USING A PITCH DETECTOR FOR ONSET DETECTION

Nick Collins
University of Cambridge
Centre for Music and Science
11 West Road, Cambridge, CB3 9DP, UK
nc272@cam.ac.uk

ABSTRACT

A segmentation strategy is explored for monophonic instrumental pitched non-percussive material (PNP) which proceeds from the assertion that human-like event analysis can be founded on a notion of stable pitch percept. A constant-Q pitch detector following the work of Brown and Puckette provides pitch tracks which are post processed in such a way as to identify likely transitions between notes. A core part of this preparation of the pitch detector signal is an algorithm for vibrato suppression. An evaluation task is undertaken on slow attack and high vibrato PNP source files with human annotated onsets, exemplars of a difficult case in monophonic source segmentation. The pitch track onset detection algorithm shows an improvement over the previous best performing algorithm from a recent comparison study of onset detectors. Whilst further timbral cues must play a part in a general solution, the method shows promise as a component of a note event analysis system.

Keywords: onset detection, pitch detection, segmentation

1 INTRODUCTION

A recent paper (Collins, 2005) compared a number of musical onset detection functions with respect to onset detection performance on sets of non-pitched percussive (NPP) and pitched non-percussive (PNP) sound files. Whilst many algorithms performed successfully at the NPP task, with few false positives for a large number of correct detections, the ability of the same algorithms to parse the PNP set was substantially reduced. The most successful attempt was that of the phase deviation algorithm (Bello et al., 2004), which uses a measure of the change of instantaneous frequency. It was proposed that this success could be linked to the use of stable pitch cues as a segmentation feature, a tactic also highlighted by Tristan Jehan in his event analysis/synthesis work (Jehan, 2004). Fundamental frequency trails have been segmentation features in work by teams from IRCAM (Rossignol et al., 1999b,a) and Universitat Pompeu Fabra (Gómez et al., 2003b,a). Whilst many signal attributes, particularly timbral descriptors, may contribute to onset detection and event parsing (Handel, 1995; Yost and Sheft, 1993; Moore, 1997), the use of a central pitch percept is investigated in this paper as one component of a plausible strategy, and a significant one for the source material tackled herein.

In this paper I attempt to explore the basis of an improved onset detection algorithm for pitched material which uses the stability of a pitch percept as the defining property of a sound event. In order to obtain a clean detection signal, the output of a pitch detection algorithm is processed in various ways, including by the suppression of vibrato, following Rossignol et al. (1999b). The choice of pitch detection algorithm is open, but the specific detector considered in this paper is Brown and Puckette's constant Q transform pitch tracker (Brown and Puckette, 1993).

The material with which I am concerned provides the hardest case of monophonic onset detection, consisting of musical sounds with slow attacks and containing vibrato, such as the singing voice (Saitou et al., 2002). Vibrato associated frequency and amplitude modulation provides problems to traditional energy based onset detectors, which tend to record many false positives as they follow the typically 4-7 Hz oscillation. For such material, the sought after performance is a segmentation as a human auditor would perceive sound events. Better than human listener performance, as possible for some high speed percussive sequences via non-real-time digital editing or by algorithm (Collins, 2005) is unlikely.

The applications of such an algorithm are multifold. Onset detection is a frontend to beat induction algorithms (Klapuri et al., 2004), empowers segmentation for rhythmic analysis and event manipulation both online and offline (Jehan, 2004; Brossier et al., 2004), and provides a basis for automatically collating event databases for compositional and information retrieval applications (Rossignol et al., 1999b; Schwarz, 2003). Extraction of note event locations from an audio signal is a necessary component of automatic transcription, and the vibrato suppression investigated here may assist clear f0 estimation. For music

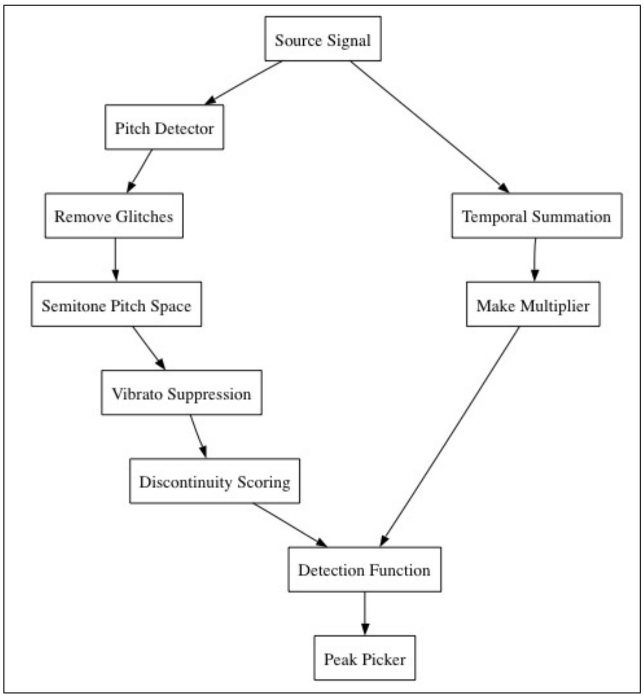

Figure 1: Overview of the algorithm

information retrieval, the ’query by humming’ approach requires the parsing of monophonic vocal melodies from audio signal alone.

2 ALGORITHM OUTLINE

Figure 1 gives an overview of the detection algorithm and the associated signal features based on the extracted fundamental frequency f_0. The following subsections will address successive stages of the onset detector.

2.1 Pitch Detection

Brown and Puckette (1993) describe an efficient FFT based pitch detection algorithm which cross correlates a harmonic template with a constant Q spectrum in a search for the best fitting fundamental frequency f_0. The form of the template is devised so as to minimise octave errors; the template consists of the first 11 harmonics, weighted from 1.0 to 0.6. A further stage evaluates phase change in the winning FFT bin to get a more accurate value for the pitch unconstrained by the limited bin resolution. Since the full details are given in their papers (Brown and Puckette, 1992, 1993) and my implementation follows that work I shall avoid a fuller discussion of this pitch detection method. Alternative pitch detection algorithms may easily be placed as front-ends to the analysis system now to be described.

The 4096 point FFT driving the pitch detector was run with a step size of 512 samples, for a frame rate of around 86 Hz (all the audio signals involved had 44100Hz sampling rate). The pitch detector output was taken from 150-2000Hz, with values outside this range shifted by octave steps into this compass, and values outside 22050Hz sent to 1 Hz, where they are easily cleaned up with the algorithm next described.

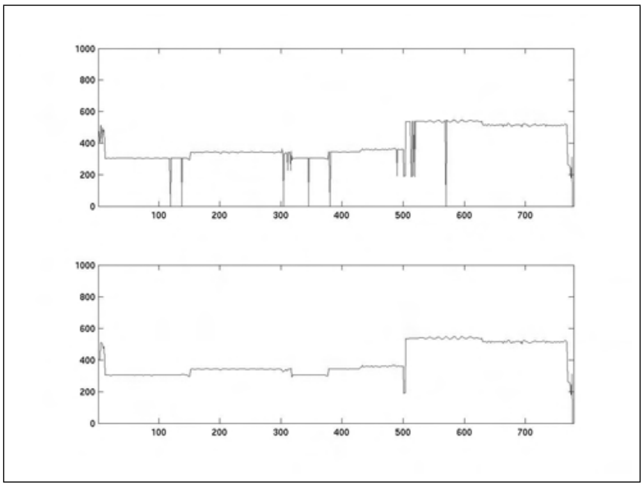

Figure 2: The upper f_0 track is cleaned up and the result is the lower track

A post processing stage was added to clean up some small blips in the signal, consisting of momentary octave errors and rogue outliers. Whilst a jump to an octave which is then maintained could indicate a true octave leap in the music, some obvious short-term octave errors were seen, with lengths of one or two frames. The original Brown/Puckette algorithm also occasionally created some strange values during otherwise relatively stable held pitches. The pseudocode in figure 3 reveals the tactic employed to clean up these short-term errors. The MATLAB indexing convention of counting from 1 is used. The two tests check against the ratio of an equal tempered semitone.

Figure 2 demonstrates the application of the algorithm on a signal which has out of bound pitches and instantaneous errors against the general trend.

It is convenient to transform the fundamental frequency track to pitch in semitones prior to vibrato suppression, as a musically normalised representation. An arbitrary reference point is selected such that 0 Hz is transformed to 0 semitones.

$$p = 12 * \log_2((f + 440)/440) \qquad (1)$$

2.2 Vibrato Suppression

The f_0 track is perturbed by vibrato, and this can be attributed as the chief cause of noise on that signal disrupting its use in segmentation. Rossignol et al. (1999b) noted this in their event segmentation paper, and sketch a vibrato suppression algorithm. Herrera and Bonada (1998) have also outlined both frequency domain and time domain vibrato suppression methods within the context of the SMS (Spectral Modeling Synthesis) framework, using an FFT to isolate 6-7Hz vibrato by analysing peaks in the frequency domain before suppression and IFFT resynthesis, and in the time domain, a 10Hz high pass filter on a 200mS window. These methods require the before application identification of the mean around which a vibrato fluctuates, and utilise fixed windows. Rossignol

```
postprocessing(arg input)
for jj= 2 to 7 {
    for ii= 1 to (length(input)-jj){
        testratio= input(ii)/input(ii+jj);
        if testratio < 1.059 AND testratio > 0.945{
            for kk=1 to (jj-1){
                mid = (input(ii)+input(ii+jj))*0.5;
                testratio2= input(ii+kk)/mid;
                if testratio2 > 1.059 OR testratio < 0.945
                    input(kk) = mid;
                }
            }
        }
    }
output=input;
```

Figure 3: Pseudocode for the outlier removal algorithm

et al. (1999a) also expands upon a selection of methods for suppression; I followed the 'minima-maxima detection' method as in common with Rossignol et al. (1999b) as the most plausible for my purposes.

Attempts to implement the Rossignol et al. (1999b) algorithm, however, were somewhat thwarted by the question of the best windowing strategy to use; their algorithm is underspecified. A vibrato suppression algorithm is described here which is inspired by their work but makes explicit how the search for regions of vibrato will take place, and uses some variation in the criteria for a vibrato detection and substituting value, along with variable window size to encompass vibrato regions.

Vibrato removal proceeds in windows of 300mS, with a step size of 100mS. If the difference of the maximum and minimum value of the input within this window is less than 1.5 semitones, a search for vibrato ensues. All maxima and minima within the (open) window range form a list of extrema. Lists of differences in time and in amplitude of the extrema are taken, and the variances of these lists calculated. Note that this is different to Rossignol et al. (1999b) where the maxima and minima lists are considered separately. The quantity *pextrema* is calculated as the proportion of the time differences between extrema that fall within the vibrato range of 0.025 to 0.175 seconds, corresponding to 2.86 to 20 Hz frequency modulation. A vibrato is detected when *pextrema* is large and the variances are sufficiently small.

Given a vibrato detected in a window, the window is now gradually extended so as to take in the whole duration of this vibrato; this guarantees that the corrections will not be piecemeal, giving rise to some erroneous fluctuations. A number of conditions are checked as the window is incrementally widened, so as not to confuse a vibrato with a jump to a new pitch. The mean of the input has been precalculated in 21 frame segments centred on each point. This mean allows a guide as to the centre point of any vibrato oscillation; if this mean changes during the window extension, it is likely that a new note event has commenced. This test was particularly important in cases of singing where the magnitude of vibrato on one tone could encompass the smaller vibrato magnitude on a succeeding tone. Secondly, the window is only extended where no value departs more than a semitone from the mean of the extrema list. The correction is applied, replacing all values in the window with the mean of the extrema list. After suppressing a vibrato, the search for vibrato recommences with the window positioned at the next frame unaffected by the changes.

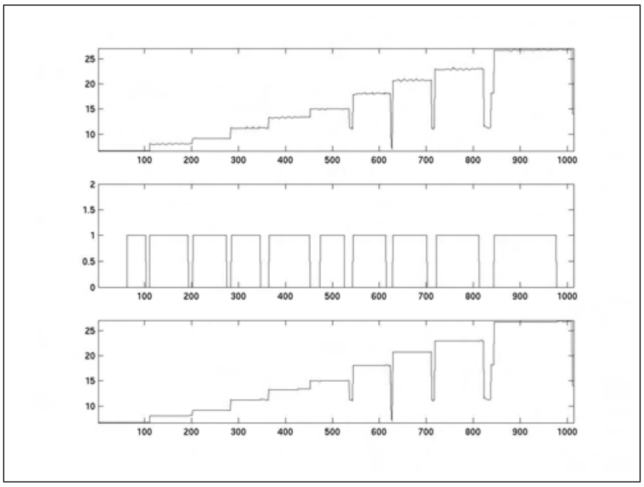

Figure 4: Vibrato suppression for an ascending arpeggiated violin signal. The FFT frames are on the abscissae, pitch in semitones or a 0/1 flag for the ordinate

Figure 4 shows an example where the vibrato suppression works effectively. The top part of the figure shows the input, the centre marks areas where vibrato was detected and shows the length of the windows after extension, and the bottom shows the vibrato suppressed output. Figure 5 shows a less clean case where the suppression does not remove all the frequency modulation. The heuristical algorithm given in this paper could likely be extended via such tactics as a cross correlation search for matches to sinusoidal variation exhaustively through appropriate frequencies or by further rules based on a study of instrumental vibrato. It works well enough, however, for evaluation purposes herein.

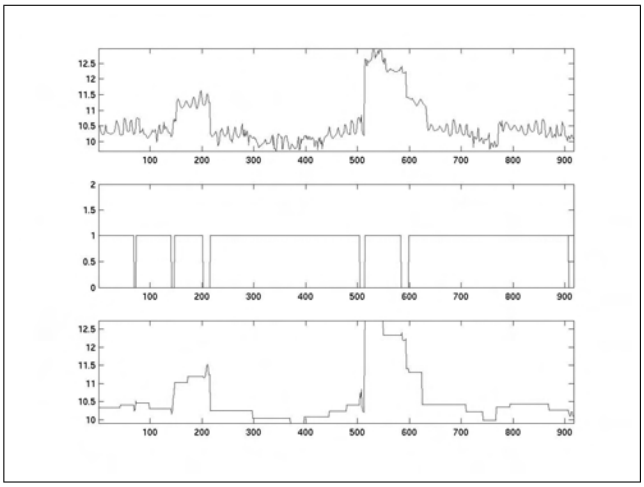

Figure 5: Vibrato suppression for a solo soprano signal. The FFT frames are on the abscissae, pitch in semitones or a 0/1 flag for the ordinate

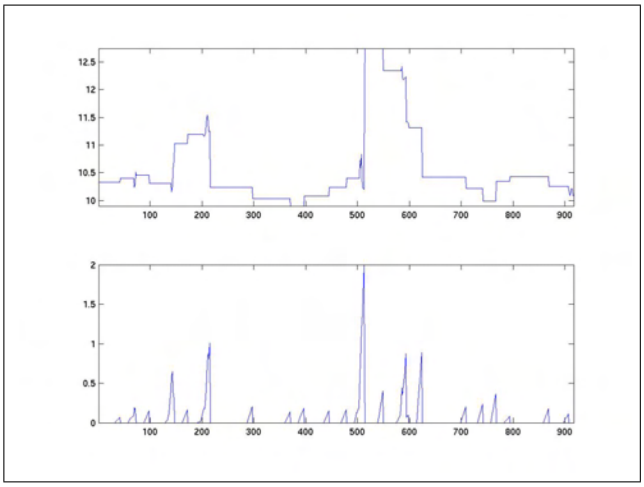

Figure 6: The upper cleaned and vibrato suppressed pitch track is converted to a detection function

2.3 Assessing Peaks of Instability

Given the vibrato suppressed pitch tracks, note events must be distinguished by jumps of pitch. A procedure is applied to rate the strength of changes in the pitch track p over time.

$$df(i) = \sum_{j=1}^{8} \min\left(|p(i) - p(i+j)|, 2\right) \qquad (2)$$

The $\min$ operator disregards the size of changes greater than a tone to avoid overly biasing the output detection function df based on the size of leap between notes involved. Figure 6 demonstrates df for a soprano signal.

Because changes are sought out, cues for multiple note events in a row of the same pitch are the most difficult case to spot (particularly questionable are the case of smooth transitions between same pitch notes- how little energy drop can a player get away with?). It is assumed that note onsets should show some slight perturbation in pitch, though the pitch integration area is around 90mS in the FFT. The pitch track test may have to be combined with other features, to be described next. However, one interesting case, that is not particularly well dealt with by the vibrato suppression stage at the present time, is that the end and restart of a vibrato itself may indicate a transition between successive notes.

2.4 Correction for Signal Power

Because the detection function did not take account of signal power, onsets would often appear at the very tails of events, for events which end in silence. To counteract this, a multiplier was introduced based on the signal power immediately following a given frame. A basic temporal integration was carried out, taking a weighted sum over 10 frames, and compressing to 1 for all reasonably large values. Small values under 0.01 of the maximum power were left unaffected and downweighted troublesome points in the pitch detector based detection function.

2.5 Peak Picking

A detection function must yield onset locations via some peak picking process. Bello et al. (2004) provide an adaptive peak picking algorithm based on a median filter on a moving window. Their peak picker was used as a common stage in the evaluation, following (Collins, 2005; Bello et al., 2004), and the algorithm is not discussed further here.

3 EVALUATION

3.1 Procedure

An evaluation of the pitch detection based onset detector was carried out using the same methodology as previous comparative studies of onset detection effectiveness (Collins, 2005; Bello et al., 2004). Pitched non-percussive (PNP) soundfiles originally prepared and annotated by Juan Bello formed the test set. 11 source files were selected, containing 129 onsets, comprising slow attack and high vibrato sounds from strings and voices. The onsets were sparse in relatively long sound files, providing a great challenge; with amplitude modulation associated with vibrato, it is unsurprising that loudness based detection functions fared so poorly in Collins (2005). The tolerance for matches between algorithm and hand-marked onsets was set at a very tolerant 100mS, though this window was small compared to the average distance between note events.

The pitch track onset detection function was compared to the phase deviation detection function with a common adaptive peak picking stage. The peak picker has a parameter δ which acts like an adaptive threshold; this was varied between -0.1 and 0.53 in steps of 0.01, giving 64 runs on the test set for each detection function. A Receiver Operating Characteristics curve was drawn out as delta is varied. This ROC curve is given in figure 7. The closest points to the top left corner indicate the better performance, with many correct detections for few false positives.

Table 1: NPP test set comparison of detection functions with Bello et al. (2004) peak picker

detection function	score (eqn 4)	CDR	Onsets Detected	False Positives	best δ
1. pitch track detection function	42.6	-17	58.1	36.4	0.13
2. phase deviation (Bello et al., 2004)	32.8	-36.4	45.0	37.0	0.13

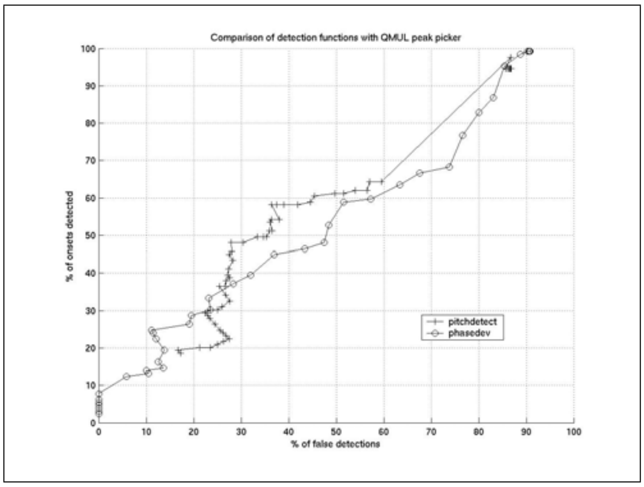

Figure 7: ROC curve of false positives against correct detections comparing phase deviation and pitch track onset detector functions over varying δ

Results for the best δ for each algorithm are given in table 1 with ratings with respect to two measures of performance. Liu et al. (2003)'s Correct Detection Ratio (CDR) is described by the equation:

$$\text{CDR} = \frac{total - missing - spurious}{total} * 100\% \qquad (3)$$

but is not constrained, however, to return values between 0-100. I also introduce therefore an evaluation formula fromDixon (2001), originally used for the assessment of beat tracking algorithm performance as an alternative scoring mechanism, combining matches m, false positives F^+ (spurious) and false negatives F^- (missing).

$$\text{score} = \frac{m}{m + F^- + F^+} * 100\% \qquad (4)$$

The denominator includes the term for the number of onsets in the trial n as $m+F^-$. These measures are the same as in (Collins, 2005).

3.2 Discussion

A small advance is shown by the pitch detection based onset detector, its performance being marginally better than the phase deviation and by extension all the energy based detection functions considered in (Collins, 2005). The success of a pitch detection cue gives corroborative evidence that note events defined by stable pitch percept are a plausible segmentation strategy. The fact that vibrato had to be suppressed for effective performance shows the importance of higher level feature extraction in human segmentation. As noted above, the onset and offset of a vibrato may be a feature that helps to segment successive notes of the same pitch. It might even be speculated that the appearance of vibrato in long notes can be linked to a human desire for stimulation over time, for the confound given by vibrato and associated amplitude modulation (often at 4-7 Hz) is comparable to new amplitude cued events at the same rate. The central pitch around which the vibrato oscillates maintains the identity of a single note event.

Various problems with the evaluation task were noted, which may have underrated the performance of the pitch detector. First, the annotations were at their most subjective for this type of note event; as Leveau et al. (2004) note, the annotation task involves some variability in decisions between human experts, particularly for complex polyphonic music and instruments with slow attacks. However, at the time of writing, the Bello database provided a larger test set (11 as opposed to 5 files), and the Leveau database could not be made to function properly within MATLAB.

Human pitch perception shows different time resolution capabilities to the computer pitch tracker used herein. Whilst the qualitative agreement of onset locations with the hand marked ones was much more impressive for the stable pitch detector than the phase deviation (for example, figure 8), these would often be early with respect to the human marked positions (though could also appear late). To compensate somewhat, a delay of 7 frames had been introduced in the detection function for the comparison test. The time resolution of the new onset detection algorithm is dependent on the lower time resolution of the pitch detection algorithm, with a 4096 point FFT (pitch detection accuracy degrades with a shorter window); the phase deviation was much less susceptible to this problem, based on a 1024 point FFT. Localisation could perhaps be improved by zero padded FFTs for the pitch detector, parallel time domain autocorrelation and timbrally motivated onset detection (differentiating transient regions from smooth wherever possible) and remains an area for further investigation.

The selection of the test set also played a role. When onsets are sparse, false positives count for proportionally more over the run. A combination of sound files requiring many onsets to be detected and those with sparse onsets is a difficult combination, for onset detectors built to risk more will score very poorly on the sparse regions. It can be speculated that additional contextual clues due to timbre and musical convention are utilised by human listeners to focus their event detection strategy. An onset detection algorithm which performed well for both NPP and PNP material would most likely require some switching mechanism based on the recognition of instrument and playing style. The evocation of a pitch percept and the detection of vibrato cues may provide knowledge for deciding the event segmentation tactic.

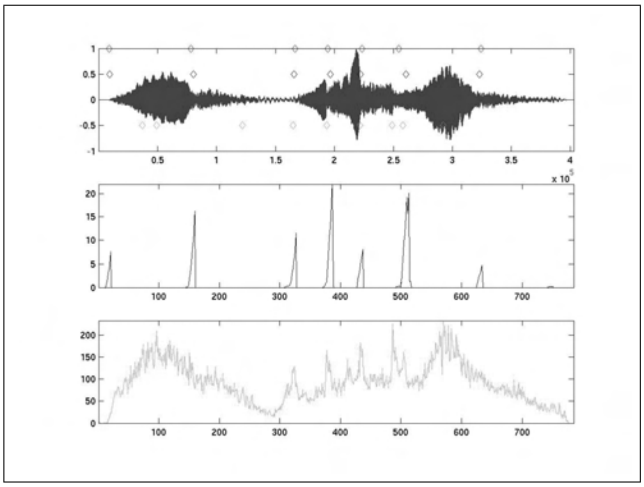

Figure 8: Comparison of pitch detector (middle) and pitch deviation (bottom) on a violin signal. The top shows the source signal with onsets marked- those on the top line show the human annotation, above the middle those due to the pitch detector algorithm and below the phase deviation

For the determination, given arbitrary material, of the best algorithm to use, a computer program might assess the stability of pitch cues (amount of fluctuation) and general inharmonicity to decide if pitched material is being targeted. Attack time cues through the file may distinguish whether to apply a combined pitch and amplitude algorithm or a pure pitch algorithm for slow attacks, and how to deal with confounds from the recognition of the specific shape of vibrato or other playing conventions (on which much further work might be done).

In testing the algorithm, it was found that the quality of pitch detection tracks was worse for lower register instruments, as for double bass or bass voice. This could be traced to inadequacies in the constant Q pitch detector for tracking fundamentals below around 150Hz. False matches to higher harmonics could skew the pitch tracks and the algorithm consistently gave the worst detection scores for such cases. Leaving these troublesome sound files out of the test set led to much improved performance. On a reduced test set of 6 files, the algorithm then achieved 58.7% correct detections for 21.4% false positives (Dixon score of 48.3, CDR 1.3) as opposed to 45.3% correct to 38.2% false positives (Dixon score 32.8, CDR -37.3) for the phase deviation.

4 CONCLUSIONS

In this paper, a pitch detection algorithm was adapted for an onset detection task on pitched non-percussive source material. This often slow attacking and vibrato-ridden monophonic music provides a challenging case for event segmentation. A very high correct identification to low false positive rate is yet to be exhibited commensurate with the success rates on the easier NPP task, but the tactic introduced shows some promise for the PNP task. It is the most promising of detection functions assessed so far, particularly by qualitative comparison of results from the new detector with that of the phase deviation algorithm.

Whilst the pitch discrimination capabilities of humans are much more refined than a semitone, a semitone has been used above as a practical working value for the size of pitch changes, as opposed to vibrato. In fact, the order of vibrato can approach that of note events, and some tighter heuristics for the vibrato suppression which take into account the nature of the vibrato percept may need to be applied.

General improvements may arise from investigating computational auditory models, for the goal on such musical material as targeted in this paper is to match a human auditor's segmentation. A better pitch detection algorithm as a frontend to event segmentation may be one modeled more thoroughly on neural coding of periodicity, with realistic pitch reaction time and stability characteristics. For example, a perceptually plausible pitch detector is proposed by Slaney and Lyon (1990).

It is likely that human auditors use instrument recognition cues to decide on a segmentation strategy. Prior knowledge of instrument timbre and associated playing conventions provide situations where human segmentation may continue to out perform machine in the near future.

ACKNOWLEDGEMENTS

Thanks are due to Juan Bello for providing the evaluation test set, and the helpful comments of four anonymous ISMIR reviewers.

REFERENCES

J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and S. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 2004.

P. Brossier, J. P. Bello, and M. D. Plumbley. Real-time temporal segmentation of note objects in music signals. In *Proc. Int. Computer Music Conference*, 2004.

J. C. Brown and M. S. Puckette. An efficient algorithm for the calculation of a constant Q transform. *J. Acoust. Soc. Am.*, 92(5):2698–701, November 1992.

J. C. Brown and M. S. Puckette. A high-resolution fundamental frequency determination based on phase changes of the Fourier transform. *J. Acoust. Soc. Am.*, 94(2):662–7, 1993.

N. Collins. A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. In *AES Convention 118*, Barcelona, May 28-31 2005.

S. Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1):39–58, 2001.

E. Gómez, M. Grachten, X. Amatriain, and J. Arcos. Melodic characterization of monophonic recordings for expressive tempo transformations. In *Proceedings of Stockholm Music Acoustics Conference 2003*, Stockholm, Sweden, 2003a.

E. Gómez, A. Klapuri, and B. Meudic. Melody description and extraction in the context of music content processing. *Journal of New Music Research*, 32(1), 2003b.

S. Handel. Timbre perception and auditory object identification. In Moore (1995), pages 425–61.

P. Herrera and J. Bonada. Vibrato extraction and parameterization in the spectral modeling synthesis framework. In *Proc. Digital Audio Effects Workshop (DAFx)*, Barcelona, 1998.

T. Jehan. Event-synchronous music analysis/synthesis. In *Proc. Digital Audio Effects Workshop (DAFx)*, Naples, Italy, Oct. 2004.

A. P. Klapuri, A. J. Eronen, and J. T. Astola. Analysis of the meter of acoustic musical signals. *IEEE Trans. Speech and Audio Processing*, forthcoming, 2004.

P. Leveau, L. Daudet, and G. Richard. Methodology and tools for the evaluation of automatic onset detection algorithms in music. In *Proc. Int. Symp. on Music Information Retrieval*, 2004.

R. Liu, N. Griffth, J. Walker, and P. Murphy. Time domain note average energy based music onset detection. In *Proceedings of the Stockholm Music Acoustics Conference*, Stockholm, Sweden, August 2003.

B. C. J. Moore, editor. *Hearing*. Academic Press, San Diego, CA, 1995.

B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, San Diego, CA, 1997.

S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J. Collette. Vibrato: Detection, estimation, extraction and modification. In *Proc. Digital Audio Effects Workshop (DAFx)*, 1999a.

S. Rossignol, X. Rodet, J. Soumagne, J.-L. Collette, and P. Depalle. Automatic characterisation of musical signals: Feature extraction and temporal segmentation. *Journal of New Music Research*, 28(4):281–95, 1999b.

T. Saitou, M. Unoki, and M. Akagi. Extraction of f0 dynamic characteristics and development of f0 control model in singing voice. In *Proc. of the 2002 Int. Conf. on Auditory Display*, Kyoto, Japan, July 2002.

D. Schwarz. New developments in data-driven concatenative sound synthesis. In *Proc. Int. Computer Music Conference*, 2003.

M. Slaney and R. F. Lyon. A perceptual pitch detector. In *Proc. ICASSP*, pages 357–60, 1990.

W. A. Yost and S. Sheft. Auditory perception. In W. A. Yost, A. N. Popper, and R. R. Fay, editors, *Human Psychophysics*, pages 193–236. Springer, New York, 1993.