NMF based Dictionary Learning for Automatic Transcription of Polyphonic Piano Music

GIOVANNI COSTANTINI[1,2], MASSIMILIANO TODISCO[1], RENZO PERFETTI[3]

[1]Department of Electronic Engineering
University of Rome "Tor Vergata"
Via del Politecnico, 1 - 00133 Rome
ITALY

[2]Institute of Acoustics and Sensors "Orso Mario Corbino"
Via del Fosso del Cavaliere, 100 - 00133 Rome
ITALY

[3]Department of Electronic and Information Engineering
University of Perugia
Via G. Duranti, 93 - 06125 Perugia
ITALY

massimiliano.todisco@uniroma2.it

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. Feature extraction is based on Sparse Nonnegative Matrix Factorization (SNMF) and Constant Q Transform (CQT), while note classification is based on Support Vector Machines (SVMs). Finally, to validate our method, we present a collection of experiments using a wide number of musical pieces of heterogeneous styles.

Key-Words: - Music transcription, classification, nonnegative matrix factorization, constant Q transform, support vector machines.

1 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 [1]. With these techniques, an excellent accuracy level has been achieved [2, 3].

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 [4] Piszczalski e Galler [5]. 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 [6] and the Sonic project [7] developed by Marolt, particularly this project makes use of classification-based approaches to transcription based on neural networks. Recent works can be found in [8, 9, 10, 11, 12].

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) and Sparse Nonnegative Matrix Factorization (SNMF).

The paper is organized as follows: in the following section the onset detection algorithm will be described; in the third section, the spectral features will be outlined; the fourth section will be devoted to the description of SNMF; in the fifth section, the classification method will be defined; in sixth section, we present the results of several experiments involving polyphonic piano music. Some comments conclude the paper.

2 Onset Detection Method

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 [13, 14, 15].

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

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

where N is the window size, h is the hop size, m∈ {0, 1, 2,…, M} is the hop number, k = 0, 1,…, 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 by means of spectral frames represented by the magnitude spectrum $|S_k(m)|$. The set of all the $|S_k(m)|$ can be packed as columns into a non-negative L×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

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

where

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

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.

The spectrogram is shown in Figure 2. Summing the elements of each column in Figure 2 we obtain the sum of rows in Figure 3 and, after a computation of the first-order difference, the onset detection function in Figure 4. The time onset resolution is 32ms. A statistical evaluation of the onset detection method will be presented in the next sections.

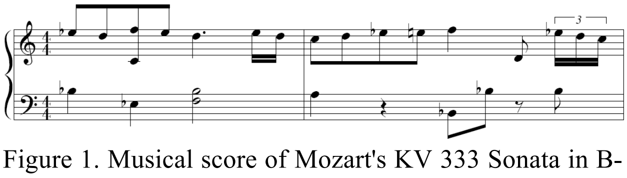

Figure 1. Musical score of Mozart's KV 333 Sonata in B-flat Major.

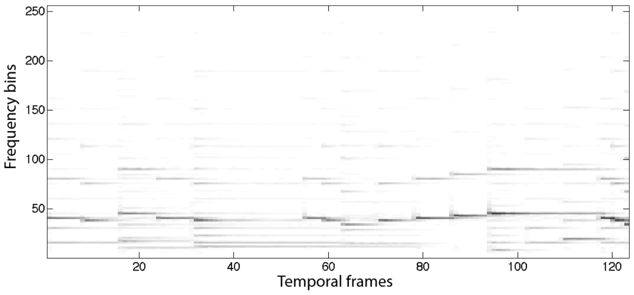

Figure 2. The spectrogram of Mozart's KV 333 Sonata in B-flat Major.

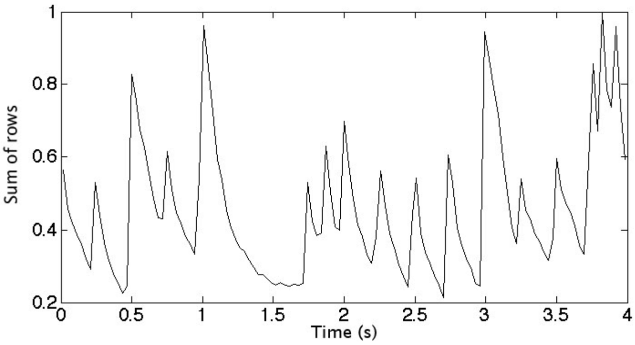

Figure 3. Normalized sum of the elements of each column of the spectrogram.

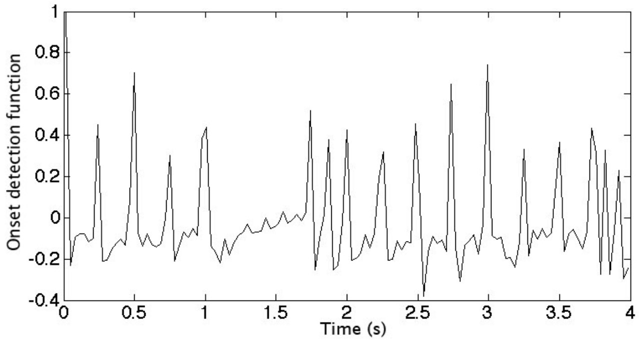

Figure 4. Onset detection function for the example in Figure 1.

3 Spectral Features based on Constant-Q Transform

A frequency analysis is performed on notes played by piano, in order to detect the signal harmonics. Using the Fast Fourier Transform (FFT) the frequency resolution could be not sufficient. In fact, a FFT with 512 temporal samples x[n] on a sound recorded with the usual sampling rate (SR) of 44100 Hz, carries out a resolution of about 86.1 Hz between two FFT 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 frequency resolution will get better if a higher number of temporal samples are used (with 8192 samples the resolution is of about 5.4 Hz), but that requires larger temporal windows for a fixed SR. In this case, the analysis of the instantaneous spectral information of the musical signal makes worse.

To solve this problem, a Constant-Q Transform (CQT) [16] is used to detect the fundamental frequency of the note. Then, the upper harmonics may be individuated easily, as they are located at frequencies nearly multiples of the fundamental frequency.

The Constant-Q Transform (CQT) 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

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

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 [17].

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 32ms. Firstly, the attack time of the note is discarded (in case of the piano, the longest attack time is equal to about 32ms). Then, after a Hanning windowing, as single CQT of the following 64ms of the audio note event is calculated. Figure 5 shows the complete process.

All the audio files that we used have a sampling rate of 8 kHz. The spectral resolution is b=372, that means 31 CQT-bins per note, starting from note C0 (~ 32 Hz) up to note B6 (~ 3951 Hz).

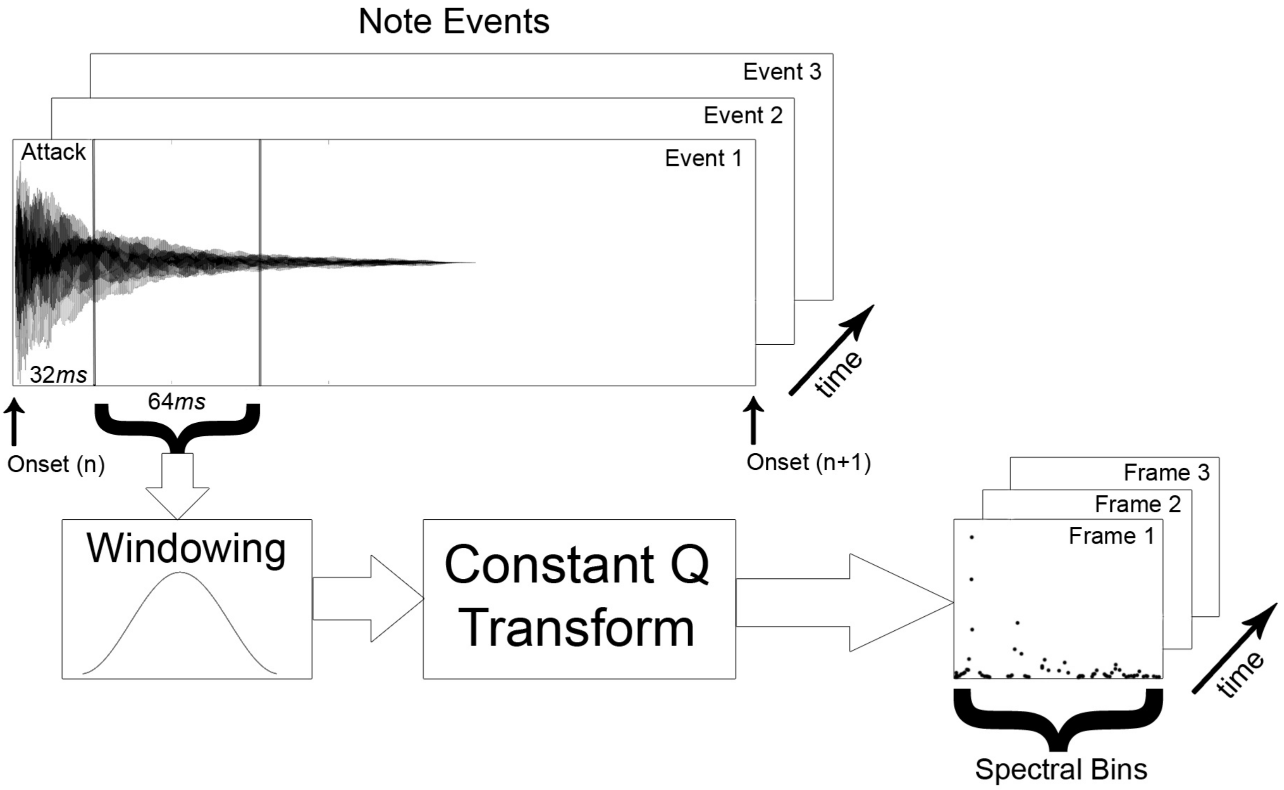

Figure 5. Spectral features extraction.

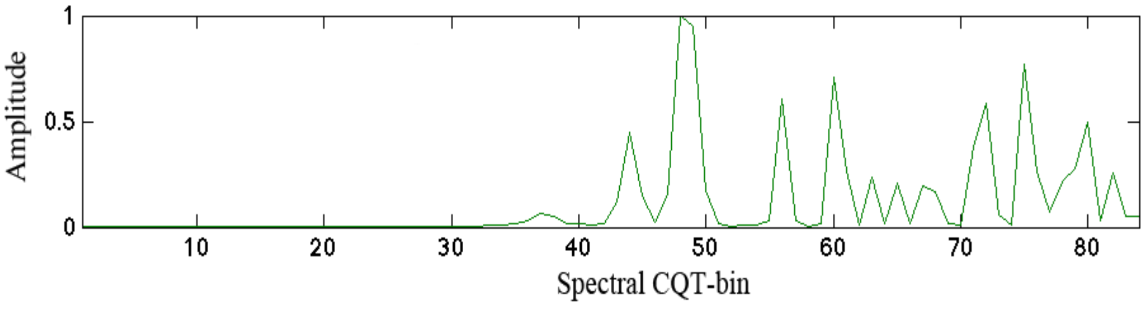

Figure 6. Feature spectral vector of MIDI note 38.

We obtain a spectral vector ***A*** composed by 2604 = 31 (CQT-bins) × 84 (musical notes). To reduce the size of the spectral vector, we operate a simple amplitude spectrum summation among the CQT-bin relative to the fundamental frequency of the considered musical note, the previous 15 CQT-bins and the subsequent 15 CQT-bins; then, we obtain a spectral vector ***B*** composed by 84 = 1 (CQT-bins) × 84 (musical notes). The scale of the values of the frequency bins is also logarithmic rescaled into a range from 0 to 1. Figure 6 shows a feature vector obtained with this method.

4 Sparse Nonnegative Matrix Factorization (SNMF)

The problem addressed by NMF [18] is as follows: given a nonnegative n×m data matrix D, find nonnegative matrices W and H in order to approximate the original matrix:

$$D \approx W \cdot H \tag{5}$$

The n×r matrix W contains the basis vectors and the r×m matrix H contains the coding vectors needed to properly approximate the columns of D as linear combinations of the columns of W. Usually, r is chosen so that (n+m)r<nm, thus resulting in a compressed version of the original data matrix.

The elements of W and H can be estimated by minimizing the Frobenius norm

$$D(V\|W,H) = \|V - WH\|_F^2 \tag{6}$$

subject to

$$W, H \geq 0$$

It may be advantageous to specify an additional constraint that will modify the representation in some way. One popular requirement is that the algorithm learns an over-complete basis by specifying additional constrain that the rows of H have a sparse activation for the basis contained in W [19].

This means that the probability of two or more activation patterns being active simultaneously is low. Thus, sparse representations lend themselves to good separability [20].

A simple way to introduce a sparseness constraint on H is to replace (6) with the following function [21]:

$$G(V\|W,H) = D(V\|W,H) + \lambda \sum_{ij} H_{ij} \tag{7}$$

where the second term enforces sparsity by minimizing the L1-norm of H. The parameter λ controls the tradeoff between sparseness and accurate reconstruction.

Function (7) is minimized by using the following update rules [15]:

$$\begin{aligned} W &= W - \mu\left(W^T \cdot H - V\right)H^T \\ H &= H.*\left(W^T \cdot V\right)./\left(W^T \cdot W \cdot H + \lambda\right) \end{aligned} \tag{8}$$

where .* and ./ are element-wise multiplication and division, respectively and μ > 0 is a small positive real number. W and H are initialized with random positive values, and alternatively updated by rules (8) until the cost function does not significantly change.

4.1 SNMF based feature extraction

The proposed method is shown schematically in Figure 7. The lower side of the figure (7b) represents the polyphonic piano signal for testing. Its decomposition onto representative templates is provided a priori to the system as basis. These bases are learned off-line, as shown on the upper side of the figure (7a), and constitute the dictionary used during decomposition.

The learning module aims at building a dictionary W of note templates onto which the polyphonic music signal is projected during the decomposition phase.

The whole polyphonic sample of notes {N0,N1,,Nn-1,Nn} that form the training dataset, from which the system learns characteristic basis template, is first processed in a short-time sound representation using the Constant Q Transform (CQT).

The representations are stacked in n matrices V{Ni}, for 0<i<n, where n is the note number and each column vj{Ni} is the sound representation of the jth time frame.

We then solve sparse NMF with V{Ni} and V{N0,…,Ni ,…,Nn}-{Ni}, this learning scheme gives two bases W{Ni} and W{N0,…,Ni ,…,Nn}-{Ni} that we stack in columns to form the dictionary W, and two activations H{Ni} and H{N0,…,Ni ,…,Nn}-{Ni} for each note sample.

The problem of the encoding phase is now to projecting the new music signal vjtest onto W.

The problem is thus equivalent to a nonnegative decomposition V ≈ WH where W is kept fixed and only H is computed.

The learned activation vectors hj provide a representation of the music signal useful for classification.

5 Multi-Class SVM Classifiers

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 [22].

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

$$f(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \qquad (9)$$

where xi is a support vector, K(xi, x) is the kernel function representing the inner product between xi and x in feature space, coefficients αi and b are obtained by solving a quadratic optimization problem in dual form [22].

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 [22, 23].

The SVMs were implemented using the software SVMlight developed by Joachims [24].

A radial basis function (RBF) kernel was used

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right), \quad \gamma > 0 \qquad (10)$$

where γ describes the width of the Gaussian function. The selection of model parameters, C and γ, was performed using a grid-search on a validation set.

For note classification, we used the one-versus-all (OVA) approach, based on 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 feature vector is built as described in Sections 3 and 4.

The presence of a note in a given audio event is detected when the discriminant function of the corresponding SVM classifier is positive. Figure 8 shows a schematic view of the complete automatic transcription process.

6 Audio Data 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/. A list of pieces can be found in [25] (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, 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.

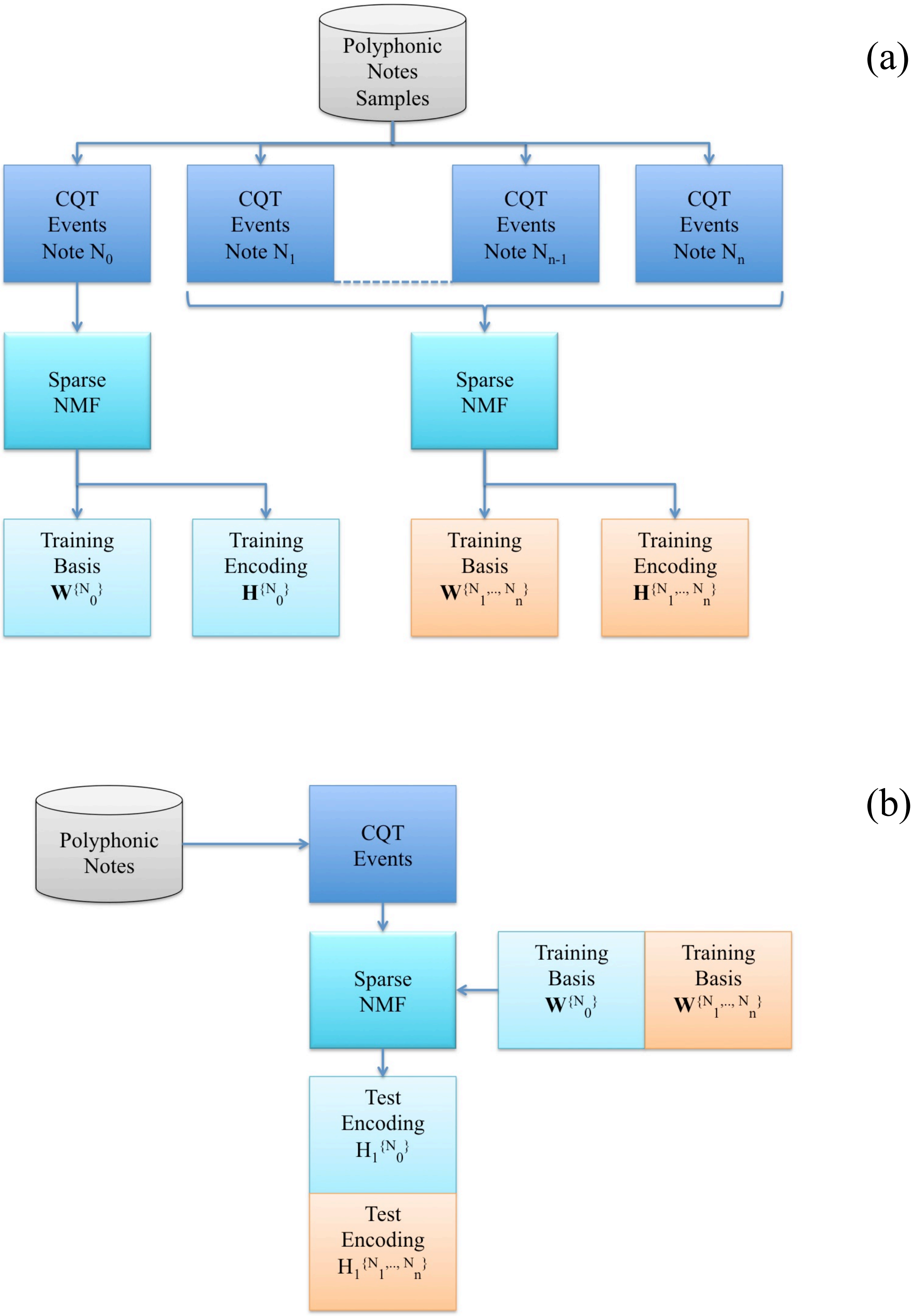
(a)
Polyphonic Notes Samples
CQT Events Note N_0
CQT Events Note N_1
CQT Events Note N_{n-1}
CQT Events Note N_n
Sparse NMF
Sparse NMF
Training Basis $\mathbf{W}^{\{N_0\}}$
Training Encoding $\mathbf{H}^{\{N_0\}}$
Training Basis $\mathbf{W}^{\{N_1,...,N_n\}}$
Training Encoding $\mathbf{H}^{\{N_1,...,N_n\}}$
(b)
Polyphonic Notes
CQT Events
Sparse NMF
Training Basis $\mathbf{W}^{\{N_0\}}$
Training Basis $\mathbf{W}^{\{N_1,...,N_n\}}$
Test Encoding $H_1^{\{N_0\}}$
Test Encoding $H_1^{\{N_1,...,N_n\}}$

Figure 7: Learning polyphonic note N0 (a). Encoding polyphonic note N0 (b).

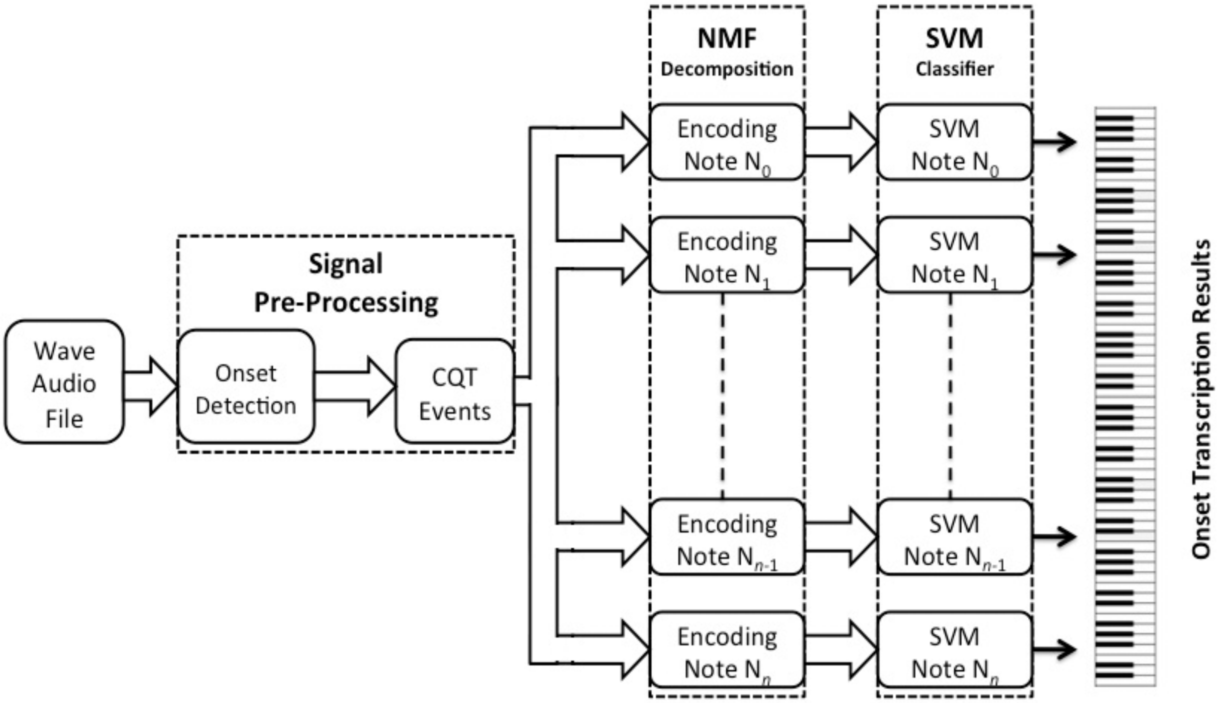

Figure 8: Schematic view of the transcription process.

Then we use a statistical evaluation of the performance of the musical notes classification by means of the metric proposed by Dixon [26], defined as Overall Accuracy

$$Acc = \frac{TP}{TP + FN + FP} \tag{11}$$

where TP ("true positives") is the correct detections, FP is the number of false positives and is the number of FN false negatives. This measure is bounded by 0 and 1, with 1 corresponding to perfect transcription.

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 onset.

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 and we have tested the system on the 24 pieces of the test set using SNMF with a factorization rank r = 50.

The accuracy results can be compared with the results in [27], where a system with a feature vector without sparse NMF was used. The accuracy in [27] was 72.3%, against an accuracy of 75.1% for the system with NMF based sparse signal representation. The results are outlined in Table II. Finally, we found that SNMF is not sufficient to be used as a clustering method for piano notes instead of an additional supervised SVM-based method.

Table I

ONSET DETECTION ACCURACY		
Precision	Recall	F-measure
96.9%	95.7%	96.3%

Table II

NOTE CLASSIFICATION ACCURACY	
System without SNMF based Feature Extraction	72.3%
System with SNMF based Feature Extraction	75.1%

7 Conclusion

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

The proposed onset detection algorithm 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.

A feature extraction based on Sparse NMF has been used for template learning method. That aims at building a dictionary of note templates onto which the polyphonic music signal is projected during the classification phase. These techniques already indicate that sparse coding is a powerful approach to automatic musical notes classification.

Finally, a wide number of musical pieces of heterogeneous styles were used to validate and test our transcription system.

The results, compared with the results obtained by a system that use a feature vector without sparse NMF, show an improvement of almost 3% in accuracy.

References:

[1] D.E. Ventzas, “Standard Signal Processing Software, Advances in Modelling & Analysis”, Series B, vol. 29, No. 4, 1994, pp. 1-10, Winter 1993-1994, AMSE

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

[3] 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.

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

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

[6] 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.

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

[8] G. Costantini, M. Todisco, R. Perfetti, R. Basili, “SVM Based Transcription System with Short-Term Memory Oriented to Polyphonic Piano Music”, 15th MELECON IEEE Mediterranean Electrotechnical Conference, Valletta, Malta, April 26-28, 2010, pp. 196-201.

[9] Reis, G., “Automatic Transcription of Polyphonic Piano Music Using Genetic Algorithms, Adaptive Spectral Envelope Modeling, and Dynamic Noise Level Estimation”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 8, pp. 2313 - 2328, Oct. 2012.

[10] G. Costantini, M. Todisco, G. Saggio, “Automatic Music Transcription Based on Non-Negative Matrix Factorization”, Proceedings of the 14th WSEAS International Conference on Systems, Corfu, Greece, July 22-24, 2010.

[11] G. Costantini, M. Todisco, G. Saggio, “A wireless glove to perform music in real time”, Proceedings of the 8th WSEAS International Conference on Applied Electromagnetics, Wireless and Optical Communications, ELECTRO '10, Penang, Malaysia, March 23-25, 2010 pp. 58-60.

[12] Hadar, O., Bykhovsky, D., Goldwasser, G., Fisher, E.A., “Musical source separation system with lyrics alignment”, WSEAS Trans. on Systems, Volume 5, Issue 10, October 2006, Pages 2464-2467.

[13] G. Costantini, M. Todisco, G. Saggio, “Musical Onset Detection by Means of Non-Negative Matrix Factorization”, Proceedings of the 14th WSEAS International Conference on Systems, Corfu, Greece, July 22-24, 2010.

[14] 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.

[15] G. Costantini, M. Todisco, R. Perfetti, G. Saggio, “On the Use of NMF for Onset Detection in Poliphonic Piano Music”, 7th WISP IEEE International Symposium On Intelligent Signal Processing, Floriana, Malta, September 19-21, 2011.

[16] 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.

[17] 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.

[18] Cichocki, A., Zdunek, R., Phan, A. H., Amari, S., Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation, Wiley, 2009.

[19] G. Costantini, M. Todisco, R. Perfetti, A. Paoloni, G. Saggio, “Single-Sided Objective Speech Intelligibility Assessment Based On Sparse Signal Representation”, 22nd IEEE International Workshop on Machine Learning for Signal Processing, Santander, Spain, September 23-26, 2012.

[20] Zhang, S., Zhao, X, Lei, B, Facial expression recognition using sparse representation”, WSEAS Trans. on Systems, Volume 11, Issue 8, August 2012, Pages 440-452.

[21] Hoyer, P. O., “Non-negative Matrix Factorization with Sparseness Constraints”, Journal of Machine Learning Research, no. 5, pp. 1457–1469, 2004.

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

[23] Trafalis, T.B., Park, J., “Uncertainty and sensitivity analysis issues in support vector machines”, WSEAS Trans. on Systems, Volume 5, Issue 9, September 2006, Pages 2086-2091.

[24] 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.

[25] 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.

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

[27] G. Costantini, R. Perfetti, M. Todisco, “Event Based Transcription System for Polyphonic Piano Music”, Signal Processing, Vol. 89, Issue 9, September 2009, pp. 1798-1811.