Learning to Detect Onsets of Acoustic Piano Tones

Simon Dixon
Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Vienna, Austria
simon@oefai.at

Abstract

In the context of a project analysing expression in musical performance, we investigate techniques for estimating the onset times of piano tones from audio signals containing polyphonic music. We aim to show that automatic extraction of timing in audio signals is sufficiently accurate to be useful in musical expression research. A computer program is described which uses a combination of time domain and frequency domain features of the signal to estimate onset times. The relative weights of the features are learnt using a genetic algorithm trained on a large database of piano music performed on a Bösendorfer SE290 computer-monitored grand piano. The alignment of features in time is achieved by correlation with known hammer-string impact times from the same data set. The resulting algorithm detects approximately 90% of onsets with an average resolution around 10ms, a sufficiently good starting point for studies of tempo and timing, for example using an interactive beat tracking system. We discuss the limitations of the current approach, and propose a score-directed system as the next step towards a fully automatic expression extraction system.

1 Introduction

Automatic content analysis of musical signals has received considerable attention in recent times, particularly with reference to the MPEG-4 standard, which provides for very high compression of signals based on representation of content at a high-level of abstraction. This paper examines one facet of music signal content analysis, namely onset detection, and presents an onset detection algorithm that is trained and optimised on piano performance data.

The motivation of this work is to analyse musical interpretation, that is, the expressive nuances which differentiate an expert human musical performance from a mechanical rendition of the musical score and from other expert performances. To be successful, we require a very high time resolution, since the differences of interest are usually of the order of tens of milliseconds.

Since the focus of this research is performance rather than perception of music, the onset time is defined as the physical onset of the musical tone rather than the perceptual onset. For example, on the piano, the physical onset time corresponds to the time of the hammer impact on the string, which occurs shortly before the perceptual onset time of the tone [23].

Other applications of onset detection include audio beat tracking [19, 12, 8], automatic music transcription [14, 18, 15] and automatic accompaniment systems [5], none of which have the same demands of time resolution as expression analysis.

In the next section, we review literature relevant to onset detection, and then discuss the reasons for the difficulty of accurate onset detection, and why standard classification algorithms are not suitable for the task. In section 3, we present our onset detection algorithm, followed by the results of testing on a large corpus of professionally performed classical piano music. The final section contains a discussion of the results and directions of further research.

2 Background

Various approaches to onset detection have been proposed, using either time domain or frequency domain signal processing algorithms, or a combination of both, such as in auditory modelling.

Time domain algorithms use signal amplitude directly to derive parameters such as average absolute amplitude, RMS amplitude, peak amplitude, or the slope of the amplitude envelope, which are calculated for overlapping windows of the data and then examined for local and global peaks. These algorithms tend to be simple and fast.

For example, Dixon [8] uses a time-domain algorithm derived from Schloss [20], which filters and smooths the signal to produce an amplitude envelope calculated from the average absolute amplitude in a 20ms window of the signal. By overlapping the windows, amplitude values are calculated at a separation of 10ms. A 4-point linear regression is used to find the slope of the amplitude envelope, and a peak-picking algorithm finds local maxima in the slope, which are taken to be onsets. This method is quite successful for drum transcription (Schloss's application), because of the sharp attack and decay of drum sounds, and for beat tracking (Dixon's application), be-

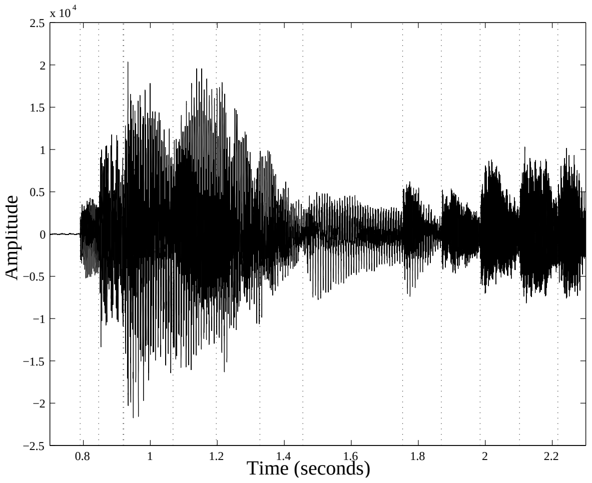

Figure 1: The beginning of Mozart's Piano Sonata in C major (K.279), with onsets marked by dotted lines. Not all onsets correspond with a sudden rise in amplitude in the time domain signal.

cause it is not necessary to detect all onsets in order to find the beat. However, in a polyphonic musical context, many onsets are masked in the time domain by sustained notes (see Figure 1), so more complex methods are required to detect all onsets.

Frequency domain algorithms first decompose the signal into distinct frequency bands and either perform time domain processing on each frequency band, or compute parameters representing spectral content. A significant time cost is incurred by the separate processing of each frequency band, as well as for the FFT, and the use of overlapping windows creates a highly redundant signal representation, which has much larger space requirements than time domain algorithms.

Klapuri [16] uses a 21-band decomposition of the audio signal, performing onset detection on each band, and then recombining the results using a psychoacoustic model of loudness to remove spurious onsets found in the separate frequency bands. Goto and Muraoka [10] use a frequency domain algorithm which focusses on the frequencies of the snare and bass drums in order to perform beat tracking on modern popular music. An extension of this work measures change in spectral content to predict higher level rhythmic boundaries [11].

Other work uses biologically-inspired models to process audio data [4]. These models are characterised by wider frequency bands but better time resolution than other frequency domain techniques. For example, Smith [21] uses a neurologically inspired model to detect onsets, by dividing the signal into third-octave frequency bands, performing full-wave rectification, summing the frequency bands and passing the resulting signal through an on-centre off-surround onset detection filter. Despite the complexity of this approach compared to time domain techniques, it reports success only with monophonic signals.

2.1 Feature Selection

In this work, we opted for a combined time and frequency domain approach, with the system being trained on part of the data for optimal detection. We view onset detection as a classification task, with the input data being processed into a series of feature vectors representing both time and frequency domain information. Knowledge representation is the key to problem solving. In the case of classification, whether based on manually or automatically generated (learnt) algorithms, success is largely dependent on the choice of feature vectors from which the classification is performed.

It is clearly beneficial to use knowledge of instrumental acoustics in selecting the features for the vectors. The acoustics of musical instruments are to a large extent well understood [3, 2, 22], and despite known variability, the basic characteristics such as amplitude envelope, spectrum and time-varying spectral characteristics have been catalogued [13, 1]. As a first approximation, acoustic piano tones have a sharp attack followed by an exponential decay, which in the frequency domain consists of a noisy transient at the beginning of the tone followed by a harmonic tone, where the relative amplitude of the harmonics is inversely proportional to the harmonic number. (More accurate models consider also the position of the hammer striking the string, the width and material of the hammer, the stiffness of the strings and the coupling between different strings and between the strings and the soundboard of the piano.)

In this work we assume a simple instrument model, and focus on detecting sudden increases of energy in the overall signal or in particular frequency bands. This is done by smoothing the signal to obtain an amplitude envelope and looking for peaks in the amplitude envelope and in the slope of the amplitude envelope, and then repeating this process for each frequency band, summing the energy across frequency bands which contain amplitude or slope peaks, and then looking for peaks in these sums.

2.2 Use of Genetic Algorithms

A number of machine learning algorithms were tested for developing a classifier for the feature vectors, but these proved to be unsuccessful. The reason that many machine learning algorithms are not suitable for classification of audio data is that they treat each vector as being independent, ignoring the context of the data. However, audio data is highly predictable in the sense that the measured features change relatively slowly between successive vectors, making context one of the most important factors in onset prediction. In fact, most features are only meaningful when viewed with respect to their local context. In more concrete terms, it is the local maxima, rather than specific values of features, which best correlate with onset times.

The other problem with classification algorithms is their means of assessment. For example, a decision tree learner presented with a set of feature vectors learns the trivial rule that no vector represents an onset (which is true about 90% of the time). A linear model using a least squares fit also fares badly, being able to classify only 65% of the vectors correctly.

The addition of context features to the vector can solve the first, but not the second problem. It is possible to weight the data set with multiple copies of positive instances, or to use a cost-based classifier with a higher penalty for false negatives, but this does not solve the basic problem that error assessment should also be context dependent. That is, the distance of the predicted onset from the actual onset time is significant. It is far better to predict an onset 10ms from the correct onset time rather than 100ms from the correct time; the learning algorithms do not represent this fact.

This problem was solved by developing a genetic algorithm to learn a linear function for classifying feature vectors. Genetic algorithms have the advantage of allowing the specification of an arbitrary evaluation function, which facilitates the solution of domain specific problems which do not fit well into other standard learning frameworks.

3 Onset Detection Algorithm

In this section we describe the specifics of the audio data, the processing of the data to extract relevant features, the learning algorithm which is used to parametrise the onset detection system, and finally the results of applying the onset detection algorithm to the remaining data.

3.1 Audio Data

The experimental data consists of 10 Piano Sonatas by W.A. Mozart (K.279–K.284 and K.330–K.333), played by a professional Viennese pianist on a Bösendorfer SE290 computer-monitored grand piano. The computer files of the performances were played back on the same piano and recorded digitally at 44100Hz sampling rate using a DAT recorder, and then transferred digitally to the computer file system. The performance files were also translated from Bösendorfer's proprietary format into MIDI format for use in the training and evaluation stages.

3.2 Feature Extraction

We use 8 features in the current system: 4 time domain features and 4 frequency domain features. To extract these features, the audio data is processed by averaging the two channels of the stereo recording, high-pass filtering, full-wave rectifying and averaging over a 20ms window. This process is repeated for overlapping windows spaced 10ms apart, so that each data point appears in exactly 2 windows. This defines the amplitude envelope, which is the first feature in the vector. The second feature is a binary feature, indicating whether the current point represents a local maximum in the amplitude envelope. A local maximum must exceed a threshold value and have no greater value within 50ms either side. The slope of the amplitude vector, calculated with a 4-point linear regression at each point, forms the third feature, and the fourth is a binary feature indicating a local maximum in the slope.

To perform frequency domain analysis, the audio signal is low-pass filtered and downsampled to 12kHz sampling rate, and a 512-point FFT is calculated at 10ms intervals. For each of the 256 frequency bands, the log power is computed and local maxima (for 50ms either side) are found and summed across frequency bands, to give the fifth feature. The sixth feature marks the peaks in the fifth. Finally, the seventh feature is found by calculating the amplitude slope in each frequency band and summing the local maxima, and the eighth feature represents peaks in the seventh.

Formulated precisely, let the audio signal be denoted $x(t)$, sampled at rate r so that $x[i] = x(rt)$ with $i = 0, 1, ..., n-1$. Let $x'(t)$ be the anti-alias filtered version of the signal x, down-sampled to rate r', so that $x'[i] = x'(r't)$ with $i = 0, 1, ..., n'-1$. Let $FFT(x, k, n)$ be the log power of the n-point discrete Fourier transform of the signal x at times $k, k+1, ..., k+n-1$. Let $w = 0.02$ be the window size and $h = 0.01$ the hop size (in seconds) used in calculating the feature vectors. Then the features $F_1, ..., F_8$ are calculated as follows for $t = hj$ with $j = 0, 1, ..., m-1$:

$$F_1[j] = \sum_{i=hjr}^{(hj+w)r} |x[i+1] - x[i]|$$

$$F_2[j] = \mathit{Peaks}(F_1, j, 5, \mathit{Threshold}_1)$$

$$F_3[j] = 0.2 \sum_{k=0}^{3} kF_1[j+k] - 0.3 \sum_{k=0}^{3} F_1[j+k]$$

$$F_4[j] = \mathit{Peaks}(F_3, j, 5, \mathit{Threshold}_2)$$

$$\begin{aligned} F_5[j] = \sum_{k=0}^{255} (&\mathit{Peaks}(FFT(x', hjr', 512)[k], j, 5, 0) * \\ &FFT(x', hjr', 512)[k]) \end{aligned}$$

$$F_6[j] = \mathit{Peaks}(F_5, j, 5, 0)$$

$$\begin{aligned} F_7[j] = \sum_{k=0}^{255} (&\mathit{Peaks}(\mathit{FreqSlope}[k], j, 5, FThreshold_k) * \\ &\mathit{FreqSlope}[k]) \end{aligned}$$

$$F_8[j] = \mathit{Peaks}(F_7, j, 5, 0)$$

where

$$\mathit{Peaks}(F_i, j, k, \mathit{Threshold}) = \begin{cases} 1, F_i[j] > \mathit{Threshold} \text{ and } F_i[j] = \max_{l=j-k}^{j+k} F_i[l] \\ 0, otherwise \end{cases}$$

$$\mathit{Threshold}_1 = \frac{0.1}{m} \sum_{l=0}^{m-1} F_1[l]$$

$$\mathit{Threshold}_2 = \frac{0.25}{m} \sum_{l=0}^{m-1} F_3[l]$$

$$FThreshold_k = \frac{0.1}{m}\sum_{l=0}^{m-1} FreqSlope(k)[l]$$

$$\begin{aligned} &FreqSlope(k)[j] = \\ &\qquad 0.2\sum_{l=0}^{3} l.FFT(x', (j+l)hr', 512)[k] - \\ &\qquad 0.3\sum_{l=0}^{3} FFT(x', (j+l)hr', 512)[k] \end{aligned}$$

3.3 Learning

Not all of the features that are used to predict onsets coincide precisely with the onsets themselves. For example, the peak amplitude occurs slightly after the physical onset of the note. In order to correct for these timing differences, correlation is performed between the features and the onset vector $Onset[j]$ derived from played note times P, where:

$$Onset[j] = \begin{cases} 1, \exists t \in P, (j-0.5)h \leq t < (j+0.5)h \\ 0, otherwise \end{cases}$$

For each feature F_i, the correlation $Corr_i(r)$ is calculated, to give a correction factor CF_i, which is used to produce a aligned set of vectors F_i'.

$$\begin{aligned} &Corr_i(r) = \sum_{j=0}^{m-1} F_i[j].Onset[j-r] \\ &CF_i = argmax_r(Corr_i(r)) \\ &F_i'[j] = F_i[j-r] \end{aligned}$$

Using the aligned feature vectors, a genetic algorithm is used to find a set of parameters $a_i, i = 1, ..., 9$ so that the prediction function $Predict[j]$ best approximates $Onset[j]$. A simplified version of the prediction function is given here.

$$Predict[j] = \begin{cases} 1, & \sum_{i=1}^{8} a_i.F_i'[j] \geq a_9 \\ 0, & otherwise \end{cases}$$

The onsets are matched and the error is calculated for each onset time t_i as follows:

$$Error[i] = \min_j |hj - t_i|, \text{ where } Predict[j] = 1$$

Various measures of error can be calculated. The average error rate, that is the average over all onsets t_i of $Error[i]$, is not an accurate measure of performance, because it treats all failures to detect an onset (false negatives) as timing errors. Although it is not possible in general to distinguish between timing errors and false negatives, we define any error greater than 70ms as being a failure to detect an onset. The remaining errors are averaged to give the average error in detection (AED). We define the predictive accuracy (PA) to be the number of matched onsets divided by the total number of onsets plus the number of false positives.

Training	Training Data		Test Data	
Set	PA	AED (s)	PA	AED (s)
K.279	90.1%	0.010	89.1%	0.011
K.280	88.8%	0.010	88.3%	0.011
K.281	86.8%	0.012	85.9%	0.011
K.282	88.5%	0.011	88.3%	0.011
K.283	90.6%	0.010	88.5%	0.011
K.284	92.2%	0.011	90.9%	0.010
K.330	91.9%	0.010	91.2%	0.011
K.331	89.3%	0.010	87.5%	0.010
K.332	87.6%	0.013	89.7%	0.011
K.333	89.4%	0.010	88.2%	0.011

Table 1: Predictive accuracy (PA) and average error in detection (AED) for the onset detection algorithm trained and tested on 10 Mozart piano sonatas.

3.4 Results

The algorithm was trained on each piano sonata separately, and then tested on the 10 complete sonatas. Predictive accuracy and average error in detection were calculated for the training data and the complete test set, and are shown in table 1. Since each sonata consists of several thousand notes, it is not surprising that the results are quite consistent across the different training sets. The best result was a 91.2% accuracy in predicting onsets across the 10 sonatas, with an average error in detection of 11ms. It is not yet known how well these parameters would apply to a different instrument, as we have no means of testing the algorithm with other data. It is expected that since the features are quite general, the approach should work well with other instruments, although training may be required to fine-tune the parameters to achieve equally good results.

4 Discussion and Further Work

The time resolution achieved in this work is sufficient for research in expressive performance, and probably more accurate than human judgement [9]. However, if every onset is required, the need to correct around 10% of onsets is still a formidable task, and limits the applicability of the system to short pieces. In further work, it is planned to use score information to guide the system to search for onsets specifically in the frequency bands of notated musical tones [17, 18]. Although this limits the system to works for which an on-line version of the score is available, it also opens the door to some optimisations which will improve the speed and the accuracy of the program. Further, it would provide precise information about which notes were delayed or anticipated with respect to the musical context, thus providing a richer source of data for use in learning experiments.

The other main application of this work is in the development of an automatic transcription system [6, 7], which extracts not just onset times from the audio data, but also pitch, amplitude and duration, which must all be interpreted in terms of musical constructs such as meter, rhythm and key. Although a complete transcription system is still a distant goal, useful interactive systems are not far off.

Acknowledgements

This research was support by the START programme of the Austrian Federal Ministry of Education, Science and Culture (grant number Y99-INF), and the EC project HPRN-CT-2000-00115 (MOSART). Thanks also to Roland Batik for permission to use his performances, and the L. Bösendorfer Company, Vienna, especially Fritz Lachnit, for providing the data.

References

[1] A. Askenfelt and E.V. Jansson. From touch to string vibrations. III: String motion and spectra. *Journal of the Acoustical Society of America*, Volume 93, Number 4, pages 2181–96, 1993.

[2] J. Backus. *The Acoustical Foundations of Music*. Norton, New York, 1977.

[3] A.H. Benade. *Fundamentals of Musical Acoustics*. Oxford University Press, New York, 1976.

[4] M.P. Cooke. *Modelling Auditory Processing and Organisation*. Cambridge University Press, 1993.

[5] R.B. Dannenberg. Recent work in real-time music understanding by computer. In J. Sundberg, L. Nord and R. Carlson (editors), *Proceedings of the International Symposium on Music, Language, Speech and Brain*, pages 194–202, Houndmills, 1991. Macmillan Press.

[6] S. Dixon. Extraction of musical performance parameters from audio data. In *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, pages 42–45, 2000.

[7] S. Dixon. On the computer recognition of solo piano music. *Mikropolyphonie*, Volume 6, 2000.

[8] S. Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, Volume 30, Number 1, 2001. To appear.

[9] S. Dixon, W. Goebl and E. Cambouropoulos. Beat extraction from expressive musical performances. In *2001 Meeting of the Society for Music Perception and Cognition (SMPC2001), Kingston, Ontario*, pages 62–63, 2001.

[10] M. Goto and Y. Muraoka. A real-time beat tracking system for audio signals. In *Proceedings of the International Computer Music Conference*, pages 171–174, San Francisco CA, 1995. International Computer Music Association.

[11] M. Goto and Y. Muraoka. Real-time rhythm tracking for drumless audio signals – chord change detection for musical decisions. In *Proceedings of the IJCAI'97 Workshop on Computational Auditory Scene Analysis*, pages 135–144. International Joint Conference on Artificial Intelligence, 1997.

[12] M. Goto and Y. Muraoka. Real-time beat tracking for drumless audio signals. *Speech Communication*, Volume 27, Number 3–4, pages 331–335, 1999.

[13] J.M. Grey. Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, Volume 61, pages 1270–1277, 1977.

[14] K. Kashino, K. Nakadai, T. Kinoshita and H. Tanaka. Organization of hierarchical perceptual sounds: Music scene analysis with autonomous processing modules and a quantitative information integration mechanism. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1995.

[15] A. Klapuri. Automatic transcription of music. Master's thesis, Tampere University of Technology, Department of Information Technology, 1998.

[16] A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Phoenix, Arizona, 1999.

[17] E.D. Scheirer. Extracting expressive performance information from recorded music. Master's thesis, Massachusetts Institute of Technology, Media Laboratory, 1995.

[18] E.D. Scheirer. Using musical knowledge to extract expressive performance information from audio recordings. In H. Okuno and D. Rosenthal (editors), *Readings in Computational Auditory Scene Analysis*. Lawrence Erlbaum, 1997.

[19] E.D. Scheirer. Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America*, Volume 103, Number 1, pages 588–601, 1998.

[20] W.A. Schloss. *On the Automatic Transcription of Percussive Music: From Acoustic Signal to High Level Analysis*. Ph.D. thesis, Stanford University, CCRMA, 1985.

[21] L.S. Smith. Sound segmentation using onsets and offsets. *Journal of New Music Research*, Volume 23, Number 1, 1994.

[22] J. Sundberg. *The Science of Musical Sounds*. Academic Press, San Diego CA, 1991.

[23] J. Vos and R. Rasch. The perceptual onset of musical tones. *Perception and Psychophysics*, Volume 29, Number 4, pages 323–335, 1981.