SIGNAL PROCESSING

An International Journal
A publication of the European Association for Signal Processing (EURASIP)

Editorial Policy. *Signal Processing* is an Interdisciplinary Journal presenting the theory and practice of signal processing. Its primary objectives are the following:

- Dissemination of research results and of engineering developments to all signal processing groups and individuals.
- Presentation of practical solutions to current signal processing problems in engineering and science.

The editorial policy and the technical content of the Journal are the responsibility of the Editor-in-Chief and the Editorial Board. The Journal is self-supporting from subscription income and contains a minimum amount of advertisements. The journal welcomes contributions from every country in the world.

Scope. *Signal Processing* incorporates all aspects of the theory and practice of signal processing (analogue and digital). It features original research work, tutorial and review articles, and accounts of practical developments. It is intended for a rapid dissemination of knowledge and experience to engineers and scientists working in signal processing research, development or practical application.

Subjects. Subject areas covered by the Journal include:
Signal Theory, Stochastic Processes, Detection and Estimation, Spectral Analysis, Filtering, Signal Processing Systems, Software Developments, Image Processing, Pattern Recognition, Optical Signal Processing, Digital Signal Processing, Multidimensional Signal Processing, Communication Signal Processing, Biomedical Signal Processing, Geophysical and Astrophysical Signal Processing, Earth Resources Signal Processing, Acoustic and Vibration Signal Processing, Data Processing, Remote Sensing, Signal Processing Technology, Speech Processing, Radar Signal Processing, Sonar Signal Processing, Special Signal Processing, Industrial Applications, New Applications.

Publication information: *Signal Processing* (ISSN 0165-1684). For 2010, Volume 90 (12 issues) is scheduled for publication. Subscription prices are available upon request from the Publisher or from the Regional Sales Office nearest you or from this journal's website (http://www.elsevier.com/locate/sigpro). Further information is available on this journal and other Elsevier products through Elsevier's website: (http://www.elsevier.com). Subscriptions are accepted on a prepaid basis only and are entered on a calendar year basis. Issues are sent by standard mail (surface within Europe, air delivery outside Europe). Priority rates are available upon request. Claims for missing issues should be made within six months of the date of dispatch.

Orders, claims, and journal enquiries: please contact the Elsevier Customer Service Department nearest you:

St. Louis: Elsevier Customer Service Department, 11830 Westline Industrial Drive, St. Louis, MO 63146, USA; phone: (877) 8397126 [toll free within the USA]; (+1) (314) 4537076 [outside the USA]; fax: (+1) (314) 5235153; e-mail: JournalCustomerService-usa@elsevier.com

Oxford: Elsevier Customer Service Department, The Boulevard, Langford Lane, Kidlington OX5 1GB, UK; phone: (+44) (1865) 843434; fax: (+44) (1865) 843970; e-mail: JournalsCustomerServiceEMEA@elsevier.com

Tokyo: Elsevier Customer Service Department, 4F Higashi-Azabu, 1-Chome Bldg, 1-9-15 Higashi-Azabu, Minato-ku, Tokyo 106-0044, Japan; phone: (+81) (3) 5561 5037; fax: (+81) (3) 5561 5047; e-mail: JournalsCustomerServiceJapan@elsevier.com

Singapore: Elsevier Customer Service Department, 3 Killiney Road, #08-01 Winsland House I, Singapore 239519; phone: (+65) 63490222; fax: (+65) 67331510; e-mail: JournalsCustomerServiceAPAC@elsevier.com

Advertising information. Advertising orders and enquiries can be sent to: Janine Castle, Elsevier Ltd., The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK; phone: (+44) 1865 843 844; fax: (+44) 1865 843 973; e-mail: j.castle@elsevier.com.

For full membership information of the Association, possibly combined with a subscription at a reduced rate, please contact: EURASIP, P.O. Box 134, CH-10000 Lausanne 13, Switzerland.

USA mailing notice: *Signal Processing* (ISSN 0165-1684) is published monthly by Elsevier B.V. (P.O. Box 211, 1000 AE Amsterdam, The Netherlands). Periodical postage rate paid at Rahway, NJ and additional mailing offices.

USA POSTMASTER: Send address changes to *Signal Processing*, Elsevier, Customer Service Department, 11830 Westline Industrial Drive, St. Louis, MO 63146, USA.

AIRFREIGHT AND MAILING in the USA by Mercury International Limited, 365, Blair Road, Avenel, NJ 07001.

♾ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Printed by Henry Ling Ltd., The Dorset Press, Dorchester, UK

Signal Processing 89 (2009) 1798–1811

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Event based transcription system for polyphonic piano music

Giovanni Costantini [a], Renzo Perfetti [b,*], Massimiliano Todisco [a]

[a] Department of Electronic Engineering, University of Rome 'Tor Vergata', Italy
[b] Department of Electronic and Information Engineering, University of Perugia, Italy

ARTICLE INFO

Article history:
Received 24 July 2008
Received in revised form
2 February 2009
Accepted 18 March 2009
Available online 5 April 2009

Keywords:
Onset detection
Music transcription
Classification
Constant Q transform
Support vector machines

ABSTRACT

Music transcription consists in transforming the musical content of audio data into a symbolic representation. The objective of this study is to investigate a transcription system for polyphonic piano, triggered by events corresponding to the played notes. The proposed method focuses on note events and their main characteristics: the attack instant, the pitch and the final instant. Onset detection exploits a binary time-frequency representation of the audio signal. Note classification and offset detection are based on constant Q transform (CQT) and support vector machines (SVMs). We present a collection of experiments using synthesized MIDI files and piano recordings, and compare the results with existing approaches.

1. Introduction

Music transcription can be considered as one of the most demanding tasks performed by our brain; not so many people are able to easily transcribe a musical score starting from audio listening, since the success of this operation depends on musical abilities, as well as on the knowledge of the mechanisms of sounds production, of musical theory and styles, and finally on musical experience and practice to listening.

Musical transcription consists in the extraction of musical content from audio data, i.e. a symbolic representation of musical notes commonly called *musical score*. A musical score contains only three basic informations: first the *note pitch*, corresponding to the fundamental frequency of sound, then its temporal features corresponding to the attack instant or *note onset* and the final instant or *note offset*; the latter feature has a lesser perceptual importance in comparison to the attack instant.

It is necessary to distinguish two cases in which the behavior of the automatic transcription systems is different: monophonic music, where notes are played one-by-one and polyphonic music, where two or several notes can be played simultaneously.

Currently, automatic transcription of monophonic music is treated in the time domain e.g. by means of zero-crossing or autocorrelation techniques, and in the frequency domain by means of discrete Fourier transform (DFT) or cepstrum. With these techniques an excellent accuracy level has been achieved [1,2].

Attempts in automatic transcription of polyphonic music have been much less successful; actually, the harmonic components of notes that simultaneously occur in polyphonic music significantly obfuscate automated transcription. The first algorithms were developed by Moorer [3,4] and Piszczalski and Galler [5]. Moorer [3] used comb filters and autocorrelation in order to perform transcription of very restricted duets. Among the most important works in this research field are the transcription systems proposed by Ryynanen and Klapuri [6], the SONIC project [7] developed by Marolt, and the transcription model of Poliner and Ellis [8]. The last two works exploit a classification approach to note transcription

* Corresponding author.
E-mail address: perfetti@diei.unipg.it (R. Perfetti).

0165-1684/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2009.03.024

based on neural networks and support vector machines (SVMs), respectively. In particular the results in [8] are very encouraging in pursuing a classification strategy to polyphonic piano transcription.

The main limitation of method proposed in [8] consists in the frame-by-frame operation, obtained applying to the audio files a sliding window. An explicit onset detection algorithm is not used, so onset detection can be performed with low precision (100 ms, equivalent to a 16th note at 80 metronome beat) and spurious onsets appear if the note is alternately detected or not detected several times in adjacent windows.

Moreover, evaluating the classification accuracy on a window basis is not meaningful from a perceptual viewpoint, since it does not take into account the temporal position of misclassifications: an error in the final part of the note is far less important than intermediate errors, since the last ones cause false onsets on the same note pitch with nasty perceptual effects. In other words, an high transcription accuracy of music frames does not guarantee good reconstruction quality.

In this paper, partly motivated by [8], we present a novel system for automatic transcription of polyphonic piano music, and elaborate on its essential features. Its main phases involve: (a) an onset detection algorithm, providing sufficient precision on note attack instant measurement; (b) recognition of note pitch for each music event corresponding to a note onset; (c) note offset detection.

The onset detection algorithm operates on a frame-by-frame basis and exploits a suitable binary time-frequency representation of the audio signal. Note classification is asynchronous, aligned with the note onsets, and is based on a bank of SVMs combined with constant Q transform (CQT) for features extraction. Finally, offset detection is obtained by checking frame-by-frame the SVM outputs, starting from a note onset.

For sake of comparison with [8], both for training and for test we used the same audio dataset, consisting of a rich collection of piano pieces of different musical styles. However, differently from [8], the evaluation of transcription accuracy has been carried out on an event basis, for the abovementioned reasons.

The rest of the paper is organized as follows. Section 2 illustrates the onset detection algorithm, while Section 3 describes the spectral features. Section 4 will be devoted to the description of the classification method. In Section 5 we present the results of a series of experiments involving synthesized MIDI files and piano recordings. Some comments conclude the paper.

2. Onset detection

Onset detection is the problem of identifying the time at which a note is sounded. These onsets are expected to emphasize the important moments of a melody and the music beats. Onset detection is a challenging problem investigated by several authors in recent years. Among the methods proposed in the literature, interesting results for piano music have been obtained using a bank of filters followed by multilayer perceptrons [9], convolutional kernels in the frequency domain [10], adaptive linear prediction [11], and machine learning techniques (neural networks and SVMs) [12]. A comparison of our method with [9–12] is presented in Section 5.1.

The proposed onset detection algorithm is based on STFT combined with a suitable binary processing in order to improve the precision of onset measurement.

Let us consider a discrete-time signal $s(n)$, whose STFT is given by

$$S_k(m) = \sum_{n=mh}^{mh+N-1} w(n-mh)s(n)e^{-j\Omega_N k(n-mh)} \quad (1)$$

where N is the window size, h is the hop size, $m \in \{0,1,2,\ldots,M\}$ the hop number, $k = 0, 1,\ldots,N-1$ is the frequency bin index, $w(n)$ is a finite-length sliding Hanning window and n is the summation variable.

We obtain a time-frequency representation of the audio signal computing its magnitude spectrum $|S_k(m)|$. The set of spectra $|S_k(m)|$ can be packed as columns into a non-negative $L \times M$ matrix, where M is the total number of spectra we computed and $L = N/2$ is the number of their frequencies. After normalization in the range from 0 to 1 we obtain a matrix $D \in [0,1]^{L\times M}$. Then, we perform a binarization of D giving the binary matrix $\overline{D} \in \{0,1\}^{L\times M}$:

$$\overline{D}(l,m) = \begin{cases} 1 & \text{if } D(l,m) > T_1 \\ 0 & \text{if } D(l,m) \leqslant T_1 \end{cases} \quad (2)$$

where T_1 is a threshold. Binarization is used to avoid the effect of spectral noise and to allow some simple 'spatial' operations, as illustrated in the following. The best threshold value can be obtained using a suitable validation set of examples, as explained in Section 5. The first two processing steps are illustrated in Fig. 1.

To detect the note onsets, we perform two further operations on the binary spectrogram $\overline{D}$. First, we set to '0' the (l,m) element of $\overline{D}$ if the previous adjacent cell $(l,(m-1))$, relative to the previous frame, is equal to '1'. This operation is adopted to point out only the spectral changes in the time-frequency representation. The result of this processing step is a new binary matrix $\overline{\overline{D}} \in \{0,1\}^{L\times M}$:

$$\overline{\overline{D}}(l,m) = \begin{cases} 0 & \text{if } \overline{D}(l,m-1) = 1 \\ \overline{D}(l,m) & \text{otherwise} \end{cases} \quad (3)$$

Then we set to '0' the (l,m) element of $\overline{\overline{D}}$ if both the previous adjacent cell $((l-1),m)$, relative to the previous frequency bin, and the subsequent adjacent cell $((l+1),m)$, relative to the subsequent frequency bin, are equal to '0'. This operation removes isolated spectral bins in the time-frequency representation. The result of this processing step is a new binary matrix $\overline{\overline{D}} \in \{0,1\}^{L\times M}$:

$$\overline{\overline{D}}(l,m) = \begin{cases} 0 & \text{if } \overline{\overline{D}}(l-1,m) = 0 \wedge \overline{\overline{D}}(l+1,m) = 0 \\ \overline{\overline{D}}(l,m) & \text{otherwise} \end{cases} \quad (4)$$

Operations (3) and (4) are summarized in Fig. 2.

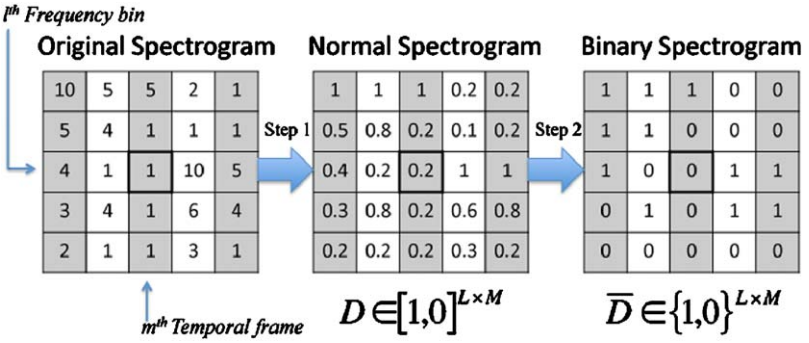

Fig. 1. Normalization and binarization with $T_1 = 0.3$.

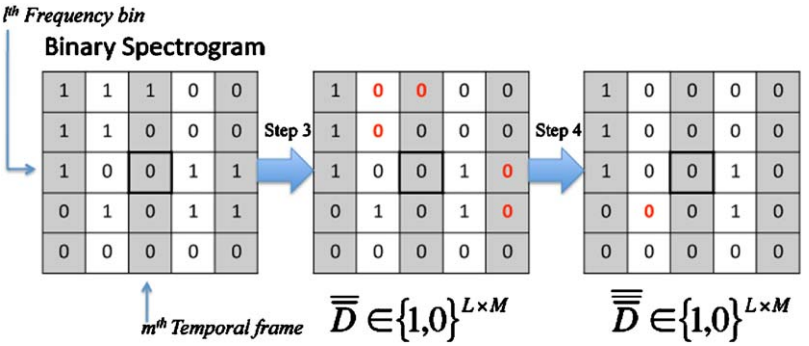

Fig. 2. Results of operations (3) and (4).

Afterwards, the rows of $\overline{\overline{D}}$ are summed, giving the following onset detection function:

$$f(m) = \sigma_m^2 \sum_{l=1}^{L} \overline{\overline{D}}(l,m) \tag{5}$$

where

$$\sigma_m^2 = \frac{1}{L}\sum_{l=1}^{L}(\overline{\overline{D}}(l,m) - \mu_m)^2 \tag{6}$$

and μ_m is the average value of entries in mth column of matrix $\overline{\overline{D}}$. The multiplication by the variance is motivated by the following facts observed in our experiments:

- in correspondence of most onsets the number of 1's and 0's in column $\overline{\overline{D}}(.,m)$ are comparable, so the variance σ_m^2 is large;
- in same onsets, most of the elements in column $\overline{\overline{D}}(.,m)$ are equal to one, so the variance is small but the sum in (5) is large;
- when the onset is absent we have small variance and small number of 1's.

In conclusion, the multiplication by the variance enhances most of true onsets while damping spurious onsets. Therefore, the peaks of $f(.)$ can be assumed to represent the times of note onsets. After peak picking, a second threshold T_2 is used to suppress spurious peaks; its value is obtained together with T_1 through a validation process (see Section 5).

To show the performance of our onset detection method, let us consider an example from real piano polyphonic music of Mozart's KV 333 Sonata in B-flat Major, Movement 3, sampled at 8 kHz and quantized with 16 bits. We will consider the second and third bars at 120 metronome beat. It is shown in Fig. 3.

We use a STFT with $N = 512$, an N-point Hanning window and a hop size $h = 256$ corresponding to 32 ms hop between subsequent frames. The original spectrogram is shown in Fig. 4. The normalized spectrogram and the binary spectrogram are shown in Figs. 5a and b, respectively (the *threshold* was set to 0.01). The last two processing steps are applied to enhance the presence of a new frequency bin and to remove isolated frequency bins. The results are shown in Figs. 6a and b. Summing the elements of each column in Fig. 6b we obtain the sum of rows in Fig. 7a and, after multiplication by the variance, the onset detection function in Fig. 7b. The time onset resolution is 32 ms. A statistical evaluation of the onset detection method will be presented in Section 5.

3. The constant Q transform and the spectral features

A frequency analysis must be performed on notes played by piano, in order to detect the signal harmonics.

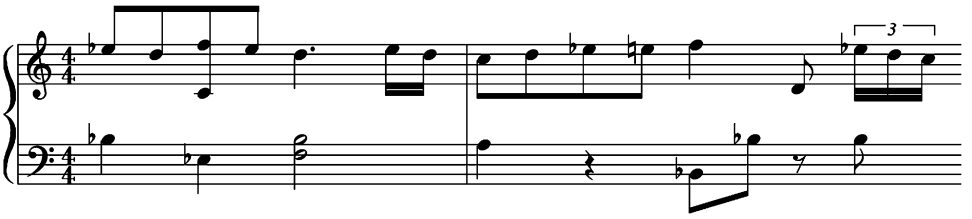

Fig. 3.

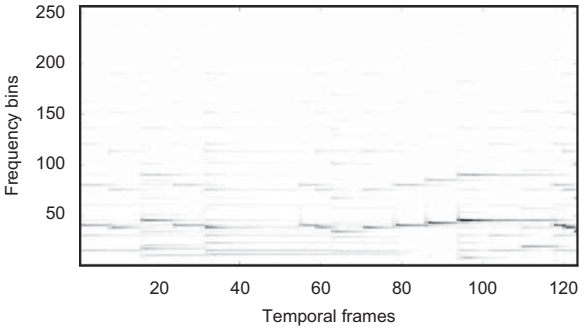

Fig. 4. Spectrogram corresponding to Fig. 3.

The frequency resolution of usual discrete Fourier transform may not be sufficient. In fact, 512 temporal samples recorded with the usual sampling rate (SR) of 8000 samples/s, correspond to a spectral separation of 15.6 Hz between two adjacent DFT samples. This is not sufficient for low frequency notes, where the distance between two adjacent semitones is about 8 Hz (C3, 131 Hz and C#3, 139 Hz). The spectral resolution can be improved using an higher number of temporal samples (with 1024 samples the resolution is about 7.8 Hz), but this requires longer temporal windows for a fixed sampling rate, worsening the time resolution. To solve this problem, a Constant Q Transform (CQT) [13–14] has been used to detect the fundamental frequency of the note. Then, the upper harmonics may be easily detected, as they are located nearly at integer multiples of the fundamental frequency. The logarithmic frequency scale provides a constant frequency-to-resolution ratio for every bin:

$$Q = \frac{f_k}{f_{k+1} - f_k} = \frac{1}{2^{1/b} - 1} \tag{7}$$

where b is the number of bins per octave and k is the frequency bin. If $b = 12$, and by choosing a suitable f_0, then k is equal to the MIDI note number (as in the equal-tempered 12-*tone-per-octave* scale).

In our system, the processing phase starts in correspondence to a note onset. First, the attack time of the note is discarded (in case of the piano, the longest attack time is equal to about 32 ms). Then, after Hanning windowing, a single CQT of the following 64 ms of the audio note event is computed. Fig. 8 shows the features extraction process. Notice that two or more notes belong to the same onset if these notes are played within 32 ms, but this is not a problem provided that the following classification stage can detect the correct notes.

All the audio files have a sampling rate of 8 kHz. We used $b = 48$, i.e. 4 *CQT-bins* per semitone, starting from note C0 (~32 Hz) up to note B6 (~3951 Hz). The output of the processing phase is a matrix with 336 columns, corresponding to the *CQT-bins*, and a number of rows equal to the total number of note events in the MIDI file. The scale of the values of the frequency bins is also logarithmic, rescaled into a range from 0 to 1.

4. Multi-class SVM classification

A SVM identifies the *optimal separating hyperplane* (OSH) that maximizes the margin of separation between linearly separable points of two classes. The data points which lie closest to the OSH are called *support vectors*. It can be shown that the solution with maximum margin corresponds to the best generalization ability [15]. Linearly non-separable data points in *input space* can be mapped into a higher dimensional (possibly infinite dimensional) *feature space* through a nonlinear mapping function, so that the images of data points become almost linearly separable. The discriminant function of a SVM has the following expression:

$$f(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{8}$$

where $\mathbf{x}_i$ is a support vector, $K(\mathbf{x}_i, \mathbf{x})$ is the kernel function representing the inner product between $\mathbf{x}_i$ and $\mathbf{x}$ in feature space, coefficients α_i and b are obtained by solving a quadratic optimization problem in dual form [15].

Usually, a soft-margin formulation is adopted where a certain amount of noise is tolerated in the training data. To this end, a user-defined constant $C > 0$ is introduced which controls the trade-off between the maximization of the margin and the minimization of classification errors on the training set [15].

SVMs were originally designed to work with dichotomies. A standard way to solve multi-class problems is to consider them as a collection of binary sub-problems, and then to combine their solutions. In this context, the one-versus-all (OVA) approach has been used. The OVA method constructs N SVMs, N being the number of classes. The ith SVM is trained using all the samples in the ith class with a positive class label and all the remaining samples with a negative class label. Our transcription system uses 84 OVA SVM note classifiers whose input is represented by a 336-element feature vector, as described in Section 3. The presence of a note in

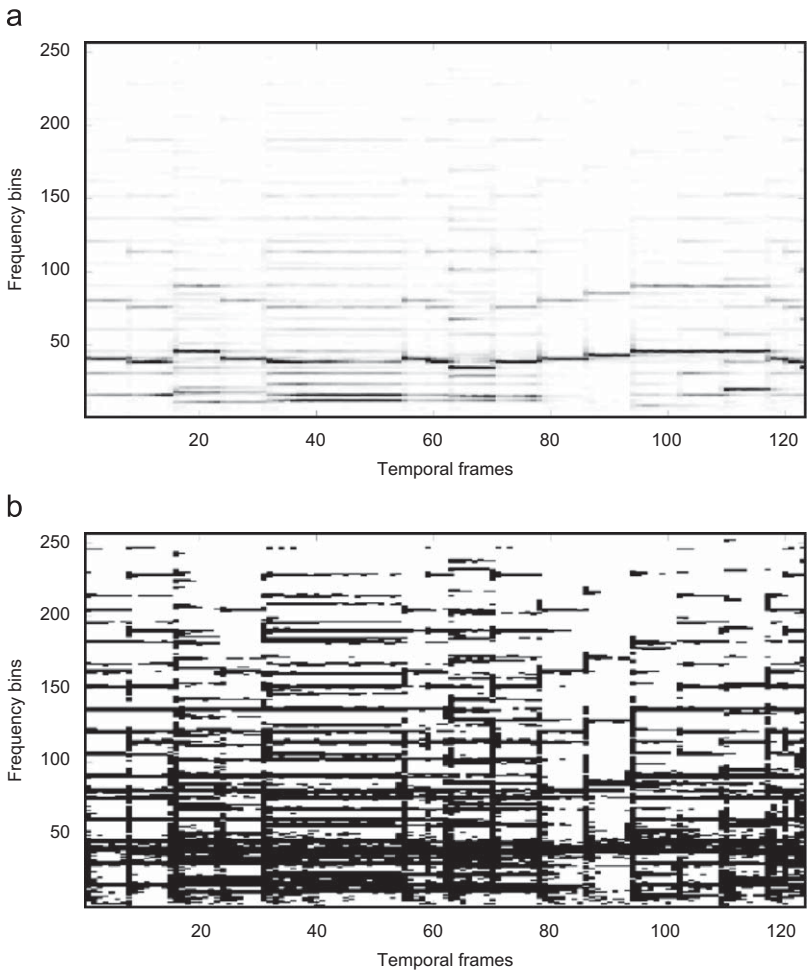

Fig. 5. Results of normalization (a) and binarization (b) of the spectrogram in Fig. 4.

a given audio event is detected if the discriminant function of the corresponding SVM classifier is positive.

SVM training was implemented using the SVM^{light} software developed by Joachims [16]. A radial basis function (RBF) kernel was adopted:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2), \quad \gamma > 0 \tag{9}$$

where γ describes the width of the Gaussian function. Using the RBF kernel two parameters, C and γ, must be fixed. To this end we looked for the best parameter values in a specific range using a grid-search on a validation set. More details will be given in the following section. Fig. 9 shows a schematic view of the complete automatic transcription process.

4.1. Offset detection

A comprehensive music transcription system should include offset detection. As said above, offset detection is less important than onset detection from a perceptual viewpoint. So, the required precision can be lower without music degradation.

In our system offset detection is obtained observing the output of the OVA classifiers over consecutive signal frames. The algorithm is as follows: (a) starting from an onset detection, and until the successive onset, we consider a sliding window of 512 samples with hop size of 256 samples, and for each window position we perform CQT and SVM classification; (b) if a note is detected up to window position m, and not detected in position m+1, we assume the offset time corresponds to window position m; (c) further detections of the same note pitch before the successive onset time are ignored. Moreover, when the note offset follows the next onset instant detected by the system, we avoid an incorrect played note ignoring the new onset.

The rationale of the proposed algorithm is that a note cannot end and restart in the interval between two consecutive onsets detected by the system. If this happens, there is an error due to one of the following causes: (1) the note has been played again; this entails an onset detection error, an hypothesis we reject having very low probability; (2) a terminated note reappear due to a pitch classification error; (3) the note is not really terminated. We assume that case (2) is the most likely, hence we ignore new note detections before the following onset time.

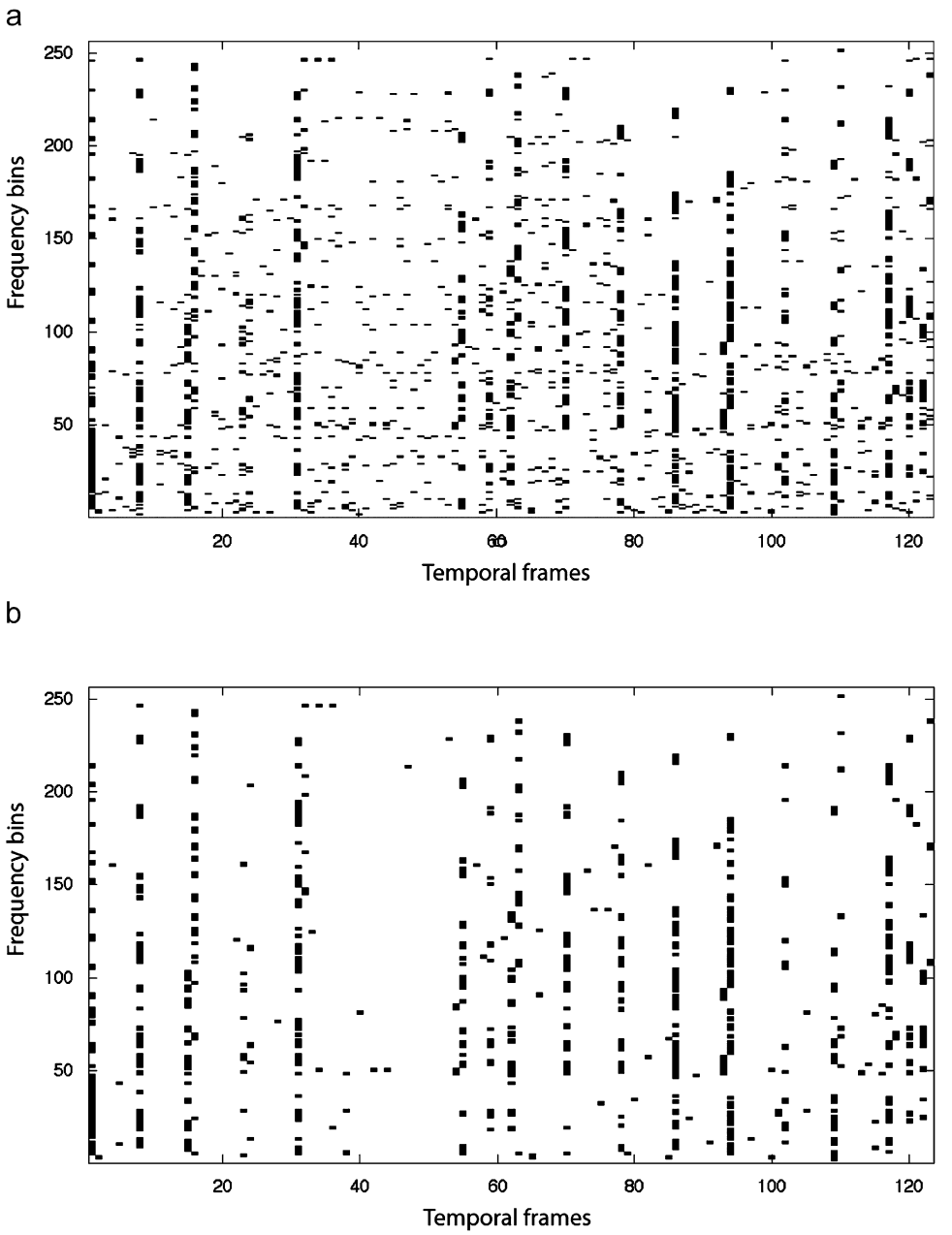

Fig. 6. Results of spatial processing operations on the binary spectrogram in Fig. 5b.

5. Simulation results

In this section we report on the simulation results of our transcription system and compare them with some existing methods. The MIDI data used in the experiments were collected from the Classical Piano MIDI Page, http://www.piano-midi.de/ [9]. A list of pieces can be found in [9, p. 8, Table 5]. The 124 pieces dataset was randomly split into 87 training, 24 testing, and 13 validation pieces. The first minute from each song in the dataset was selected for experiments, providing a total of 87 min of training audio, 24 min of testing audio, and 13 min of audio for parameter tuning (validation set). This amounted to 22 680, 6142, and 3406 note onsets in the training, testing, and validation sets, respectively. The distribution of notes in the three datasets are shown in Fig. 10.

5.1. Onset detection

First, we performed a statistical evaluation of the performance of the onset detection method. The results are summarized by three statistics: the precision *P*, the

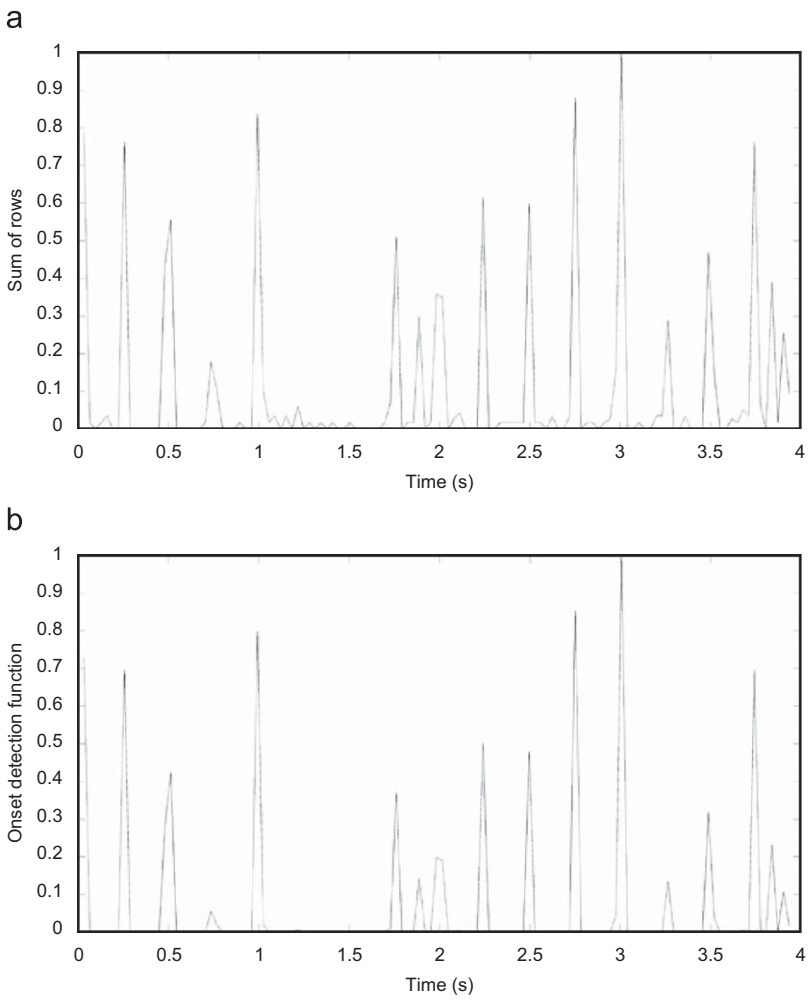

Fig. 7. (a) Sum of rows and (b) onset detection function for the example in Fig. 3.

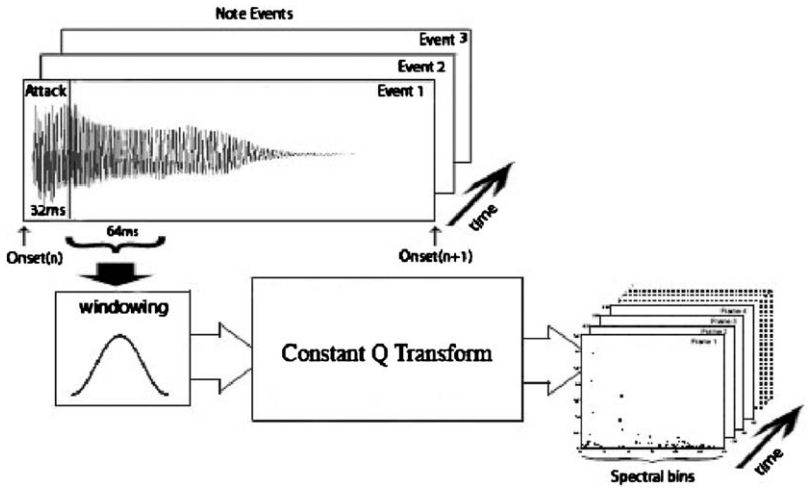

Fig. 8. Spectral features extraction.

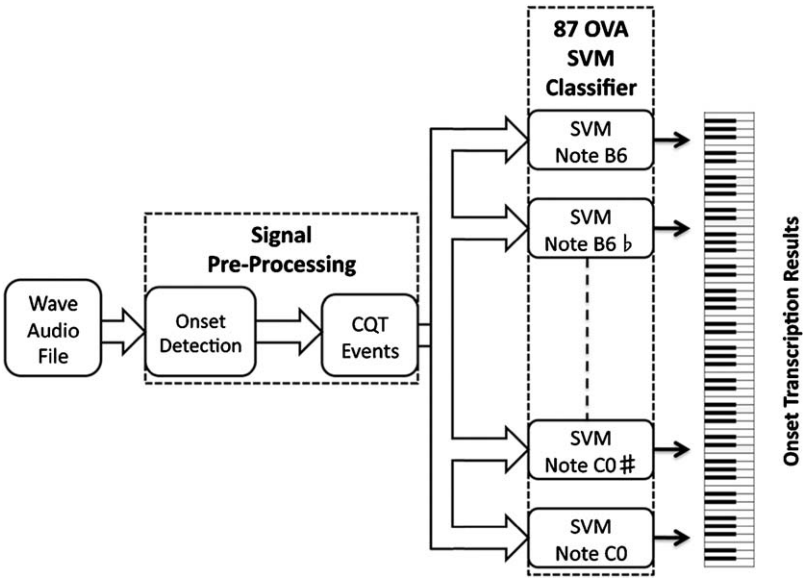

Fig. 9. Scheme of the proposed event-based transcription system.

recall R and the F-measure, which are given by

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F\text{-}measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{12}$$

In the above formulas TP is the number of correct detections, FP is the number of false positives and FN is the number of false negatives. *Precision* represents the percentage of correct positive predictions in the identification of an example. *Recall* represents the capacity of the onset detector to identify the positive examples. The global variable *F-measure* is the harmonic mean of *Precision* and *Recall*.

The onset detection algorithm requires two thresholds: T_1, for spectrogram binarization, and T_2, to suppress spurious peaks in the onset detection function. The threshold values have been obtained through maximization of the F-measure value on the validation dataset. Fig. 11 shows the contour plot of F-measure vs. threshold values. The optimal values are $T_1 = 0.0073$ and $T_2 = 0.0215$. Fig. 12 shows Precision, Recall and F-measure values versus T_1 while T_2 is held fixed to the optimal value. Finally, Table 1 illustrates the good performance of the method on the test set (including 6142 onsets).

We can try a comparison with some of the proposed methods for onset detection of piano music. In [9] the onset detector, based on a bank of filters and multilayer perceptrons, is trained using synthesized piano. The test set is represented by only 6 pieces (3 real piano, 3 synthesized piano), and the accuracy is 98%. The accuracy of our method on the test set (24 pieces) is 98.1%. The method in [10] has been tested both on jazz and classic pieces (3+3). The metrics adopted to evaluate the performance is the percentage of correct onset detection: $\mathrm{POD} = (N_{cd} - E_d)/N_{tot}$, where N_{cd} is the number of correct onset detections, E_d is the number of erroneous detections and N_{tot} is the total number of actual onsets (manually labelled). For the method in [10] it is POD = 90.7% on jazz pieces and POD = 94.3% on classic pieces. With our method we obtain POD = 93.3% on a test set of 24 pieces of classic piano. The method in [11] has been tested on only 205 onsets of piano music giving precision 98%, recall 99% and F-measure 98%. Finally, in [12] the applicability of neural networks and SVMs for onset recognition is investigated. Both for training and test, the J.S. Bach's 24 preludes for well tempered harpsichord have been used. In particular, the test set consists of the first 32 beats of each prelude, and the training set consists of the rest. The best results have been obtained using SVMs: precision 88%, recall 85%, F-measure 86.5%. In conclusion, notwithstanding its simplicity, our method gives better or comparable performance on a wider test set.

5.2. Note classification

After detection of the note onsets, we trained the SVMs on the 87 pieces of the training set and we tested the system on the 24 pieces of the test set. The results are outlined in Table 2.

Similar to [8], in addition to the synthesized audio, 19 training files and 10 test files piano recordings were made from a subset of the MIDI files using a Yamaha Disklavier playback grand piano. The results obtained using both full-synthesized and recorded data are outlined in Table 3.

The entries in Table 3 concerning the methods in [6–8] are taken from ([8], Table 4). It is worth noting that the

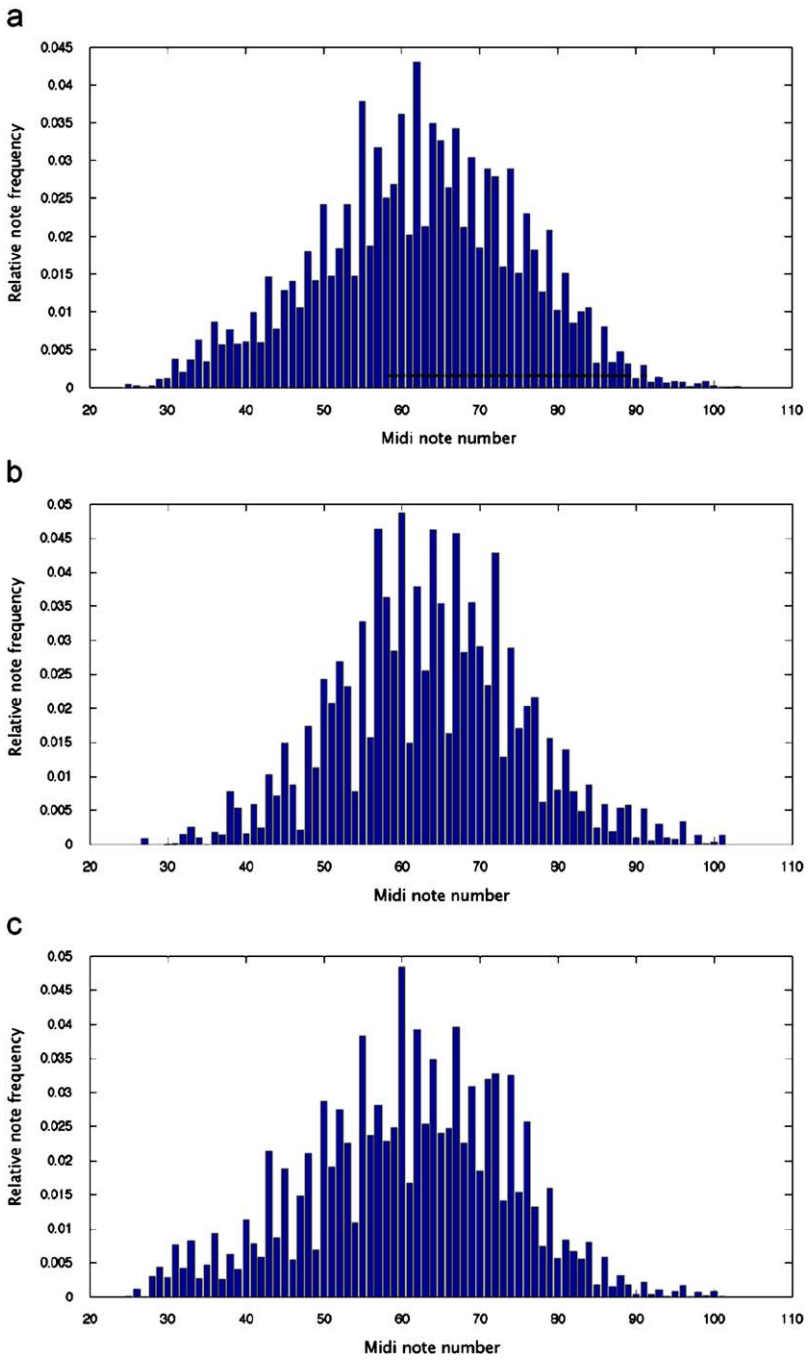

Fig. 10. Distribution of notes in the (a) training, (b) validation and (c) test sets.

comparison can be made only with note onset transcription results; the remaining performance results in [8] cannot be compared with ours since the percentages were computed on a frame-basis.

In Tables 2 and 3, *Acc* denotes the accuracy metric proposed by Dixon [17], E_{tot} is the *transcription error score* defined by NIST (National Institute of Standards Technology) [18] for evaluations of 'who spoke when' in recorded meetings, E_{subs} is the percentage of substitution errors, while E_{miss} and E_{fa} denote the percentages of 'miss' and 'false alarm' errors. The exact definition of the above-mentioned quantities can be found in [8], where the number of frames T must be replaced with the number of note onset events.

The results in Table 3 reveal some general trends with respect to existing methods: the overall accuracy is increased while the total error is lower, even if the substitution error is increased with respect to [6,8]. We

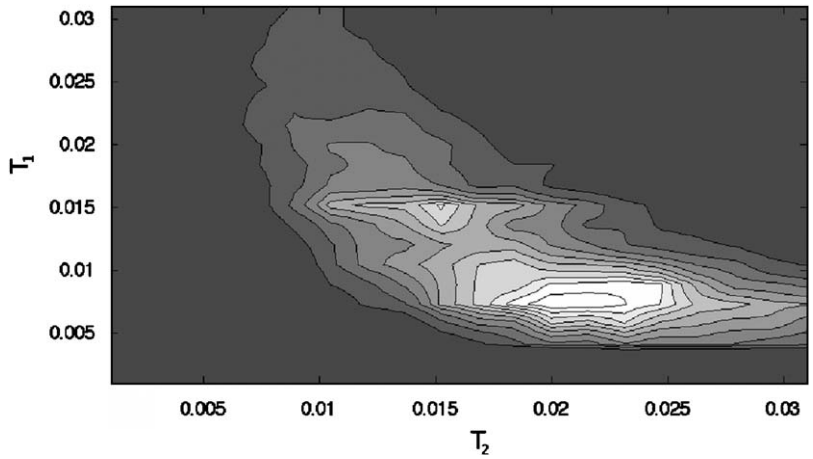

Fig. 11. *F-measure* vs. threshold values (brighter pixels correspond to larger values of *F-measure*).

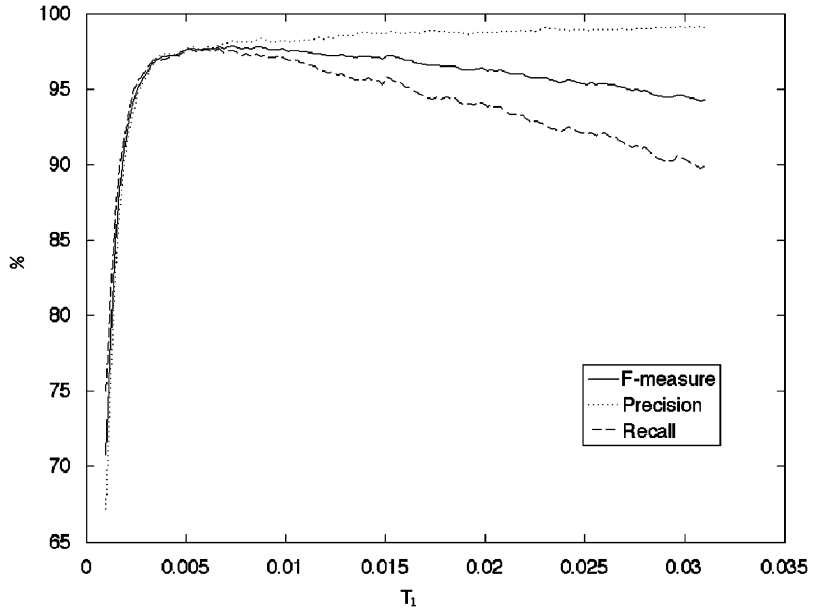

Fig. 12. *Precision*, *Recall* and *F-measure* for the validation set versus threshold T_1 ($T_2 = 0.0215$).

Table 1
Test results of the onset detection algorithm.

Precision	98.0%
Recall	97.6%
F-measure	97.8%

Table 2
Note onset transcription results on full synthesized audio test set.

	Acc (%)	E_{tot} (%)	E_{subs} (%)	E_{miss} (%)	E_{fa} (%)
Proposed method	72.3	20.1	10.5	9.4	0.02

Table 3
Note onset transcription results on full synthesized-plus-recorded test set.

	Acc (%)	E_{tot} (%)	E_{subs} (%)	E_{miss} (%)	E_{fa} (%)
Proposed method	68.0	24.6	11.3	13.2	0.04
Poliner and Ellis	62.3	43.2	4.5	16.4	22.4
Ryynanen and Klapuri	56.8	46.0	6.2	25.3	14.4
Marolt	30.4	87.5	13.9	41.9	31.7

note a drastic decrease of E_{fa} (percentage of false alarms) in both experiments, compared to other classification-based methods. We suppose this is due to the spectral features we have adopted, based on the constant Q transform applied on temporal windows where piano

sound wave is almost periodic. Finally, let us stress that in the classification results presented in [8] a correct detection happens if the system 'switches on' a note of the correct pitch within 100 ms of the ground-truth onset. Instead, our algorithm has a 32 ms time onset resolution.

An analysis of missing errors has been carried out to ascertain the causes of missing detections (the results are summarized in Table 4). We found that 89.3% of missing errors are characterized by the presence of a correctly detected note at an octave interval separation from the missing note. Moreover, 8.4% (1.8%) of missing errors are characterized by the presence of a correctly detected note at a fifth (third) interval separation from the note which has not been recognized. So, in most cases the missing note shares all or most of its harmonics with a second correctly detected note. Only 0.5% of missing errors does not belong to the previous case. Furthermore, we found that in 93.3% of missing errors the note has been played before the instant of missing detection; namely, in most cases the missing errors concern note 'tails'.

Finally, we present the results of two further experiments. Table 5 shows the performance on only the recorded test set without the synthesized one (we remember that our results are not directly comparable to [8], as explained above).

Table 6 shows the results obtained using pianos with different timbres for training and test. As can be expected, the performance is worse with respect to Table 2; this can be easily explained by Fig. 13, showing the spectra of note C2 ($\approx$131 Hz) for both pianos: we can see that the amplitudes of harmonics are very different. Probably, better generalization can be expected by training the SVMs using different timbres.

Table 4
Analysis of missing errors.

Separation interval	Presence of a note with overlapping harmonics (%)
Octave	89.3
Fifth	8.4
Third	1.8
Other	0.5

Table 5
Note onset transcription results on recorded test set.

	Acc (%)	E_{tot} (%)	E_{subs} (%)	E_{miss} (%)	E_{fa} (%)
Proposed method	59.2	33.3	12.9	20.0	0.4
Poliner and Ellis	56.5	46.7	10.2	15.9	20.5

Table 6
Training using iTunes_Sinth (grand piano); test using iTunes_Sinth (acoustic piano).

	Acc (%)	E_{tot} (%)	E_{subs} (%)	E_{miss} (%)	E_{fa} (%)
Proposed method	64.0	33.3	6.8	21.9	4.6

5.3. Offset detection

A statistical evaluation of the offset detection method was carried out with the same audio dataset used for onset detection. The results of this test are shown in Table 7, where we consider as correct the offset detected within 100 or 200 ms of the ground-truth offset.

5.4. Piano transcription example

We show now a specific example of piano music transcription concerning Chopin Opus 10, n. 1. We consider the fourth fifth and sixth bar at 60 metronome beat. The notes of these three bars have a wide range pitch

Table 7
Test results of the offset detection algorithm.

	100 ms (%)	200 ms (%)
Precision	60.4	76.3
Recall	69.0	73.7
F-measure	64.4	75.0

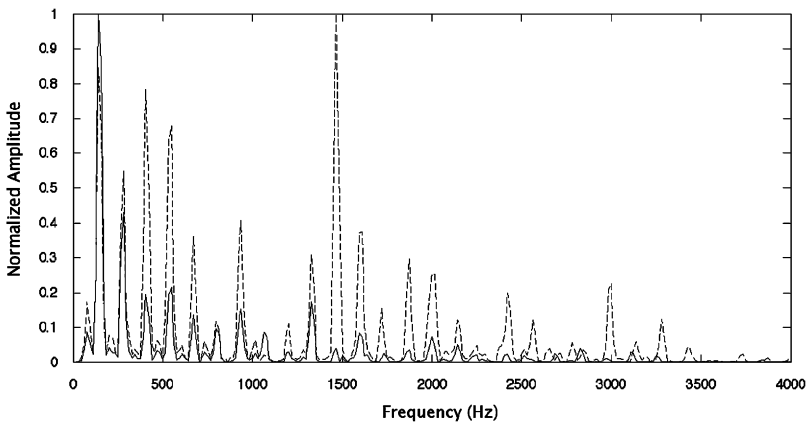

Fig. 13. Normalized amplitude spectra of note C2: iTunes_Sinth (grand piano) (continuous line); iTunes_Sinth (acoustic piano) (dashed line).

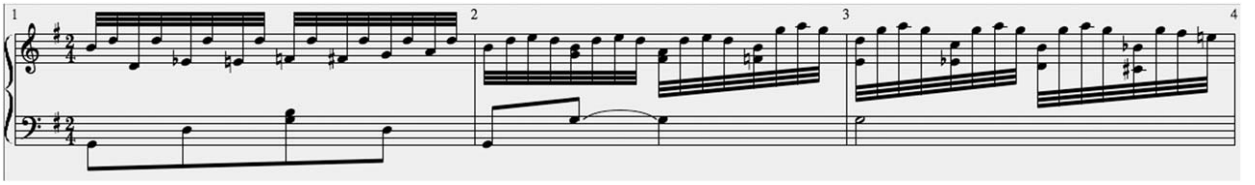

Fig. 14. Original piano score of Chopin Opus 10 n. 1.

Fig. 15. Original midi piano roll corresponding to Fig. 14.

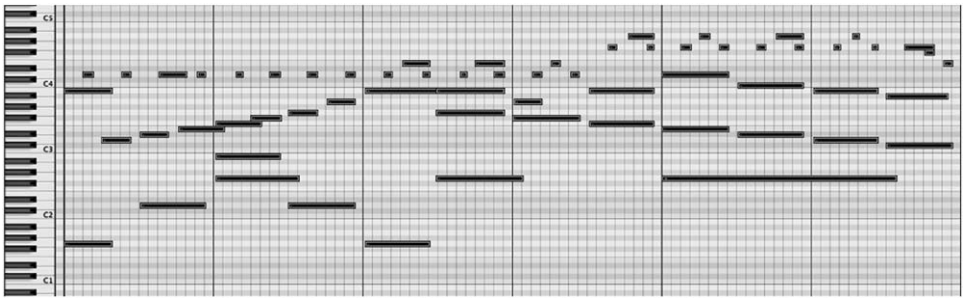

Fig. 16. Transcription result corresponding to the midi piano roll in Fig. 14.

(from G2 to A5) and duration (from 0.125 to 2 s). In this example all 63 notes were successfully detected. In Figs. 14–16 the original piano score, the original midi piano roll and the transcription result midi piano roll are shown. We note some differences between the original and transcribed piano roll, mainly for the offset time, but this not affects the quality of the reconstruction. Finally, in Fig. 17 we show the distribution of the difference between ground-truth and detected onset time and offset time, respectively.

6. Conclusions

In this study, we presented a polyphonic piano transcription system based on the characterization of note events. As a result of the underlying procedures, for each event we measure both the note onset and the note offset, the last with lower precision, and then we classify the note pitch. These informations can be directly used to reconstruct the piano roll. We pursued an event-based analysis of errors, since frame-level performance scores do not take into account the perceptual effects due to the temporal position of missing or uncorrect detections.

It has been shown that the proposed onset detection is helpful in the determination of note attacks with modest computational cost and good accuracy. It has been found that the choice of CQT for spectral analysis plays a pivotal role in the performance of the transcription system.

Actually, the most time-consuming task of the proposed system is the grid-search for optimal SVM parameters (C,γ) which are data dependent. On the other hand this cost is unavoidable to obtain satisfactory recognition accuracy. Further refinements one could envision may concern some guidelines for SVM parameter tuning, restricting the search to a limited interval around 'good' initial points. A further improvement could be obtained by modifying the training of SVMs to take into account the imbalanced number of positive and negative examples [19,20].

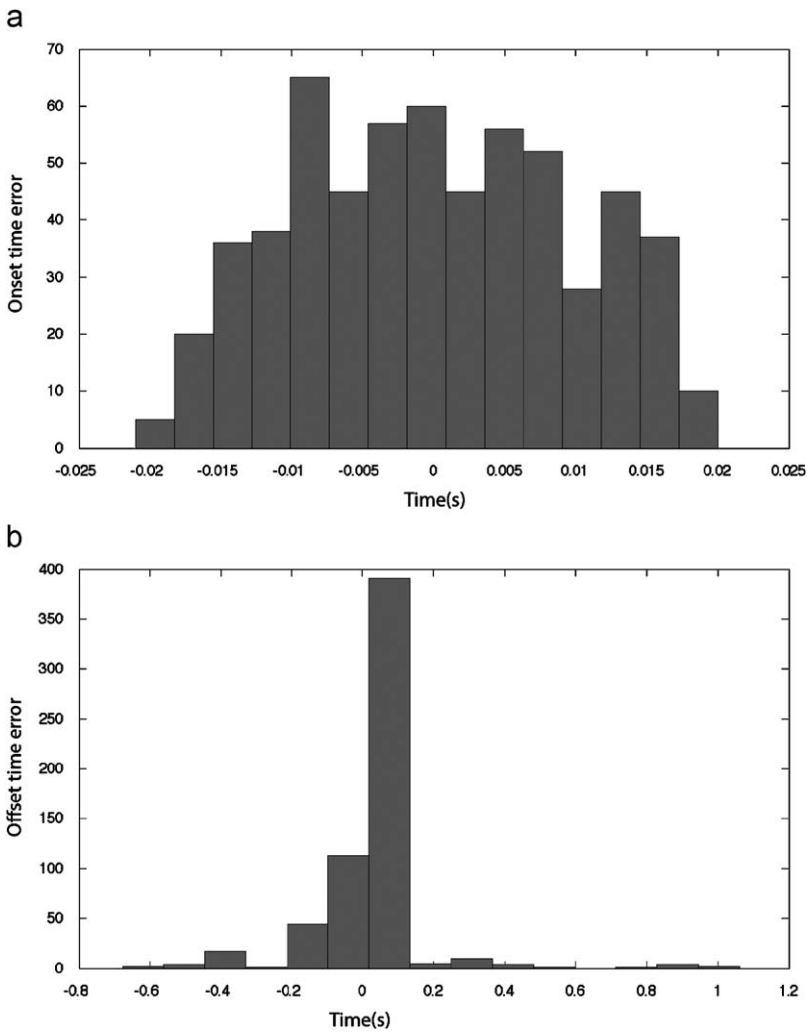

Fig. 17. Onset time (a) and offset time (b) error distribution for the piano transcription example.

References

[1] J.C. Brown, B. Zhang, Musical frequency tracking using the methods of conventional and "narrowed" autocorrelation, J. Acoust. Soc. Am. 89 (5) (1991) 2346–2354.

[2] J.C. Brown, Musical fundamental frequency tracking using a pattern recognition method, J. Acoust. Soc. Am. 92 (3) (1992) 1394–1402.

[3] J.A. Moorer, On the segmentation and analysis of continuous musical sound by digital computer, Ph.D. Thesis, Department of Music, Stanford University, Stanford, CA, May 1975.

[4] J.A. Moorer, On the transcription of musical sound by computer, Comput. Music J. 1 (4) (1977) 32–38.

[5] M. Piszczalski, B. Galler, Automatic music transcription, Comput. Music J. 1 (4) (1977) 24–31.

[6] M. Ryynanen, A. Klapuri, Polyphonic music transcription using note event modeling, in: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '05), New Paltz, NY, USA, October 2005.

[7] M. Marolt, A connectionist approach to automatic transcription of polyphonic piano music, IEEE Trans. Multimedia 6 (3) (2004) 439–449.

[8] G. Poliner, D. Ellis, A discriminative model for polyphonic piano transcription, EURASIP J. Adv. Signal Process (2007) 1–9 Article ID 48317.

[9] M. Marolt, A. Kavcic, M. Privosnik, S. Divjak, On detecting note onsets in piano music, IEEE Melecon 2002, Cairo, Egypt, 2002, pp. 385–389.

[10] G.P. Nava, H. Tanaka, I. Ide, A convolutional-kernel based approach for note onset detection in piano-solo audio signals, in: International Symposium Musical Acoustical ISMA 2004, Nara, Japan, 2004, pp. 289–292.

[11] W.C. Lee, C.C.J. Kuo, Musical onset detection based on adaptive linear prediction, in: IEEE International Conference on Multimedia and Expo, ICME 2006, Toronto, Canada, 2006, pp. 957–960.

[12] C.-H. Chuan, E. Chew, Audio onset detection using machine learning techniques: the effect and applicability of key and tempo information, Computer Science Department Technical Report No. 08-895, University of Southern California, 2008. Available: 〈http://www-rcf.usc.edu/~echew/bibliography/index3.html〉.

[13] J.C. Brown, Calculation of a constant Q spectral transform, J. Acoust. Soc. Am. 89 (1) (1991) 425–434.

[14] J.C. Brown, M.S. Puckette, An efficient algorithm for the calculation of a constant Q transform, J. Acoust. Soc. Am. 92 (5) (1992) 2698–2701.

[15] J. Shawe-Taylor, N. Cristianini, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, 2000.

[16] T. Joachims, Making large-scale SVM learning practical, in: B. Schölkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods—Support Vector Learning, MIT Press, Cambridge, MA, 1999.

[17] S. Dixon, On the computer recognition of solo piano music, in: Proceedings of Australasian Computer Music Conference, Brisbane, Australia, July 2000, pp. 31–37.

[18] National Institute of Standards Technology, Spring 2004 (RT-04S) rich transcription meeting recognition evaluation plan, 2004. 〈http://nist.gov/speech/tests/rt/rt2004/spring/〉.
[19] N. Japkowicz, Learning from imbalanced data sets: a comparison of various strategies, in: N. Japkowicz (Ed.), Learning from Imbalanced Data Sets: Papers from the AAAI Workshop (Austin, Texas, 2000), AAAI Press, Menlo Park, CA, 2000, pp. 10–15.
[20] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: Proceedings of the 14th ICML, Nashville, Tennessee, USA, 1997, pp. 179–186.