Techniques for Automatic Music Transcription

Juan P. Bello, Giuliano Monti and Mark Sandler

Department of Electronic Engineering, King's College London,
Strand, London WC2R 2LS, UK
juan.bello_correa@kcl.ac.uk, giuliano.monti@kcl.ac.uk

Introduction

Musical transcription of audio data is the process of taking a sequence of digital data corresponding to the sound waveform and extracting from it the symbolic information related to the high-level musical structures that might be seen on a score [1]. This paper gives an overview of systems that are being developed at King's College London.

Monophonic transcription with autocorrelation

The scheme of the monophonic transcription system implemented here, is illustrated in figure 1. If the fundamental frequency of a harmonic signal is calculated, and the resulting track is visualised, it can be noticed that, for most of the duration of the notes, the pitch remains approximately constant. This relation, so clear to the eye, requires some comment. In order to implement some grouping criteria and rules for sounds, emphasis should be given to the similarity in human perception between image and sound [2]. Important clues can be obtained by observing carefully the plot of the pitch track.

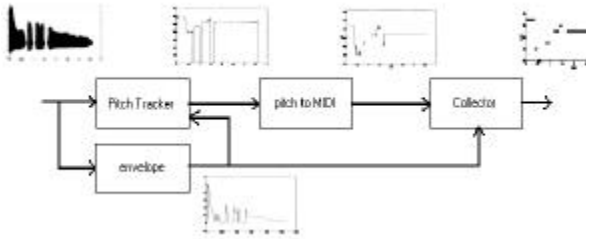

Figure 1: Scheme of the transcription system.

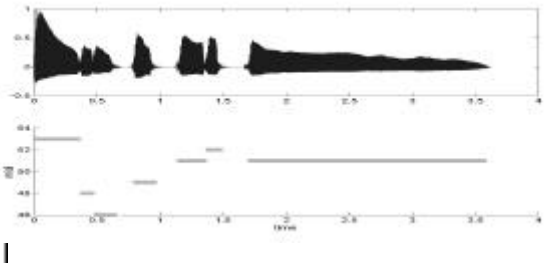

Figure 2: *Original music (top) and score file (below)*

Autocorrelation Tracking: In order to estimate the pitch in the musical signal, autocorrelation pitch tracking has been chosen [3], showing good detection and smooth values during the steady part of a note. The steady part of a note is just after the attack, where all the harmonics become stable and clearly marked in the spectrum. Peaks in the autocorrelation function correspond to the lags where periodicity is stronger. The first peak in the autocorrelation function, after the zero lag value, is a measure of the inverse of the fundamental frequency. The implementation takes advantage of some algorithms from the Auditory toolbox for MATLAB [4].The envelope block calculates the envelope of the signal. This goes to the pitch tracker, in order to skip the pitch calculation when the signal energy falls below the audibility threshold.

The Collector: The collector extracts the score, considering the pre-elaborated track and the signal amplitude, sorting: onsets, pitch and offsets. It recognises when the pitch maintains the same value, and proposes a note onset at the beginning of the first value of the sequence. The onset is confirmed when the pitch lasts for the minimum note duration accepted. The termination of a note is determined by the start of a new note or by the recognition of silence. The *minimum note duration* is the main parameter in the collector. By modifying its value, the system adapts to the speed of the music, improving the performance of the transcription. The minimum duration parameter controls also the memory of the system: when a note is detected, the pitch can vary inside the memory window before having again the same value, to be considered part of the same note. This is very similar to the consideration taken in sound restoration [5]: the human brain takes information from the cochlea, and interprets it with the knowledge of the previous samples; this behaviour is called streaming or integration process in psychoacoustics [2].

Figure 2 shows the plot of the original waveform and the final score. The onsets and the pitches are well recognised. The instruments under test were part of the brass family (saxophone, flugel horn): many complicated riffs (smooth or fast note changing) gave correct results. Although in order to synthesize the original music, envelope and timbre are determinant parameters to extract. To verify that the pitch has been correctly tracked and the melody of the original file

has not been modified, the system writes a CSOUND [6] score file. Transcription reveals correct recognition for pitches covering the B1-E6 note range.

Simple polyphonic transcription: Blackboard system and neural net

The blackboard system is a relatively complex problem-solving model prescribing the organisation of knowledge and data, and the problem-solving behaviour within the overall organisation[7]. The architecture of the blackboard system is shown in figure 3 [8]. Due to its open architecture, different knowledge can easily be integrated into the system, allowing the utilisation of various areas of expertise. The basic structure of a blackboard system consists of three fundamental parts: *the blackboard* or the global database where the hypotheses are proposed and developed; *the scheduler* or opportunistic control system, which determines how these hypotheses will be developed and by who; and *the knowledge sources*, which are the modules that execute the developing actions.

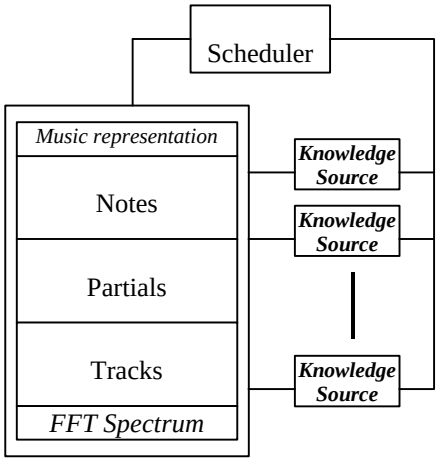

Figure 3: The control structure and the data hierarchy of the blackboard system

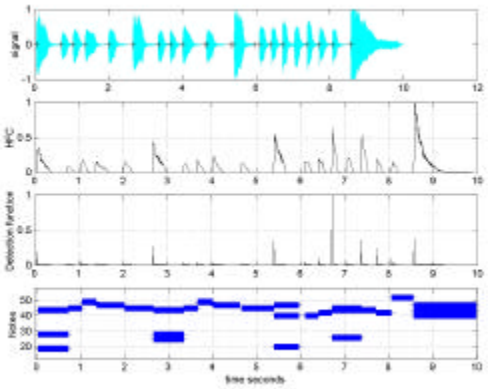

Figure 4: Example of automatic transcription of a simple polyphonic piano song of four bars.

Previous implementations [8][9] ignored the experience-based knowledge used by humans for the music transcription task. Those works were based on bottom-up processing. In contrast, the approach when the different levels of the system are determined by predictive models of the analysed object or by previous knowledge of the nature of the data is known as top-down processing [10]. In this paper, the top-down processing is achieved through the implementation of a neural network.

Implementation: Whilst analysing the running spectrum of the sound it is possible to notice that when an onset occurs, the high frequency content is relevant [11]. This property is exploited to determine the onset time. The segmentation is performed by averaging the signal's STFT between onsets. The blackboard is arranged in a hierarchy with three levels: tracks, partials and notes. *Tracks* are the input data (peaks of the averaged STFT in a given segment). The *partials* create a link between tracks and notes. *Notes* represent the high-level musical structures the system aims to extract. The type of the neural net implemented is known as feed-forward network. The input pattern consists of the spectrogram of a piano signal's segment (a note or a chord). The target output is represented by the absence "0" or presence "1" of a chord in the sample. Offline learning is used.

When the system is running, the network receives as input the STFT data the blackboard system analyses. The network's output changes the performance of the system allowing multiple note hypotheses to survive if necessary. The general output is given in the form of a piano roll and a CSOUND [6] score file.

Example: An example is shown here to illustrate how the system is working. In figure 4, a four bar section of a piano song including several chords, is represented. Twelve notes of the piece and the last chord were correctly identified by the system; however, mistakes are made in the transcription of the first three chords, where correct note hypotheses were discarded by the system in favour of their lower octave equivalents. The same error was detected in the note before the last chord. This octave error is recurrent in the test performed by the system.

Conclusions and next steps

On the first system, the good performance on monophonic input suggests the application of the heuristic criteria to a more complex representation of the signal, to deal with polyphonic music. Advanced signal analysis techniques (cochlear sub-band decomposition, multiresolution analysis) will be used to provide the necessary analysis resolution.

For the simple polyphonic system new knowledge sources should be added to the blackboard based on musical knowledge to overcome the octave detection problem. The architecture needs to be modified, incorporating dynamic structures to handle different sized hypotheses, e.g. chords of more than three notes. Also, the training space of the network has to be expanded to all the octaves of the piano. To define the next steps towards the handling of different instruments, more extensive testing has to be performed.

As a first approach, the results depicted here are very encouraging showing that further development of these ideas could be the way for more robust and general results.

Suggested Readings

[1] Eric Scheirer. "Extracting expressive performance information from recorded music". Master's thesis, MIT, 1995.

[2] Bregman A., "Auditory Scene Analysis", MIT Press, 1990.

[3] Brown , "Musical frequency tracking using the methods of Conventional and Narrowed Autocorrelation", JASA, 1991.

[4] Slaney M. "Auditory Toolbox for Matlab" available at URL http://www.interval.com/papers/1998-010/

[5] Ellis D, "*Hierarchic models of sound for separation and restoration*",Proc. IEEE Mohonk Workshop, 1993.

[6] CSOUND web page URL: http://music.dartmouth.edu/~dupras /wCsound/ csoundpage.html.

[7] R.S Engelmore and A.J. Morgan. "Blackboard Systems". Addison-Wesley publishing, 1988.

[8] Keith Martin. "A Blackboard system for Automatic Transcription of Simple polyphonic Music". MIT Media Lab, Technical Report # 385, 1995.

[9] Daniel Ellis. "Mid-level Representation for computational auditory scene analysis". In Proc. Of the Computational Auditory Scene Analysis Workshop; 1995 International Joint Conference on Artificial intelligence, Montreal, Canada, August 1995.

[10] Anssi Klapuri. "Automatic Transcription of Music". MSc Thesis, Tampere University of Technology, 1998.

[11] P. Masri and A. Bateman. "Improved Modelling of Attack Transient in Music Analysis-Resynthesis". University of Bristol. 1996.

[12] Stuttgart Neural Network Simulator. User Manual, version 4.1. University of Stuttgart, Institute for Parallel and Distributed High Performance Systems. Report No. 6/95.