Beat Extraction from Expressive Musical Performances

Simon Dixon, Werner Goebl and Emilios Cambouropoulos
Austrian Research Institute for Artificial Intelligence,
Schottengasse 3, A-1010 Vienna, Austria.
{simon,wernerg,emilios}@ai.univie.ac.at

August 27, 2001

Abstract

In order to analyse timing in musical performance, it is necessary to develop reliable and efficient methods of deriving musical timing information (e.g. tempo, beat and rhythm) from the physical timing of audio signals or MIDI data. We report the results of an experiment in which subjects were asked to mark the positions of beats in musical excerpts, using a multimedia interface which provides various forms of audio and visual feedback. Six experimental conditions were tested, which involved disabling various parts of the system's feedback to the user. Even in extreme cases such as no audio feedback or no visual feedback, subjects were often able to find the regularities corresponding to the musical beat. In many cases, the subjects' placement of markers corresponded closely to the onsets of on-beat notes (according to the score), but the beat sequences were much more regular than the corresponding note onset times. The form of feedback provided by the system had a significant effect on the chosen beat times: visual feedback encouraged a closer alignment of beats with notes, whereas audio feedback led to a smoother beat sequence.

1 Introduction

Beat extraction involves finding the times of beats in a musical performance, which is often done by subjects tapping or clapping in time with the music (Drake et al., 2000; Repp, 2001). Sequences of beat times generated in this way represent a mixture of the listeners' perception of the music with their expectations, since for each beat they must make a commitment to tap or clap before they hear any of the musical events occurring on that beat. This type of beat tracking is causal (the output of the task does not depend on any future input data) and predictive (the output at time t is a predetermined estimate of the input at t).

However, in studies of expressive timing, the aim is to investigate production rather than perception of timing, that is, independently of the listeners' expectations. To some extent, expressive timing is precisely these differences in timing between the

performance and the listeners' expectations. Therefore it is necessary to use a different approach for beat extraction: the events which occur on musical beats are identified (often they are explicitly determined by the musical score), and the sequence of beat times are generated from the onsets of these events. In this study we examine the methodology of this second task, which is non-predictive and non-causal (the choice of a beat time can be affected by events occurring later in time). In particular, we examine the roles of visual and auditory feedback and estimate the bias induced by the feedback and the precision of this type of beat extraction.

The subjects were trained to use a computer program for labelling the beats in an expressive musical performance. The program provides a multimedia interface with several types of visual and auditory feedback which assists the subjects in performing beat labelling. This interface, built as a component of a tool for the analysis of expressive performance timing (Dixon, 2001b), provides a graphical representation of both audio and symbolic forms of musical data. Audio data is represented as a smoothed amplitude envelope with detected note onsets optionally marked on the display, and symbolic (e.g. MIDI) data is shown in piano roll notation. The user can then add, adjust and delete markers representing the times of musical beats. The time durations between adjacent pairs of markers is then shown on the display. At any time, the user can listen to the performance with or without an additional percussion track representing the currently chosen beat times. The musical data used for this study are excerpts of Mozart piano sonatas played by a professional pianist on Bösendorfer SE290 computer-monitored grand piano.

This investigation examines the precision obtainable with the use of this tool under various conditions of disabling parts of the visual or auditory feedback provided by the system. We determine the influence of the various representations of data (the amplitude envelope, the onset markers, the inter-beat times, and the auditory feedback) on both the precision and the smoothness of beat sequences, and evaluate the differences between these beat times to the onset times of corresponding 'on-beat' notes. Finally, we discuss the significance of these differences in the analysis of expressive performance timing.

2 Aims

The interactive beat tracking and visualisation system (Dixon, 2001a,b) is being developed as part of a project using artificial intelligence methods to extract models of expressive music performance from human performance data (Widmer, 2001a,b). The system is being used to preprocess the performance data (audio or MIDI data) into a form that a suitable for higher level analysis using machine learning and data mining algorithms. This experiment was formulated as a pilot study to examine the role of the user interface in this task, and to estimate the precision and any bias of using such a tool.

3 Method

A group of 6 musically trained and computer literate subjects were paid to participate in the experiment. They had an average age of 27 years (range 23–29) and an average of 13 years of musical instruction (range 7–20). They were informed that they were to use a computer program to mark the times of each beat in a number of musical excerpts (according to their own perception of the beat), after being given detailed instructions on the use of the program.

The program displays the input data as either onset times, piano roll notation or amplitude envelope (Figures 1 – 4), and the mouse is used to add, delete or move markers representing the perceived times of musical beats. Audio feedback is given in the form of the original input data accompanied by a percussion instrument sounding at the selected beat times.

The experiment consisted of six conditions, relating to the type of audio and visual feedback provided by the system to the user. For each condition, three musical excerpts of approximately 15 seconds each were used. The excerpts were taken from performances of Mozart piano sonatas by a professional Viennese pianist: K331, 1st movement, bars 1–4; K281, 3rd movement, bars 8–17; and K284, 3rd movement, bars 35–42. The excerpts were chosen on the basis of their having large local tempo deviations, and the existence of data from a listening experiment performed using the same excerpts (Cambouropoulos et al., 2001).

The experiment was performed in 2 sessions of approximately 3 hours each. Each session tested 3 experimental conditions with each of three excerpts. It was attempted to design the experiment to minimise any carry-over (memory) effect for the pieces between conditions. In each session, the first condition provided audio-only feedback, the second provided visual-only feedback, and the third condition provided a combination of audio and visual feedback.

Condition 1 provided the user with no visual representation of the input data. Only a time line, the locations of user-entered beats and the times between beats (inter-beat intervals) were shown on the display (figure 1). The lack of visual feedback forced the user to rely on the audio feedback to position the beat markers.

Condition 2 tested whether a visual representation alone provided sufficient information to detect beats. The audio feedback was disabled, and only the onset times of notes were marked on the display (figure 2). The subjects were told that the display represented a musical performance, and that they should try to infer the beat visually from the patterns of note onset times.

Condition 3 tested the normal operation of the beat visualisation system using MIDI data. The notes were shown in piano-roll notation (figure 3), with the onset times marked underneath as in condition 2.

Condition 4 was identical with condition 1, except that the inter-beat intervals were not displayed. This was designed to test whether subjects made use of these numbers in judging beat times.

Condition 5 repeated the display in piano-roll notation as in condition 3, but this time with audio feedback disabled as in condition 2.

Finally, condition 6 tested the normal operation of the beat visualisation system using audio data. A smoothed amplitude envelope (10ms resolution with 50% overlap)

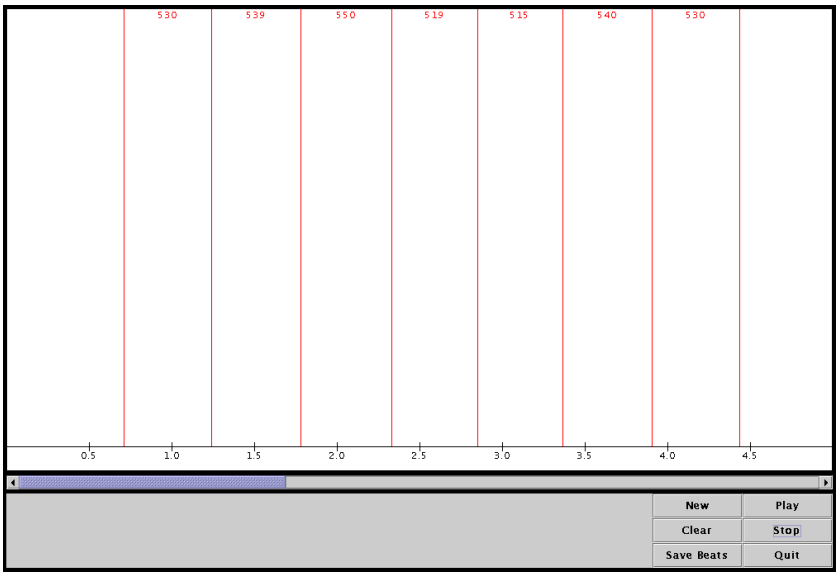

Figure 1: Screen shot of the beat visualisation system with visual feedback disabled. The beat times are shown as vertical lines, and the inter-beat intervals are marked between the lines at the top of the figure.

was displayed (figure 4), and audio feedback was enabled.

4 Results

From the selected beat times, the inter-beat intervals were calculated, as well as the difference between the beat times and the corresponding performed notes which are notated as being on the beat (the *performed beat times*). For each beat, the performed beat time was taken to be the onset time of the highest pitch note which is on that beat according to the score. Where no such note existed, linear interpolation was performed between the nearest pair of surrounding on-beat notes. The performed beat can be computed at various metrical levels (e.g. half note, quarter note, eighth note levels). We call the metrical level defined by the denominator of the time signature the *default metrical level*, which was the level that subjects were instructed to use in performing the experiment.

We say that the subject marked the beat successfully if the chosen beat times corresponded reasonably closely to the performed beat times, specifically if the greatest difference was less than half the average IBI, and the average absolute difference was less than one quarter of the IBI. Figure 5 shows the number of successfully marked excerpts at the default metrical level for each condition. The following results and most of the graphs use only the successfully marked data.

The first graphs show the effect of condition on the beat times for the excerpts from

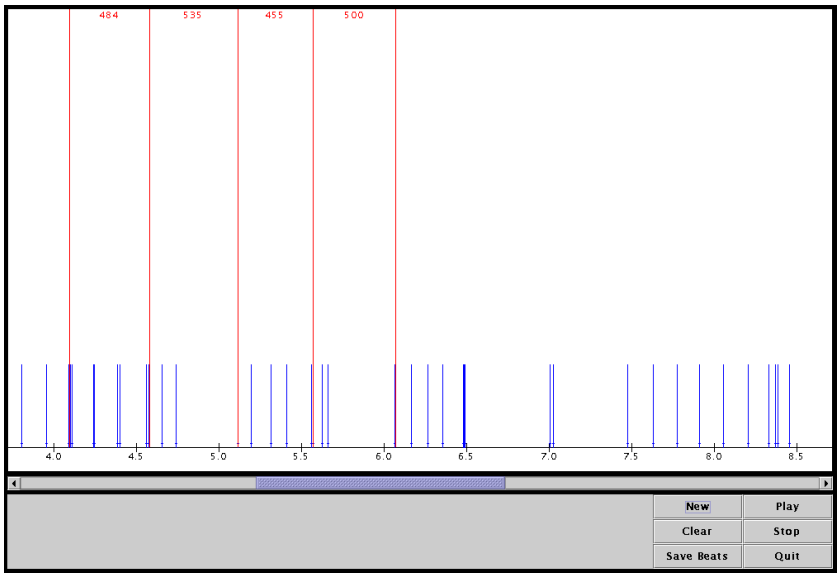

Figure 2: Screen shot of the beat visualisation system showing the note onset times as short vertical lines.

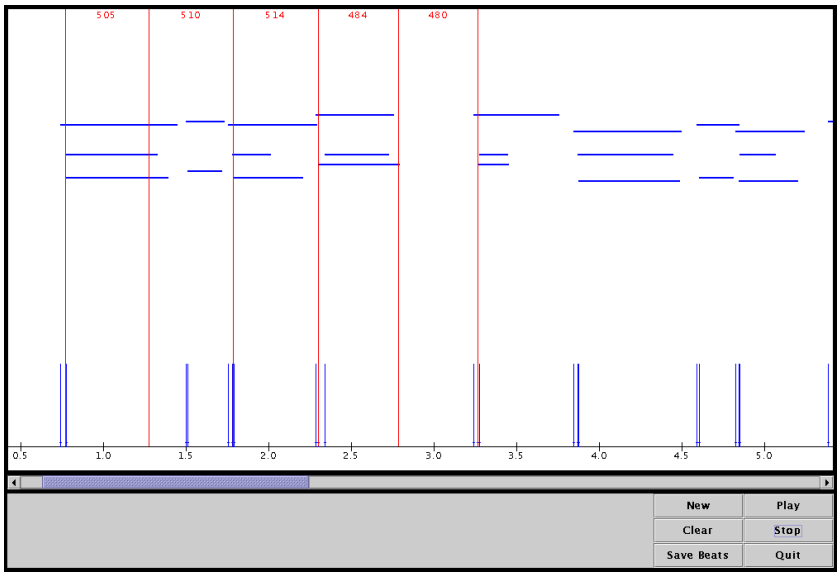

Figure 3: Screen shot of the beat visualisation system showing MIDI input data in piano roll notation, with onset times marked underneath.

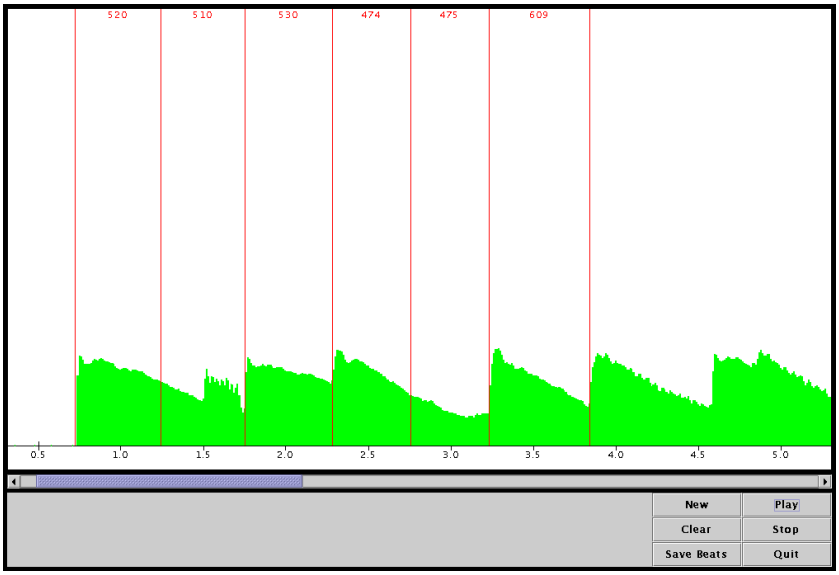

Figure 4: Screen shot of the beat visualisation system showing the acoustic waveform as a smoothed amplitude envelope.

Excerpt	Condition						Total
	1	2	3	4	5	6	
K331	3	0	6	5	4	4	22
K284	1	1	2	2	2	3	11
K281	4	4	5	4	4	4	25
Total	8	5	13	11	10	11	58

Figure 5: Number of subjects who successfully marked each excerpt for each condition (at the default metrical level).

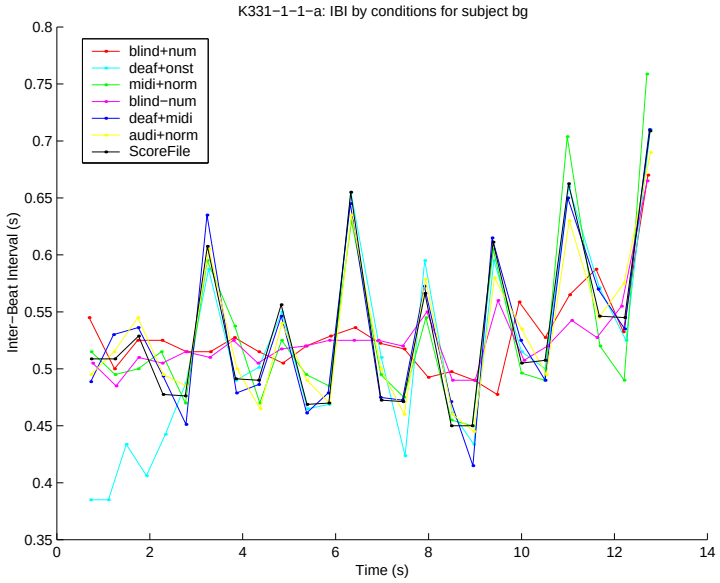

Figure 6: Inter-beat intervals from one subject for the K331 excerpt. The black line (SF) is the inter-beat intervals of performed notes.

K331 (figure 6), K284 (figure 7) and K281 (figure 8), shown for 3 different subjects. In each of these cases, the beat was successfully labelled. The notable features of these graphs are that the two audio-only conditions have a much smoother sequence of beat times than the conditions which gave visual feedback. This is also confirmed by the standard deviations of the inter-beat intervals (figure 9), which are lowest for conditions 1 and 4.

Another observation from figure 9 is found by comparing conditions 1 and 4. The only difference in these conditions is that the inter-beat intervals were not displayed in condition 4, which shows that these numbers are used, by some subjects at least, to adjust beats to make the beat sequence more regular than if attempted by listening alone.

The next set of graphs shows differences between subjects for the same conditions. Figure 10 shows that while all subjects follow the basic shape of the tempo changes, they prefer differing amounts of smoothing of the beat relative to the performed onsets. Figure 11 shows the differences in onset times between the chosen beat times and the performed beat times. The fact that some subjects remain mostly on the positive side of the graph, and others mostly negative, suggests that some prefer a lagging click track, and others a leading click track (Prögler, 1995), or at least that subjects are more sensitive to the tempo than the absolute onset times. This effect is much stronger in the cases without visual feedback (figure 12), where there is no visual cue to align the beat sequences with the performed music. The effect can be explained by auditory streaming (Bregman, 1990), which predicts that the difficulty of judging the relative timing of two sequences increases with differences in the sequences' properties such

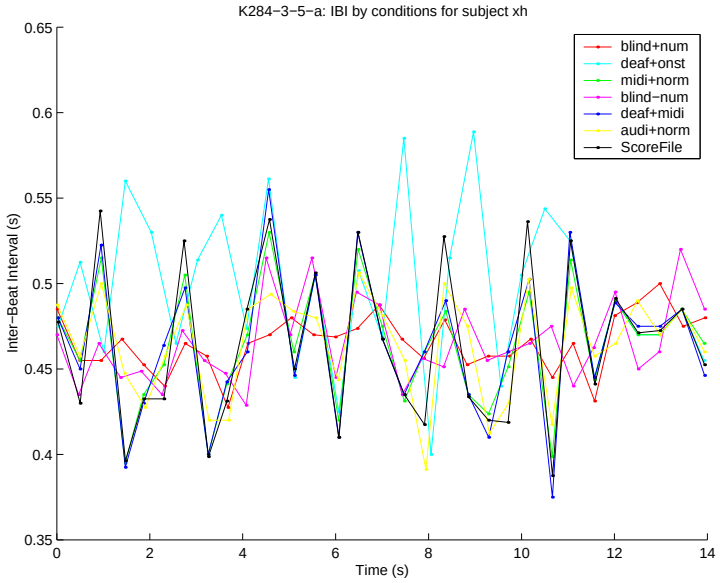

Figure 7: Inter-beat intervals from one subject for the K284 excerpt.

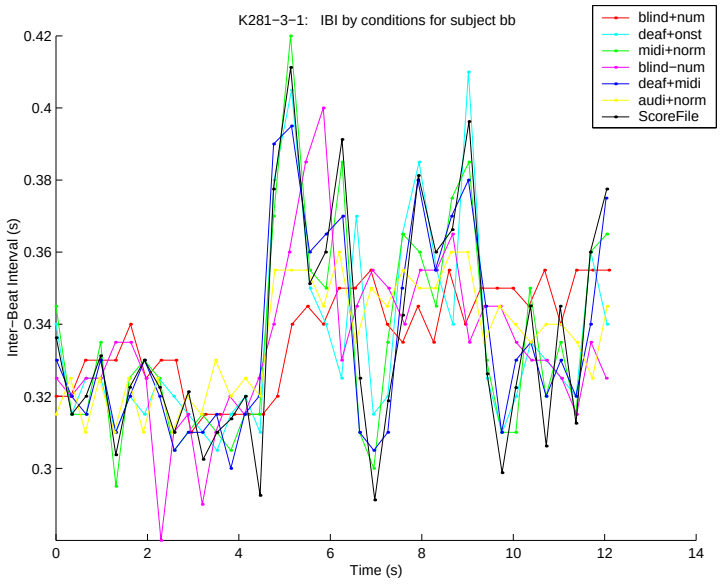

Figure 8: Inter-beat intervals from one subject for the K281 excerpt.

Excerpt	Condition						All	Performed
	1	2	3	4	5	6		
K331	35	–	59	43	68	56	53	72
K284	17	68	26	22	44	27	32	47
K281	18	29	28	22	31	25	26	31
All	24	37	42	32	48	37	37	50

Figure 9: Standard deviations of inter-beat intervals (in ms), averaged across subjects, for excerpts marked successfully at the default metrical level.

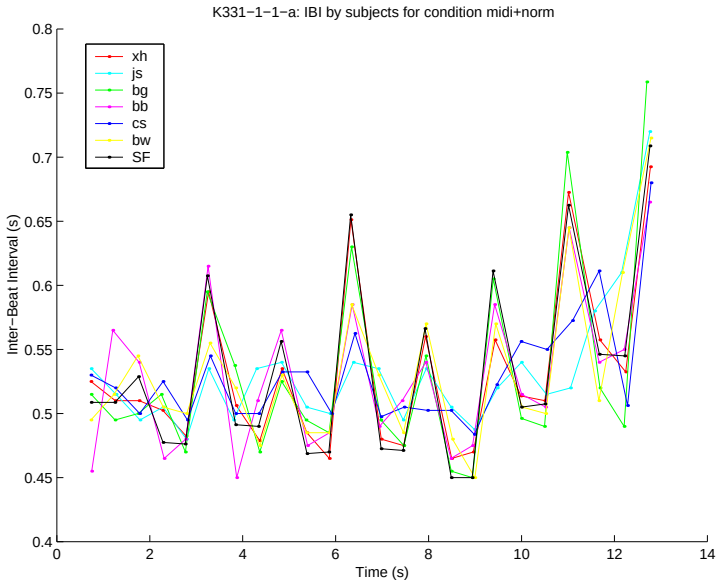

Figure 10: Comparison by subject of inter-beat intervals for the K331 excerpt.

as timbre, pitch and spatial location.

The next set of graphs show that even without hearing a musical performance, it is possible to see patterns in the timing of note onsets, and infer regularities corresponding to the beat. It was noticeable from the results that by disabling audio feedback there is more variation in the choice of metrical level. Figures 13 and 14 show successfully marked excerpts for conditions 2 and 5 respectively. Particularly in the latter case it can be seen that without audio feedback, subjects do not perform nearly as much smoothing of the beat (compare with figure 10).

Finally, we compare the presentation of visual feedback in audio and MIDI formats. Clearly the MIDI (piano roll) format provides more high-level information than the audio (amplitude envelope) format. For some subjects this made a large difference in the way they performed beat tracking (figure 15), whereas for others, it made very little difference at all (figure 16). This may be related to the familiarity of the subjects with

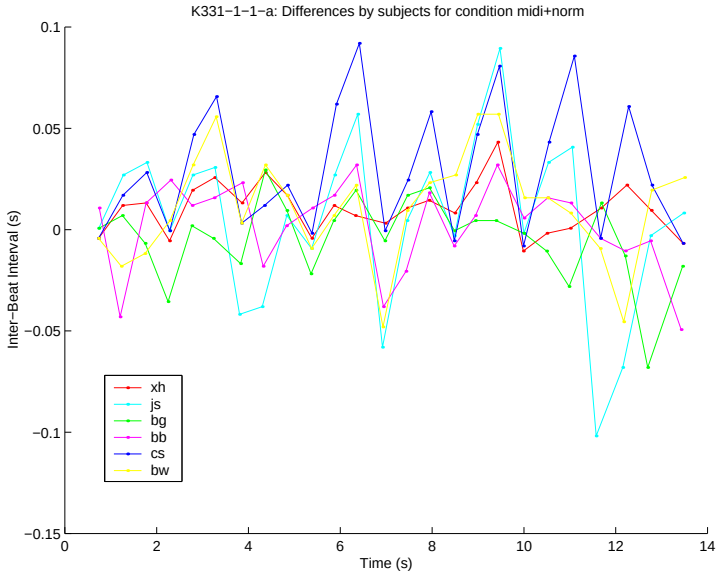

Figure 11: Beat times relative to performed notes. Differences are mostly under 50ms, with some subjects lagging, others leading the beat.

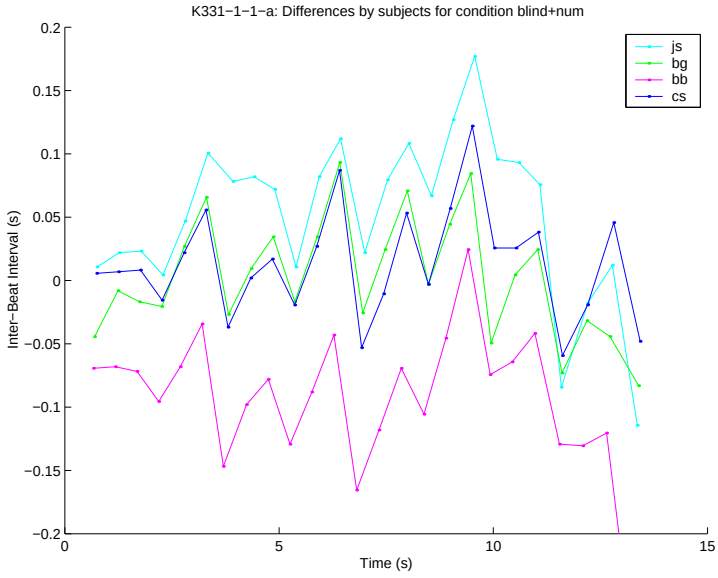

Figure 12: Beat times relative to performed notes for a test with no visual feedback. Subjects appear to be better at estimating the tempo than aligning the beats with performed notes.

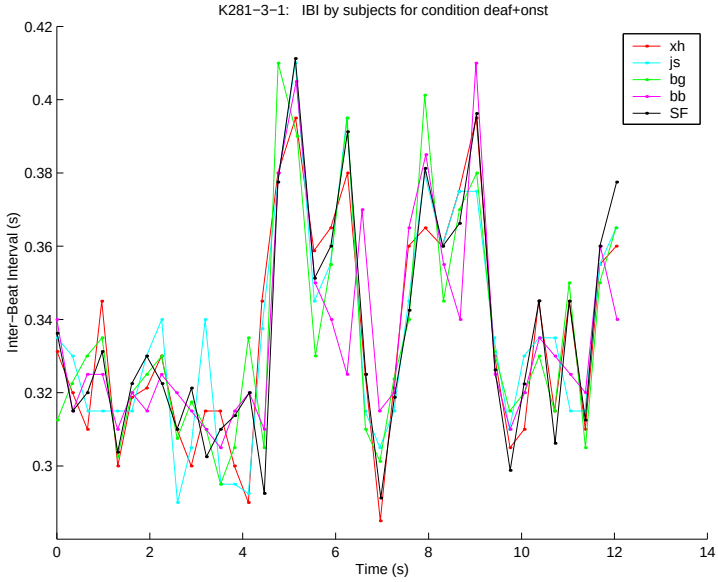

Figure 13: A beat can be found from just a visual pattern of onset times.

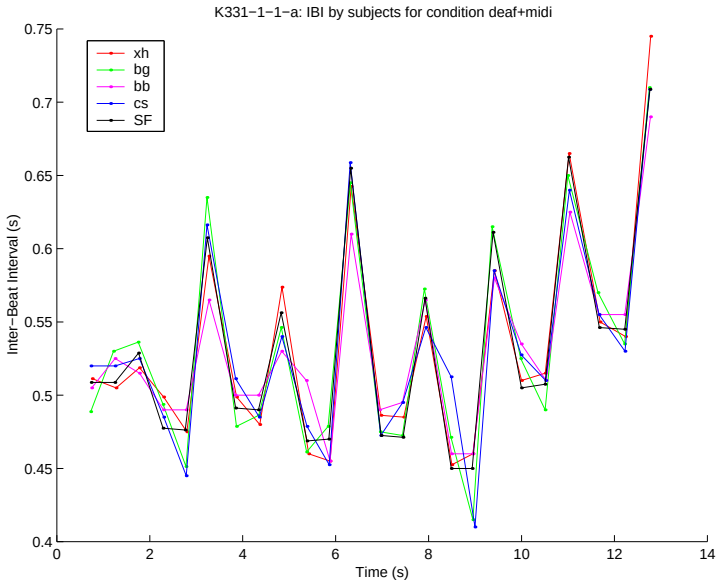

Figure 14: Marking beats on piano roll notation with no audio feedback. Without audio feedback, there is much less smoothing of the beat.

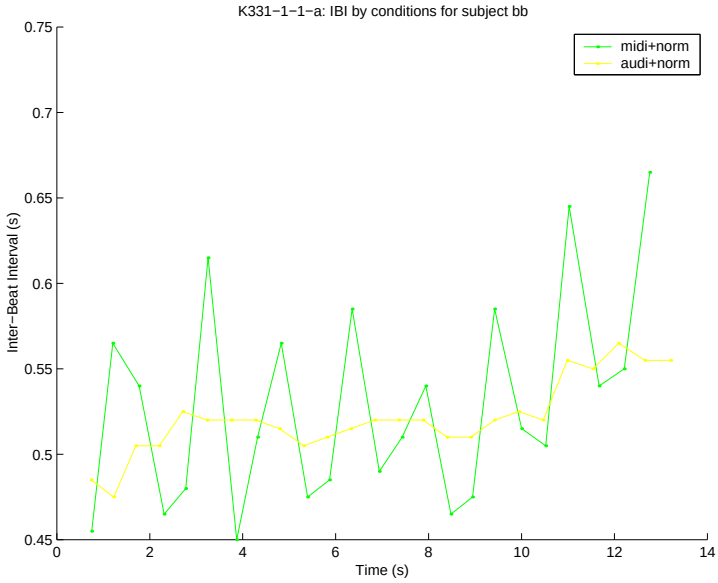

Figure 15: The difference between two types of visual feedback (piano roll notation and amplitude envelope) for one subject. The piano roll notation assists the subject in following local tempo changes.

this type of data, which varied greatly.

5 Conclusions

Although this experiment is only a pilot study, a number of observations can be made from this experiment about the perception of beat and the beat labelling interface. Firstly, even in the extreme conditions without visual or audio feedback, it is possible to find regularities corresponding to the notated musical beat, although this task becomes more difficult as feedback is reduced.

Secondly, subjects appear to prefer a beat that is smoother than the onset times of the performed on-beat notes. This is in agreement with a recent listening test with artificially generated beat sequences using the same musical excerpts (Cambouropoulos et al., 2001). In almost all cases, the beat sequences were smoother (as measured by standard deviation of inter-beat intervals) than the performed notes, but the amount of smoothing varied greatly with the feedback provided by the user interface. When users had explicit access to note onset times, the beat times were chosen mostly to correspond to these times, but without this information, the beat times reflected more of an average than an instantaneous tempo.

No numerical analysis of significance has been performed with this data yet; it is planned first to extend the study to a larger set of subjects and then perform further analysis. Further work is also required to assess the utility of the graphical interface

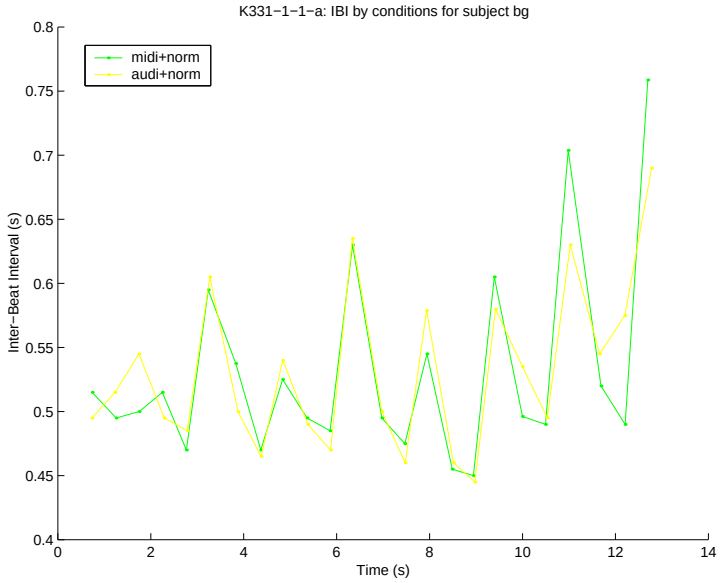

Figure 16: This subject does not seem to be influenced by the difference between two types of visual feedback (compare with previous figure).

for performing beat extraction, particularly from audio data. It is not yet clear to what extent this tool could be used in expressive performance research, that is, whether it provides sufficient precision to capture the features of interest in a performance.

Acknowledgements

This research is part of the project Y99-INF, sponsored by the Austrian Federal Ministry of Education, Science and Culture (BMBWK) in the form of a START Research Prize. The BMBWK also provides financial support to the Austrian Research Institute for Artificial Intelligence. We thank Roland Batik for permission to use his performances, and the L. Bösendorfer Company, Vienna, especially Fritz Lachnit, for providing the data.

References

Bregman, A. (1990). *Auditory Scene Analysis: The Perceptual Organisation of Sound*. MIT Press, Cambridge MA.

Cambouropoulos, E., Dixon, S., Goebl, W., and Widmer, G. (2001). Computational models of tempo: Comparison of human and computer beat-tracking. In *Proceedings of VII International Symposium on Systematic and Comparative Musicology*

and III International Conference on Cognitive Musicology, pages 18–26, University of Jyväskylä, Finland.

Dixon, S. (2001a). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1). To appear.

Dixon, S. (2001b). An interactive beat tracking and visualisation system. In *Proceedings of the International Computer Music Conference*, San Francisco CA. International Computer Music Association. To appear.

Drake, C., Penel, A., and Bigand, E. (2000). Tapping in time with mechanically and expressively performed music. *Music Perception*, 18(1):1–23.

Prögler, J. (1995). Searching for swing: Participatory discrepancies in the jazz rhythm section. *Ethnomusicology*, 39(1):21–54.

Repp, B. (2001). The embodiment of musical structure: Effects of musical context on sensorimotor synchronization with complex timing patterns. In Prinz, W. and Hommel, B., editors, *Attention and Performance XIX: Common Mechanisms in Perception and Action*.

Widmer, G. (2001a). Inductive learning of general and robust local expression principles. In *Proceedings of the International Computer Music Conference*. International Computer Music Association.

Widmer, G. (2001b). Using AI and machine learning to study expressive music performance: Project survey and first report. *AI Communications*, 14.