Perceptual Smoothness of Tempo in Expressively Performed Music

Simon Dixon
Austrian Research Institute for Artificial Intelligence, Vienna, Austria

Werner Goebl
Austrian Research Institute for Artificial Intelligence, Vienna, Austria

Emilios Cambouropoulos
Department of Music Studies, Aristotle University of Thessaloniki, Greece

We report three experiments examining the perception of tempo in expressively performed classical piano music. Each experiment investigates beat and tempo perception in a different way: rating the correspondence of a click track to a musical excerpt with which it was simultaneously presented; graphically marking the positions of the beats using an interactive computer program; and tapping in time with the musical excerpts. We examine the relationship between the timing of individual tones, that is, the directly measurable temporal information, and the timing of beats as perceived by listeners. Many computational models of beat tracking assume that beats correspond with the onset of musical tones. We introduce a model, supported by the experimental results, in which the beat times are given by a curve calculated from the tone onset times that is smoother (less irregular) than the tempo curve of the onsets.

Received June 7, 2004, accepted October 3, 2004

Perceptual Smoothness of Tempo in Expressively Performed Music

Tempo and beat are well-defined concepts in the abstract setting of a musical score, but not in the context of analysis of expressive musical performance. That is, the regular pulse, which is the basis of rhythmic notation in common music notation, is anything but regular when the timing of performed notes is measured. These deviations from mechanical timing are an important part of musical expression, although they remain, for the most part, poorly understood. In this study we report on three experiments using one set of musical excerpts, which investigate the characteristics of the relationship between performed timing and perceived local tempo. The experiments address this relationship via the following tasks: rating the correspondence of a click track to a musical excerpt with which it was simultaneously presented; graphically marking the positions of the beats using an interactive computer program; and tapping in time with the musical excerpts.

Theories of musical rhythm (e.g., Cooper & Meyer, 1960; Yeston, 1976; Lerdahl & Jackendoff, 1983) do not adequately address the issue of expressive performance. They assume two (partially or fully) independent components: a regular periodic structure of beats and the structure of musical events (primarily in terms of phenomenal accents). The periodic temporal grid is fitted onto the musical structure in such a way that the alignment of the two structures is optimal. The relationship between the two is dialectic in the sense that quasi-periodical characteristics of the musical material (patterns of accents, patterns of temporal intervals, pitch patterns, etc.) induce perceived temporal periodicities while, at the same time, established periodic metrical structures influence the way musical structure is perceived and even performed (Clarke, 1985, 1999).

Computational models of beat tracking attempt to determine an appropriate sequence of beats for a given musical piece, in other words, the best fit between a regular sequence of beats and a musical structure. Early work took into account only quantized representations of musical scores (Longuet-Higgins & Lee, 1982; Povel & Essens, 1985; Desain & Honing, 1999), whereas modern beat tracking models are usually applied to performed music, which contains a wide range of expressive timing deviations (Large & Kolen, 1994; Goto & Muraoka, 1995; Dixon, 2001a). In this article this general case of beat tracking is considered.

Music Perception volume 23, issue 3, pp. 195-214, issn 0730-7829, electronic issn 1533-8312

Many beat-tracking models attempt to find the beat given only a sequence of onsets (Longuet-Higgins & Lee, 1982; Povel & Essens, 1985; Desain, 1992; Cemgil, Kappen, Desain, & Honing, 2000; Rosenthal, 1992; Large & Kolen, 1994; Large & Jones, 1999; Desain & Honing, 1999), whereas some recent attempts also take into account elementary aspects of musical salience or accent (Toiviainen & Snyder, 2003; Dixon & Cambouropoulos, 2000; Parncutt, 1994; Goto & Muraoka, 1995, 1999). An assumption made in most models is that a preferred beat track should contain as few empty positions as possible, that is, beats on which no note is played, as in cases of syncopation or rests. A related underlying assumption is that musical events may appear only on or off the beat. However, a musical event may both correspond to a beat but at the same time not coincide precisely with the beat. That is, a nominally on-beat note may be said to come early or late in relation to the beat (a *just-off-the-beat* note). This distinction is modeled by formalisms that describe the local tempo and the timing of musical tones independently (e.g., Desain & Honing, 1992; Bilmes, 1993; Honing, 2001; Gouyon & Dixon, 2005).

The notion of just-off-the-beat notes affords beat structure a more independent existence than is usually assumed. A metrical grid is not considered as a flexible abstract structure that can be stretched within large tolerance windows until a best fit to the actual performed music is achieved but as a rather more robust psychological construct that is mapped to musical structure whilst maintaining a certain amount of autonomy.

It is herein suggested that the limits of fitting a beat track to a particular performance can be determined in relation to the concept of tempo smoothness. Listeners are very sensitive to deviations that occur in isochronous sequences of sounds. For instance, the relative JND constant for tempo is 2.5% for inter-beat intervals longer than 250 ms (Friberg & Sundberg, 1995). For local deviations and for complex real music, the sensitivity is not as great (Friberg & Sundberg, 1995; Madison & Merker, 2002), but it is still sufficient for perception of the subtle variations characteristic of expressive performance. It is hypothesized that listeners prefer relatively smooth sequences of beats and that they are prepared to abandon full alignment of a beat track to the actual event onsets if this results in a smoother beat flow.

The study of perceptual tempo smoothing is important as it provides insights into how a better beat tracking system can be developed. It also gives a more elaborate formal definition of beat and tempo that can be useful in other domains of musical research (e.g., in studies of musical expression, additional expressive attributes can be attached to notes in terms of being early or delayed with respect to the beat).

Finding the times of perceived beats in a musical performance is often done by participants tapping or clapping in time with the music (Drake, Penel, & Bigand, 2000; Snyder & Krumhansl, 2001; Toiviainen & Snyder, 2003), which is to be distinguished from the task of synchronization (Repp, 2002). 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*).

Real-time beat prediction implicitly performs some kind of smoothing, especially for ritardandi, as a beat tracker has to commit itself to a solution before seeing any of the forthcoming events—it cannot wait indefinitely before making a decision. In the example of Figure 1, an on-line beat tracker cannot produce the intended output for both cases, since the input for the first four beats is the same in both cases, but the desired output is different. The subsequent data reveals whether the fourth onset was displaced (i.e., just off the beat, Figure 1a) or the beginning of a tempo change (Figure 1b). It is herein suggested that a certain amount of a posteriori beat correction that depends on the forthcoming musical context is important for a more sophisticated alignment of a beat track to the actual musical structure.

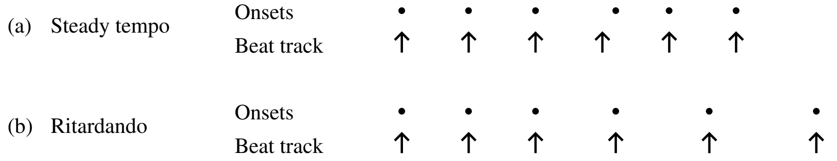


FIG. 1. Two sequences of onsets and their intended beat tracks: (a) the tempo is constant and the fourth onset is displaced so that it is just off the beat; (b) the tempo decreases from the fourth onset, and all onsets are on the beat. The sequences are identical up to and including the fourth beat, so the difference in positioning the beats can only be correctly made if *a posteriori* decisions are allowed.

Some might object to the above suggestion by stating that human beat tracking is always a real-time process. This is in some sense true, however, it should be mentioned that previous knowledge of a musical style or piece or even a specific performance of a piece allows better time synchronization and beat prediction. Tapping along to a certain piece for a second or third time may enable a listener to use previously acquired knowledge about the piece and the performance for making more accurate beat predictions (Repp, 2002).

There is a vast literature about finger-tapping, describing experiments requiring participants either to synchronize to an isochronous stimulus (sensorimotor synchronization) or to tap at a constant rate without any stimulus (see Madison, 2001). At average tapping rates between 300 and 1000 ms per tap, the reported variability in tapping interval is 3-4%, increasing disproportionately above and below these boundaries (Collyer, Horowitz, & Hooper, 1997). This variability is about the same as the JND for detecting small perturbations in an isochronous sequence of sounds (Friberg & Sundberg, 1995). In these tapping tasks, a negative synchronization error was commonly observed, that is, participants tend to tap earlier than the stimulus (Aschersleben & Prinz, 1995). This asynchrony is typically between −20 and −60 ms for metronomic sequences (Wohlschläger & Koch, 2000), but is greatly diminished when dealing with musical sequences, where delays between −6 and +16 ms have been reported (Snyder & Krumhansl, 2001; Toiviainen & Snyder, 2003). Recent research has shown that even subliminal perturbations in a stationary stimulus (below the perceptual threshold) are compensated for by tappers (Thaut, Tian, & Sadjadi, 1998; Repp, 2000).

However, there are very few attempts to investigate tapping along with music (either deadpan or expressively performed). One part of the scientific effort is directed to investigate at what metrical level and at what metrical position listeners tend to synchronize with the music and what cues in the musical structure influence these decisions (e.g., Parncutt, 1994; Drake et al., 2000; Snyder & Krumhansl, 2001). These studies did not analyze the timing deviations of the taps at all. Another approach is to systematically evaluate the deviations between taps and the music. In studies by Repp (1999a, 1999b, 2002), participants tapping in synchrony with a metronomic performance of the first bars of a Chopin study showed systematic variation that seemed to relate more closely to the metrical structure of the excerpt, although the stimulus lacked any timing perturbations. In other conditions of the studies, pianists tapped to different expressive performances (including their own). It was found that they could synchronize well with these performances, but they tended to underestimate long inter-beat intervals, compensating for the error on the following tap.

Definitions

In this article, we define *beat* to be a perceived pulse consisting of a set of *beat times* (or *beats*) which are approximately equally spaced in time. More than one such pulse can coexist, where each pulse corresponds with one of the *metrical levels* of the musical notation, such as the quarter note, eighth note, half note or the dotted quarter note level. The time interval between two successive beats at a particular metrical level is called the *inter-beat interval* (IBI), which is an inverse measure of instantaneous (local) tempo. A more global measure of tempo is given by averaging IBIs over some time period or number of beats. The IBI is expressed in units of time (per beat); the tempo is expressed as the reciprocal, beats per time unit (e.g., beats per minute).

In order to distinguish between the beat times as marked by the participants in Experiment 2, the beat times as tapped by participants in Experiment 3, and the timing of the musical excerpts, where certain tones are notated as being on the beat, we refer to these beat times as *marked, tapped*, and *performed beat times*, respectively, and refer to the IBIs between these beat times as the *marked IBI* (m-IBI), the *tapped IBI* (t-IBI) and the *performed IBI* (p-IBI). 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). For each excerpt, a suitable metrical level was chosen as the *default metrical level*, which was the quarter note level for 4/4 and 2/2 time signatures, and the eighth note level for the 6/8 time signature. (The default levels agreed with the rates at which the majority of candidates tapped in Experiment 3.) More details of the calculation of performed beat times are given in the description of stimuli for Experiment 1.

Outline

Three experiments were performed which were designed to examine beat perception in different ways. Brief reports of these experiments were presented previously by Cambouropoulos, Dixon, Goebl, and Widmer (2001), Dixon, Goebl, and Cambouropoulos

(2001), and Dixon and Goebl (2002), respectively. Three short (approximately 15-second) excerpts from Mozart's piano sonatas, performed by a professional pianist, were chosen as the musical material to be used in each experiment. Excerpts were chosen which had significant changes in tempo and/or timing. The excerpts had been played on a Bösendorfer SE275 computer-monitored grand piano, so precise measurements of the onset times of all notes were available.

In the first experiment, a listener preference test, participants were asked to rate how well various sequences of clicks (*beat tracks*) correspond musically to simultaneously presented musical excerpts. (One could think of the beat track as an intelligent metronome which is being judged on how well it keeps in time with the musician.) For each musical excerpt, six different beat tracks with different degrees of smoothness were rated by the listeners.

In the second experiment, the participants' perception of beat was assessed by beat marking, an off-line, nonpredictive task (that is, the choice of a beat time could be revised in light of events occurring later in time). The participants were trained to use a computer program for labeling the beats in an expressive musical performance. The program provides a multimedia interface with several types of visual and auditory feedback, which assists the participants in their task. This interface, built as a component of a tool for the analysis of expressive performance timing (Dixon, 2001a, 2001b), provides a graphical representation of both audio and symbolic forms of musical data. Audio data are represented as a smoothed amplitude envelope with detected note onsets optionally marked on the display, and symbolic (e.g., MIDI) data are 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 are 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.

We investigated the beat tracks obtained with the use of this tool under various conditions of disabling parts of the visual and/or auditory feedback provided by the system, in order to determine the bias induced by 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 examine the differences between these beat times and the onset times of corresponding on-beat notes. We discuss the significance of these differences for the analysis of expressive performance timing.

In the third experiment, participants were asked to tap in time with the musical excerpts. Each excerpt was repeated 10 times, with short pauses between each repeat, and the timing of taps relative to the music was recorded. The repeats of the excerpts allowed the participants to learn the timing variations in the excerpts, and adjust their tapping accordingly on subsequent attempts.

We now describe each of the experiments in detail and then conclude with a discussion of the conclusions drawn from each and from the three together.

Experiment 1: Listener Preferences

The aim of the first experiment was to test the smoothing hypothesis directly, by presenting listeners with musical excerpts accompanied by a click track and asking them to rate the correspondence of the two instruments. The click tracks consisted of a sequence of clicks played more or less in time with the onsets of the tones notated as being on a downbeat, with various levels of smoothing of the irregularities in the timing of the clicks. A two-sided smoothing function (i.e., taking into account previous and forthcoming beat times) was applied to the performance data in order to derive the smoothed beat tracks.

It was hypothesized that a click track which is fully aligned with the onsets of notes which are nominally on the beat sounds unnatural due to its irregularity, and that listeners prefer a click track which is less irregular, that is, somewhat smoothed. At the same time, it was expected that a perfectly smooth click track which ignores the performer's timing variations entirely would be rated as not matching the performance.

PARTICIPANTS

Thirty-seven listeners (average age 30) participated in this experiment. They were divided into two groups: 18 musicians (average 19.8 years of music training and practice), and 19 nonmusicians (average 2.3 years of music training and practice).

STIMULI

Three short excerpts of solo piano music were used in all three experiments, taken from professional performances played on a Bösendorfer SE275 computer-monitored grand piano by the Viennese pianist Roland Batik (1990). Both the audio recordings and precise measurements (1.25 ms resolution) of the timing of each note were available for these performances. The excerpts were taken from Mozart's piano sonatas K.331, K.281, and K.284, as shown in Table 1. (The fourth excerpt in the table, K284:1, was only used in Experiment 3.)

TABLE 1. Stimuli used in the three experiments. The tempo is shown as performed inter-beat interval (p-IBI) and in beats per minute (BPM), calculated as the average over the excerpt at the default metrical level (ML).

Sonata : Movement	Bars	Duration	p-IBI	BPM	Meter	ML
K331:1	1-8	25 s	539 ms	111	6/8	1/8
K281:3	8-17	13 s	336 ms	179	2/2	1/4
K284:3	35-42	15 s	463 ms	130	2/2	1/4
K284:1	1-9	14 s	416 ms	144	4/4	1/4

For each excerpt, a set of six different beat tracks was generated as follows. The unsmoothed beat track (U) was generated first, consisting of the performed beat times. For this track, the beat times were defined to coincide with the onset of the corresponding on-beat notes (i.e., according to the score, at the default metrical level). If no note occurred on a beat, the beat time was linearly interpolated from the previous and next beat times. If more than one note occurred on the beat, the melody note (highest pitch) was assumed to be the most salient and was taken as defining the beat time. The maximum asynchrony between voices (excluding grace notes) was 60 ms, the average was 18 ms (melody lead), and the average absolute difference between voices was 24 ms.

A difficulty occurred in the case that ornaments were attached to on-beat melody notes, since it is possible that either the (first) grace note was played on the beat, so as to delay the main note to which it is attached, or that the first grace note was played before the beat (Timmers, Ashley, Desain, Honing, & Windsor, 2002). It is also possible that the beat is perceived as being at some intermediate time between the grace note and the main note; in fact, the smoothing hypothesis introduced above would predict this in many cases.

Excerpt K284:3 contains several ornaments, and although it seems clear from listening that the grace notes were played on the beat, we decided to test this by generating two unsmoothed beat tracks, one corresponding to the interpretation that the first grace note in each ornament is on the beat (K284:3a), and the other corresponding to the interpretation that the main melody note in each case is on the beat (K284:3b). The listener preferences confirmed our expectations; there was a significant preference for version K284:3a over K284:3b in the case of the unsmoothed beat track U. In the remainder of the article, the terms "performed beat times" and "p-IBIs" refer to the interpretation K284:3a. In the other case of grace notes (in excerpt K281:3), the main note was clearly played on the beat. The resulting unsmoothed IBI functions are shown aligned with the score in Figure 2.

The remaining beat tracks were generated from the unsmoothed beat track U by mathematically manipulating the sequence of inter-beat intervals. If U contains the beat times t_i:

$$U = \{t_1, t_2, ..., t_n\}$$

then the IBI sequence is given by:

$$d_i = t_{i+1} - t_i \circ\circ \text{for } i = 1, ..., n - 1$$

A smoothed sequence $Dw = \{d_1^w, ..., d_{n-1}^w\}$ was generated by averaging the inter-beat intervals with a window of $2w$ adjacent inter-beat intervals:

$$d_i^w = \sum_{j=-w}^{w} \frac{d_{i+j}}{2w+1} \circ\circ \text{for } i = 1, ..., n - 1$$

where w is the smoothing width, that is, the number of beats on either side of the IBI of beats t_i, t_{i+1} which were used in calculating the average. To correct for missing values at the ends, the sequence $\{d_i\}$ was extended by defining:

$$d_{1-k} = d_{1+k}$$

and

$$d_{n-1+k} = d_{n-1-k}$$

where $k = 1, \ldots, w$. Finally the beat times for the smoothed sequences are given by:

$$t_i^w = t_1 + \sum_{j=1}^{i-1} d_j^w$$

Modifications to these sequences were obtained by reversing the effect of smoothing to give the sequence Dw-R:

$$r_i^w = t_i - (t_i^w - t_i) = 2t_i - t_i^w \circ\circ \text{for } i = 1, ..., n$$

and by adding random noise to give the sequence Dw-Nρ:

$$n_i^w = t_i^w + \frac{\sigma_i}{1000}$$

where σ is a uniformly distributed random variable in the range $[-\rho, \rho]$. These conditions were chosen to verify that manipulations of the same order of magnitude

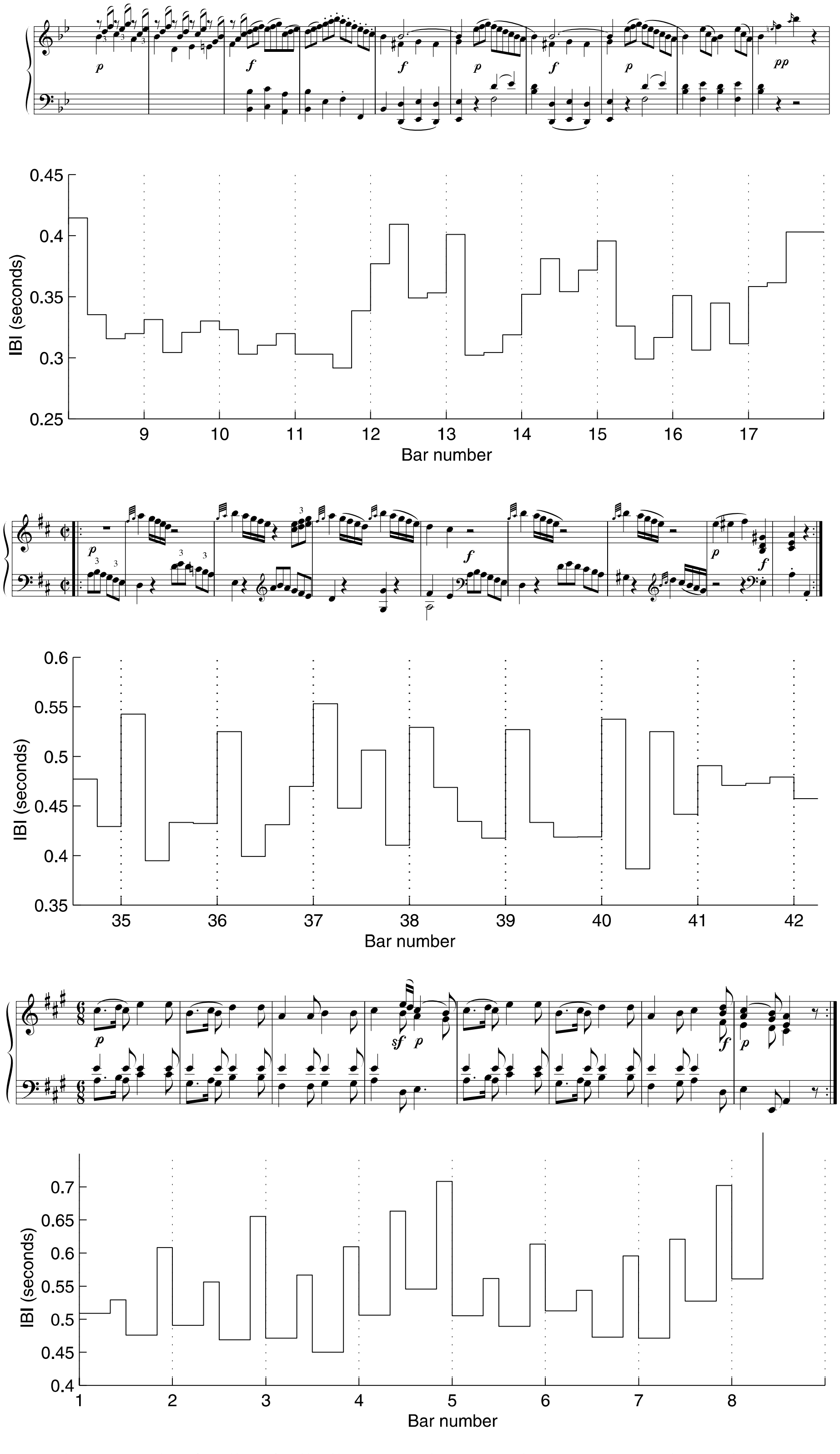


FIG. 2. The score and IBI functions for the three excerpts K281:3 (above), K284:3 (center), and K331:1 (below).

TABLE 2. Stimuli for Experiment 1: beat tracks generated for each excerpt.

Beat Track	w	Direction	Noise ρ
U	0	None	0
D1	1	Normal	0
D3	3	Normal	0
D5	5	Normal	0
D1-R	1	Reverse	0
D1-N30	1	Normal	30 ms

as those produced by the smoothing functions could be unambiguously detected.

Table 2 summarizes the six types of beat tracks used for each excerpt in this experiment.

PROCEDURE

Each beat track was realized as a sequence of woodblock clicks, which was mixed with the recorded piano performance at an appropriate loudness level. Five groups of stimuli were prepared: two identical groups using excerpt K281:3, two groups using excerpt K284:3, and the final group using excerpt K331:1. One of the two identical groups (using K281:3) was intended to be used to exclude any participants who were unable to perform the task (i.e., shown by inconsistency in their ratings). This turned out to be unnecessary. The two groups using excerpt K284:3 corresponded respectively to the two interpretations of grace notes, as discussed above. For each group, the musical excerpt was mixed with each of the six beat tracks and the resulting six stimuli were recorded in a random order, with the tracks from each group remaining together. Three different random orders were used for different groups of participants, but there was no effect of presentation order.

The stimuli were presented to the listeners, who were asked to rate how well the click track corresponded musically with the piano performance. This phrasing was chosen so that the listeners made a musical judgment rather than a technical judgment (e.g., of synchrony). The participants were encouraged to listen to the tracks in a group as many times as they wished, and in whichever order they wished. The given rating scale ranged from 1 (best) to 5 (worst), corresponding to the grading system in Austrian schools.

RESULTS

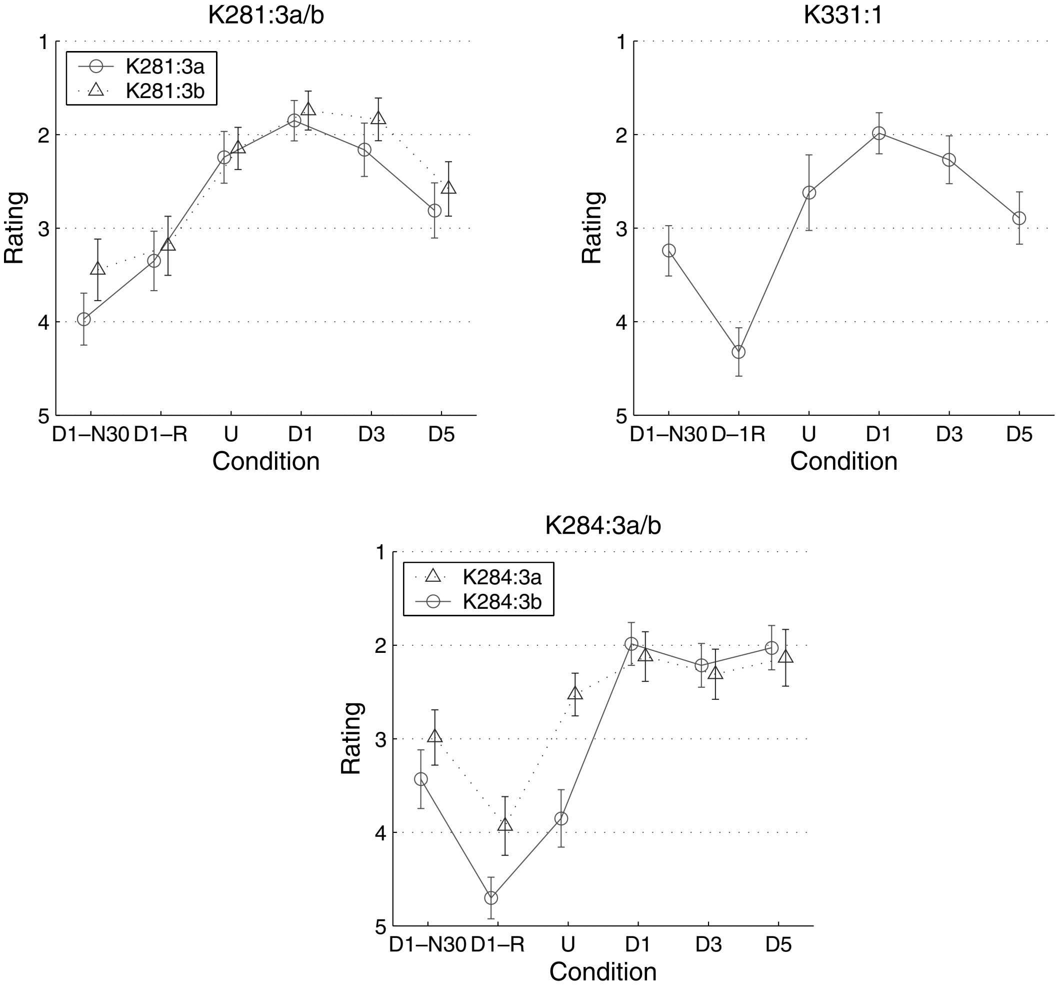
The average ratings of all participants are shown in Figure 3. As the range of ratings is small, participants tended to use the full range of values. The average ratings for the two categories of participant (musicians and nonmusicians) are shown in Figure 4. The two groups show similar tendencies in rating the excerpts, with the nonmusicians generally showing less discernment between the conditions than the musicians. One notable difference is that the musicians showed a much stronger dislike for the click sequences with random perturbations (D1-N30). Further, in two pieces the musicians showed a stronger trend for preferring one of the smoothed conditions (D1 or D3) over the unsmoothed (U) condition.

A repeated-measures analysis of variance was conducted for each excerpt separately, with condition (see Table 2) as a within-subject factor and skill (musician, nonmusician) as a between-subject factor. For excerpts K281:3 and K284:3, repetition (a, b) was also a within-subject factor. The analyses revealed a significant effect of condition in all cases: for excerpt K281:3, $F(5, 175) = 40.04$, $\epsilon_{G.G.} = .79$, $p_{adj} < .001$; for excerpt K331:1, $F(5, 175) = 26.05$, $\epsilon_{G.G.} = .63$, $p_{adj} < .001$; and for excerpt K284:3, $F(5, 175) = 59.13$, $\epsilon_{G.G.} = .66$, $p_{adj} < .001$. There was also a significant interaction between condition and skill in each case, except for excerpt 331:1, where the Greenhouse-Geisser corrected p-value exceeded the 0.05 significance criterion: for excerpt K281:3, $F(5, 175) = 8.04$, $\epsilon_{G.G.} = .79$, $p_{adj} < .001$; for excerpt K331:1, $F(5, 175) = 2.48$, $\epsilon_{G.G.} = .63$, $p_{adj} < .06$; and for excerpt K284:3, $F(5, 175) = 4.96$, $\epsilon_{G.G.} = .66$, $p_{adj} < .002$.

All participants were reasonably consistent in their ratings of the two identical K281:3 groups (labeled K281:3a and K281:3b, respectively, to distinguish the two groups by presentation order). There was a small but significant tendency to rate the repeated group slightly lower (i.e., better) on the second listening [$F(1, 35) = 9.49$, $p < .004$]. It is hypothesized that this was due to familiarity with the stimuli—initially the piano and woodblock sound strange together.

For the excerpt K284:3, it is clear that the grace notes are played on the beat, and the ratings confirm this observation, with those corresponding to the on-beat interpretation (K284:3a) scoring considerably better than the alternative group (K284:3b) [$F(1, 35) = 25.41$, $p < .001$]. This is clearly seen in the unsmoothed condition U in Figure 3 (below). However, it is still interesting to note that simply by applying some smoothing to the awkward sounding beat track, it was transformed into a track that sounds as good as the other smoothed versions (D1, D3, and D5). In the rest of the analysis, the K284:3b group was removed.

A post hoc Fischer LSD test was used to compare pairs of group means in order to assess where significant differences occur (Table 3). Some patterns are clear for all pieces: the conditions D1-R and D1-N30 were


K281:3a/b

K331:1

K284:3a/b

FIG. 3. Average ratings of the 37 listeners for the six conditions for the three excerpts. The error bars show 95% confidence intervals.

TABLE 3. *p*-values of differences in means for all pairs of smoothing conditions (post hoc Fischer LSD test).

K281:3	D1-R	.07				
	U	.00	.00			
	D1	.00	.00	.10		
	D3	.00	.00	.42	.40	
	D5	.00	.02	.04	.00	.00
K331:1	D1-R	.00				
	U	.01	.00			
	D1	.00	.00	.01		
	D3	.00	.00	.13	.22	
	D5	.13	.00	.24	.00	.01
K284:3	D1-R	.00				
	U	.04	.00			
	D1	.00	.00	.07		
	D3	.00	.00	.33	.39	
	D5	.00	.00	.08	.95	.43
		D1-N30	D1-R	U	D1	D3

rated significantly worse than the unsmoothed and two of the smoothed conditions (D1 and D3). Although the D1 condition was rated better than the unsmoothed condition for each excerpt, the difference was only significant for K331:1 ($p = .01$); for the other excerpts, the p-values were .10 and .07, respectively. There was no significant difference between the D1 and D3 conditions, but the D5 condition was significantly worse than D1 and D3 for two of the three excerpts.

Experiment 2: Beat Marking

In the second experiment, participants were asked to mark the positions of beats in the musical excerpts, using a multimedia interface that provides various forms of audio and visual feedback. One aim of this experiment was to test the smoothing hypothesis in a

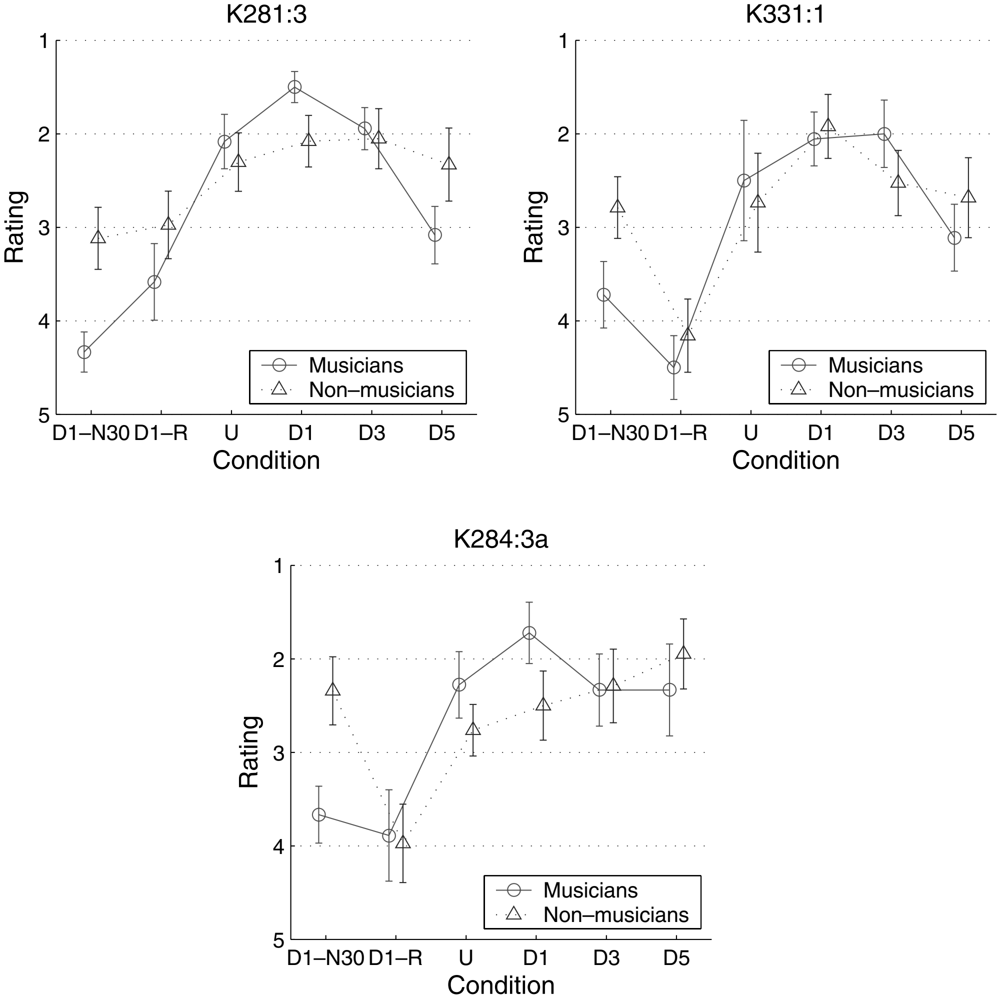


FIG. 4. Average ratings of the 18 musicians and 19 nonmusicians for the six conditions for the three excerpts. The ratings for K281:3a and K281:3b are combined, but the ratings for K284:3b are not used.

context where the participants had free choice regarding the times of beats and where they were not restricted by real-time constraints such as not knowing the subsequent context. Another motivation was to test the effects of the various types of feedback. Six experimental conditions were chosen, in which various aspects of the feedback were disabled, including conditions in which no audio feedback was given and in which no visual representation of the performance was given.

PARTICIPANTS

Six musically trained and computer literate participants took part in the experiment. They had an average age of 27 years and an average of 13 years of musical instruction. Because of the small number of participants, it was not possible to establish statistical significance.

STIMULI

The stimuli consisted of the same musical excerpts as used in Experiment 1 (K331:1, K281:3, and K284:3), but without the additional beat tracks.

EQUIPMENT

The software BeatRoot (Dixon, 2001b), an interactive beat tracking and visualization program, was modified for the purposes of this experiment. The program can display the input data as onset times, amplitude envelope, piano roll notation, spectrogram, or a combination of these (see Figure 5). The user places markers representing the times of beats onto the display, using the mouse to add, move, or delete markers. Audio feedback is given in the form of the original input data accompanied by a sampled metronome tick sounding at the selected beat times.

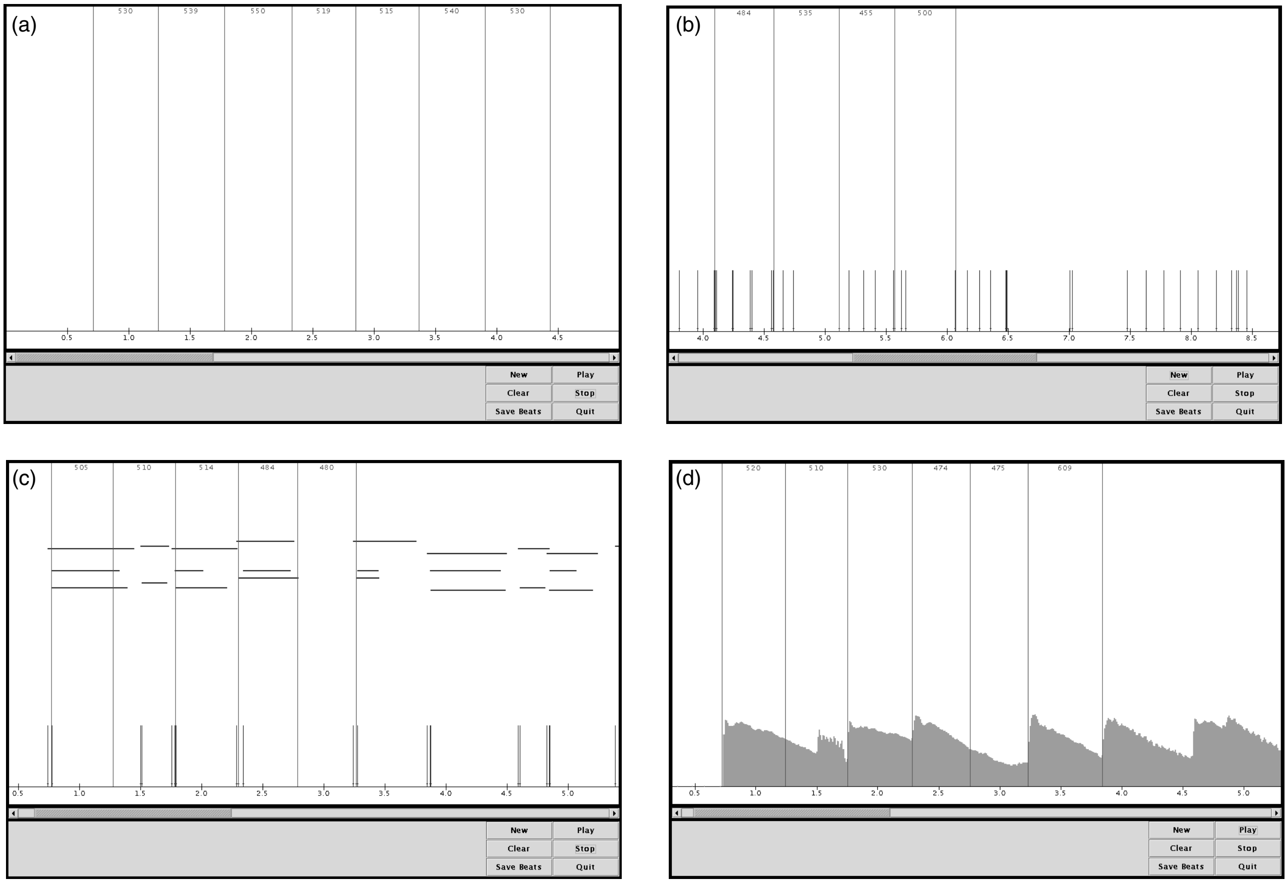


FIG. 5. Screen shots of the beat visualization system, showing: (a) Condition 1, visual feedback disabled: the beat times are shown as vertical lines, and the inter-beat intervals are marked between the lines at the top of the figure; (b) Condition 2, the note onset times as short vertical lines; (c) Conditions 3 and 5, MIDI input data in piano roll notation, with onset times marked underneath; (d) Condition 6, the acoustic waveform as a smoothed amplitude envelope. Condition 4 is like Condition 1, but with the IBIs removed.

PROCEDURE

The participants were shown how to use the software and were instructed to mark the times of perceived musical beats. The experiment consisted of six conditions related to the type of audio and visual feedback provided by the system to the user. For each condition and for each of the three musical excerpts, the participants used the computer to mark the times of beats and adjust the markers based on the feedback until they were satisfied with the results.

The experiment was performed in two sessions of approximately three hours each, with a break of at least a week between sessions. Each session tested three experimental conditions with each of the three excerpts. The excerpts for each condition were presented as a block, with the excerpts being presented in a random order. The otherwise unused excerpt K284:1 was provided as a sample piece to help the participants familiarize themselves with the particular requirements of each condition and ask questions if necessary. The presentation order was chosen to minimize any carryover (memory) effect for the pieces between conditions. Therefore the order of conditions (from 1 to 6, described below) was not varied. 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.

The six experimental conditions are shown in Table 4. 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, as in Figure 5a. 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, as shown in

TABLE 4. **Experimental conditions for Experiment 2.**

	Visual Feedback				Audio
Condition	Waveform	Piano Roll	Onsets	IBIs	Feedback
1	no	no	no	yes	yes
2	no	no	yes	yes	no
3	no	yes	yes	yes	yes
4	no	no	no	no	yes
5	no	yes	yes	yes	no
6	yes	no	no	yes	yes

Figure 5b. The participants 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 visualization system using MIDI data. The notes were shown in piano-roll notation as in Figure 5c, with the onset times marked underneath as in Condition 2, and audio feedback was enabled.

Condition 4 was identical with Condition 1, except that the inter-beat intervals were not displayed. This was designed to test whether participants 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 visualization system using audio data. Audio feedback was enabled, and a smoothed amplitude envelope, calculated as an RMS average of a 20 ms window with a hop size of 10 ms (50% overlap), was displayed as in Figure 5d.

BeatRoot allows the user to start and stop the playback at any point in time. The display initially shows the first 5 s of data, and users can then scroll the data as they please, where scrolling has no effect on playback.

RESULTS

From the marked beat times, the m-IBIs were calculated as well as the difference between the marked and performed beat times, assuming the default metrical level (ML) given in Table 1. We say that the participant marked the beat successfully if the marked beat times corresponded reasonably closely to the performed beat times, specifically if the greatest difference was less than half the average IBI (that is, no beat was skipped or inserted), and the average absolute difference was less than one quarter of the IBI. Table 5 shows the number of successfully marked excerpts at the default metrical level for each condition. The following results and graphs (unless otherwise indicated) use only the successfully marked data.

TABLE 5. **Number of participants who successfully marked each excerpt for each condition (at the default metrical level).**

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

TABLE 6. **Standard deviations of inter-beat intervals (in ms), averaged across participants, for excerpts marked successfully at the default metrical level. The rightmost column shows the standard deviations of p-IBIs for comparison.**

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

The low success rate is due to a number of factors. In some cases, participants marked the beat at a different metrical level than the default level. Since it is not possible to compare beat tracks at different metrical levels, it was necessary to leave out the results which did not correspond to the default level. The idea of specifying the desired metrical level had been considered and rejected, as it would have contradicted one goal of the experiment, which was to test what beat the participants perceived. Another factor was that two of the subjects found the experimental task very difficult and were only able to successfully mark respectively four and five of the 18 excerpt-condition pairs.

Figure 6 shows the effect of condition on the inter-beat intervals for each of the three excerpts, shown for three different participants. In each of these cases, the beat was successfully labeled. The notable features of these graphs are that the two audio-only conditions (1 and 4) have a much smoother sequence of beat times than the conditions in which visual feedback was given. This is also confirmed by the standard deviations of the inter-beat intervals (Table 6), which are lowest for Conditions 1 and 4.

Another observation from Table 6 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

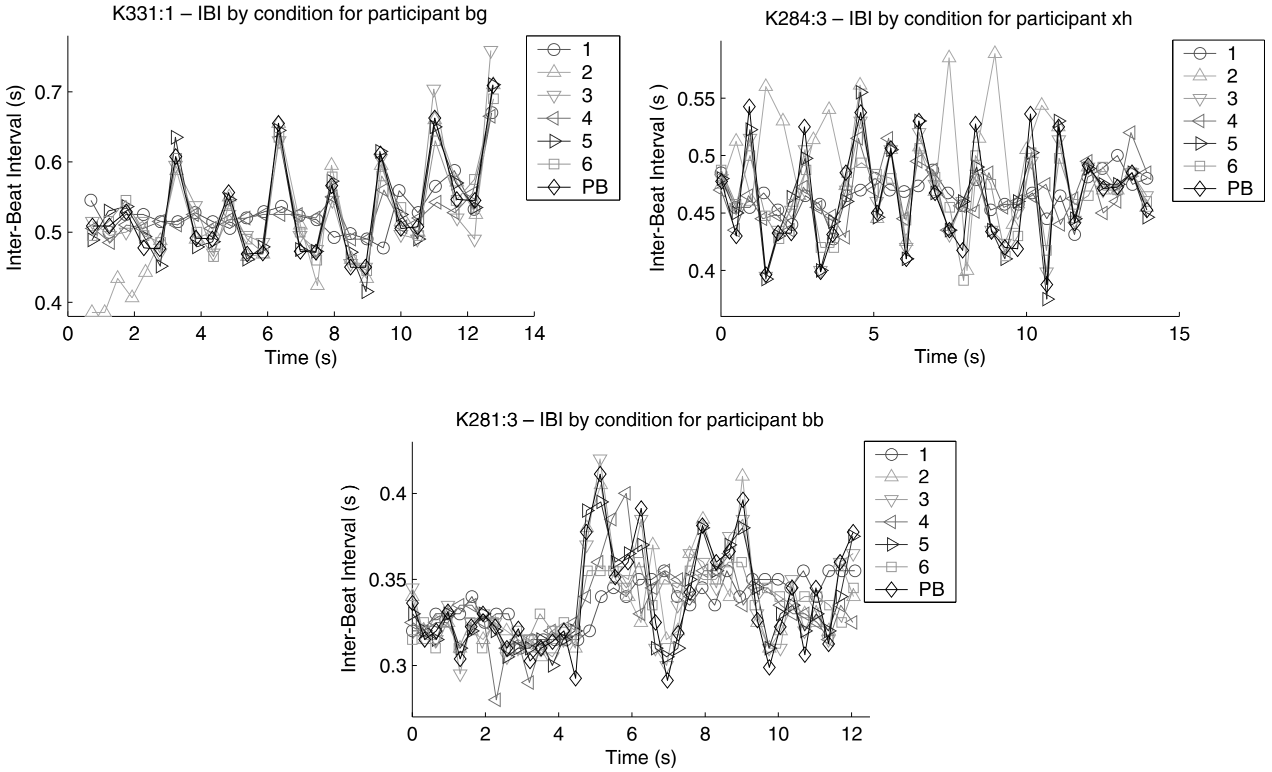


FIG. 6. Inter-beat intervals by condition for one participant for each excerpt. In this and following figures, the thick dark line (marked PB, performed beats) shows the inter-beat intervals of performed notes (p-IBI).

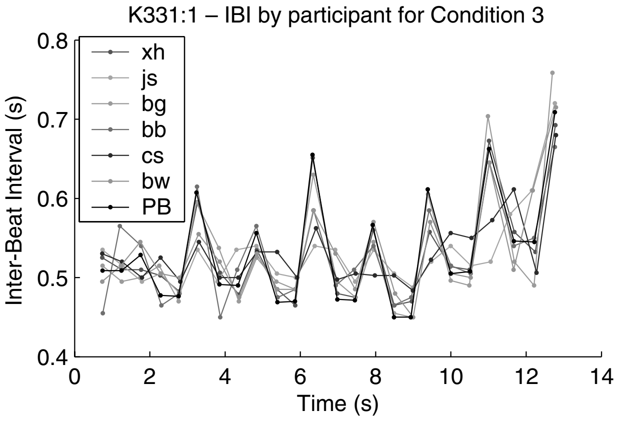


FIG. 7. Comparison by participant of inter-beat intervals for excerpt K331:1, Condition 3.

are used, by some participants at least, to adjust beats to make the beat sequence more regular than if attempted by listening alone. This suggests that participants consciously attempted to construct smooth beat sequences, as if they considered that a beat sequence *should* be smooth.

The next three figures show differences between participants within conditions. Figure 7 illustrates that for Condition 3, all participants follow the same basic shape of the tempo changes, but they exhibit differing amounts of smoothing of the beat relative to the performed onsets. In this case, the level of smoothing is likely to have been influenced by the extent of use of visual feedback.

Figure 8 shows the differences in onset times between the chosen beat times and the performed beat times for Conditions 1 and 3. The fact that some participants 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. Similar inter-participant differences in synchronization offset were found in tapping studies (Friberg & Sundberg, 1995) and in a study of the synchronization of bassists and drummers playing a jazz swing rhythm (Prögler, 1995). This asynchrony is much stronger in the conditions without visual feedback (Figure 8, left), where there is no visual cue to align the beat sequences with the performed music.

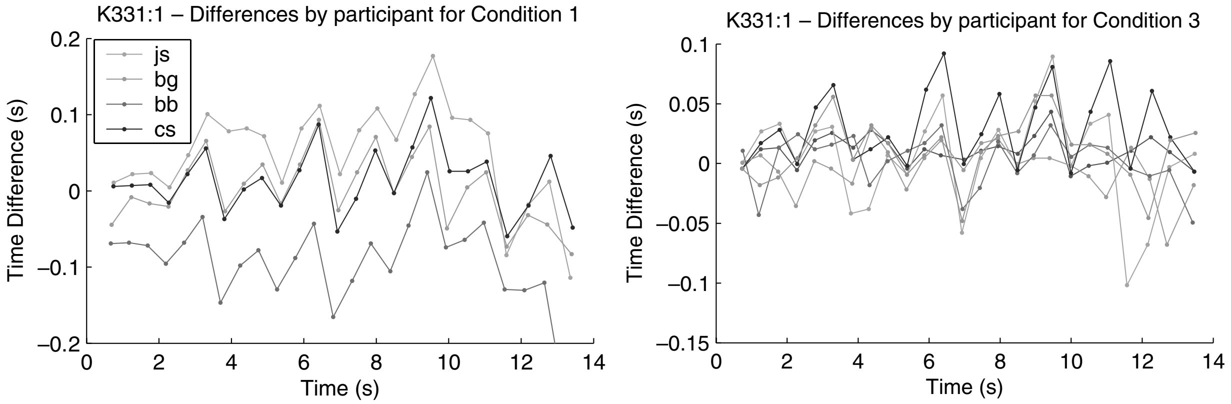


FIG. 8. Beat times relative to performed notes for Conditions 1 (left) and 3 (right). With no visual feedback (left), participants follow tempo changes, but with differences of sometimes 150 ms between the marked beats and corresponding performed notes, with some participants lagging and others leading the beat. With visual feedback (right), differences are mostly under 50 ms.

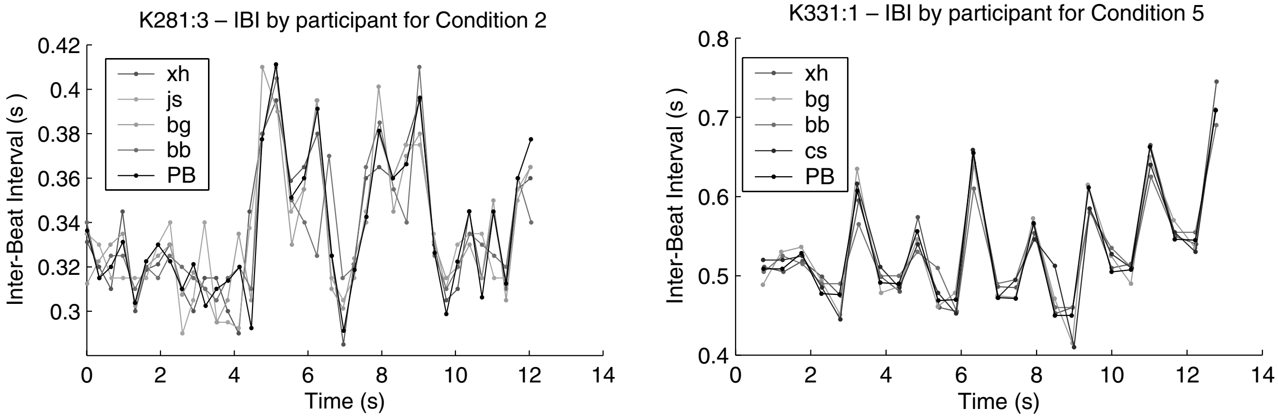


FIG. 9. IBI for Conditions 2 (left) and 5 (right), involving visual feedback but no audio feedback. The visual representations used were the onsets on a time line (left) and a standard piano roll notation (right).

It might also be the case that participants are more sensitive to tempo changes than to the synchronization of onset times. Research on auditory streaming (Bregman, 1990) predicts that the difficulty of judging the relative timing between two sequences increases with differences in the sequences' properties such as timbre, pitch, and spatial location. In other words, the listeners may have heard the click sequence as a separate stream from the piano music, and although they were able to perceive and reproduce the tempo changes quite accurately within each stream, they were unable to judge the alignment of the two streams with the same degree of accuracy.

Figure 9 shows successfully marked excerpts for Conditions 2 (left) and 5 (right). Even without hearing the music, these participants were able 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. Particularly in Condition 5 it can be seen that without audio feedback, participants do not perform nearly as much smoothing of the beat (compare with Figure 7).

Finally, in Figure 10, we compare the presentation of visual feedback in two different formats: as the amplitude envelope, that is, the smoothed audio waveform (Condition 3), and as piano roll notation (Condition 6; see Figure 5). Clearly the piano roll format provides more high-level information than the amplitude envelope, since it explicitly shows the onset times of all notes. For some participants this made a large difference in the way they performed beat tracking (e.g., Figure 10, left), whereas for others, it made very little difference (Figure 10, right). The effect of the visual feedback is thus modulated by inter-participant differences. The participant who showed little difference between the two visual representations has extensive experience with analysis and production of digital audio, which enabled him to align beats with onsets visually. The alternative explanation that he did not use

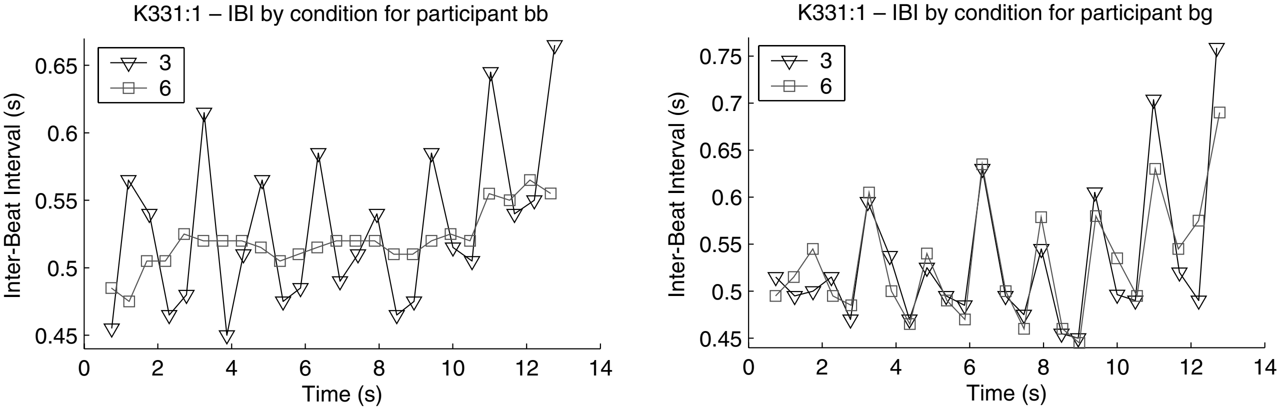


FIG. 10. The difference between two types of visual feedback (Condition 3, piano roll notation, and Condition 6, amplitude envelope) are shown for two participants (left and right). One participant (left) used the piano roll notation to align beats, but not the amplitude envelope, whereas the other participant (right) used both types of visual feedback to place the beats.

the visual feedback in either case is contradicted by comparison with the audio-only conditions (1 and 4) for this participant and piece (Figure 6, top), which are much smoother than the other conditions.

Experiment 3: Tapping

In this experiment, the participants were asked to tap the beat in time to a set of musical excerpts. The aim was to investigate the precise timing of taps and to test whether spontaneously produced "beats" coincide with listening preferences (Experiment 1) and beats produced in an off-line task, where corrections could be performed after hearing the beats and music together (Experiment 2).

PARTICIPANTS

The experiment was performed by 25 musically trained participants (average age 29 years). The participants had played an instrument for an average of 19 years; 19 participants studied their instrument at university level (average length of study 8.6 years); 14 participants play piano as their main instrument.

STIMULI

Four excerpts from professional performances of Mozart's piano sonatas were used in the experiment, summarized in Table 1. These are the same three excerpts used in Experiments 1 and 2, plus an additional excerpt, chosen as a warm-up piece, which had less tempo variation than the other excerpts. Each excerpt was repeated 10 times with random duration gaps (between 2 and 5 s) between the repetitions and was recorded on a compact disk (total duration 13 minutes 45 seconds for the 40 trials).

EQUIPMENT

Participants heard the stimuli through AKG K270 headphones and tapped with their finger or hand on the end of an audio cable. The use of the audio cable as tapping device was seen as preferable to a button or key, as it eliminated the delay between the contact time of the finger on the button and the electronic contact of the button itself. The stimuli and taps were recorded to disk on separate channels of a stereo audio file, through an SB128 sound card on a Linux PC. The voltage generated by the finger contact was sufficient to determine the contact time unambiguously with a simple thresholding algorithm. The participants also received audio feedback of their taps in the form of a buzz sound while the finger was in contact with the cable.

PROCEDURE

The participants were instructed to tap in time with the beat of the music, as precisely as possible, and were allowed to practice tapping to one or two excerpts, in order to familiarize themselves with the equipment and clarify any ambiguities in instructions. The tapping was then performed, and results were processed using software developed for this experiment. The tap times were automatically extracted with reference to the starting time of the musical excerpts, using a simple thresholding function.

In order to match the tap times to the corresponding musical beats, the performed beat times were extracted from the Bösendorfer piano performance data, as described in Experiment 1. A matching algorithm was developed which matched each tap to the nearest played beat time, deleting taps that were more than 40% of the average p-IBI from the beat time or that matched to a beat which already had a nearer tap matched to it.

The metrical level was then calculated by a process of elimination: metrical levels that were contradicted by at least three taps were deleted, which always left a single metrical level and phase if the tapping was performed consistently for the trial. The initial synchronization time was defined to be the first of three successive beats which matched the calculated metrical level and phase. Taps occurring before the initial synchronization were deleted. If no such three beats existed, we say that the tapper failed to synchronize with the music.

RESULTS

Table 7 shows for each excerpt the total number of repetitions that were tapped by the participants at each metrical level and phase. The only surprising results were that two participants tapped on the second and fourth quarter note beats of the bar (level 2, out of phase) for several repetitions of K281:3 and K284:3. The three failed tapping attempts relate to participants tapping inconsistently; that is, they changed phase during the excerpt. For each excerpt, the default metrical level (given in Table 1) corresponded to the tapping rates of the majority of participants.

Table 8 shows the average beat number of the first beat for which the tapping was synchronized with the music. For each excerpt, tappers were able to synchronize on average by the third or fourth beat of the excerpt, despite differences in tempo and complexity. This is similar to other published results (e.g., Snyder & Krumhansl, 2001; Toiviainen & Snyder, 2003).

In order to investigate the precise timing of taps, the t-IBIs of the mean tap times were calculated, and these are shown in Figure 11, plotted against time, with the p-IBIs shown for comparison. (In this and subsequent results, only the successfully matched taps are taken into account.) Two main factors are visible from these graphs: that the t-IBIs describe a smoother curve than the p-IBIs of the played notes, and the following of tempo changes occurs after a small time lag. These effects are examined in more detail below.

TABLE 7. Number of excerpts tapped at each metrical level and phase (in/out), where the metrical levels are expressed as multiples of the default metrical level (ML) given in Table 1.

	Metrical level (phase)					
Excerpt	1	2 (in)	2 (out)	3 (in)	3 (out)	Fail
K284:1	250	0	0	0	0	0
K331:1	164	0	0	86	0	0
K281:3	220	16	11	0	0	3
K284:3	153	89	8	0	0	0

In order to test the smoothing hypothesis more rigorously, we calculated the distance of the tap times from the performed beat times and from smoothed versions of the performed beat times. The distance was measured by the root mean squared (RMS) time difference of the corresponding taps and beats. This was calculated separately for each trial, and the results were subsequently averaged. (The use of average tap times would have introduced artifacts due to the artificial smoothing produced by averaging.) Four conditions are shown in Table 9: the unsmoothed beat times (U); two sets of retrospectively smoothed beats (D1 and D3; see Table 2), created by averaging each p-IBI with one or three p-IBI(s) on each side of it; and a final set of predictively smoothed beats (S1) created using only the current and past beat times, according to the following equation, where $x[n]$ is the unsmoothed p-IBI sequence, and $y[n]$ is the smoothed sequence:

$$y[n] = \frac{x[n] + y[n-1]}{2}$$

Table 9 shows the average RMS distance between the smoothed beat times and the taps. For each excerpt, at least one of the smoothed tempo curves gives beats closer to the tap times than the original tempo curve. For excerpt K331:1, only the D1 smoothing produces beat times closer to the taps. The reason for this can be understood from Figure 11 (top right): the tempo curve is highly irregular due to relatively long pauses, which are used to emphasize the phrase structure, and if these pauses are spread across the preceding or following beats, the result contradicts musical expectations.

On analyzing these results, it was found that part of the reason that smoothed tempo curves model the tapped beats better is that the smoothing function creates a time lag similar to the response time lag found in the tapping. To remove this effect, we computed a second set of differences using p-IBIs and t-IBIs instead of onset times and tap times. The results, shown in Table 10, confirm that even when synchronization is factored out, the tap sequences are closer to the smoothed tempo curves than to the performance data.

TABLE 8. Average synchronization time (i.e., the number of beats until the tapper synchronized with the music).

Excerpt	Synchronization time (in beats)
K284:1	3.29
K331:1	3.46
K281:3	3.88
K284:3	3.82

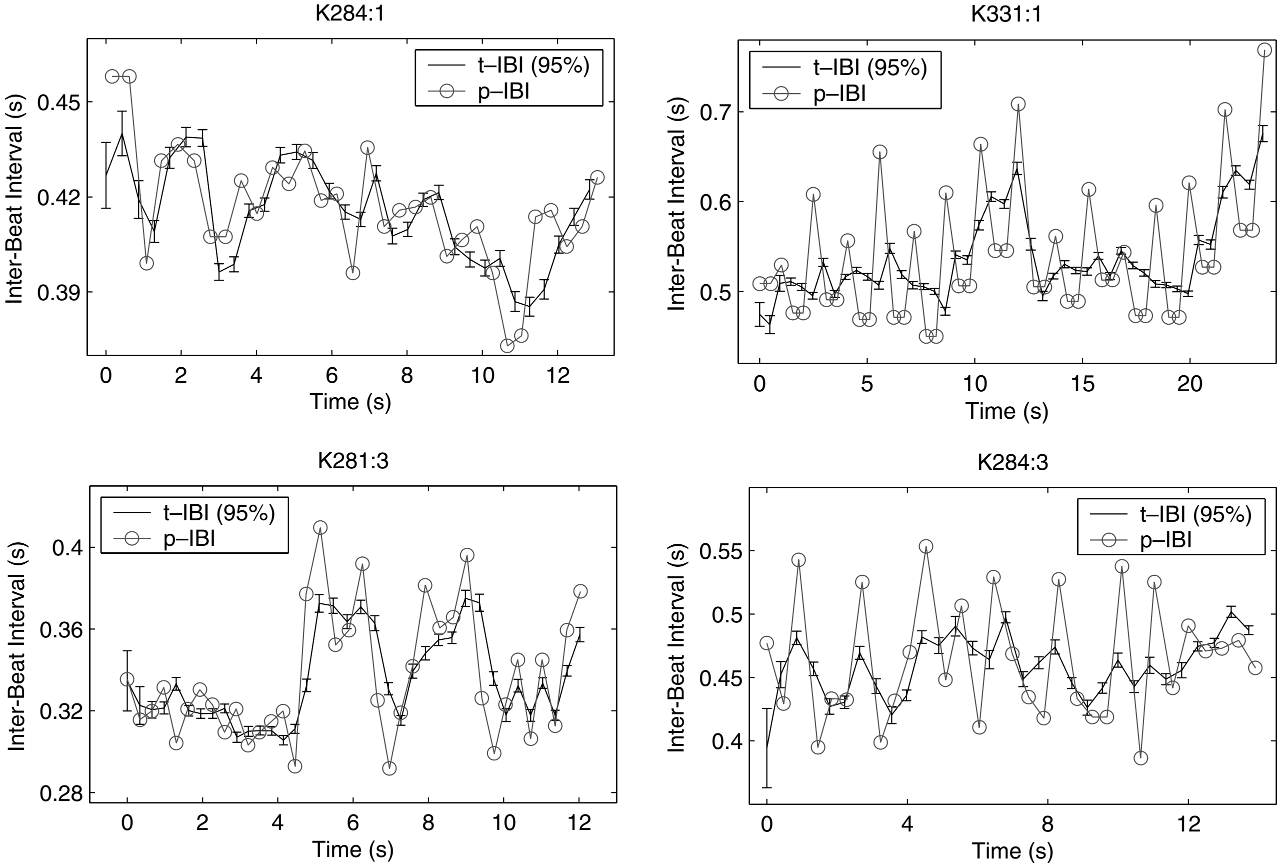


FIG. 11. Dark line: t-IBIs calculated from tap times averaged over all participants and trials. Error bars show double standard error (approx. 95% confidence). Light line: p-IBIs calculated from performed note times.

TABLE 9. Average RMS difference between taps and smoothed beats (in ms) for various smoothing conditions.

Smoothing	Excerpt				
Condition	K284:1	K331:1	K281:3	K284:3	Average
U	33	65	41	53	51
D1	32	56	39	49	46
D3	33	63	40	49	49
S1	38	89	33	44	59
Average	34	69	39	49	51

TABLE 10. Average RMS difference between t-IBIs and smoothed p-IBIs (in ms) for various smoothing conditions.

Smoothing	Excerpt				
Condition	K284:1	K331:1	K281:3	K284:3	Average
U	26	73	31	54	51
D1	24	41	26	41	34
D3	24	42	27	41	35
S1	23	45	24	38	35
Average	24	52	27	44	39

The data were also tested for two types of learning effect. The first test was whether the participants' tapping changed over the repetitions from an initially smooth sequence of taps to a sequence fitting closer to the unsmoothed data as the participants learned the tempo changes. It was found that the distances decreased with repetition, but the ranked order of distance by condition remained as shown in Tables 9 and 10. Thus there was no evidence that the smoothing was due to unfamiliarity with the tempo changes.

The second test for a learning effect was to test whether the time taken to react to tempo changes decreased with repetition. To find the time lag between tempo changes and changes in tapping rate, the p-IBI and t-IBI sequences were cross-correlated, and the lags corresponding to the highest correlation were found for each repetition. Table 11 shows for each lag how often this lag gave the best correlation. The results show that the lag of one tap is most common, that is, participants respond to a tempo change on the tap after it occurs. It was expected that with repetition, the lag would

TABLE 11. Analysis of time lags of responses to tempo changes, measured by correlation of t-IBIs and p-IBIs, shown as percentages of repetitions for which each lag had the highest correlation.

	Lag			
Excerpt	0	1	2	3
K284:1	10.0	64.4	9.2	5.2
K331:1	45.1	43.3	2.4	0.6
K281:3	31.4	57.7	7.3	2.7
K284:3	58.2	13.7	3.9	10.5

TABLE 12. Analysis of time lags of responses to tempo changes, showing the effects of learning on the lag 0 and lag 1 percentages.

	Rpt 1–3		Rpt 5–7		Rpt 8–10	
Excerpt	Lag 0	Lag 1	Lag 0	Lag 1	Lag 0	Lag 1
K284:1	9.3	61.3	10.7	66.7	8.0	66.7
K331:1	26.0	60.0	52.9	35.3	54.2	33.3
K281:3	13.4	65.7	33.3	56.1	47.0	50.0
K284:3	55.1	10.2	62.2	13.3	56.8	18.2

decrease, as the participants would remember and predict the tempo changes in their tapping. In fact, some participants expressed an awareness of the learning effect; one commented after the experiment, “It was like a chamber music rehearsal—you get it right after the third time.”

Table 12 shows a learning effect for excerpts K331:1 and K281:3, where with increasing repetitions, the 0 lag has the best correlation more frequently (Repp, 2002). For the other two excerpts, no learning trend is seen; K284:3 has a high correlation at lag 0 from the outset, and K284:1 has much smaller tempo deviations, to which, it appears, the participants are able to respond but not to learn. It may be the case that such learning requires conscious recognition of timing fluctuations, or a greater number of repetitions.

Discussion and Conclusion

The first experiment used a restricted choice scenario and showed for one excerpt a significant preference of listeners for beat sequences that are smoother than the onset times of the corresponding musical notes. For the other excerpts, the same trend was seen, but the results were just below significance. In no case was the unsmoothed track preferred over the beat track created with the D1 smoothing function. This effect was slightly stronger for musicians than nonmusicians. It is not certain whether the nonsignificance of the results is due to a lack of precision in perception (participants do not hear the difference due to smoothing) or a lack of preference (they perceive the difference but rate both smooth and unsmoothed beat tracks equally). The participants did show a significant dislike for reverse smoothing and for random deviations in the beat times, even though they were smaller than many of the differences in IBIs, and for smoothing functions with a too large window (D5). This indicates that the closeness of some of the ratings is more likely to be due to a lack of preference. It is clear that the current smoothing functions are much too simple, since they do not take musical context into account, so it is possible that artifacts of the chosen smoothing functions counteract some of their effectiveness in modeling tempo perception.

The results from the second experiment provide several inferences about the perception of beat and the beat marking software. Participants marked sequences of beat times that were smoother than the performed beat times, supporting the results of the first experiment, where listeners preferred artificially smoothed beat sequences to unsmoothed sequences. In almost all conditions, the beat sequences were smoother (as measured by standard deviation of the m-IBIs) than the performed beat times (measured by the standard deviation of the p-IBIs), but the amount of smoothing varied greatly with the feedback provided by the user interface.

When users had an explicit visual representation of the note onset times, the beat times were chosen mostly to correspond to these times, but when such information was not provided, the beat times were more evenly spaced, reflecting the local average tempo rather than the instantaneous tempo. This seems to correspond to other known effects, such as the McGurk effect (McGurk & MacDonald, 1976), where contradictory visual information modifies the perception of auditory stimuli. Even in the extreme conditions, where feedback was restricted to either visual or audio, but not both, participants were often able to find regularities corresponding to the notated musical beat, although this task obviously becomes more difficult as feedback is reduced.

An interesting question is the extent to which the beat marking software (BeatRoot) could be used in expressive performance research, that is, whether it provides sufficient precision to capture the features of interest in a performance. The results indicated large inter-participant differences and various biases introduced by certain forms of feedback. In order to use the system in practical situations, it is important to be clear about the specific aims of the analysis. For example, for performance research, one might be more concerned about what was played rather than how it was perceived. In such a case, the aim is to mark the note onsets rather than the perceived beat times. In this case the visual feedback is valuable, whereas it can be argued that it is a distraction when the perceived beat is sought. We consider that specific training would be useful to reduce inter-participant differences. BeatRoot has in fact been used in a number of studies of piano performance (Widmer, Dixon, Goebl, Pampalk, & Tobudic, 2003; Pampalk, Goebl, & Widmer, 2003; Goebl, Pampalk, & Widmer, 2004).

The results from the tapping experiment also support the hypothesis that the perceived beat is smoother than the played notes would indicate. This agrees with the findings of Repp (1999a) that tappers underestimate timing changes. The nature and extent of the smoothing that occurs is still unclear. We observed that applying rather arbitrarily chosen smoothing functions to the note onset times gave a closer match to the tapping times than the onset times themselves gave. But different functions perform better for different excerpts, and there is clearly a dependence on musical context which is not modeled by a simple smoothing function. It remains to be shown whether more accurate models can be found. We proposed a simple moving average function as a first approximation; this could be generalized by finding the FIR filter parameters that provide an optimal fit to the data, which might indicate some constraints of the smoothing process. However, we do not think that such fine tuning is warranted until more important aspects such as musical structure are modeled.

This experiment also showed a lag between tempo changes and changes in tapping rate, which reduced with repetition, that is, as the participants learned to anticipate the changes. For one of the excerpts, the lag remained at one beat; that is, participants seemed unable to learn the tempo changes, perhaps because in this case the tempo changes were small and therefore below the threshold for perception.

We now compare the results of the three experiments, which used the same excerpts (except for K284:1). In Experiment 1, the greatest preference, particularly of musically trained listeners, was for the click sequence corresponding to the D1 condition (smoothing with one beat each side of the current beat), which agrees with the tapping results in Experiment 3. Similarly, in the beat marking experiment, the sequences of beats chosen were smoother than the performed beat times, but this effect was greatly reduced for most participants when they could see the onset times on the display and align the beats with onsets visually. Further work is required to ascertain whether the off-line nature of the task influenced the results as compared to an on-line task such as tapping.

Although the experiments are not broad enough to suggest a complete model of beat perception, the evidence from all of the experiments supports the hypothesis that perceived beat sequences are smoother than the timing of the performed notes. This implies that timing fluctuations are not necessarily perceived as tempo changes. Beat perception shows a resistance to change and to random fluctuations; it is only when timing changes persist that one perceives an intended tempo change.

One possible explanation for the smoothing effect is that nominally on-beat notes may be perceived as anticipating or following the beat, rather than defining the beat. Another explanation is provided by categorical perception (Clarke, 1987). That is, the durations of p-IBIs are perceived in musical units, based on the current tempo percept, so that small deviations do not affect the categorization of the interval, and thus the perceptual system reduces the effect of deviations from strictly metrical time. Madison and Merker (2002) showed that if these deviations from regularity are too large, the sequence is no longer perceived as a sequence of beats. Another factor to consider is the role of higher metrical levels, which tend to be more stable, and thus override the irregularities in lower metrical levels. Further work is required to analyze these aspects of beat perception more precisely and to investigate further the relationship between tempo and timing (period and phase correction, Repp, 2002), as well as the specific effects of modality on the results from the three experiments.

The results from all three experiments support the smoothing hypothesis, that the perceived beat is a compromise between the two extremes of a perfectly regular, isochronous pulse and the irregular timing of the performed notes. The complexities of musical structure and performance tempo make it difficult to quantify this effect. The current results indicate that a moving average smoothing function with a small window size (around one beat) provides a reasonable model for the smoothing behavior.

Author Note

This research was supported by the project Y99-INF of the Austrian Federal Ministry of Education, Science and Culture in the form of a START Research Prize. The first author was also supported by the EU project FP6-507142 (SIMAC). The Austrian Research Institute for Artificial Intelligence acknowledges the support of the Austrian Federal Ministry for Education, Science and Culture and the Austrian Ministry for Transport, Innovation and Technology. We also thank Prof. Roland Batik for permission to use the performance data, the L. Bösendorfer Company, Vienna, especially Fritz Lachnit, for providing the data; the participants who took part in the experiments; and Renee Timmers, Bruno Repp, and the anonymous reviewers for comments on earlier versions of this article.

Address correspondence to: Dr. Simon Dixon, Austrian Research Institute for Artificial Intelligence, Freyung 6/6, Vienna 1010, Austria. E-MAIL *simon.dixon@ofai.at*

References

ASCHERSLEBEN, G., & PRINZ, W. (1995). Synchronizing actions with events: The role of sensory information. *Perception and Psychophysics, 57*, 305-317.

BATIK, R. (1990). *Wolfgang Amadeus Mozart: The complete piano sonatas*. (Gramola 98701-705).

BILMES, J. (1993). *Timing is of the essence: Perceptual and computational techniques for representing, learning and reproducing expressive timing in percussive rhythm*. Unpublished master's thesis, MIT, School of Architecture and Planning.

BREGMAN, A. (1990). *Auditory scene analysis: The perceptual organisation of sound*. Cambridge, MA: MIT Press.

CAMBOUROPOULOS, E., DIXON, S., GOEBL, W., & 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* (pp. 18-26). Jyväskylä, Finland: University of Jyväskylä.

CEMGIL, A., KAPPEN, B., DESAIN, P., & HONING, H. (2000). On tempo tracking: Tempogram representation and Kalman filtering. In *Proceedings of the 2000 international computer music conference* (pp. 352-355). San Francisco, CA: International Computer Music Association.

CLARKE, E. (1985). Structure and expression in rhythmic performance. In P. Howell, I. Cross, & R. West (Eds.), *Musical structure and cognition* (pp. 209-236). London: Academic Press.

CLARKE, E. (1987). Categorical rhythm perception: An ecological perspective. In *Action and perception in rhythm and music* (pp. 19-34). Sweden: Royal Swedish Academy of Music.

CLARKE, E. (1999). Rhythm and timing in music. In D. Deutsch (Ed.), *The psychology of music* (pp. 473-500). San Diego, CA: Academic Press.

COLLYER, C., HOROWITZ, S. B., & HOOPER, S. (1997). A motor timing experiment implemented using a musical instrument digital interface (MIDI) approach. *Behavior Research Methods, Instruments and Computers, 29*, 346-352.

COOPER, G., & MEYER, L. (1960). *The rhythmic structure of music*. Chicago: University of Chicago Press.

DESAIN, P. (1992). A (de)composable theory of rhythm perception. *Music Perception, 9*, 439-454.

DESAIN, P., & HONING, H. (1992). Tempo curves considered harmful: A critical review of the representation of timing in computer music. In *Music, mind and machine: Studies in computer music, music cognition and artificial intelligence*. Amsterdam: Thesis Publishers.

DESAIN, P., & HONING, H. (1999). Computational models of beat induction: The rule-based approach. *Journal of New Music Research, 28*, 29-42.

DIXON, S. (2001a). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research, 30*, 39-58.

DIXON, S. (2001b). An interactive beat tracking and visualisation system. In *Proceedings of the international computer music conference* (pp. 215-218). San Francisco, CA: International Computer Music Association.

DIXON, S., & CAMBOUROPOULOS, E. (2000). Beat tracking with musical knowledge. In *ECAI 2000: Proceedings of the 14th European conference on artificial intelligence* (pp. 626-630). Amsterdam: IOS Press.

DIXON, S., & GOEBL, W. (2002). Pinpointing the beat: Tapping to expressive performances. In *7th international conference on music perception and cognition* (ICMPC7) (pp. 617-620). Adelaide, Australia: Causal Productions.

DIXON, S., GOEBL, W., & CAMBOUROPOULOS, E. (2001). *Beat extraction from expressive musical performances* (Tech. Rep. No. 2001-22). Vienna, Austria: Austrian Research Institute for Artificial Intelligence. Presented at 2001 Meeting of the Society for Music Perception and Cognition (SMPC2001), Kingston, Ontario.

DRAKE, C., PENEL, A., & BIGAND, E. (2000). Tapping in time with mechanically and expressively performed music. *Music Perception, 18*, 1-23.

FRIBERG, A., & SUNDBERG, J. (1995). Time discrimination in a monotonic, isochronous sequence. *Journal of the Acoustical Society of America, 98*, 2524-2531.

GOEBL, W., PAMPALK, E., & WIDMER, G. (2004). Exploring expressive performance trajectories: Six famous pianists play six Chopin pieces. In S. Lipscomp, R. Ashley, R. Gjerdingen, & P. Webster (Eds.), *Proceedings of the 8th international conference on music perception and cognition* (ICMPC8) (pp. 505-509). Adelaide, Australia: Causal Productions.

Goto, M., & Muraoka, Y. (1995). A real-time beat tracking system for audio signals. In *Proceedings of the international computer music conference* (pp. 171-174). San Francisco, CA: International Computer Music Association.

Goto, M., & Muraoka, Y. (1999). Real-time beat tracking for drumless audio signals. *Speech Communication, 27*, 311-335.

Gouyon, F., & Dixon, S. (2005). A review of automatic rhythm description systems. *Computer Music Journal, 29*(1), 34-54.

Honing, H. (2001). From time to time: The representation of timing and tempo. *Computer Music Journal, 25*(3), 50-61.

Large, E., & Jones, M. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review, 106*, 119-159.

Large, E., & Kolen, J. (1994). Resonance and the perception of musical meter. *Connection Science, 6*, 177-208.

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Longuet-Higgins, H., & Lee, C. (1982). The perception of musical rhythms. *Perception, 11*, 115-128.

Madison, G. (2001). *Functional modelling of the human timing mechanism*. Uppsala, Sweden: Uppsala University.

Madison, G., & Merker, B. (2002). On the limits of anisochrony in pulse attribution. *Psychological Research, 66*, 201-207.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

Pampalk, E., Goebl, W., & Widmer, G. (2003). Visualizing changes in the inherent structure of data for exploratory feature selection. In *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 157-166). Washington, DC: ACM.

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception, 11*, 409-464.

Povel, D., & Essens, P. (1985). Perception of temporal patterns. *Music Perception, 2*, 411-440.

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

Repp, B. (1999a). Control of expressive and metronomic timing in pianists. *Journal of Motor Behaviour, 31*, 145-164.

Repp, B. (1999b). Detecting deviations from metronomic timing in music: Effects of perceptual structure on the mental timekeeper. *Perception and Psychophysics, 61*, 529-548.

Repp, B. (2000). Compensation for subliminal timing perturbations in perceptual-motor synchronization. *Psychological Research, 63*, 106-128.

Repp, B. (2002). The embodiment of musical structure: Effects of musical context on sensorimotor synchronization with complex timing patterns. In W. Prinz & B. Hommel (Eds.), *Attention and performance XIX: Common mechanisms in perception and action* (pp. 245-265). Oxford, UK: Oxford University Press.

Rosenthal, D. (1992). Emulation of human rhythm perception. *Computer Music Journal, 16*(1), 64-76.

Snyder, J., & Krumhansl, C. (2001). Tapping to ragtime: Cues to pulse-finding. *Music Perception, 18*, 455-489.

Thaut, M., Tian, B., & Sadjadi, M. A. (1998). Rhythmic finger tapping to cosine-wave modulated metronome sequences: Evidence of subliminal entrainment. *Human Movement Science, 17*, 839-863.

Timmers, R., Ashley, R., Desain, P., Honing, H., & Windsor, L. (2002). Timing of ornaments in the theme of Beethoven's Paisiello Variations: Empirical data and a model. *Music Perception, 20*, 3-33.

Toiviainen, P., & Snyder, J. (2003). Tapping to Bach: Resonance-based modeling of pulse. *Music Perception, 21*, 43-80.

Widmer, G., Dixon, S., Goebl, W., Pampalk, E., & Tobudic, A. (2003). In search of the Horowitz factor. *AI Magazine, 24*(3), 111-130.

Wohlschläger, A., & Koch, R. (2000). Synchronisation error: An error in time perception. In P. Desain & W. L. Windsor (Eds.), *Rhythm perception and production* (pp. 115-127). Lisse: Swets and Zeitlinger.

Yeston, M. (1976). *The stratification of musical rhythm*. New Haven, CT: Yale University Press.