Static and Dynamic Classification Methods for Polyphonic Transcription of Piano Pieces in Different Musical Styles

GIOVANNI COSTANTINI$^{1,2}$, MASSIMILIANO TODISCO$^1$, MASSIMO CAROTA$^1$, DANIELE CASALI$^1$

$^1$Department of Electronic Engineering
University of Rome “Tor Vergata”
Via del Politecnico, 1 - 00133
ITALY

$^2$Institute of Acoustics “O. M. Corbino”
Via del Fosso del Cavaliere, 100 – 00133 ROMA
ITALY

Abstract: In this paper, we present two methods based on neural networks for the automatic transcription of polyphonic piano music. The input to these methods consists in piano music recordings stored in WAV files, while the pitch of all the notes in the corresponding score forms the output. The aim of this work is to compare the accuracy achieved using a feed-forward neural network, such as the MLP (MultiLayer Perceptron), with that supplied by a recurrent neural network, such as the ENN (Elman Neural Network). Signal processing techniques based on the CQT (Constant-Q Transform) are used in order to create a time-frequency representation of the input signals. Since large scale tests were required, the whole process (synthesis of audio data generated starting from MIDI files, comparison of the results with the original score) has been automated. Test, validation and training sets have been generated with reference to three different musical styles respectively represented by J.S Bach’s inventions, F. Chopin’s nocturnes and C. Debussy’s preludes.

Key-Words: Automatic piano music transcription, MultiLayer Perceptron, Elman Neural Network, Constant-Q Transform

1 Introduction

The target of this work dealt with the problem of extracting musical content (i.e., a symbolic representation of musical notes) from audio data, particularly with reference to polyphonic piano music. 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.

Currently, automatic transcription of monophonic music, where notes are played one-by-one, 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, where some notes are played simultaneously, 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] Piszczalski e Galler [5]. Moorer (1975) used comb filters and autocorrelation in order to perform transcription of very restricted duets.

The most important work in this research field is the Sonic project [6] developed by Marolt; this project makes use of classification-based approaches to transcription based on neural networks.

In this paper, we compare the results obtained with two different classification-based methods for automatic piano music transcription. In particular, we propose two supervised classification methods that infer the correct note labels based only on training with labeled examples. These methods performs polyphonic transcription via a system of MultiLayer Perceptron (MLP) classifiers and a system of Elman Neural Network (ENN) 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 the generation of data set and spectral features will be described; in the third section, the structure of a MultiLayer Perceptron and an Elman Neural Network will be introduced; the fourth section will be devoted to the description of the classification methods; finally, the results concerning the transcription of the notes in the score will be shown.

2 Audio Data and Spectral Features

The most outstanding problem in the classification-based approach concerns the collection of suitable training data. In this work, in order to collect the training, testing and validation data sets, we have used synthesized audio generated by supplying properly chose MIDI files to a widely used software synthesizer. The same MIDI data have
also been used as a reference for the estimation of the classification results.

2.1 Audio Data
In order to investigate the influence of musical styles on classification systems, we used three different MIDI files containing works by three composers lived in different ages: J.S Bach (three part inventions), F. Chopin (nocturnes) and C. Debussy (preludes).

<table>
<thead>
<tr>
<th>Bach’s &amp; Chopin’s Pieces</th>
<th>Bach’s &amp; Chopin’s Pieces</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major Tonality</td>
<td>Minor Tonality</td>
</tr>
<tr>
<td>Training</td>
<td>Training</td>
</tr>
<tr>
<td>Validation</td>
<td>Validation</td>
</tr>
<tr>
<td>Bach’s three part invention in Cmajor</td>
<td>Bath’s three part invention in Dmajor→Cmajor</td>
</tr>
<tr>
<td>Chopin’s Nocturn in Abmajor→Cmajor</td>
<td>Chopin’s Nocturn in Gmaj→Aminor</td>
</tr>
</tbody>
</table>

Table 1: Training-validation audio data set.

We have trained the neural networks with five different training-validation data sets: two major tonalities and two minor tonalities, taken from Bach and Chopin repertoires, and one taken from Debussy’s pieces (Table 1); besides, with the aim of removing the influence due to different song tonalities, we have translated, when needed, all the major pieces in Cmajor and all the minor pieces in Aminor.

The input to the classification system is formed by digital audio files generated by playing MIDI files with the Edirol Virtual Sound Canvas (VSC) [7], a widely used high quality software synthesizer. All the audio files have been converted to monaural audio files with a sampling rate of 44.1 kHz.

In this work, the classification system had to accomplish the task to extract from the audio file the so called note events, where every note event starts with a midi onset and ends with a midi offset.

2.2 Spectral Features
The Constant-Q Transform 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 [8].

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

\[ Q = \frac{f_k}{f_{k+1} - f_k} = \left(2^{\frac{1}{b}} - 1\right)^{-1} \]

where \( b \) is the number of bins per octave and \( k \) the frequency bin. If \( b = 12 \), and by choosing a particular \( f_0 \), 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 [9].

Provided that the MIDI file and the correspondent WAV file are correctly aligned (no latency, between a MIDI note event and the audio note event), the onset time of a note in the audio file is given by the time of the correspondent NOTE ON in the MIDI file.

The processing phase starts in correspondence to a note onset. 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. Fig. 1 shows the complete process.

In our work, we used \( b = 48 \), that means 4 CQT-bins per semitone, starting from note C0 (~ 32 Hz) up to note C7 (~ 4186 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.

3 Artificial Neural Network Structure
An artificial neural network consists of a pool of simple processing units which communicate by sending signals to each other over a large number of weighted connections. The weights on these connections encode the knowledge of a neural network [10].

In supervised learning, the training set, formed by input/desired-output couples, convey all the necessary information needed to train the network. The nonlinear nature of the neural network processing elements, called neurons, provides the system with lots of flexibility to achieve practically any desired input/output map.
In this context, we consider both feed-forward neural networks, where signals propagate from input to output layer with no loops back to previous neurons, as well as recurrent neural network architectures, that do contain feedback connections; contrary to feed-forward networks, the dynamical properties of recurrent networks are very important.

We use the most popular back-propagation learning algorithm for both feed-forward and recurrent neural networks.

3.1 MultiLayer Perceptron
The MultiLayer Perceptron (MLP) [10] consists of several layers of nodes, interconnected through weighted arcs that start from a layer and end to the layer that follows, without lateral or feedback connections.

The MLP, however, has the major limitation that it can only learn an input – output mapping which is static [11]. Thus it can be used to perform a nonlinear prediction of a stationary only time series.

3.2 Elman Neural Network
Unlike feed-forward networks, the recurrent architecture of Elman Neural Networks allows to include context information about past frames in the form of hidden layer states in the recurrent connections.

Particularly, the network has an input, a hidden and an output layers. Special units, called context units, store previous output values of hidden layer neurons and thus they serve as additional inputs to the network. Moreover, context units keep exponentially decreasing trace of past hidden neuron output values [11].

4 Classification Methods
Separate one-versus-all (OVA) feed-forward Neural Network and Elman Neural Network binary classifiers were trained on the spectral features of the most important degrees in the major and minor scales.

In major and minor tonalities, we consider only the I, II, III, V, VI, VII degrees of a scale, while, in case of Debussy’s pieces, where tonality concept is less present, we consider the I, III, V, VI, VII, IX degrees of a scale.

Our classification method is based on 6 OVA binary note classifiers with 50 neurons in the hidden layer, for both MLP networks and Elman networks, and it can detect the presence of a note in a given audio event, where each event is represented by a 336-element feature vector, as described in Section 2. Fig. 2 shows the complete automatic transcription process.

5 Note Transcription Results
The results are evaluated in terms of false positives (FP = number of notes returned by the classification method that were not present in the MIDI file) and true positives (TP = number of correctly identified notes). We can now define two variables:

\[
Precision = \frac{TP}{TP + FP}
\]

\[
Recall = \frac{TP}{TP + FN}
\]

Fig. 2: The complete automatic transcription process
“Precision” represents the percentage of precision in the identification of an example as positive. “Recall” represents the capacity of the classifier for the identification of the greatest number of positive examples. We consider, as a measure of the accuracy of our method, the global variable $f_{\text{measure}}$ that is the harmonic mean of precision and recall:

$$f_{\text{measure}} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

In the following tables, $f_{\text{measure}}$ transcription results of MLP and ENN classifiers tested on all Bach’s Inventions in three parts, all Chopin’s Nocturnes and all Debussy’s Preludes are reported.

### Training on Bach’s Invention in Cmajor

<table>
<thead>
<tr>
<th></th>
<th>MLP $f_{\text{measure}}$</th>
<th>ENN $f_{\text{measure}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test on Bach’s Cmajor Inventions</td>
<td>75.02%</td>
<td>84.63%</td>
</tr>
<tr>
<td>Test on Chopin’s Cmajor Nocturnes</td>
<td>54.83%</td>
<td>60.74%</td>
</tr>
<tr>
<td>Test on Debussy’s Preludes</td>
<td>39.60%</td>
<td>42.91%</td>
</tr>
</tbody>
</table>

### Training on Bach’s Invention in Aminor

<table>
<thead>
<tr>
<th></th>
<th>MLP $f_{\text{measure}}$</th>
<th>ENN $f_{\text{measure}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test on Bach’s Aminor Inventions</td>
<td>79.08%</td>
<td>88.54%</td>
</tr>
<tr>
<td>Test on Chopin’s Aminor Nocturnes</td>
<td>65.95%</td>
<td>69.45%</td>
</tr>
<tr>
<td>Test on Debussy’s Preludes</td>
<td>54.83%</td>
<td>57.08%</td>
</tr>
</tbody>
</table>

### Training on Chopin’s Nocturne in Cmajor

<table>
<thead>
<tr>
<th></th>
<th>MLP $f_{\text{measure}}$</th>
<th>ENN $f_{\text{measure}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test on Chopin’s Cmajor Nocturnes</td>
<td>73.87%</td>
<td>80.22%</td>
</tr>
<tr>
<td>Test on Bach’s Cmajor Inventions</td>
<td>70.04%</td>
<td>74.85%</td>
</tr>
<tr>
<td>Test on Debussy’s Preludes</td>
<td>52.95%</td>
<td>54.48%</td>
</tr>
</tbody>
</table>

### Training on Chopin’s Nocturne in Aminor

<table>
<thead>
<tr>
<th></th>
<th>MLP $f_{\text{measure}}$</th>
<th>ENN $f_{\text{measure}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test on Chopin’s Aminor Nocturnes</td>
<td>74.00%</td>
<td>79.34%</td>
</tr>
<tr>
<td>Test on Bach’s Aminor Inventions</td>
<td>73.05%</td>
<td>76.03%</td>
</tr>
<tr>
<td>Test on Debussy’s Preludes</td>
<td>54.21%</td>
<td>56.24%</td>
</tr>
</tbody>
</table>

### Training on Debussy’s Prelude n.1

<table>
<thead>
<tr>
<th></th>
<th>MLP $f_{\text{measure}}$</th>
<th>ENN $f_{\text{measure}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test on Debussy’s Preludes</td>
<td>70.63%</td>
<td>68.93%</td>
</tr>
<tr>
<td>Test on Chopin’s Cmajor Nocturnes</td>
<td>56.23%</td>
<td>48.17%</td>
</tr>
<tr>
<td>Test on Bach’s Cmajor Inventions</td>
<td>52.50%</td>
<td>48.15%</td>
</tr>
<tr>
<td>Test on Chopin’s Aminor Nocturnes</td>
<td>59.43%</td>
<td>55.80%</td>
</tr>
<tr>
<td>Test on Bach’s Aminor Inventions</td>
<td>41.38%</td>
<td>46.52%</td>
</tr>
</tbody>
</table>

### 6 Conclusion and Discussion

We have presented two binary classifier methods, based on static and dynamic neural networks, for polyphonic piano transcription in various musical styles. Audio events, identified by MIDI onsets, are processed with Constant-Q Transform (CQT). The input to the proposed transcription systems is a 336 CQT-bin feature vector, while the detected notes range from note C0 to note C7.

We have compared the $f_{\text{measure}}$ rate results returned by a feed-forward neural network, such as a MLP (MultiLayer Perceptron), and by a recurrent neural network, such as an ENN (Elman Neural Network).

We have considered musical pieces composed in three different historical periods (Bach’s Three Part Inventions, Chopin’s Nocturnes and Debussy’s Preludes) and taken into account the most important degrees of the major and minor scales in order to generate training, validating and testing data sets to be used for the training of the transcription systems.

Fig. 3 shows transcription accuracy difference between ENN $f_{\text{measure}}$ and MLP $f_{\text{measure}}$ decreasing with respect to musical styles from Bach to Chopin to Debussy.
The success of dynamic neural network on Bach’s Inventions and on Chopin’s Nocturnes is due to the composition method, based on classical harmony, adopted by these two authors. This happens because every musical note is highly correlated to the notes that precede and that follow in the composition. Therefore, automatic transcription results of a dynamic classifier, such as an ENN (Elman Neural Network), have a significant improvement. On the contrary, in Debussy’s compositions, all the passage from a musical note to another is equally likely. In this case, the performances of the dynamic classifier and the static classifier are comparable.

References:


