slope and average separation of partials were measured and the values that minimize the least squared error in the convolution are taken to form the kernel that will be used for that sub-band.

With the corresponding kernel for each band, the two dimensional convolution with the magnitude spectral power matrix is performed as follows:

$$h_{SB}(i,j) = \sum_{k_1=1}^{m} \sum_{k_2=1}^{n} \left|S(k_1,k_2)\right| K(i-k_1, j-k_2)$$

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, $i = 1, 2, \ldots, M\text{-}m$, and $j = 1, 2, \ldots, N\text{-}n$.

The resulting convolved matrices h_{SB} of the sub-bands are then normalized so that the magnitudes range in the interval 0-1, and passed to the onset labeling stage.

2.2. Onset hypotheses labeling

Before starting the labeling of onsets, the normalized magnitudes of the h_{SB} matrices are summed along the frequency axis f. This function generates a one-dimensional signal with prominent peaks corresponding to the possible onsets. Then, by peak-picking, a binary signal is generated, indicating the presence or absence of possible onsets that should be evaluated at the probabilistic network. In Figure 3, an example of the spectrogram of a piano signal processed in the lowest frequency sub-band is shown in a), together with the resulting convolution with the kernel in b), and the one-dimensional signal of the energy integration function, c).

2.3. The probabilistic evaluation network

The train of pulses generated in the previous stage represents labels where hypothetically an onset was found. However, the final decision about whether they represent true onsets or not is taken based on a probabilistic evaluation. From each hypothetic onset signal of the sub-bands, the labels are observed and the data are fed to the network nodes shown in Figure 4.

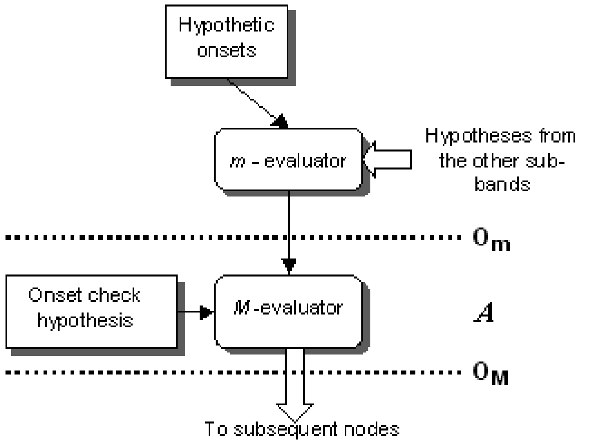

Figure 4. A single node for the probabilistic evaluation of the onset hypotheses.

Let us consider that we want to find the probability of the true onset inferred in A. Then letting O_m the hypothesis of an onset detected, and O_M be the hypothesis of the true onset detected and that will be propagated to the next nodes. This can be written as:

$$P(A) = P(A \mid O_M , O_m)$$

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

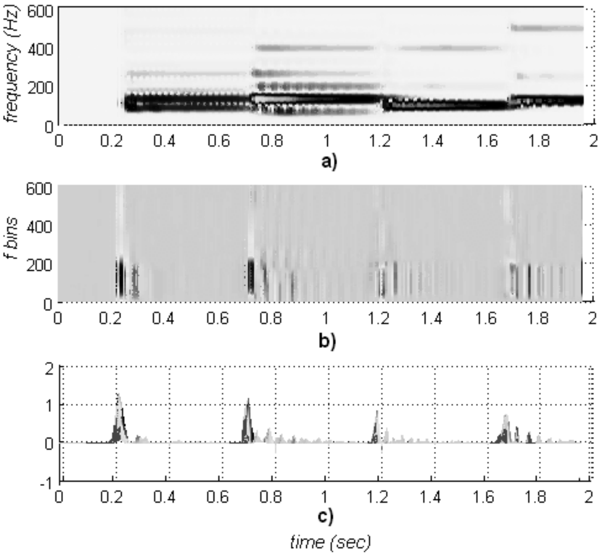

Figure 3. Example of a jazz piano 2 sec segment processed at the lowest frequency analysis sub-band. a) Signal spectrogram, b) Matrix after convolution, c) The prominent peak signal.

By using Bayes theorem, the probabilities can be expressed as:

$$P(A) = P(O_m \mid A) P(O_M \mid A)$$

which corresponds to the probabilities computed by the m and M evaluator after having observed the evidences in the hypotheses labels.

The hypotheses with their corresponding probabilities are then integrated at the onset information integrator and the final labels for true onsets are assigned.

3. Experiments

We performed preliminary experiments for the overall system evaluation. First, the samples used to train and construct the kernels were single piano notes sounds obtained from a standard music database [8], which contain sounds of some instruments recorded at a sampling rate of 44.1 kHz and 16 bits of resolution. The notes recorded were played with variations in articulation (Normal, Stacatto, Pedal), and dynamics (Forte, Mezzo, Piano), summing a total of 30 notes that were analyzed to train each kernel. In the CD labeled as RWC-MDB-I-2001-W01 of [8], the pictures of the pianos used to record the database are presented (showing a YAMAHA piano and a STEINWAY & SONS among them).

On the other hand, tests done until now, consist on 30 sec. segments of piano-solo performances recorded at 44.1 kHz and 16 bits of resolution, taken from different CD's and tracks of the same database. Although until now few classic and jazz piano samples have been tested with our system, more samples from commercial CD's and in other styles are being considered to verify the performance of the system. Table 1 resumes the sample segments used for preliminary tests.

The onsets of the sample segments were first labeled manually and then compared to those detected by the system. Then we estimate the overall performance as follows:

$$POD = \frac{Ncd - Ed}{Tot} \times 100\%$$

where:

- *POD* = Percentage of correct onset detection.
- *Ncd* = number of onsets correctly detected.
- *Ed* = erroneous detections.
- *Tot* = number of the total actual onsets (labeled manually for the corresponding segments).

These partial evaluations indicated average rates of 90.7% of onset detection for piano-solo in jazz style, and 94.3% for classic style, rates that are a meaningful indicative of the viability for improvements on this method.

Classic	
Piece	**# of segments tested**
Mozart, "Piano Sonata in A major, K. 331/300i"	1
Bethoven, "Piano Sonata no. 23 in F major"	1
Choppin, "Nocturne no. 2 in Eb major", "Etude in E major"	2
Total	**4**
Jazz	
Piece	**# of segments tested**
Jive (Piano solo)	1
For two (Piano solo)	1
Crescent Serenade (Piano solo)	1
Total	**3**

Table 1. Sample segments used for evaluation.

In this first approach to piano onset detection by using a convolutional kernel, we have modeled the kernel with parameters that defined at a first glance the transients of the piano onsets. However, the system has shown to be robust to spurious onset detections with the aid of the probability network. In further studies, the kernel can be refined to consider more complex characteristics of the piano onset transients and with this achieve higher levels of correct detection for most of the piano performance styles. Another extension that is considered for this method is the implementation of new kernels modeled from onsets of other instruments, so that the system will be theoretically able to detect onsets not only from piano sounds but also from other instruments in polyphonic signals.

4. Conclusions

We presented a new approach to piano onset detection based on a convolutional kernel modeled from the piano onset transients. In fact, the convolutional method used here is a particularized version of an image processing technique used to detect edges in gray-scaled images. The particularization resides in the core of this technique, the convolutional kernel. This method has proved to be useful to aid onset detection when applied to the analysis of the spectrogram of audio signals. Therefore, we can expect that, combining and updating this approach with more advanced and efficient techniques from image processing and pattern recognition, better results can be achieved.

References

[1] Scheirer E., "Tempo and beat analysis of acoustic musical signals", *Journal of Acoustic Soc. of America*, 103(1), 1998.

[2] Goto M., and Muraoka, "Real-time beat tracking for drumless audio signals"*, Speech Communications,* 27(3-4):331-335, 1999.

[3] Dixon S., Goebl M. and Widmer G., "Real-time tracking and visualization of musical expression", *Proceedings of the 2nd International Conference ICMAI 2002*, UK, 2002.

[4] Dubuxury C., Bello J., Davies M., and Sandler M., "Complex domain onset detection for musical signals", *Proc. of the 6th Intl. Conf. on Digital Audio Effects (DAFX-03)*, London, UK, 2003.

[5] Marolt M., "Adaptative oscillator networks for partial tracking and piano music transcription", *Proc. of the 2000 Intl. Computer Music Conf.*, Berlin, Germany, 2000.

[6] Marolt M., "SONIC: Transcription of polyphonic piano music with neural networks", *Workshop on CurrentRresearch Directions in Computer Music"*, Barcelona, Spain, 2001.

[7] Dixon S., "Learning to detect onsets of acoustic piano tones", *MOSART Workshop on Current Research Directions in Computer Music*, Barcelona, Spain, 2001.

[8] RWC Music Database, 2003.03, Real World Computer Partership, C MUSIC Corporation, Japan.