Neural Networks for Note Onset Detection in Piano Music

Article · July 2003

Source: CiteSeer

CITATIONS
24

READS
926

3 authors, including:

Matija Marolt
University of Ljubljana
118 PUBLICATIONS **753** CITATIONS
SEE PROFILE

Alenka Kavčič
University of Ljubljana
23 PUBLICATIONS **317** CITATIONS
SEE PROFILE

Neural Networks for Note Onset Detection in Piano Music

Matija Marolt, Alenka Kavcic, Marko Privosnik

Faculty of Computer and Information Science, University of Ljubljana
email: matija.marolt@fri.uni-lj.si

Abstract

This paper presents a brief overview of our researches in the use of connectionist systems for transcription of polyphonic piano music and concentrates on the issue of onset detection in musical signals. We propose a technique for detecting onsets in a piano performance, based on a combination of a bank of auditory filters, a network of integrate-and-fire neurons and a multilayer perceptron. Such structure has certain advantages over the more commonly used peak-picking methods and we present its performance on several synthesized and real piano recordings. Results show that our approach represents a viable alternative to existing onset detection algorithms.

1 Introduction

Transcription of polyphonic music (polyphonic pitch recognition) is a process of converting an acoustical waveform into a parametric representation, where notes, their pitches, starting times and durations are extracted from the waveform. Transcription is a difficult cognitive task and is not inherent in human perception of music. It is also a very difficult problem for current computer systems. Separating notes from a mixture of other sounds, which may include notes played by the same or different instruments or simply background noise, requires robust algorithms with performance that should degrade gracefully when noise increases.

In recent years, several transcription systems have been developed. Some of them are targeted to transcription of music played on specific instruments (Rossi 1998; Sterian 1999; Dixon 2000), while others are more general transcription systems (Klapuri 1997). Onset detection is an integral part of these transcription systems, as it helps to determine exact onset times of notes in the transcribed piece. Some authors use implicit onset detection schemes (Rossi 1998; Sterian 1999), while others, including us, chose to implement a separate onset detection algorithm to improve the accuracy of onset times. We present our approach to onset detection in this paper.

2 Piano Music Transcription

Our transcription system, called SONIC, is a system for transcription of polyphonic piano music. It takes an acoustical waveform of a piano recording (44.1 kHz sampling rate, 16 bit resolution) as its input and produces a MIDI file containing the transcription as its result. Notes, their starting times, durations and loudness' are extracted from the signal in this process. Besides the piano being the only instrument in the transcribed signal, the system imposes no other limitations on the type of signal being transcribed, such as minimal note length, maximal polyphony, style of transcribed music, etc.

The structure of SONIC is similar to most other transcription systems. The main distinction to existing approaches is that we use neural networks to perform tasks such as partial tracking and note formation. These parts of the system have already been presented elsewhere (Marolt 2000; Marolt 2001) and will not be discussed in this paper. We dedicate the next two sections to the onset detection subsystem implemented within SONIC, and present its performance on several synthesized and real piano recordings.

3 Onset Detection

3.1 Overview

Note onsets play an important role in the perception of music. Studies showed that onsets play a pivotal role in the perception of timbre, as it is much more difficult to recognize the timbre of tones with removed onsets (Martin 1999). Onsets also make it easier to detect new information in music; we can detect tones with pronounced onsets well before we can determine their pitch.

In a music transcription system, an onset detection algorithm is needed to correctly determine the starting times of notes in the transcribed signal. Several authors use implicit onset detection schemes in their systems and make the onset time of a note equal to the time of its finding. At first, we used a similar solution, but later abandoned it as it did not produce accurate results, especially for notes in lower octaves, where delays of several 10ths of milliseconds were very common. Such timing deviations led to unpleasant effects when listening to resynthesized transcriptions, and also made performance evaluation of the entire system very difficult, as one had to take such deviations into

consideration. We have therefore decided to add a separate onset detection algorithm to our system.

Detection of onsets in a monophonic signal is not a difficult problem, especially if onsets are prominent, as is the case with piano tones. Onsets in a monophonic piano signal could be calculated with high accuracy by simply locating peaks in the amplitude envelope of the input signal. In polyphonic music, such an approach fails, because the amplitude envelope of an entire signal reveals little of what is going on in individual frequency regions of the signal, where note onsets and offsets may occur.

Many researches in onset detection have been made in the field of beat and rhythm tracking. Unfortunately, these algorithms are usually not accurate enough to be used in transcription systems, as they only discover very prominent onsets in a signal. Better approaches were used in some transcription systems (Klapuri 1999; Scheirer 1995). There, the signal is first split into several frequency bands. A relative difference function is then calculated on the amplitude envelope of each frequency band and peaks above a certain threshold are taken as onset candidates. Peaks across all bands are then merged together, their new amplitudes calculated and all new peaks that fall below a certain new threshold removed. The remaining peaks are considered to represent onsets in the signal. Such approaches tend to be very sensitive to the choice of threshold values; if they are set too low, many spurious onsets are detected and vice versa; high threshold values produce many missing onsets. We have therefore chosen to take a somewhat different approach to onset detection.

3.2 Onset Detection in SONIC

Our onset detector is based on a model for segmentation of speech signals, as proposed by Smith (Smith 1996). The model is founded on psychoacoustic findings and is based on a network of integrate-and-fire neurons that detects possible onsets in the input signal. We extended Smith's model with a multilayer perceptron neural network to improve the reliability of onset detection.

The first phase of the model splits the signal into several frequency bands with a bank of auditory filters. These are bandpass IIR filters, their parameters were calculated from psychoacoustic findings (Patterson and Holdsworth 1990). We use these filters to split the signal into 22 overlapping frequency bands, each covering half an octave.

The signal in each of the 22 resulting frequency bands is full-wave rectified and processed with a difference filter that calculates the difference between two amplitude envelopes; one calculated with a first order IIR smoothing filter with a time constant between 6 and 20 ms (depending on the center frequency of the band), and the other by smoothing the signal with a longer time constant (20-40 ms). Output of the difference filter has positive values when the signal rises and negative otherwise.

Figure 1 shows the output of the difference filter on an excerpt taken from Glenn Gould's interpretation of Bach's Two-part Invention No. 8 (Sony 6622). The upper left part of the figure shows the acoustical waveform of the entire signal, vertical lines show note onsets. The right part of the figure shows two amplitude envelopes, calculated in the frequency band that covers frequencies between pitches of notes Gb4 and B4. Envelopes were calculated with a different amount of smoothing (time constants were 6 ms and 20 ms) and the difference is clearly visible. Output of the difference filter is shown in the lower left part of the figure. One can see that peaks of the filter output correspond to onsets of notes that fall within the Gb4-B4 frequency band. The last note (D4) falls outside of this frequency range, so its peak is not very prominent.

The main task of the onset detector is to determine which peaks in outputs of difference filters correspond to note onsets and which are the result of various noises or beating in the signal. Our onset detector performs this task with a combination of a network of integrate-and-fire neurons and a multilayer perceptron neural network. Outputs of all 22 difference filters are first fed into a fully connected network of integrate-and-fire neurons. Each integrate-and-fire neuron i in the network

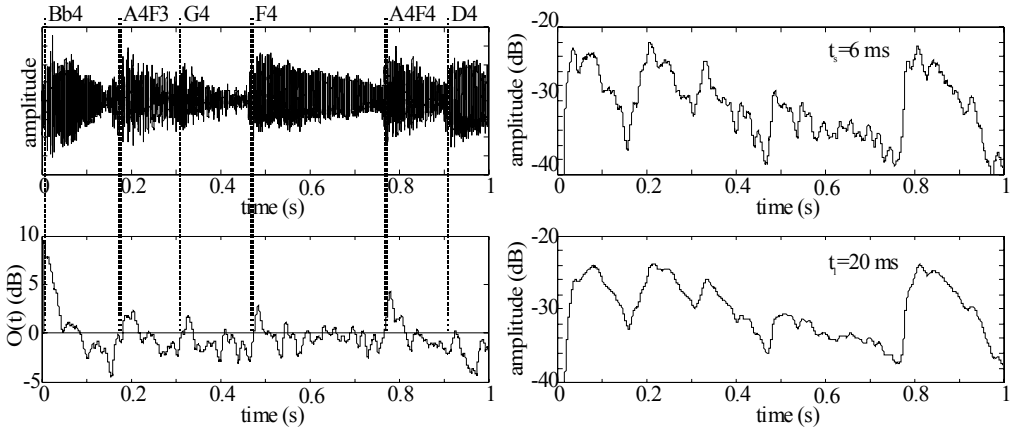

Figure 1: Acoustical waveform, two amplitude envelopes and the output of the difference filter $O(t)$

changes its activity A_i (initially set to 0) according to:

$$\frac{dA_i}{dt} = O_i(t) - \gamma A_i,$$

where $O_i(t)$ represents the output of the i-th difference filter, and γ the leakiness of integration. When A_i reaches a threshold, the neuron fires (emits an output pulse), and its activity A_i is reset to 0. After firing, there is a period of insensitivity to input, called the refractory period (50 ms in our model). Firings of neurons provide indications of amplitude growths in frequency channels. Neurons are connected to all other neurons in the network with excitatory connections and the firing of a neuron raises activities of all other neurons in the network and thereby accelerates their firing, if imminent.

Onset discovery with a network of integrate-and-fire neurons provides two main advantages over classical peak-picking algorithms. Network connections cluster neuron firings, which may otherwise be dispersed in time, while at the same time the refractory period prevents neurons from generating a series of impulses at each onset. Connections also improve the detection of weak onsets, as they encourage firings of neurons that are close to the firing threshold, but would not fire without additional help.

A network of integrate-and-fire neurons outputs a series of impulses indicating the presence of onsets in the signal. Not all impulses represent onsets, since various noises and beating can cause amplitude oscillations in the signal. We use a multilayer perceptron (MLP) neural network to decide which impulses represent onsets. Inputs of the MLP consist of activities A_i of integrate-and-fire neurons and several other parameters, such as amplitudes of individual frequency bands. The MLP only has one output, which indicates whether an onset has occurred in the input signal. The MLP was trained to recognize note onsets on a set of synthesized piano pieces and tested on a mixture of synthesized and real piano recordings. The performance of the entire onset detection system is presented in the next section.

4 Performance Evaluation

We tested our algorithm on a set of synthesized and real piano pieces. On average, the system correctly found around 98% of all onsets, together with 2% of spurious onsets (onsets not present in the input signal). We present results on three synthesized and three real piano recordings in table 1.

piano piece	no. of onsets	missed onsets	spurious onsets
1	4793	51 = 1.1%	3 = 0.1%
2	1305	37 = 2.8%	3 = 0.2%
3	963	10 = 1.0%	2 = 0.2%
4	786	25 = 3.1%	13 = 1.6%
5	206	13 = 6.3%	6 = 2.9%
6	556	0	8 = 1.4%

Table 1: Performance statistics on three synthesized and three real piano recordings

The synthesized pieces are: (1) J.S. Bach, Partita no. 4, BWV828 (Fazioli piano), (2) P.I. Tchaikovsky: Miniature Overture from The Nutcracker, (Bösendorfer piano), (3) S. Joplin in S. Hayden: Kismet Rag (Steinway D40 piano). Real recordings are: (4) J.S. Bach: English suite no. 5 (BWV810), 1. movement, performer Murray Perahia (Sony Classical SK 60277), (5) F. Chopin, Nocturne Op. 9/2, performer Artur Rubinstein (RCA: 60822), (6) S. Joplin, The Entertainer, performer unknown (MCA 11836).

Results on synthesized recordings are generally better than those on real recordings. A large number of missed notes are notes played in very fast passages or in ornamentations such as thrills and fast arpeggios (most missed notes in Bach's Partita (1)). The main cause of such misses is the refractory period of integrate-and-fire neurons, which prevents them from firing and thus detecting onsets in very fast pace. The system also often misses quietly played notes, which are masked by other louder notes or chords occurring shortly before or after the missed onset.

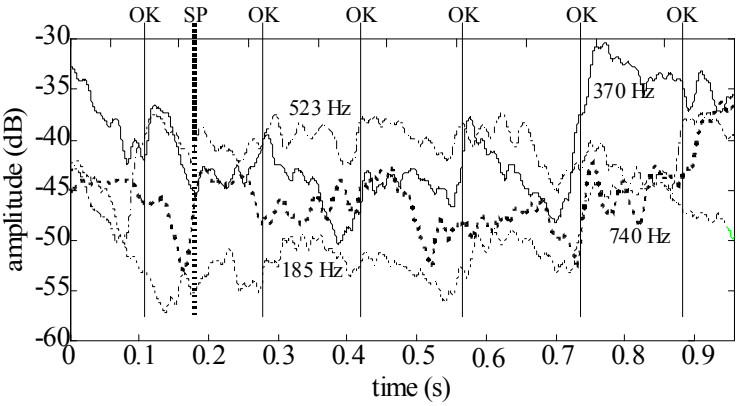

Figure 2: Detection of a spurious onset

Poorer onset detection accuracy on real recordings is a consequence of several factors. Recordings contain reverberation and more noise, while the sound of real pianos includes beating and sympathetic resonance. Furthermore, performances of piano pieces are much more expressive, they contain increased dynamics, more arpeggios and pedaling. All of these factors make onset detection more difficult. Still, the overall performance is satisfying. The causes of missed notes are similar to the ones we mentioned when looking at synthesized recordings; increased dynamics of performances is the main factor that contributes to a larger percentage of missed notes in real recordings. A good example of this is Chopin's Nocturne (5^{th} row in table 1), where a distinctive melody is played over very quiet, sometimes barely audible left hand chords, which are often missed.

The larger percentage of spurious notes in real recordings is a result of more noise and piano imperfections, such as beating and unpredictable partial behavior. An example of a spurious note detection is given in figure 2. The figure represents amplitude envelopes of four frequency bands calculated on a one second excerpt of Bach's English suite (4). Vertical lines represent onsets found by the system. Six onsets were correctly found (OK), together with one spurious onset (SP). The spurious onset occurred because of a large increase in the amplitude envelope of the 740 Hz frequency band (for which there is no obvious explanation).

5 Conclusion

In this paper, we presented our approach to detection of note onsets in a polyphonic piano performance. The approach is based on a connectionist paradigm and employs a bank of auditory filters and a network of integrate-and-fire neurons, coupled with a multilayer perceptron neural network. By using a connectionist approach to onset detection, we tried to avoid threshold problems that occur with standard “peak picking” algorithms. We presented performance statistics of our system on several synthesized and real piano recordings. Overall, we are satisfied with the results obtained; the presented onset detector brought a substantial improvement in the overall performance of our transcription system. The algorithm shows that connectionist approaches represent a good alternative in building onset detection systems and should be further studied.

References

Klapuri, A. 1997. *Automatic Transcription of Music.* M.Sc. Thesis, Tampere University of Technology, Finland.

Rossi, L. 1998. *Identification de Sons Polyphoniques de Piano*. Ph.D. Thesis, L'Universite de Corse, France.

Sterian, A.D. 1999. *Model-based Segmentation of Time-Frequency Images for Musical Transcription*. Ph.D. Thesis, Univesity of Michigan.

Dixon, S. 2000. "On the computer recognition of solo piano music." *Proceedings of Australasian Computer Music Conference*. Brisbane, Australia.

Marolt, M. 2000. "Adaptive oscillator networks for partial tracking and piano music transcription", *Proceedings of the 2000 International Computer Music Conference,* Berlin, Germany.

Marolt, M., 2001. "SONIC : transcription of polyphonic piano music with neural networks." *Workshop on Current Research Directions in Computer Music*, Barcelona, Spain.

Martin, K.D. 1999. *Sound-Source Recognition: A Theory and Computational Model*. Ph.D. Thesis, MIT, USA.

Klapuri, A. 1999. "Sound Onset Detection by Applying Psychoacoustic Knowledge." *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*.

Scheirer, E.D. 1995. *Extracting expressive performance information from recorded music.* M.Sc. Thesis, MIT, USA.

Smith, L.S. 1996. "Onset-based Sound Segmentation," *Advances in Neural Information Processing Systems 8.* Touretzky, Mozer and Haselmo (eds.). Cambridge, MA: MIT Press.

R. D. Patterson, J. Holdsworth. 1990. "A functional model of neural activity patterns and auditory images," *Advances in speech, hearing and auditory images.* W.A. Ainsworth (ed.). London: JAI Press.