Automatic Bass Line Transcription from Polyphonic Music

Stephen W. Hainsworth, Malcolm D. Macleod
Signal Processing Laboratory, Department of Engineering
University of Cambridge.
{swh21, mdm}@eng.cam.ac.uk

Abstract

This paper examines the transcription of just a single instrument from a rich, polyphonic audio stream. The 'bass line' in jazz and rock recordings was chosen for this as a reasonably separable instrument, given several assumptions. An algorithm is presented which performs the following, sequential steps: note onset location; note hypothesis generation; a hypothesis tracker; and finally a hypothesis resolver. Results are presented for examples taken from commercially available CDs which achieve a note recognition rate of 78.7%, using the metric of Kashino and Murase (1998).

1 Introduction

Computer transcription of polyphonic music, the process of attaining a score or equivalent representation from an audio signal, has long been an active area of research but results have so far been limited. Monophonic transcription has been thoroughly investigated, mainly because of its application to speech coding, but analysis of polyphonic signals is much harder. This is especially the case in music where the multiple notes are usually consonant and hence share harmonics. This leads to ambiguity in assigning harmonics to notes and is a difficult problem to solve.

Traditionally, simple examples have been used as tests, such as piano music, eg Martin (1996), where all the notes are generated by a single instrument, or reduced polyphony using instruments with very different characteristics, eg Kashino and Murase (1998). Even with these limitations, the methods proposed still do not perform as well as a trained human, missing notes or making pitch errors. Sterian and Wakefield (2000) give a good review of research in the field.

However, to the best of the authors' knowledge, no-one has looked at the problem of transcribing just a single instrument from a rich, polyphonic sound scene. The nearest is Goto and Hayamizu (1999) who find the predominant fundamental frequency in the bass range and also in the melody. This, while producing a good indication of the notes in the bass and melody lines, does not actually produce a transcription as the identity of the instrument(s) playing the notes is not considered and neither is the problem of repeated notes or rests in the line.

1.1 Bass Line Transcription

The double bass or bass guitar (hereafter referred to as 'the bass line') was chosen for a first attempt at this single instrument transcription task for a number of reasons:

- it almost always only plays one note at a time
- it is usually the lowest pitched instrument in an ensemble
- the amplitude of bass notes is usually quite high.

These prior assumptions provide a number of powerful processing methods which allow data from other instruments to be ignored.

Jazz and rock music have been used to test the algorithm, because this is the field in which an automatic transcriber program would have most benefit. Classical music is performed from a pre-existing score, whereas jazz and rock are much more aural traditions, the performance either being spontaneous or communicated to the performer verbally.

A transcriber could help musicians learn how others played; it could aid musicologists in the study of these types of music and also has a commercial application in that it could be used to produce music for sale in shops automatically from the recording. This could even possibly include being used to transcribe individual jazz solos upon request, which would be an otherwise uneconomical task for a publishing company.

2 Algorithm Description

The program can be split into four independent stages:

1. Note onset locator
2. Frequency estimator and hypothesis generator
3. Hypothesis tracker
4. Hypothesis resolver

These process the digital audio signal independently and sequentially. Hypotheses and data found from each stage are then passed onto the subsequent one in a hierarchical fashion. Throughout, the algorithm uses a frame based Fourier Transform approach (often referred to as the STFT).

The main philosophy behind the program is that it is better to propose multiple hypotheses and then to select between competing ones, when more data is available. This hopefully then limits the chance of missing relevant data because one stage assigns it a low probability of being correct. The only exception to this philosophy was the first stage which was optimised because it was discovered that it was hard to reliably remove hypotheses generated by spurious onsets. Similarly, if an onset was missed, then the program structure meant that it was impossible to detect the note associated with that onset.

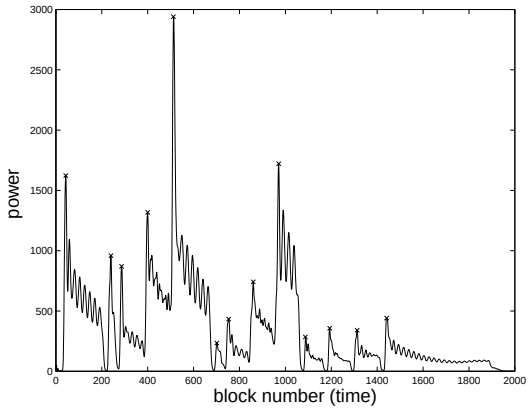

Figure 1: Output of stage 1; crosses indicate located onsets.

2.1 Note onset location

The first step is to extract the onset times of any bass notes (as opposed to notes played by other instruments). Time resolution is of the essence here so short, 6ms frames were used, with an overlap of half a frame. This gives a theoretical onset resolution of about 3ms, which is less than the temporal resolution usually attributed to the ear (Scheirer 1995).

The aim is to extract only onsets from bass notes which are generally the lowest notes played in a piece of music. Hence a low pass filtering method would seem sensible. Generally, bass note fundamental frequencies usually fall below 200Hz and was found that this frequency range also gave a good approximation of the temporal envelope of the bass notes. Therefore the data was low-pass filtered at 200 Hz to remove the power of higher pitched notes.

The total power in this range was found and time smoothed using a Gaussian kernel, to remove spurious peaks, and then a peak detection algorithm applied. A number of different schemes were tried for this task, with the simplest and generally most effective being a simple test such that if the ratio of a maxima was greater than three times the previous minima, an onset was placed at the maxima.

Next, a data interpolation scheme was used to re-estimate the time location of the peak before the smoothing was applied (as the smoothing process naturally shifts an asymmetric peak towards the higher weighted side) and finally any onsets with very low amplitude (less than 1/66th of the maximum onset amplitude) were ignored. Figure 1 shows the output of the note onset locator for an example which is a bass line played in isolation.

2.2 Frequency Estimation

The next stage is to propose hypotheses for the fundamental frequency (F0) of the notes determined by the onsets found in section 2.1. This task is a frequency critical one, so long frames were used. In fact, the frame was tailored to start just after a note onset so as to not include the transient and then to be as long as possible before the next onset time. A Hanning window was applied and then an algorithm for detecting sinusoids developed by Macleod (1998) was used to get a high accuracy frequency estimate of any tones found[1].

Multiple frequency hypotheses were then generated for each onset by considering any tone under 200Hz to be an F0 candidate and any tone under 400Hz to be a potential first harmonic candidate. The harmonic series associated with the hypotheses are found using a harmonic comb and a confidence measure S_j also assigned to each as follows:

$$S_j = \sum_i \lambda_i a_i \qquad (1)$$

where a_i is the amplitude of the i^{th} harmonic in the located series and λ_i is a weighting function: λ_1 = 20, λ_{odd} = 3 and λ_{even} = 2. Thus, S_j can also be thought of as a measure of power in the series.

2.3 Hypothesis tracking

The third stage tracks the frequency hypotheses proposed by the second stage over time, starting from the frame of onset. Hanning windowed frames of 2048 samples overlapped by 50% were used for this as a compromise between time and frequency resolution. A simplified version of the Macleod algorithm described in section 2.2 [2], was again used for a high accuracy estimate of the sinusoidal frequencies.

[1] As a quick summary, this algorithm performs a tone by tone amplitude and frequency estimation by fitting a quadratic to the largest amplitude peak before subtracting it from the spectrum and moving on to the next highest peak. Detection is stopped at a noise floor and then each peak is sequentially reinserted to the spectrum, re-estimated and then removed again.

[2] with the re-estimation stage was turned off

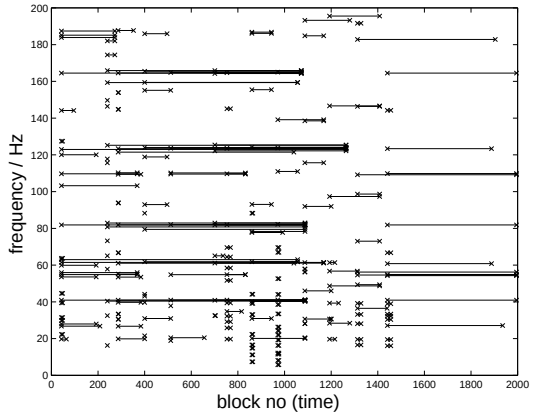

Figure 2: Output of stage 3.

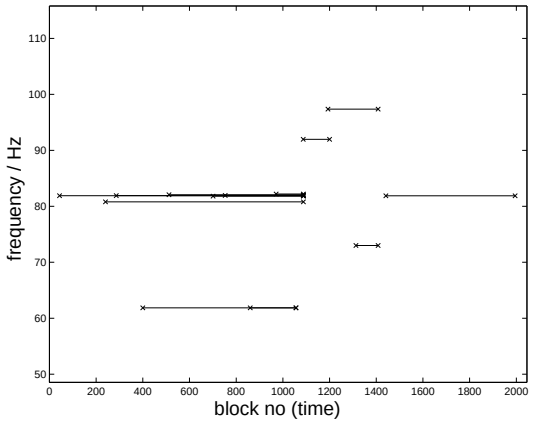

Figure 3: Output from hypothesis selector

For each hypothesis, the first four partials were tracked using a harmonic comb type method. A 3% error was allowed, justified by the fact that this is both the maximum error allowed before the pitch is quantised to the next semitone and also that it is the psychoacoustic limit before a harmonic becomes separated from the note percept (Moore 1997).

A hypothesis was defined to have no support when there was no energy found in any of the first four partials. At this stage, it was moved to a list of dying harmonics for one frame, from which it could be resurrected if further data was found. Otherwise, it was permanently moved to a list of completed notes. Figure 2 shows the output of this stage for the example of figure 1. Some hypotheses have continued to find support for long periods and hence there is some overlap.

2.4 Hypothesis trimming

Finally, there is the complicated process of determining which is the correct hypothesis at any given point in the sample. The first stage of this is to determine which is the most likely hypothesis for each onset. This is done by finding a likelihood as follows:

$$\mathrm{P} = \mathrm{P}_l \times \mathrm{P}_s \quad (2)$$

$$\mathrm{P}_s = \frac{S_n}{S_{max}} \quad (3)$$

$$\mathrm{P}_l = \begin{cases} \frac{L_n}{L_{pk}} & \text{if } \frac{L_n}{L_{pk}} < 1 \\ 1 & \text{otherwise} \end{cases} \quad (4)$$

where L_n is the duration of hypothesis, n; L_{pk} is the duration from the considered onset to the subsequent one; S_n is the note confidence; and S_{max} is the maximum note confidence for the onset considered.

To avoid octave errors, the measure P for any subharmonics of a hypothesis is subtracted from the considered hypothesis. Then, for each onset, the hypothesis with the highest value, P is selected. Figure 3 shows a sample output from this stage for the previous example.

The final stage could be considered to be a tidying up process; there are several situations that occur which are undesirable. Firstly, there is the case of spurious onsets; these could either be instances where there is actually no bass note at all or where there is a note continuing from before the spurious onset. Another case is that a note might continue to have evidence after a new (genuine) onset, perhaps due to reverberation or it being an undamped open string. Also, there is the issue of repeated notes at the same pitch, where the first note will continue to have found evidence over the duration of the second.

To address the last of these, when there is an overlap, if the pitch of the second note is the same as the first and if the power in any of the first four partials rose by a factor of 2 or more, then the first note is deemed to have ended at the second onset, and the new note of the same pitch started.

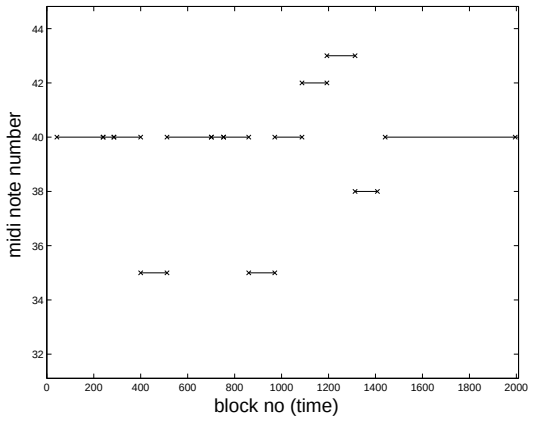

Figure 4: Final Output from Algorithm

name	N	N_f	N_c	N_p	N_m	N_x	R_k	R_h
Climb	6	6	5	1	0	0	83.3	83.3
Sstion (a)	20	21	19	0	1	2	90.4	86.4
Sstion (b)	15	19	14	1	0	4	73.7	73.7
Rocky	8	8	6	1	1	1	75.0	66.7
Surrey	17	18	17	0	0	1	94.4	94.4
Fever	18	19	14	1	3	4	73.7	63.6
Fame	7	7	7	0	0	0	100	100
LCB	40	35	27	6	7	2	77.1	64.3
Selota	23	26	16	5	2	5	61.5	57.1
Cantaloup	15	15	12	2	1	1	80.0	75.0

Table 1. Results for a variety of examples

To address the other issues, it was decided to err on the side of caution and only remove notes which had a very low confidence. A note was cut and the previous note left unchanged if the value of a measure, $\mathrm{P}_{cut} < 0.05$.

$$\mathrm{P}_{cut} = M_c \times M_p \quad (5)$$

$$M_c = \frac{S_n}{S_{n-1}} \quad (6)$$

$$M_p = \frac{2a_n}{a_{n-1} + a_{n+1}} \quad (7)$$

M_c is therefore a measure of the note confidence relative to the previous peak; M_p is a measure of the onset amplitude, a, relative to the surrounding peaks. Otherwise, the initial note is deemed to have ended at the new onset.

3 Results

The example shown throughout section 2 is for a real bass line but played in isolation and is hence a somewhat contrived example. A real example, shown in figure 5, is Miles Davis' rendition 'Surrey With a Fringe On Top', consisting of upright bass, drums, piano and trumpet. Here, the bass line has been correctly identified for each note and there is a single spurious onset which is due to a piano chord at that moment.

Table 1 presents results for a number of real world examples[3]. N is the actual number of notes; N_f, the number found; N_c, those found correctly; N_p, number with pitch errors; N_m the number missing; and N_x the number of notes added extraneously.

Using the metric of Kashino and Murase (1998):

$$R_k = \frac{\text{\# notes found correctly}}{\text{\# notes found in total}} \quad (8)$$

these examples give a 78.7% success rate. This is comparable with the best figures quoted by other transcription methods, though it should be noted that the task attempted here is not to identify every note sounded, but just those from a given instrument. A metric which takes account of the notes for which the onsets are missed is proposed below. This gives a success rate of 72.5%.

$$R_h = \frac{\text{\# notes found correctly}}{\text{\# notes found in total} + \text{\# notes missed}} \quad (9)$$

[3] see http://www-sigproc.eng.cam.ac.uk/∼swh21/icmcbass.html or the ICMC conference proceedings CDrom for plots and audio signals.

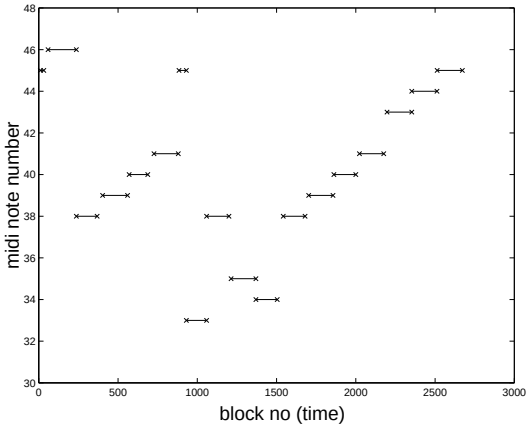

Figure 5: Output for 'Surrey With a Fringe on Top'

4 Conclusions

An algorithm to transcribe bass lines from audio recordings is presented. This has four stages which operate sequentially upon the data. Results are show for a monophonic bass line in isolation and then for a number of real world examples, with a high success rate.

References

Goto, M. and S. Hayamizu (1999). A real-time music scene description system: detecting melody and bass lines in audio signals. In *Proc. IJCAI Workshop on CASA*, pp. 31–40.

Kashino, K. and H. Murase (1998). Music recognition using note transition context. In *Proc. ICASSP*, pp. VI 3593–6.

Macleod, M. (1998, January). Nearly fast ML estimation of the parameters of real or complex single tones or resolved multiple tones. *IEEE Trans. Signal Processing 46*(1), 141–148.

Martin, K. (1996). A blackboard system for automatic transcription of simple polyphonic music. Technical Report 385, Media Laboratory, MIT.

Moore, B. (1997). *An introduction to the psychology of Hearing* (4th ed.). Academic Press.

Scheirer, E. (1995, September). Extracting expressive performance information from recorded music. Master's thesis, Media Lab, MIT.

Sterian, A. and G. Wakefield (2000). Music transcription systems: from sound to symbol. In *Proc. AAAI-2000 Workshop on Aritificial Intelligence and Music*.