On the Use of Memory for Detecting Musical Notes in Polyphonic Piano Music

Giovanni Costantini[1,2], Massimiliano Todisco[1]
[1]Department of Electronic Engineering
University of Rome "Tor Vergata"
[2]Institute of Acoustics "O. M. Corbino"
Rome, Italy

Renzo Perfetti
Department of Electronic and Information Engineering
University of Perugia
Perugia, Italy

***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. The proposed method focuses on temporal musical structures, note events and their main characteristics: the attack instant and the pitch. Onset detection exploits a time-frequency representation of the audio signal. Note classification is based on constant Q transform (CQT) and support vector machines (SVMs). Finally, to validate our method, we present a collection of experiments using a wide number of musical pieces of heterogeneous style.**

I. INTRODUCTION

Music transcription can be considered as one of the most demanding activities 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.

In fact, be necessary discern two cases in what 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 time domain by means of zero-crossing or auto-correlation techniques and in 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] Piszczalski e Galler [4]. Moorer (1975) used comb filters and autocorrelation in order to perform transcription of very restricted duets.

The most important works in this research field is the Ryynanen and Klapuri transcription system [5] and the Sonic project [6] developed by Marolt, particularly this project makes use of classification-based approaches to transcription based on neural networks.

The target of our work dealt with the problem of extracting musical content or a symbolic representation of musical notes, commonly called musical score, from audio data of polyphonic piano music.

In this paper, an algorithm and model for automatic transcription of piano music are presented. The solution proposed is based on the onsets detection algorithm based on Short Time Fourier Transform (STFT) and a classification-based algorithm to identify the note pitch. In particular, we propose a supervised classification method that infers the correct note labels based only on training with tagged examples. This method performs polyphonic transcription via a system of Support Vector Machine (SVM) classifiers that have been trained starting from spectral features obtained by means of the well-known Constant-Q Transform (CQT).

The paper is organized as follows: in the following section our onsets detection algorithm will be described; in the third section, the spectral features will be formulated; the fourth section will be devoted to the description of the classification methods; in final section, we present and show the results of a series of experiments involving polyphonic piano music. Some comments conclude the paper.

II. ONSET DETECTION

The aim of note onset detection is to find the starting time of each musical note. Several different methods have been proposed for performing onset detection [7, 8].

Our method is based on STFT and, notwithstanding its simplicity, it gives better or equal performance compared to other methods [7, 8].

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

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

where N is the window size, h is the hop size, $m \in \{0, 1, 2, \cdots, M\}$ the hop number, $k = 0, 1, \cdots, N\text{-}1$ 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 by means of spectral frames represented by the magnitude spectrum .

The set of all the can be packed as columns into a non-negative $L \times M$ matrix, where M is the total number of spectra we computed and $L=N/2+1$ is the number of their frequencies.

Afterwards, the rows of S are summed, giving the following onset detection function based on the first-order difference

$$f_{onset}(m) = \frac{df(m)}{dm} \tag{2}$$

where

$$f(m) = \sum_{l=1}^{L} S(l,m) \tag{3}$$

Therefore, the peaks of the function fonset can be assumed to represent times of note onsets. After peak picking, a threshold T is used to suppress spurious peaks; its value is obtained through a validation process as explained in the next sections.

To demonstrate the performance of our onset detection method, let us show 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 bar at 120 metronome beat. It is shown in Figure 1.

We use a STFT with N=512, an N-point Hanning window and a hop size h=256 corresponding to a 32 *milliseconds* hop between successive frames.

Figure 2 shows the onset detection function.

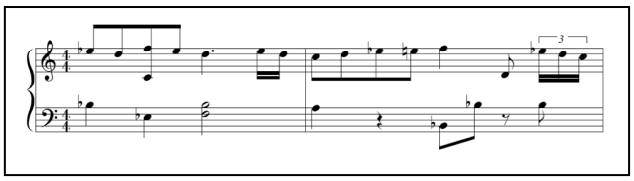

Figure 1. Musical score of Mozart's KV 333 Sonata in B-flat Major.

III. The Constant-Q Trasform ad the Spectral Features

The Constant-Q Transform (CQT) [9] is similar to the Discrete Fourier Transform (DFT) but with a main difference: it has a logarithmic frequency scale, since a variable width window is used. It suits better for musical notes that are based on a logarithmic scale.

The logarithmic frequency scale provides a constant frequency-to-resolution ratio for every bin

$$Q = \frac{f_k}{f_{k+1} - f_k} = \frac{1}{2^{1/b} - 1} \tag{4}$$

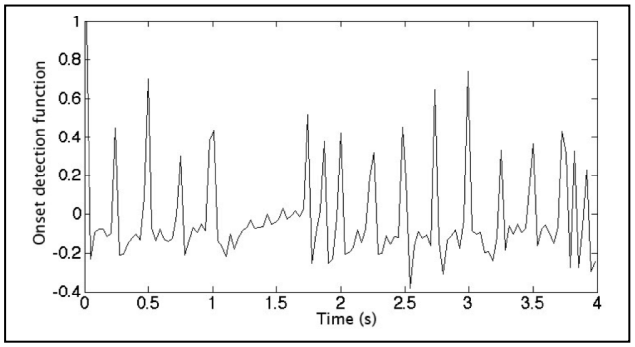

Figure 2. Onset detection function for the example in Figure 1.

where b is the number of bins per octave and k the frequency bin. If b=12, and by choosing a particular, then k is equal to the MIDI note number (as in the equal-tempered 12-tone-per-octave scale). There is an efficient version of the CQT that's based on the FFT and on some tricks, as shown in [10].

In our work, the processing phase starts in correspondence to a note onset. Notice that two or more notes belong to the same onset if these notes are played within 32*ms*. Firstly, 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 a Hanning windowing, as single CQT of the following 64*ms* of the audio note event is calculated. Figure 3 shows the complete process. All the audio files have a sampling rate of 8 *kHz*. We used b=48, that means 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 that's 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.

We note that melodic and harmonic structures depend on the composition method adopted by the compositor, this means that every musical note is highly correlated to the preceding and following notes in the composition. In our system, we take account of this assumption, consequently the feature vector is composed by 672 bins that coincide with the CQT-bins of the considered note and the CQT-bins of the previous note. Figure 4 shows the feature vector.

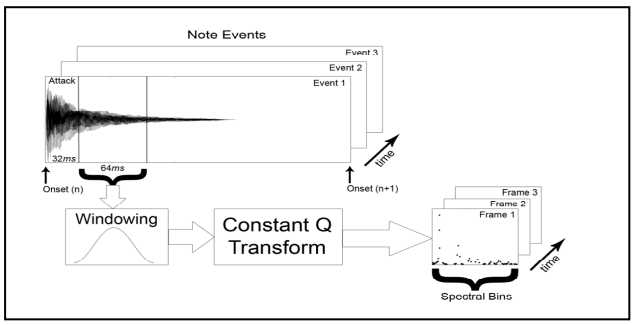

Figure 3. Spectral features extraction.

IV. 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 [11]. 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

$$f(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \qquad (5)$$

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 [11]. 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 [11].

The SVMs were implemented using the software SVMlight, developed by Joachims [12]. A radial basis function (RBF) kernel was used

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right), \quad \gamma > 0 \qquad (6)$$

where γ describes the width of the Gaussian function. For SVMs with RBF kernel two parameters, C and γ, need to be determined. 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 Section 5.

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 i^{th} SVM is trained using all the samples in the i^{th} 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 672-element feature vector, as described in Section 3. The presence of a note in a given audio event is detected when the discriminant function of the corresponding SVM classifier is positive. Figure 5 shows a schematic view of the complete automatic transcription process.

V. AUDIO DATA SET AND EXPERIMENTAL 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/ [13]. A list of pieces can be found in [13] (p. 8, Table 5).). The 124 pieces dataset was randomly split into 87 training, 24 testing, and 13 validation pieces.

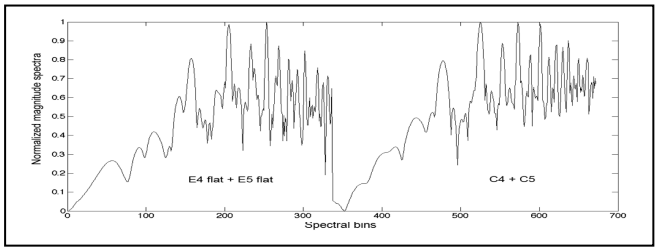

Figure 4. Normalized magnitude spectrum of the notes E4 flat + E5 flat and the previous notes C4 + C5.

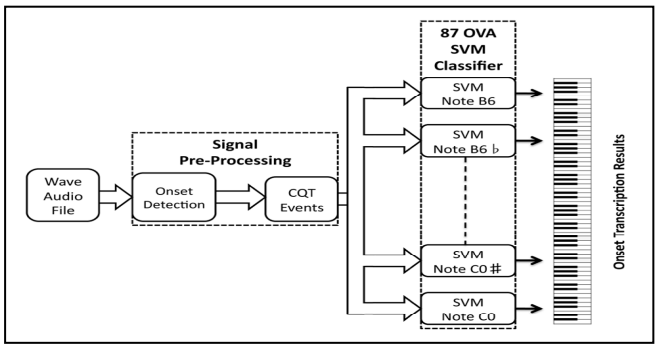

Figure 5. Schematic view of the complete automatic transcription process.

The first minute from each song in the dataset was selected for experiments, which provided us with a total of 87 minutes of training audio, 24 minutes of testing audio, and 13 minutes of audio for parameter tuning (validation set). This amounted to 22680, 6142, and 3406 note onsets in the training, testing, and validation sets, respectively.

First, we performed a statistical evaluation of the performance of the onset detection method. The results are summarized by three statistics: the *Precision*, the *Recall* and the *F-measure*.

When running the onset detection algorithm we experimented with the threshold value for peak picking. We consider as correct the onset detected within 32 *milliseconds* of the ground-truth offset. The results reported here were obtained using the threshold value 0.01; it was chosen to maximize the *F-measure* value regarding the 13 pieces of validation dataset. Table I quantifies the performance of the method on the test set (including 6142 onsets). After detection of the note onsets, we have trained the SVMs on the 87 pieces of the training set for each musical instrument and we have tested the system on the 24 pieces of the test set. Besides, to compare the accuracy of our system with a system with no memory, a second trial was performed on the same data set, using the 336 CQT-bins. The results are outlined in Table II.

TABLE I. ONSET DETECTION RESULTS

Precision	96.9%
Recall	95.7%
F-measure	**96.3%**

TABLE II. NOTE TRASCRIPTION RESULTS

	System With Memory 672 CQT-bins	System Without Memory 336 CQT-bins
F-measure	**85.5%**	**82.7%**
E_{tot}	**20.4%**	**22.0%**
E_{subs}	10.5%	11.5%
E_{miss}	9.9%	10.3%
E_{fa}	0.02%	0.02%

In addition of the *F-measure* metric, a different metric was used to evaluate the accuracy of our transcription system. In particular, the transcription error score defined by NIST for evaluations of "who spoke when" in recorded meetings, National Institute of Standards Technology [14]. Specifically, the total error score is given by

$$E_{tot} = \frac{\sum_{t=1}^{T} \max\left(N_{ref}(t), N_{sys}(t)\right) - N_{corr}(t)}{\sum_{t=1}^{T} N_{ref}(t)} \quad (7)$$

where T is the number of total time frames, N_{sys} is the number of notes detected by the system, N_{ref} is the number of notes in the musical pieces (ground-truth) and N_{corr} is the number of notes correctly detected by the system. Transcription error score is a sum of three components. The first is substitution error; the second and third components are miss error and false alarms error.

VI. CONCLUSIONS

In this study, we have discussed a polyphonic piano transcription system based on the characterization of note events. We focus our attention on temporal musical structures to detect notes.

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.

We compare two systems based on different featured vectors of 336 CQT-bins and 672 CQT-bins that represent the system without memory and the system with memory, respectively.

Musical note recognition system used 87 OVA binary classifiers based on SVM. A wide number of musical pieces of heterogeneous styles we used to validate and test our sensor interface.

REFERENCES

[1] J. C. Brown, "Musical fundamental frequency tracking using a pattern recognition method", Journal of the Acoustical Society of America, vol. 92, no. 3, 1992.

[2] J. C. Brown and B. Zhang, "Musical frequency tracking using the methods of conventional and narrowed autocorrelation", Journal of the Acoustical Society of America, vol. 89, no. 5, 1991.

[3] Moorer, "On the Transcription of Musical Sound by Computer". Computer Music Journal, Vol. 1, No. 4, Nov. 1977.

[4] M. Piszczalski and B. Galler, "Automatic Music Transcription", Computer Music Journal, Vol. 1, No. 4, Nov. 1977.

[5] M. Ryynanen and 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.

[6] M.Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," IEEE Transactions on Multimedia, vol. 6, no. 3, 2004.

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

[8] G.P. Nava, H. Tanaka, I. Ide, "A convolutional-kernel based approach for note onset detection in piano-solo audio signals", Int. Symp. Musical Acoust. ISMA 2004, Nara, Japan, pp. 289-292, 2004.

[9] J. C. Brown, "Calculation of a constant Q spectral transform", Journal of the Acoustical Society of America, vol. 89, no. 1, pp. 425–434, 1991.

[10] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," Journal of the Acoustical Society of America, vol. 92, no. 5, pp. 2698–2701, 1992.

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

[12] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[13] G. Poliner and D. Ellis "A Discriminative Model for Polyphonic Piano Transcription," EURASIP Journal of Advances in Signal Processing, vol. 2007, Article ID 48317, pp. 1-9, 2007.

[14] National Institute of Standards Technology, Spring 2004 (RT-04S) rich transcription meeting recognition evaluation plan, 2004. http://nist.gov/speech/tests/rt/rt2004/spring/.