Timing-improved Guitar Loop Pedal based on Beat Tracking

Daniel Rudrich

Institute of Electronic Music and Acoustics
University of Music and Performing Arts Graz

Supervisor: **Univ.Prof. Dipl.-Ing. Dr.techn. Alois Sontacchi**

This thesis is submitted for the degree of
Master of Science

January 2017

I dedicate this thesis to my dad, who would have loved this work, combining his biggest hobby (playing the guitar) with his expertise (engineering).

Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Daniel Rudrich
January 2017

Abstract

Loop pedals become more and more popular due to their growing features and capabilities - not only in live performances, but also as a rehearsal tool. These pedals are effect units which record a played phrase and play it back unchanged and repetitiously. However, when the start and stop cues of the recording are not entered with right timing, an audible gap may occur at every iteration. This thesis proposes an algorithm, which analysis the recorded phrase and aligns start and stop positions to beats found by the analysis. The audibility of said gaps were investigated within two listening tests. Subjects with musical background performed significantly better in detecting rhythm perturbances. The average threshold of musicians found was 5.4 % of the inter onset interval of tone bursts, outperforming the non-musicians (9.0 %) remarkably. The found thresholds were used to evaluate the algorithm's performance.

Kurzfassung

Gitarren-Looper werden durch ihre wachsenden Ausstattungen und Möglichkeiten immer beliebter und das nicht nur bei Live-Performances, sondern auch zu Übungszwecken. Diese Effektgeräte zeichnen eine eingespielte Phrase auf und spielen sie anschließend wiederholend ab. Werden jedoch Start- und Stopzeitpunkte der Aufnahme nicht korrekt eingegeben, kann bei jedem Durchlauf ein hörbarer Sprung entstehen. Diese Arbeit stellt einen Algorithmus vor, welcher die eingespielte Phrase analysiert und die Start- und Stopzeitpunkte an die Stellen der gefunden Taktschläge verschiebt. Die Hörbarkeit besagter Sprünge wurde mittels zweier Hörversuche untersucht. ProbandInnen mit musikalischer Ausbildung erzielten bei der Detektion von rhythmischen Störungen signifikant bessere Ergebnisse. Die durchschnittliche Hörbarkeitsschwelle der MusikerInnen lag bei 5.4 % der Inter-Onset-Intervalle von Tone Bursts und damit bemerkenswert niedriger als die der Nicht-MusikerInnen (9.0 %). Diese Schwellen wurden weitergehend genutzt, um den Algorithmus zu evaluieren.

Contents

Chapter 1

Introduction

> Repetition is important.
> Repetition is important in music.
> Repetition is important in music in several aspects.
> Repetition is important.

1.1 Repetition in Music

Every tonal sound, whose pitch can be perceived by humans, consists of between 20 and 20 000 repetitions per second. This embodies the lowest level of repetition in music. But also if examined on a bigger scale, music in general is full of repetitions: A repetition of a single note can guide attention to it. Like in Ludwig van Beethoven's 5th symphony, the famous motif (which gets repeated hundreds of times in the entire piece) would sound less interesting and dramatic, if the first note would not be repeated two times before dropping down a major third. A more abstract form is the repetition of short melody phrases as in the children's song *Frère Jacques*, where every bar gets repeated before continuing to the next part of the melody. Also the repetition of whole parts is very common in both classical and pop music (verse, chorus). [19]

But there is also music which not only uses repetitions but is based on them, e.g. canons. A very famous example is Johann Pachelbel's *Canon in D*, in which several voices play the same part over and over. Fig. 1.1 shows the tonal based self similarity matrix of said canon. The depicted data shows the correlation of tonal components between different analysis blocks. It can be read by choosing one time block on the abscissa and imaging a vertical line. The data on that line represents the correlation of the tonal components of that specific block with all the other blocks of the audio file.

In other words: darkly represented values point out a tonal repetition of one specific block on the x-axis. This is of course given when comparing one block with itself (black diagonal line in the middle). Any parallel line depicts a repetition with its temporal offset being proportional to the offset of the two lines.

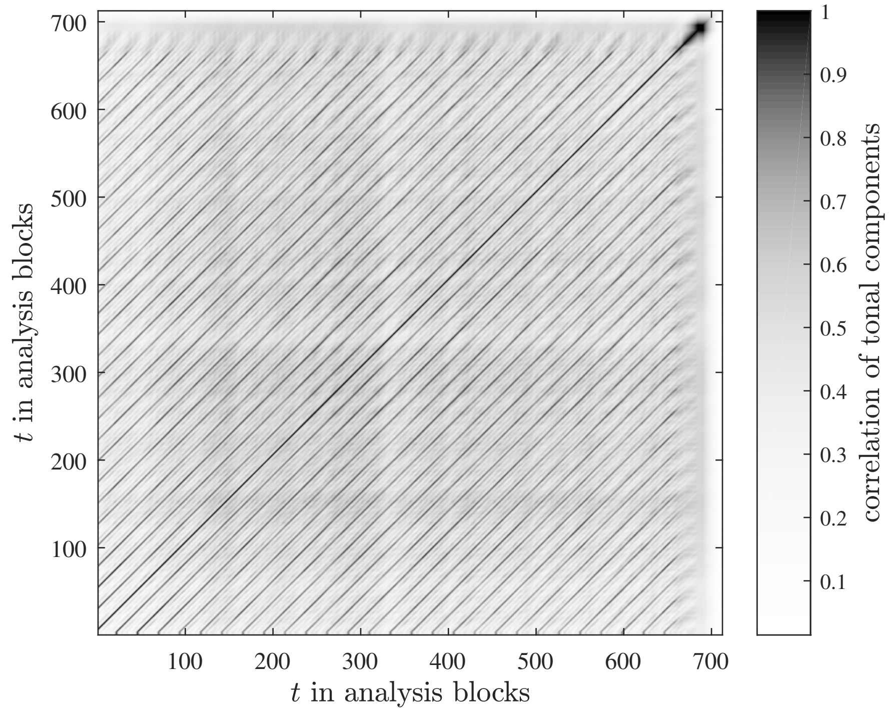


Figure 1.1 Self-smilarity matrix of Johann Pachelbel's *Canon in D*. Created with [23].

It is no surprise that repetition appears so frequently throughout a variety of music genres as several studies [24, 21, 18] showed that music is more liked when it is repeated. According to the authors Peretz et al. and Margulis the *mere exposure effect* is one of the underlying reasons. This effect states that "repeated exposure of the individual to a stimulus object enhances his [or her] attitude toward it." [33, p. 1].

Margulis [18] asked subjects to rate short excerpts of contemporary music by Luciano Berio and Elliott Carter in terms of enjoyability, interest and artistry. The subjects listened to both unaltered and modified versions. The latter were altered such they consist of repetitions. Stimuli with segments that were repeated either immediately or with a delay after they appeared first were rated significantly higher in terms of enjoyment and artistry.

In alignment, Rickard [27] used an algorithm[1], to write a completely pattern-free music. Neither rhythm nor melody consists of repetitions. Rickard himself describes it as “the worlds ugliest music”.

As the first five lines of this chapter show, repetition can be used as an emphasis. But not only an emphasis of the redundant, recurrent part but furthermore the differential information, which is added or omitted between two iterations. But even if no further information was added physically, the listener can experience something new as Deleuze writes in his work *Difference and Repetition*:

> The role of the imagination, or the mind which contemplates in its multiple and fragmented states, is to draw something new from repetition, do draw difference from it. [6, p. 97]

Apart from the developements and possibilities due to digital signal processing[2], this all together could be the reason, why live loop performances enjoy bigger popularity. Not only is it based on repetition, but also the audience can experience the development from e.g. a bass line or simple chords at the beginning to a more and more elaborated sound as further phrases get overlaid by the musician. Fig. 1.2 shows the self similarity matrix of a guitar loop played by the author. The distinct lines with an offset of about 18 blocks from the diagonal in the middle show very similar repetitions of the previously played phrases. The higher the offset the more diverse are the analyzed blocks as the lines get washed out towards the upper left or lower right corner. This happens as more and more phrases get overlaid to the initial phrase. The less marked lines between the distinct ones point out that the main phrase consists of two similar parts.

Loop pedals in general can not only be used for live performances but also for rehearsal of scales, melodies or the like. Repetitio mater studiorum est. Repetition is the mother of all learning. To directly see if a scale works in a musical context, many musicians use so called backing tracks if practicing alone. To create one’s own backing track a guitar looper can be used. It enables self accompanying as a harmonic phrase can be recorded and gets repeated over and over again.

However, especially when new to looping it is difficult to find the right timing as start and stop of the recording have to be controlled by foot. If both cues do not

[1] The algorithm was actually used in the development of the best sonar ping without any repetitions.

[2] With the analog predecessor of digital guitar loop pedals - the *Time Lag Accumulator*[1] - it was almost impossible to set the loop length while playing. The later use of digital memory instead of tape loops first made this possible leading to a substantially different sound and use of live looping. [13]

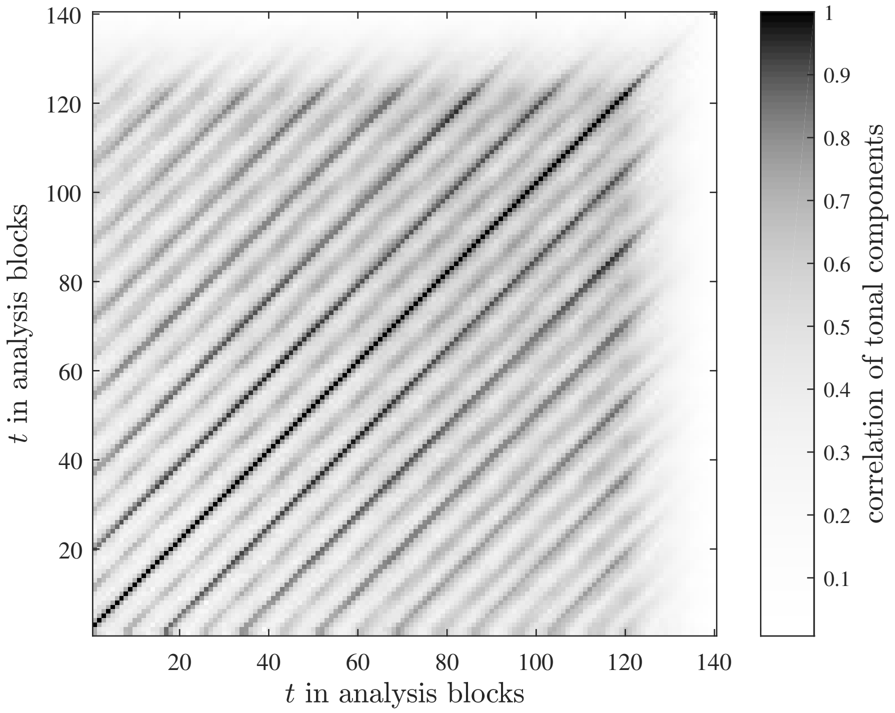


Figure 1.2 Self-smilarity matrix of a guitar loop performance by the author. Created with [23].

coincide on the same musical measure this timing mistake creates a gap, which also gets repeated over and over again.

The objective of this thesis is to develop an algorithm, which analyses the recorded phrase and supports the musician by aligning the start and stop cues such that the gap is reduced and not perceivable any more. To validate the algorithms performance, the perception of these *rhythm perturbances* is also part of this work.

1.2 Outline

Chapter 2 gives a brief introduction to loop pedals with an overview of the looping process. Also this thesis' problem is stated in detail.

In Chapter 3 the algorithm for solving the problem is described and its behavior is shown with different examples.

Chapter 4 deals with the audibility of rhythm perturbances. Two listening tests were conducted to obtain perception thresholds used for the validation of the algorithm's performance.

Chapter 2

Guitar Looper and Problem Description

2.1 About Guitar Loopers

A guitar looper (or a looper in general) is a device, with which a played phrase can be recorded and played back repetitiously. Also, more and more sound layers can be added to create an elaborated sound to which the musician can play/sing along. Therefore, the looper gets added between the signal source (e.g. guitar) and playback system (e.g. guitar amp).

The following section shows the individual states of a basic looper. More sophisticated devices consist of a higher functionality like different effects which can be added to the signal or the loop itself (e.g. octave effect for bass simulation, reverse or half-speed).

2.2 States of Looping

This section describes the phases of live looping with a guitar looper equipped with two foot pedals: REC and START/STOP. For devices with only one foot pedal, these commands are generally encoded with gestures like double-tap or hold for a specific time. Fig. 2.1 shows a state diagram of the looping process. The transition from one state to another is initiated by pressing the REC or START/STOP button and can vary between different loop devices, however, the states themselves are the same.

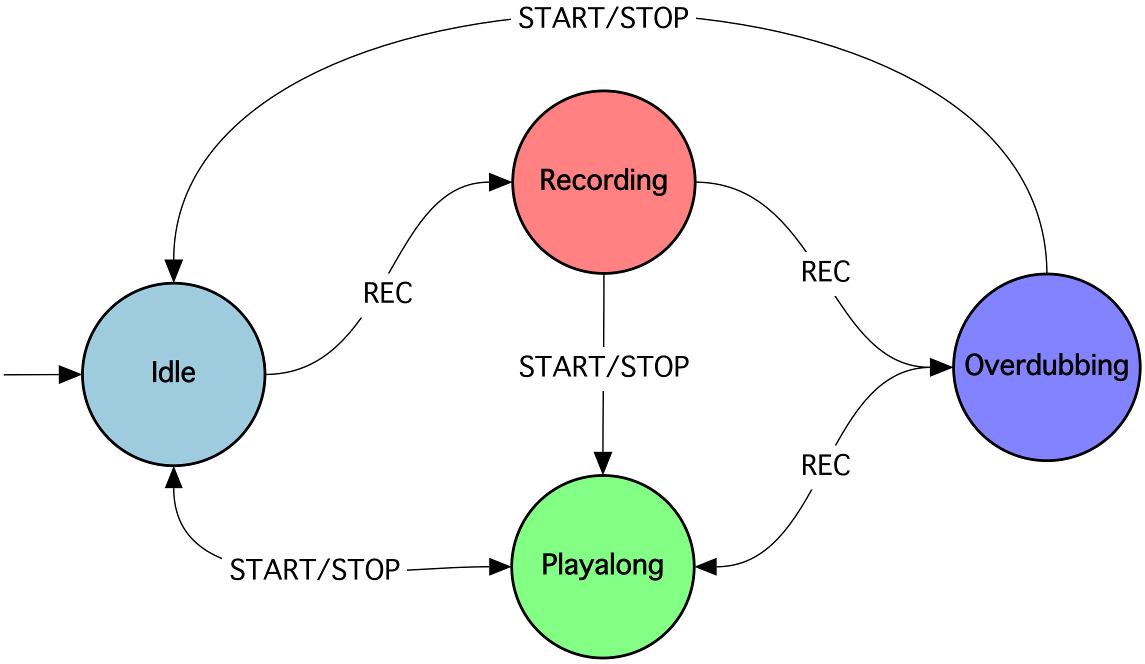


Figure 2.1 State diagram of the looping process.

Idle

In general, the *Idle*-state is the starting point. The playback is stopped but the memory can already be filled with a recorded phrase with or without additional layers. In that case, pressing the START/STOP button will start the playback. By pressing the REC button instead, the *Recording*-state will be activated.

Recording

Once reaching this state, the memory gets erased and the device starts recording the input signal. In other words, a new phrase gets recorded. The recording gets stopped by pressing one of the two buttons, switching to the corresponding state.

Overdubbing

When reaching the *Overdubbing*-state from the *Recording*-state, the device starts the playback of the recording in the memory. Otherwise (coming from *Playalong*-state), the playback continues. In both cases, the input signal gets recorded and added to the already existing recording in the memory. Also when reaching the end of the recording, the playback continues from the start (looping). This state can be used to add new sound layers leading to a more dense and elaborated sound.

Playalong

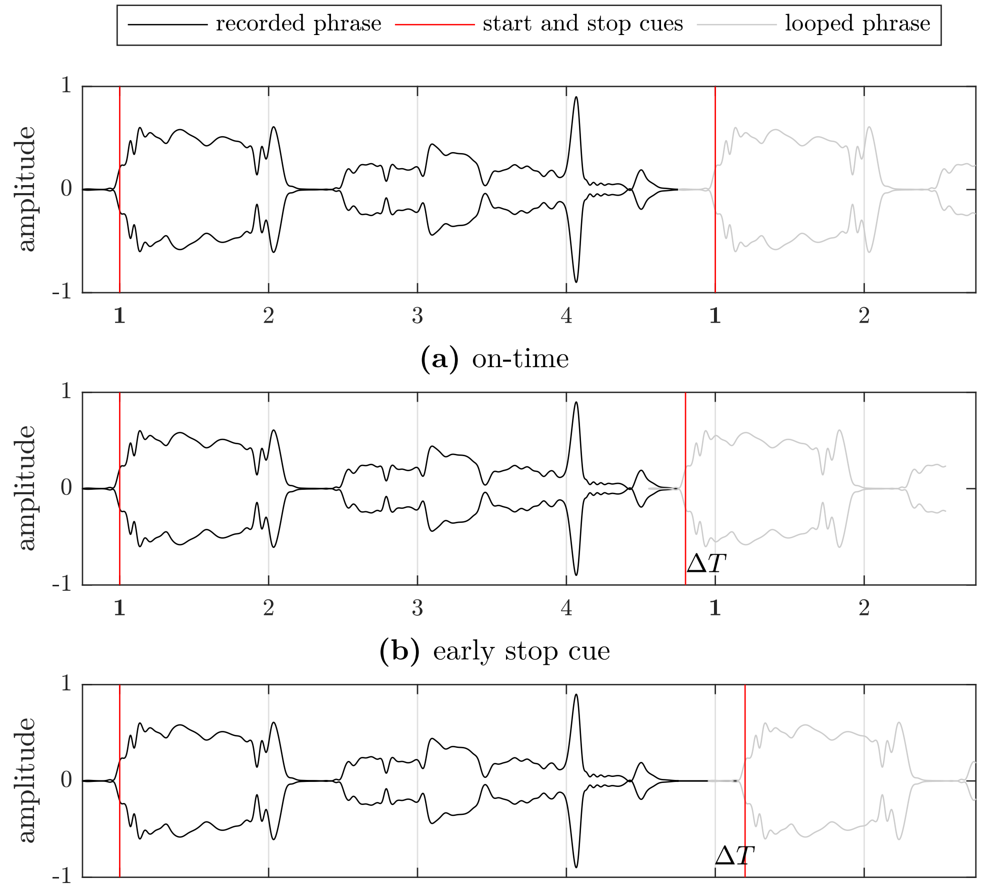
In this state, the device plays back the content of the memory either from start (from *Recording*- or *Idle*-state) or the previous playback position (*Overdubbing*). Once reaching the end, the playback continues from the start (looping). The input signal gets directly routed to the output without recording. This state is used for playing or singing along to the loop.

2.3 Problem Description

A looped phrase gets repeated seamlessly when start and stop cues (pressing the REC button) of the recording are consistent in terms of musical timing. This is visualized in Fig. 2.2a. The musical beats are generally not known by the system and are depicted only for visualization of the problem. Here, the start and stop cues are consistent as both occur on beat one.

When hitting the stop button too early also the looped phrase sets in too early, leading to a skipped part of length ΔT between the actual cue and the intended cue (Fig. 2.2b). As the phrase gets shortened, this kind of gap is called *negative* gap throughout this thesis. Alternatively, a *positive* gap occurs when the stop cue comes late (Fig. 2.2c), meaning an unwanted pause was added to the loop. Both cases also happen, if the timing of the start cue is off, or a combination of both. Nevertheless, when both cues have the same offset, the loop gets repeated without any pauses or skips.

This gap, whether positive or negative, can be audible and annoying if it gets repeated all the time. Some loop effect pedals solve this issue with a quantization of the start and stop cues to a given metronome click. However, the objective of this work is to develop a solution which does not require any tempo settings in advance.



(a) on-time

(b) early stop cue

(c) late stop cue

Figure 2.2 Examples of a one-bar-phrase (four-four time) getting looped with different stop-cue timings. The abscissa holds the musical beats. The black line represents the envelope of the recorded signal. The red lines indicate the start and stop cues. The grey line represents the repeated (looped) phrase setting in directly after the stop cue. Each figure depicts one of the three timing cases: (a) on time (b) early and (c) late. For the latter two a gap of ΔT occurs.

Chapter 3

Algorithm

This chapter depicts the approach to the problem described in section 2.3. Therefore, the basic approach is introduced as well as its realization with a detailed view of the individual processing blocks. The chapter also covers the real-time requirements and shows examples of the algorithm in use.

3.1 Tatum

> When I asked Barry Vercoe if this concept had a term, he felicitously replied "Not until now. Call it *temporal atom* or *tatom*." So, in honor of Art Tatum, whose tatum was faster than all others, I chose the word *tatum*.
>
> — Jeffrey Adam Bilmes, *Timing is of the Essence*

Bilmes coined the term *tatum* in his thesis *Timing is of the Essence* [2]. It describes the lowest metrical level in a musical piece like a fine underlying grid to which every note and rest can be assigned to. However, this does not generally mean that the tatum equals the smallest existing note or rest value, rather the greatest common divisor of all values, as can be seen in example 3.4.3 on page 22. In this thesis, the tatum is used as a context-free measure of tempo, as it does not require a musical reference value as the musical measure *beats per minute* (bpm) does. For the algorithm it is not possible - and then again, not necessary - to find the right musical reference point. However, the tatum embodies the time interval of the underlying grid.[1] Lower values describe a higher tempo and vice versa.

[1] With assigning a musical note value to it, it can be converted to bpm.

3.2 Basic Approach

Fundamentally, the goal is to align the loop's start and stop cues in a way that no pauses or skips occur when the loop gets repeated. In other words: the musical end position of the loop matches perfectly with the start position. The basic idea is to find the tatum of a recorded signal and align the start and stop cues to the corresponding beats. Therefore, the tatum has to be estimated. Basically, this is done by calculating an onset-detection-function (ODF), placing different tatum-grids over it (see Fig. 3.1) and picking the one which fits best. Due to a possible variability of the tempo within the recorded phrase, the tatum estimation can't be done for the whole sequence at once, but rather in blocks. Otherwise, the averaged tatum would be right, but there might be an offset especially at the beginning and the end of the sequence. This is delicate due to the beat alignment taking place at these locations. A more precise description of the implemented algorithm is shown below.

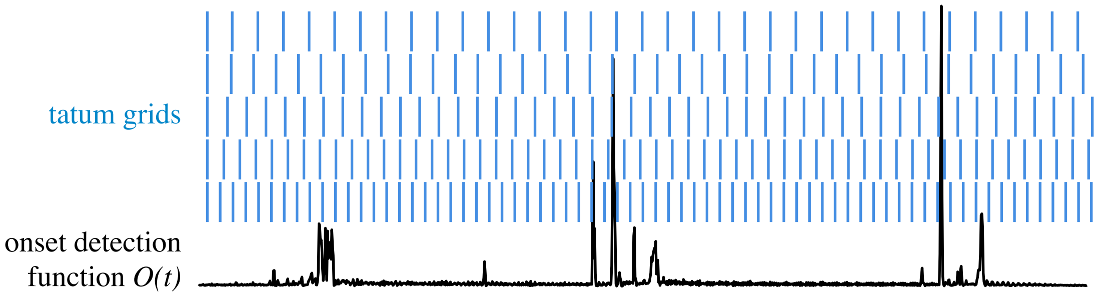


Figure 3.1 Qualitative illustration of the basic approach: Different tatum grids get placed over the onset detection function $O(t)$. The degree of matching provides information about the presence of different tatums in the signal.

3.3 Detailed View

3.3.1 Onset Detection: Spectral Flux Log Filtered

Böck et al. proposed an onset detection method which is called *Spectral flux log filtered* [3] and, as the name implies, is an advancement of the spectral flux algorithm by Masri [20]. The main difference is the processing in Mel frequency bands and usage of the logarithm for the magnitudes. This leads to an algorithm which involves the psychoacoustic characteristics of the human hearing. Instead of magnitude differences, it evaluates the magnitude ratios in each frequency band, as the human perception is ratio based [8]. This also compensates the temporal variability (dynamic) of the audio signal, as parts with high amplitudes do not get emphasized in comparison to parts with lower amplitudes.

Additionally, *Adaptive Whitening* suggested by Stowell and Plumbley [29] is used for compensating variability in both time and frequency domain. It normalizes the magnitude of each STFT bin (here: of each frequency band) in respect to a preceding maximum value and, therefore, compensates spectral differences as higher frequencies often hold lower energy values (spectral roll-off). Without the spectral whitening, lower frequencies would drown out the higher ones, which also posses important onset information, as the surprisingly good results of the very simple HFC (high frequency content) onset detector demonstrate [20]. The effects of adaptive whitening are shown in Fig. 3.3b on page 13.

Short Time Fourier Transform (STFT)

Prior to the transformation into the frequency domain the signal $x(t)$ gets segmented into N overlapping frames with a length of $K = 2048$ samples and the hopsize $h = 480$ samples, resulting in an overlap of roughly 77 %. With $f_s = 48\,000\,\text{Hz}$ two subsequent frames are 10 ms apart (resulting in an ODF sampling frequency of $f_{s,O} = 100\,\text{Hz}$). Each frame gets windowed with a Hann window $w(t)$. The segmentation and windowing-process is depicted in Fig. 3.2. The windowed signals $x_n(t)$ can be calculated with

$$x_n(t) = x(t+nh)w(t), \quad n \in [0, N-1],\ t \in [0, K-1]. \tag{3.1}$$

Each frame then gets transformed into the frequency domain using the discrete Fourier transform:

$$X(n,k) = \sum_{t=0}^{K-1} x_n(t) \cdot e^{-\mathrm{i}\frac{2\pi kt}{K}}, \quad k \in [0, K-1] \tag{3.2}$$

with n denoting the frame and k the frequency bin.

Adaptive Whitening

The so called *peak spectral profile* $P(n,k)$ is used to whiten the signal prior to the actual onset detection. The whitening process can be expressed as

$$P(n,k) = \begin{cases} \max\{|X(n,k)|, r\} & \text{for } n = 0, \\ \max\{|X(n,k)|, r, \mu P(n-1,k)\} & \text{otherwise} \end{cases} \tag{3.3}$$

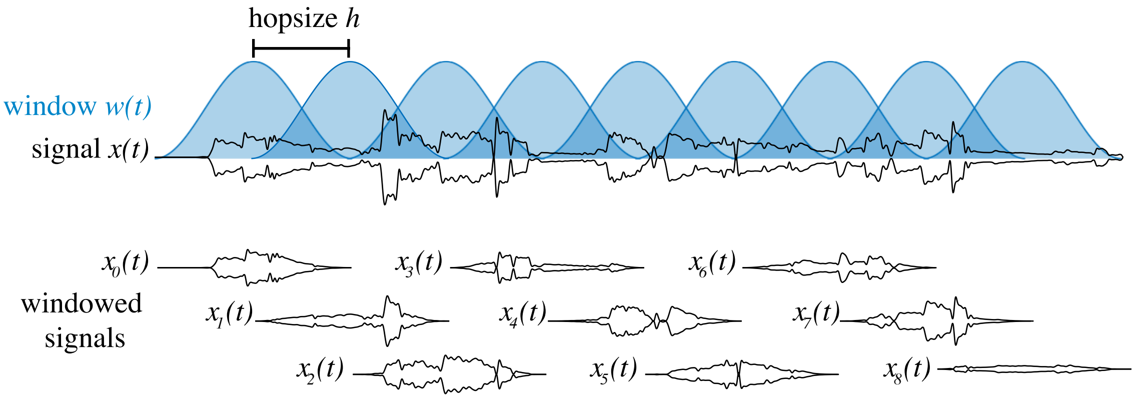


Figure 3.2 Qualitative illustration of the windowing process. The signal $x(t)$ gets segmented and weighted with the window function $w(t)$ resulting in the windowed signals $x_n(t)$.

$$X(n,k) \leftarrow \frac{X(n,k)}{P(n,k)} \tag{3.4}$$

where μ is the forgetting factor and r the floor parameter.[29] $\mu = 0.997$ and $r = 0.6$ have proved suitable for this application.

Logarithmic Filtered Spectrum

The main difference to the regular spectral flux onset detection is the analysis in sub-bands. Therefore, the magnitude spectrum $|X(n,k)|$ runs through a filter bank $F(k,b)$ with $B = 50$ overlapping filters with center frequencies from 94 Hz to 15 375 Hz. Each band has the same width on the Mel-scale. The filters are not normalized to constant energy which emphasizes higher frequencies (like a high frequency content HFC onset detector). The filtered spectrum $X_{filt}(n,b)$ is given by:

$$X_{filt}(n,b) = |X(n,k)| \cdot F(k,b) \tag{3.5}$$

with b denoting the sub-band number. The logarithmic filtered magnitude spectrum is obtained by applying the following equation to the filtered spectrum:

$$X_{filt}^{log}(n,b) = log(\lambda \cdot X_{filt}(n,b) + 1) \tag{3.6}$$

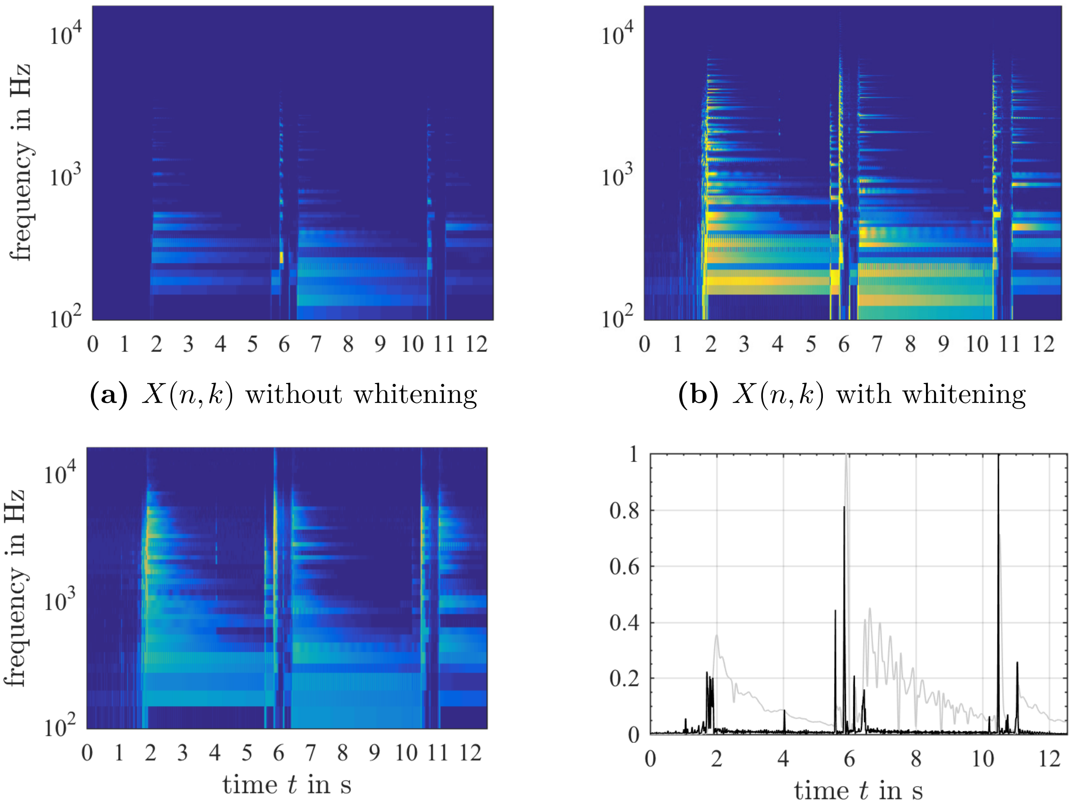
with the compression parameter λ. A value of $\lambda = 2$ yielded good results. The additive term $+1$ assures only positive values.

Difference

The final step to derive the onset detection function $O(n)$ is to calculate the spectral difference between the frame n and its previous frame $n-1$ with a subsequent summation of all spectral bins. The half-wave rectifier function $H(x) = \frac{x+|x|}{2}$ ensures that only onsets and not offsets get summed up. That all together can be expressed by the following equation:

$$O(n) = \sum_{b=1}^{B} H(X_{filt}^{log}(n,b) - X_{filt}^{log}(n-1,b)) \tag{3.7}$$

Fig. 3.3 depicts the just described steps for the calculation of the onset detection function.



(a) $X(n,k)$ without whitening

(b) $X(n,k)$ with whitening

(c) $X_{filt}^{log}(n,b)$

(d) $O(n)$

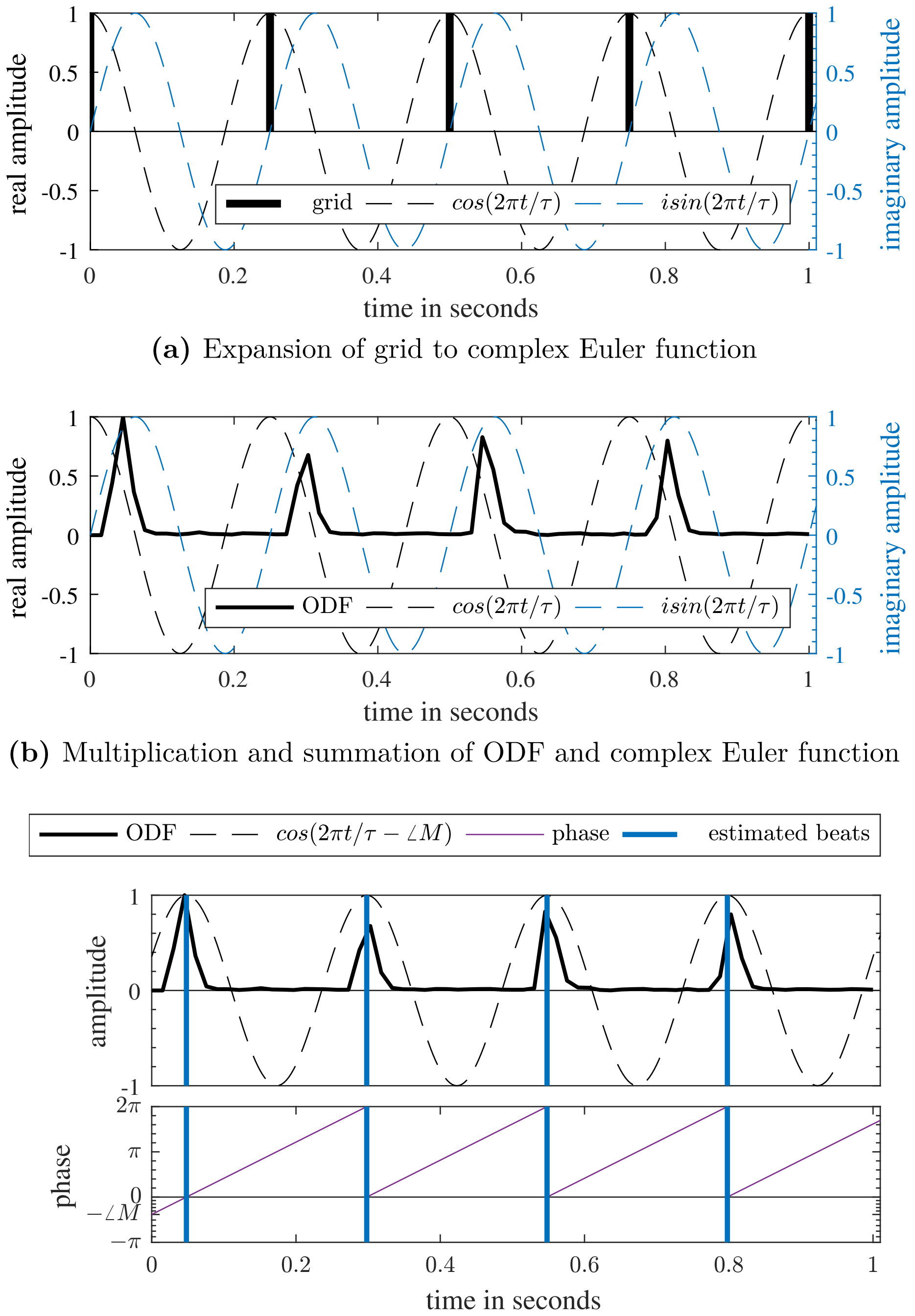
Figure 3.3 Steps of calculation the *Spectral Flux Log Filtered* ODF: a) spectrogram b) whitening c) log filtered spectrogram d) final ODF (black) and the audio signal (gray). The data of all four figures is normalized to a maximum value of 1.

3.3.2 Beat Estimation

For many methods it is common to implement a peak picking after computing the onset detection function. This leads to a sharpening of the autocorrelation of the onset function for periodicity extraction or is even necessary for the calculation of onset time differences to derive the tatum out of it. Here, the raw onset detection function is used for further processing out of the following reasons: If single peaks are picked, a lot of data gets discarded, which - even with low amplitudes - can contribute important information about the phase, which plays an important role in this algorithm. Also, with reducing the amount of data to just a few onsets, the effect of a single onset onto the tempo estimation increases. Hence, a wrongly played onset might interfere more severely when there are less anchor points to compensate this deviation.

For a simple tempo grid analysis, in which grids with different spacings (tatums) are getting matched with the ODF or their autocorrelation function (as in [5]), the grid or the picked peaks have to get expanded to make them insensitive to small deviations in the ODF, and shifted to cover phase shifts of a periodicity. The basic idea behind this thesis' beat estimation approach is to expand the grid to a cosine function, where every grid line coincides with the cosines maxima. The signal gets multiplied with this grid and summed up. This yields a measure for how good the grid fits to the ODF. The use of the cosine function has two advantages: firstly, noisy areas between two peaks are getting suppressed due to the negative half waves of the cosine. This reduces the need of a peak picking and also procures, that wrong grid spacings result into smaller magnitude values as strong onsets between the grid get summed up with negative coefficients. Secondly, with extending the cosine to the complex Euler function, the grid does not have to be shifted anymore. The phase shift of the periodicity can be extracted by calculating the phase of the now complex fitting measure. As a matter of fact, this phase information can directly be used to place beats to the signal as beats occur at every maxima of the phase-corrected cosine function and, therefore, every time the phase strides a multiple of 2π. The whole approach is illustrated in Fig. 3.4.

Wu et al. [32] used a similar approach for tempo estimation, which was used as a basis. In their proposal, a STFT of the ODF is calculated for a subset of frequencies. This results in the tempogram $M(j,\tau)$, which holds a measure for the presence of a tatum τ in each ODF frame (j). This embodies the above described method. Instead of picking the tatum with the highest magnitude for each block, a dynamic programming approach is used to find the optimum tempo path: a utility function is maximized with a gain for high magnitudes and a penalty for high tatum differences between two



(a) Expansion of grid to complex Euler function

(b) Multiplication and summation of ODF and complex Euler function

(c) Utilize phase information for beat estimation

Figure 3.4 Visualization of the beat estimation approach. After expanding the tatum grid (tick black lines) to complex Euler functions (dashed lines) (a), the onset detection function gets weighted with it (b). The summation yields the complex matching measure M, whose phase $\angle M$ (purple line) is equivalent to the phase offset of the corresponding cosine grid (dashed line) (c). Whenever this cosine grid reaches its maximum, a beat is estimated. This is equivalent to the cosine's phase striding a multiple of 2π.

consecutive frames. This prevents paths with unlikely high changes in tempo. The resulting tempo path is used to perform a beat search in the ODF.

However, this thesis' approach focuses on phase information and, therefore, modifications were made regarding the enhancement of the tempogram and usage of the extracted information:

Modified tempogram The tempogram gets enhanced by the information of how well the phase of a frame conforms with its expected value, calculated by the time difference between two consecutive frames and the tatum value.

Phase information In contrast to the tatum information of the optimum path, the phase of the path is used. It contains all the information needed to calculate the positions of the beats, as described above. Especially for an insufficient resolution of observed tatums, the phase can be used to calculate the actual tatum more precisely.

Phase reliability A measure for the phase reliability is used to spot frames, in which the phase information can not be trusted. This happens when there are no significant onsets. For these frames the placed beats get discarded and new beats get inserted by interpolation.

Tempogram

As described, the tempogram $M(j,\tau)$ is obtained by an STFT of $O(n)$, evaluated for specific frequencies, corresponding to a set of tatums. This can be expressed as:

$$M(j,\tau) = \sum_{n=0}^{L-1} O(n+j,\tau)w(n)e^{-\mathrm{i}\dfrac{2\pi n}{\tau f_{s,O}}}, \quad j \in [0, J-1] \tag{3.8}$$

with j being the ODF frame index, L the ODF frame length, $w(n)$ the Hann window function, $f_{s,O}$ denoting the ODF sampling frequency (here 100 Hz) and $\tau \in \mathrm{T}$ denoting the tatum out of a set of different tatum values between 60 and 430 ms. An ODF frame length of $L = 150$ yielded good results, meaning that one tempogram value represents a 1.5 s frame and the beginning of one frame is $\frac{1}{f_{s,O}} = 10\,\mathrm{ms}$ apart from the beginning of its predecessor. The total number of time-steps j can be expressed as $J = N - L + 1$.

To emphasize on phase continuity the phase difference between two consecutive frames $d\Phi(j,\tau)$ is calculated and compared to the expected value $\hat{d\phi}(\tau)$. The difference of both results into the phase deviation matrix:

$$\Delta\Phi(j,\tau) = d\Phi(j,\tau) - \hat{d\phi}(\tau) \tag{3.9}$$

with

$$d\Phi(j,\tau) = \angle M(j,\tau) - \angle M(j-1,\tau) \quad \text{and} \tag{3.10}$$

$$\hat{d\phi}(\tau) = \frac{2\pi}{\tau f_{s,O}} \tag{3.11}$$

By adding π, applying a 2π modulo operation, subtracting π again and dividing by π, the values are getting mapped to a range of $[-1,1]$, with a value of 0 indicating a perfect phase deviation of 0. The modified tempogram $M'(j,\tau)$ gets calculated as follows:

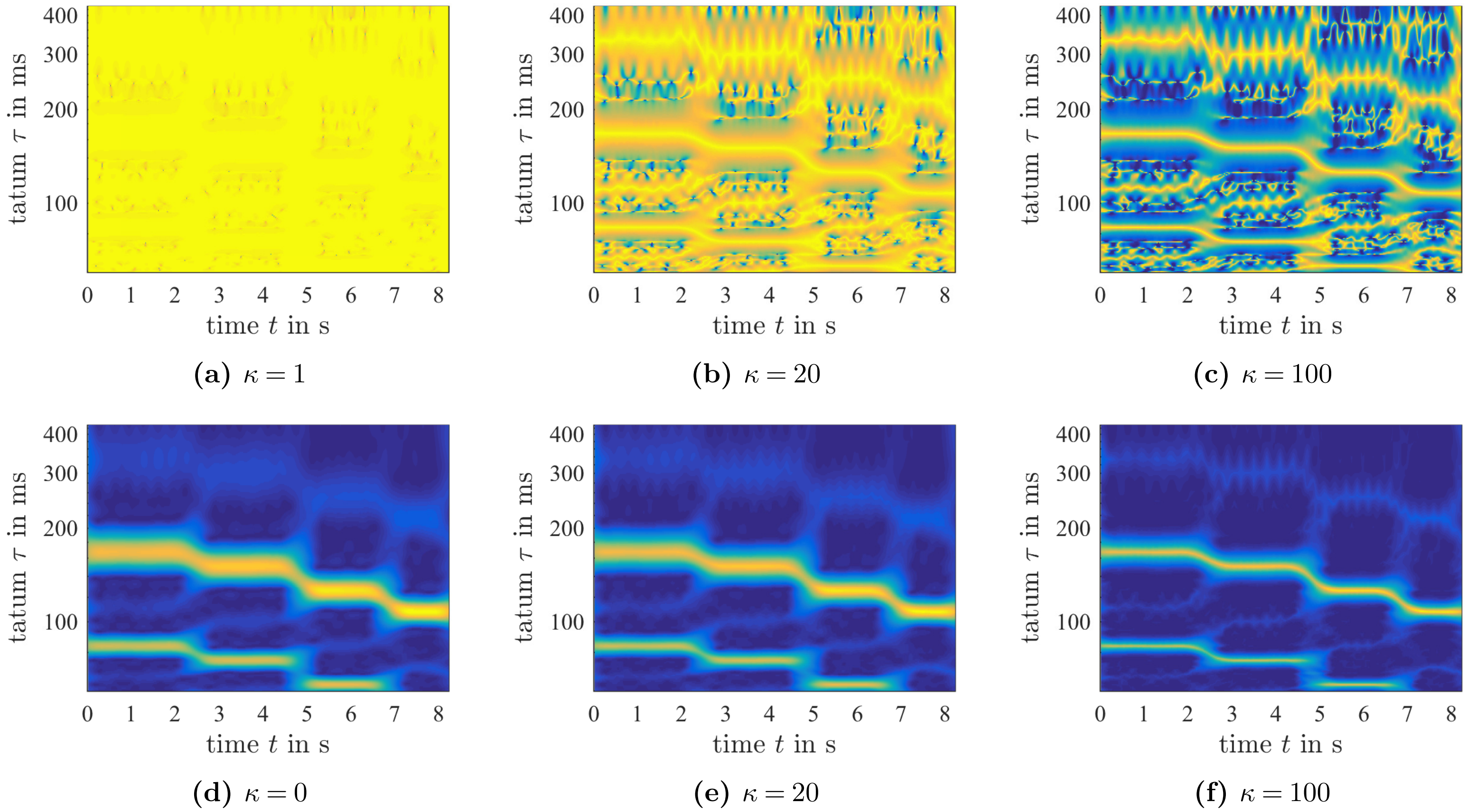
$$M'(j,\tau) = M(j,\tau) \cdot (1 - |\Delta\Phi_{mapped}(j,\tau)|)^{\kappa} \tag{3.12}$$

κ denotes the degree of factoring in the phase conformance. A value of $\kappa = 100$ was ascertained experimentally and suited well for this application. The effect of this modification is depicted in Fig. 3.5 on page 18. The used signal is a hi-hat sequence with leaps in tempo. As can be seen, the modification sharpens the initial tempogram. With κ the amount of sharpening can be adjusted. [2]

As a last step, $M'(j,\tau)$ gets normalized to a maximum absolute value of 1:

$$M'(j,\tau) \leftarrow \frac{M'(j,\tau)}{\max\limits_{j,\tau} |M'(j,\tau)|} \tag{3.13}$$

[2] In general, for a coarse sampled set of tatums a lower κ value should be chosen. Otherwise, the phase nonconformance as a consequence of a non-sampled tatum would lead to a falsification of the tempogram.



(a) $\kappa = 1$

(b) $\kappa = 20$

(c) $\kappa = 100$

(d) $\kappa = 0$

(e) $\kappa = 20$

(f) $\kappa = 100$

Figure 3.5 Effect of different κ values for the modified tempogram. The range of the depicted values is between 0 and 1, with dark blue representing 0, yellow representing a value of 1. a)-c) show $(1-|\Delta\Phi_{mapped}(j,\tau)|)^{\kappa}$. d) shows the initial tempogram $M(j,\tau)$ $(\kappa=0)$, e) and f) the modified version $M'(j,\tau)$.

Optimum Tatum Path

The optimum tatum path can be extracted out of the modified tempogram by maximizing the utility function $U(\boldsymbol{\tau},\theta)$. This function is designed in order that high absolute tempogram values $|M'(j,\tau)|$ (tatum conformance) are advantaged and high leaps in tempo/tatum will result in a penalty (second term in equation 3.14). The goal is to find a series of tatum values $\boldsymbol{\tau} = [\tau_0,...\tau_j,...\tau_{J-1}]$, with τ_j the tatum value for ODF frame j, which maximizes the utility function

$$U(\boldsymbol{\tau},\theta) = \sum_{j=0}^{J-1} |M'(j,\tau_j)| - \theta \sum_{j=1}^{J-1} |\frac{1}{\tau_{j-1}} - \frac{1}{\tau_j}|, \tag{3.14}$$

with θ denoting the penalty factor for a tatum difference between two consecutive frames. With $\theta = 0$ the maximization could be replaced by picking the tatum with the highest absolute tempogram value. The higher θ the smoother the tempo path due to a higher penalty for tempo changes. A value of $\theta = 20$ is best suited for this application.

The search for the maximum of the utility function can be done efficiently with dynamic programming. Therefore, the maximum can be written as:

$$\max_{\boldsymbol{\tau}} U(\boldsymbol{\tau},\theta) = \max_{\tau} D(J-1,\tau) \tag{3.15}$$

with the recurrent equation

$$D(j,\tau) = \begin{cases} |M'(0,\tau)| & \text{if } j = 0, \\ |M'(j,\tau)| + \max\limits_{\tau_{j-1}} \left(D(j-1,\tau) - \theta |\frac{1}{\tau_{j-1}} - \frac{1}{\tau_j}| \right) & \text{otherwise} \end{cases} \tag{3.16}$$

Basically, after initialization the first frame $j = 0$, for every tatum τ_j of the frame j the algorithm looks for that tatum τ_{j-1} of the previous frame which yields the most rewarding transition $\tau_{j-1} \rightarrow \tau_j$. With memorizing $\tau_{j-1,max}$ for every $D(j,\tau)$, the optimum path can be found by backtracking $\tau_{j-1,\mathrm{max}}$ starting with the tatum $\arg\max_{\tau} D(J-1,\tau)$ of the last frame.

The optimum path extracted for the previous shown tempogram is depicted in Fig. 3.6. The path (red line) follows the leaps in tempo.

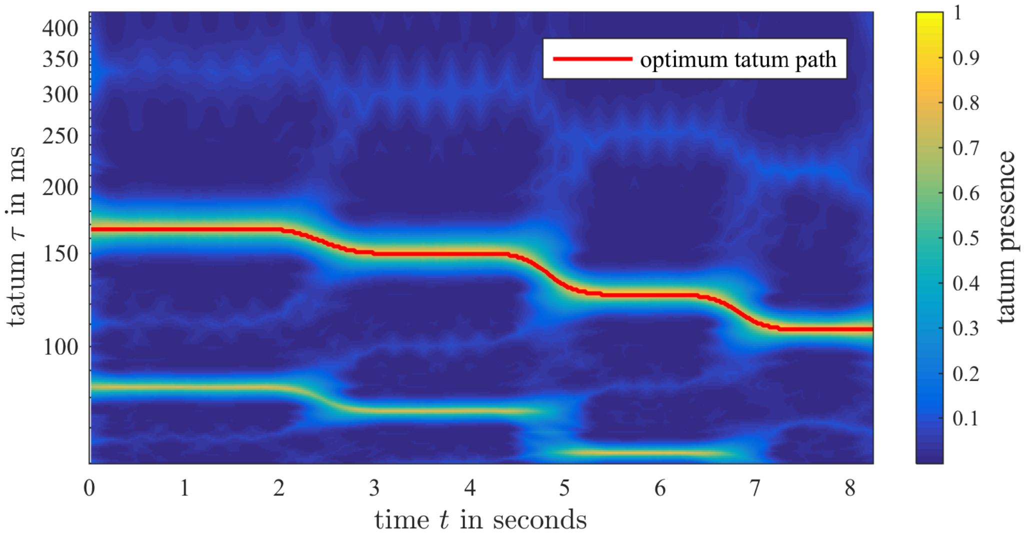


Figure 3.6 Resulting optimum path for a hi-hat signal with leaps in tempo.

3.3.3 Beat Placement and Start/Stop Alignment

As described above, the beat placement uses the phase $\phi(n)$ of the optimum tatum path $\boldsymbol{\tau} = [\tau_0, ...\tau_j, ...\tau_{J-1}]$. The phase can be obtained from the tempogram $M(j, \tau_j)$ [3] by calculating the angle of the complex values. To find a phase value for every time step n of the ODF, the tempogram time steps j have to be mapped to n:

$$j \rightarrow n = j + \frac{L}{2} \tag{3.17}$$

The offset of $\frac{L}{2}$ issues from the phase information being calculated for a frame with length L (see equation (3.8)). So the center of each frame was chosen for the mapping. Nevertheless, the phase information is valid for the beginning of each frame, therefore, the phase itself also has to be adjusted to the frame center by the expected amount $\frac{L}{2}\hat{d\phi}(\tau)$. So the extraction of the phase can be formulated as follows:

$$\phi(n) = \phi(j + \frac{L}{2}) = \angle M(j, \tau_j) + \frac{L}{2}\hat{d\phi}(\tau_j), \quad \text{for } j \in [0, J-1] \tag{3.18}$$

The remaining phase information for $n < \frac{L}{2}$ and $n \geq J + \frac{L}{2} = N - \frac{L}{2} + 1$ has to be derived by extrapolation. The phase then can be used to place beats to the ODF. A

[3] Also the modified tempogram $M'(j, \tau)$ can be used here, as it holds the same phase information.

beat occurs every time the phase strides a multiple of 2π (see Fig. 3.4c on page 15). This search is equal to the calculation of positive[4] zero crossings of $sin\,(\phi(n))$.

Differentiating $\phi(n)$ yields the current phase step, which can be transformed to the current tatum analogous to equation (3.11):

$$\tau(n) = \frac{2\pi}{d\phi(n) f_{s,O}} \tag{3.19}$$

The $\tau(n)$ values are not bound to those of the tatum set T and, as a consequence, are sampled with a higher precision[5]. Averaging these values results into the mean tatum $\bar{\tau}$, which can be used for the phase extrapolation and interpolation of beat gaps (described hereafter).

Additionally to the angle of the tempogram, the magnitude is used as a measure of phase reliability. If the magnitude is lower than a predefined threshold, the phase information can not be trusted. Low tempogram magnitudes can occur in frames with only few or no onsets. In that case, the phase information gets discarded and the resulting gaps get filled with equally spaced beats corresponding to the mean tatum. An example of interpolated beats is shown in Fig. 3.7.

The last step is to align the start and stop cues to the estimated beat positions. This is easily be down by shifting each cue the closest beat.

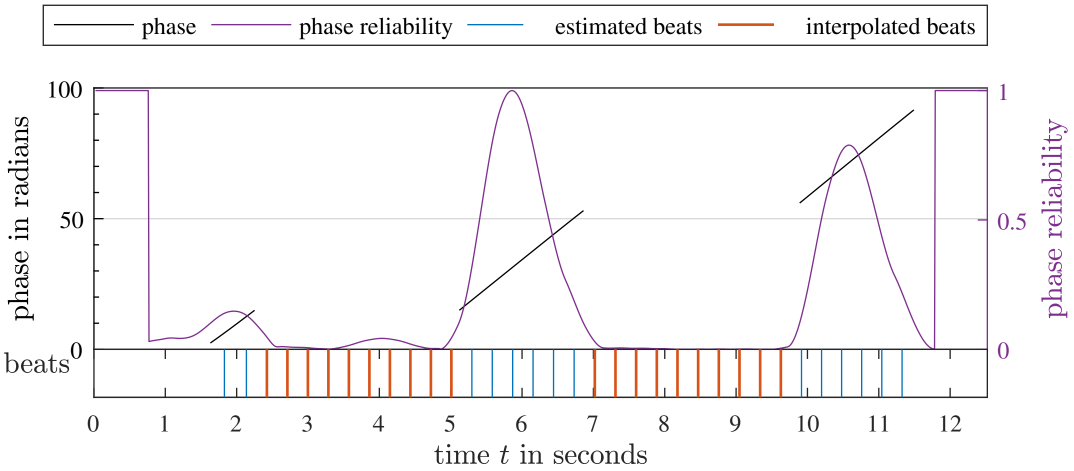


Figure 3.7 Example of filling gaps with low phase reliability. The black line represents the phase with gaps as the phase reliability (purple) drops below 0.1. The estimated beats (blue) are gathered with the phase information, the interpolated beats (orange) are filled in by interpolation. The impression of the phase keeping its value right before and after a gap is an artifact of phase-unwrapping.

[4] transition from negative to positive values

[5] see the example in section 3.4.2 for demonstration

3.4 Examples

3.4.1 Beat Interpolation of Low Magnitude Gaps

Fig. 3.8 shows ODF and tempogram of an audio signal with sustaining chords - hence, only a few onsets exist. As both ODF and tempogram reveal, there are only few anchor points for the pathfinding process at around $t = 5$ and 10 s. Even the start of the signal does not hold any useful tatum information. Nevertheless, the algorithm found the optimum path, which fits the audio signal best. Due to the vanishing magnitude of the tempogram between $t = 2$ and 4 s and between $t = 6$ and 9 s, the phase reliability measure is not high enough to place reliable beats. As a consequence, two gaps emerge after discarding unreliable beats, which get filled with interpolated beats as can be seen in Fig. 3.7.

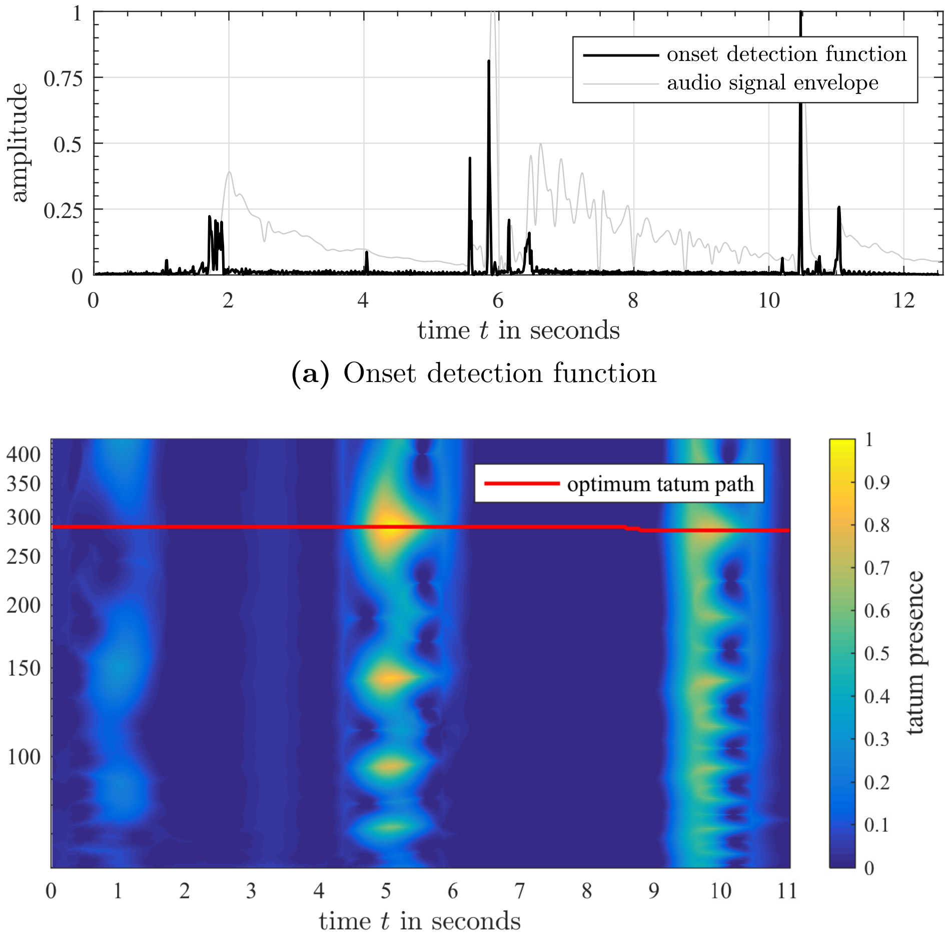
3.4.2 High Tatum Precision despite Coarse Sampling

To demonstrate the higher tatum precision than the sampled tatum grids due to factoring in phase information, a semiquaver hi-hat signal at tempo 93 bpm is used. The resulting tatum is $\tau_0 = 161.29\,\text{ms}$. The used tatum grids of the tempogram stage are sampled with a coarse resolution: the examined tatum values next to τ_0 are 155.01 and 165.89 ms. The corresponding tempogram is depicted in Fig. 3.9.

The optimum path search yielded a constant tatum of $\tau = 165.89\,\text{ms}$, as being closest to τ_0. However, the phase information yielded an average tatum of $\bar{\tau} = 161.40\,\text{ms}$, which is remarkably closer, with a difference of just 0.11 ms instead of 4.6 ms.

3.4.3 Greatest Common Divisor (GCD) of Existing Tatums

As described earlier, the algorithm tries to find that tatum, which fits best to all occurring tatums. This mechanism is demonstrated with the hi-hat signal depicted in Fig. 3.10. It consists of two different note values, quavers and quaver triplets. At tempo 100 bpm, the time difference between two successive quaver notes is $\tau_1 = 300\,\text{ms}$ and for quaver triplets $\tau_2 = 200\,\text{ms}$, respectively. As expected, these tatum values also occur in the tempogram with a high presence measure (see Fig. 3.11). However, the optimum tatum path was found for a tatum τ_3, which does not occur explicitly, but is implied by the two tatums τ_1 and τ_2 by being the greatest common divisor $\tau_3 = \gcd(\tau_1, \tau_2) = 100\,\text{ms}$.



(a) Onset detection function

(b) Tempogram

Figure 3.8 Onset detection function (a) and tempogram (b) of an audio file with only a few onsets.

3.4.4 1/f Noise Jitter

Human rhythm production underlies timing fluctuations [15]. This example demonstrates the effect of 1/f-noise (pink noise) jitter within a semiquaver hi-hat sequence on the tempogram. Fig. 3.12a shows the tempogram without jitter. The presence of a tatum of $\tau = 150\,\text{ms}$ can be observed which corresponds to a tempo of 100 bpm[6]. After applying a 1/f noise jitter with a standard deviation of 500 samples (about 10 ms), noise occurs in the tempogram, as can be seen in Fig 3.12b. Also the optimum path underlies this fluctuation and, therefore, impacts the algorithms performance (for quantitative values see section 4.3 on page 53). Note, this is an exaggerated example, as the used jitter with a standard deviation of 10 ms is very well perceivable.

[6] if one hi-hat hit equals a semiquaver note

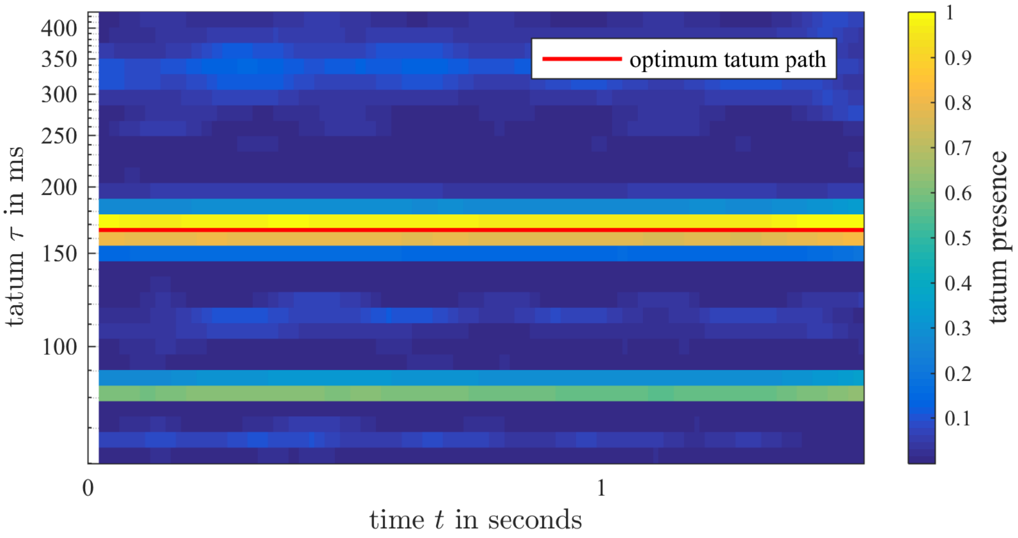


Figure 3.9 Tempogram with low tatum sampling.

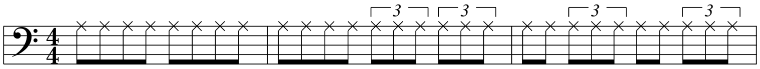


Figure 3.10 Score of the complex to rhythm.

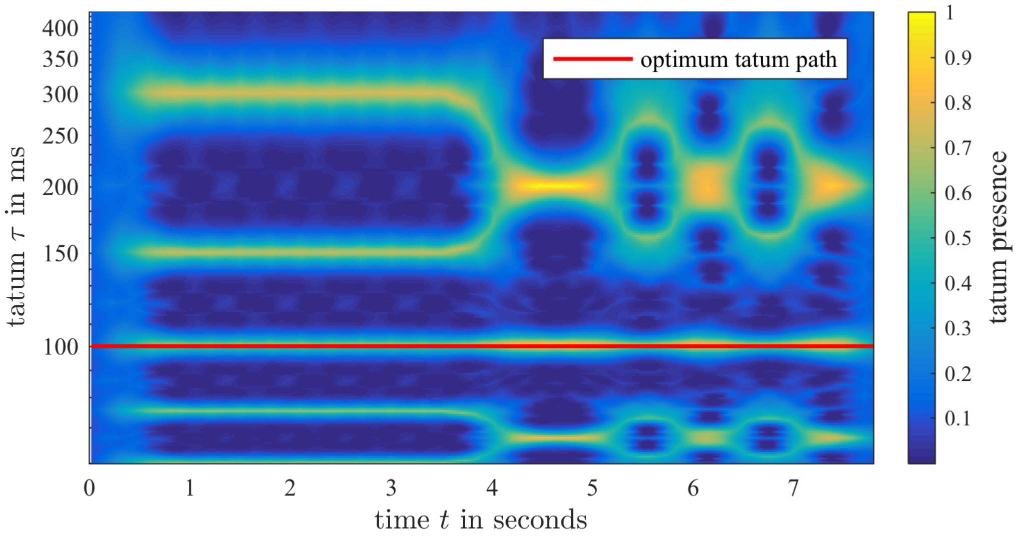
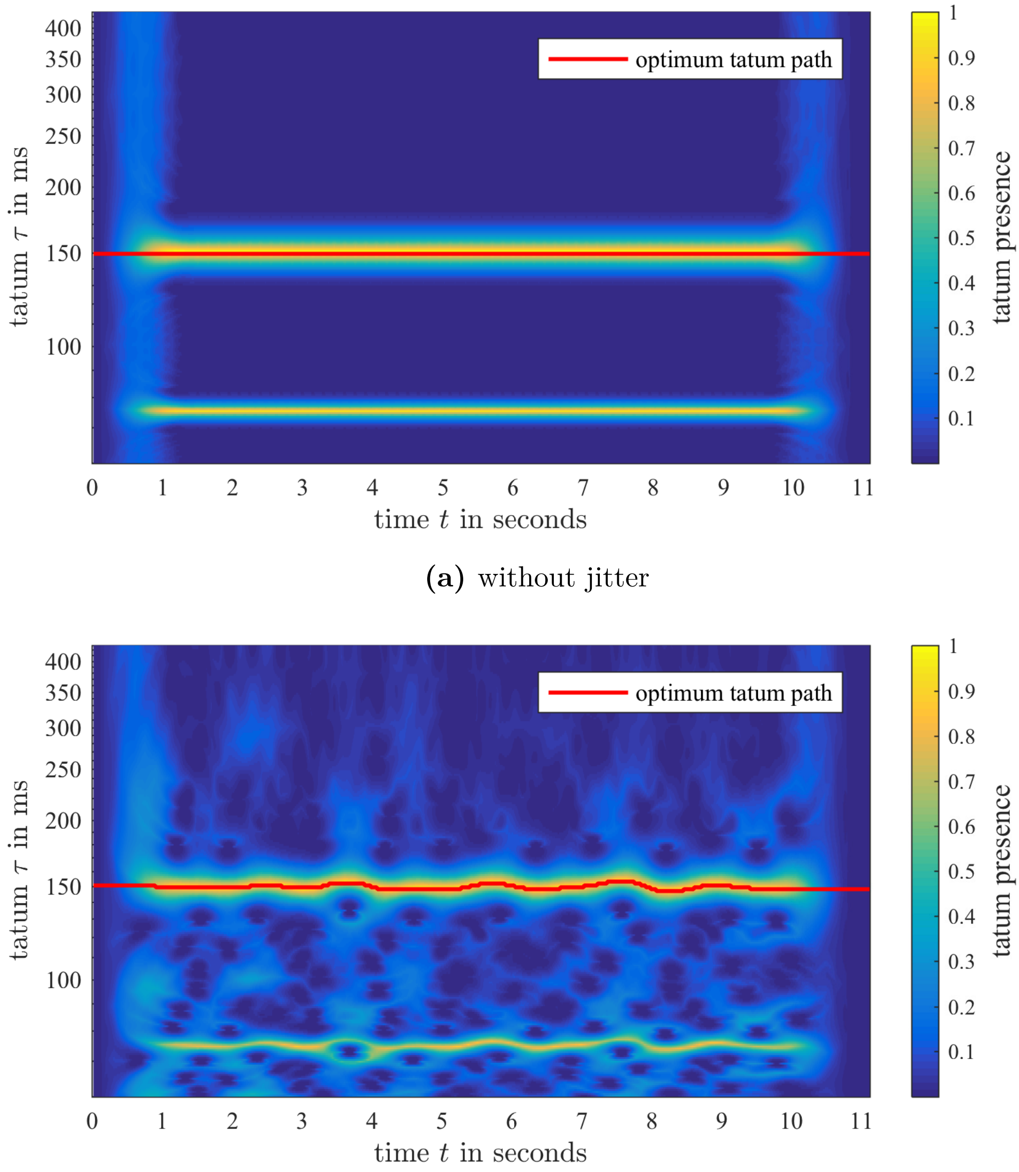


Figure 3.11 Effects of complex rhythms on the algorithm's performance.



(a) without jitter

(b) with 1/f-noise jitter

Figure 3.12 Effect of 1/f-noise jitter with a standard deviation of 500 samples on the tempogram.

3.5 Realtime Capability

The above described algorithm only embodies a useful tool for live looping if it is real-time capable. At a first glance this means, that all the calculations have to be completed when the musician stops the recording and expects the phrase being repeated seamlessly. This actually is not possible as the back tracking process of the optimum tatum path estimation (Section 3.3.2) can not start until the last frame, in which the stop cue occurs, was processed.

Fortunately, by lowering the requirement of the alignment to being completed before the second iteration starts (instead of the first), real-time capability can be easily achieved: now the algorithm has to be completed within the duration of one phrase, which is usually around 3 s to 10 s. However, a possibly perceivable gap at the first iteration has to be accepted.

Table 3.1 shows the needed time for computations in percentage of the duration of two different audio signals. The data shows that the algorithm's performance in regards of computation time depends strongly on the number of evaluated tatums as the computing effort of the tempogram increases linearly and that of the optimum path search quadratically with $n_\mathcal{T}$. The rather constant relative overall time shows a linear relationship with the signal's duration. Note that these results are gathered by a single execution of the algorithm for each combination of audio signal and number of tatums and, therefore, may be affected by different CPU loads. Also these results were gathered in an offline version of the algorithm. By computing the ODF and parts of the tempogram and the optimum path search during the recording phase, these values can even be lowered. It can therefore be concluded that the algorithm is real-time capable and only needs a fraction of the recorded signal's duration for computation.

Table 3.1 Results of time profiling of two audio files with different durations and different number of tatums. The algorithm was implemented in Matlab and executed on a 2.3 GHz Intel Core i7 processor.

duration (in s)	$n_{\mathcal{T}}$	time needed for computation (in % of duration)				
		ODF	tempogram	path	beat estimation	overall
9.3	60	1.76	0.08	2.51	0.10	4.45
9.3	120	1.83	0.16	5.93	0.14	8.06
9.3	240	1.76	0.26	16.48	0.14	18.64
22.2	60	1.02	0.09	2.85	0.07	4.03
22.2	120	1.02	0.17	6.80	0.06	8.05
22.2	240	1.00	0.23	17.02	0.07	18.32

Chapter 4

Validation

4.1 Obstacles to an Ideal Performance

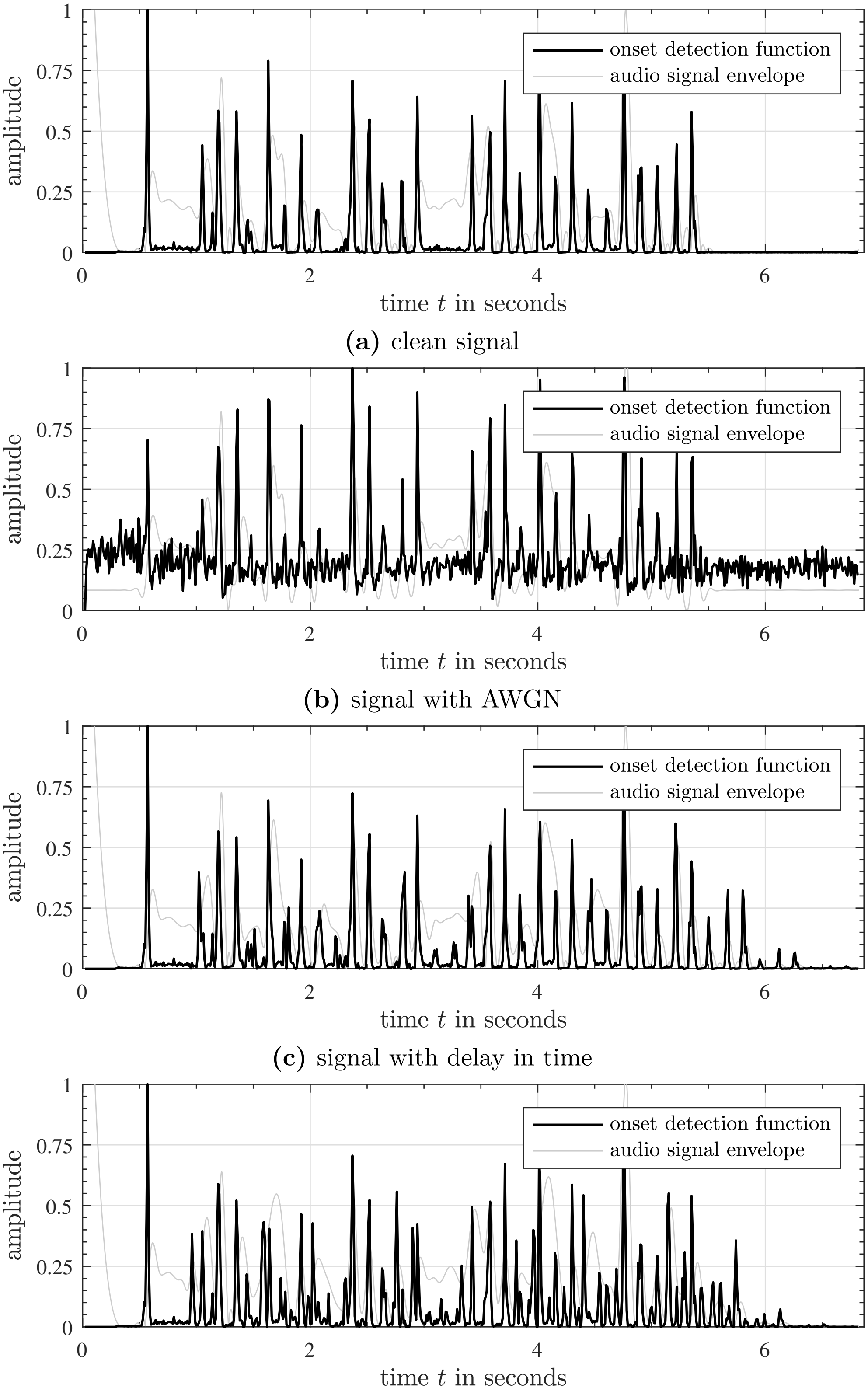
Ideally, the algorithm detects the tatum correctly and changes the start and stop cues to make them align perfectly. The loop would be seamlessly continuous. Due to noise in the signal chain or timing fluctuations of the recorded phrase, this goal is hard to achieve. In the following, the effect of these circumstances is shown.

4.1.1 Noise

Not only noise introduced by the signal chain (guitar, cable, amplifier) can be seen as interference but also sound effects applied to the guitar signal. To investigate the effects of noise on the performance of the algorithm, additive white gaussian noise (AWGN) and delay effects were added to a recorded guitar signal with strumming chords at a tempo of 100 bpm. The smallest existing tatum is $\tau = 150\,\text{ms}$. Two different delay times were used: $t_1 = 450\,\text{ms}$ and $t_2 = 386\,\text{ms}$ leading to one delay being synchronous with the tempo and one being asynchronous.

Fig. 4.1a shows the onset detection function of the clean audio signal without any form of noise. As for the case of AWGN (Fig. 4.1b), a noise floor was also added to the onset detection function. Nevertheless, this hardly affected the tempogram and the optimum path search, as can be seen in Fig. 4.2.

When the delay effect is added to the signal, the two cases of being in and out of time affect the performance differently. For a delay time in sync with the tempo, new onsets were added to the initial ODF (see Fig. 4.1c). With the delay time being a multiple of the tatum ($t_1 = 3\tau$), these new onsets fit the underlying grid perfectly. As a result, the tempogram depicted in Fig. 4.3a is not affected in a negative way, if



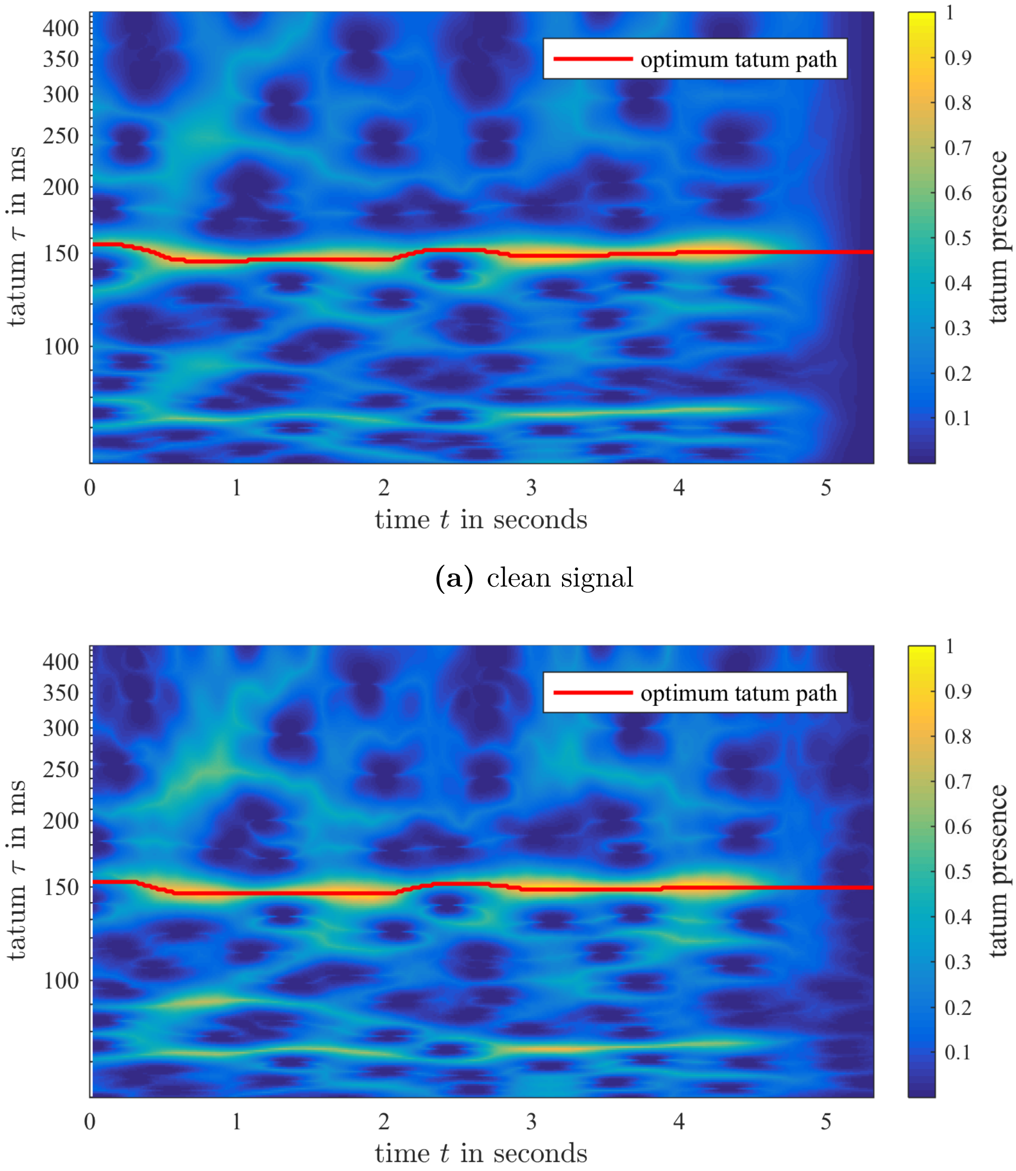
(a) clean signal

(b) signal with AWGN

(c) signal with delay in time

(d) signal with delay out of time

Figure 4.1 Effects of additive white gaussian noise (AWGN) and delay effects (both in and out of time) on the onset detection function.

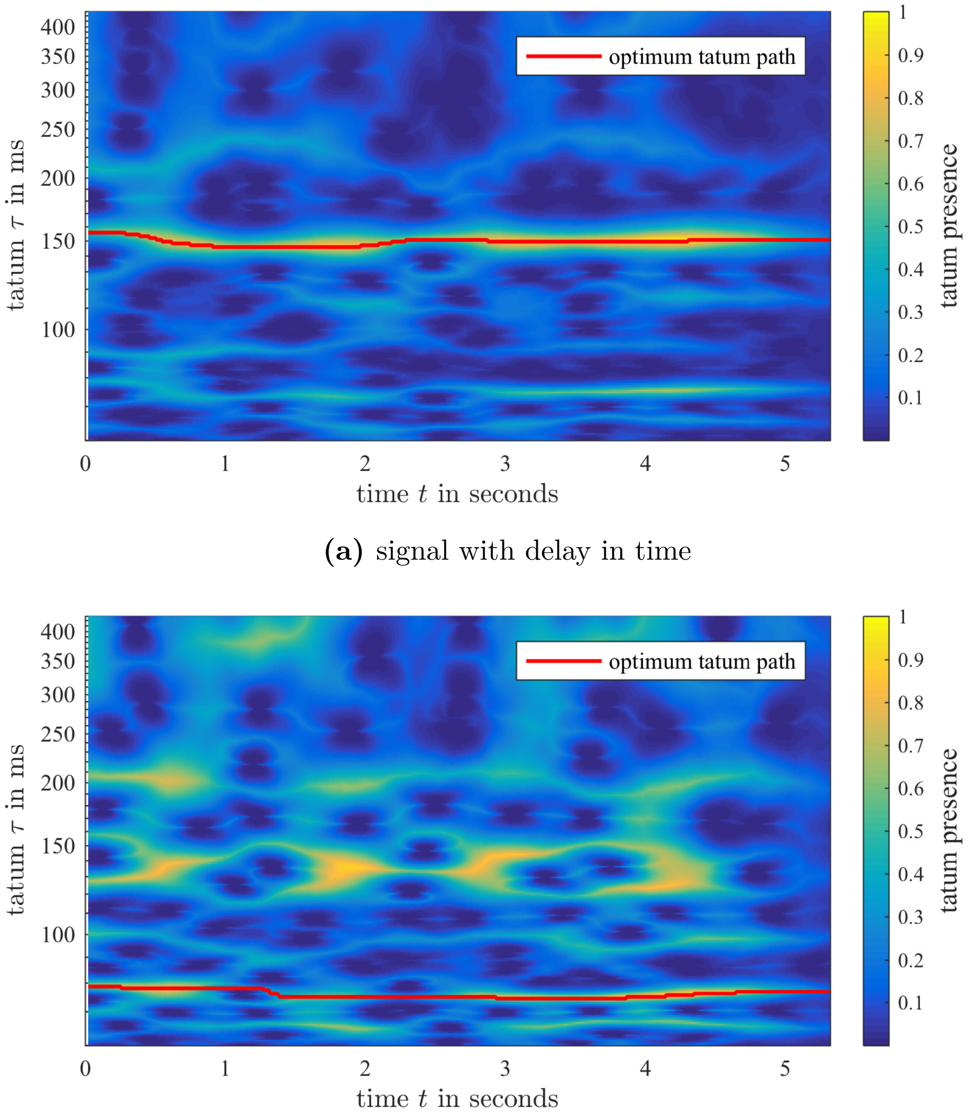


(a) clean signal

(b) signal with AWGN

Figure 4.2 Effect of additive white gaussian noise (AWGN) on the tempogram.

not even the optimum path has become more distinct. However, with delay times not fitting the tempo, the ODF gets interfered by new onsets being out of time, as can be seen in Fig. 4.1d. Now the timing of the added onsets does not correspond to the initial onsets and new inter onset intervals (IOI) occur in the ODF. Similar to the example *Greatest Common Divisor* in Sec. 3.4.3 on page 22, the optimum path search finds that path wich fits all the existing tatums (Fig. 4.3b).



(a) signal with delay in time

(b) signal with delay out of time

Figure 4.3 Effect of noise on the tempogram.

However, the negative effects of noise (whether AWGN or sound effects) can be averted by splitting the signal into a clean signal for the beat analysis and an effect signal for recording. As long as there is no long delay between these signals due to the effect processing - which is usually the case - the analysis fits the timing of the record.

4.1.2 Timing Fluctuations

The production of rhythm is subject to the inaccuracy of the human motor action. As a consequence, inevitable timing fluctuations occur, making the rhythm deviate from the ideal beat pattern. This actually leads to a preferred listening experience, which is why so called “humanizer” are used to simulate human fluctuations in computer-generated music [16, 26]. Hennig et al. investigated the fluctuation of human rhythm production. They found long-range correlations in the deviations of produced beats from metronome clicks similiar to $1/f^{\beta}$-noise, with $0.2 < \beta < 1.3$.

The effects of $1/f$-fluctuations in timing on the algorithm’s performance are qualitatively depicted in Fig. 3.12 on page 25 of the previous chapter. Quantitative results are given in the *Validation* section at the end of this chapter.

Fortunately, in the practical application the loop does not have to be perfectly coherent as long as the gap between start and stop is small enough to be not audible. Thus, to be able to validate the algorithm it is obligatory to take a look at the perception of gaps in rhythmic phrases - both negative and positive - hereinafter *rhythm perturbances.*

4.2 Audibility of Rhythm Perturbances

4.2.1 Classical Threshold Theory and Used Terms

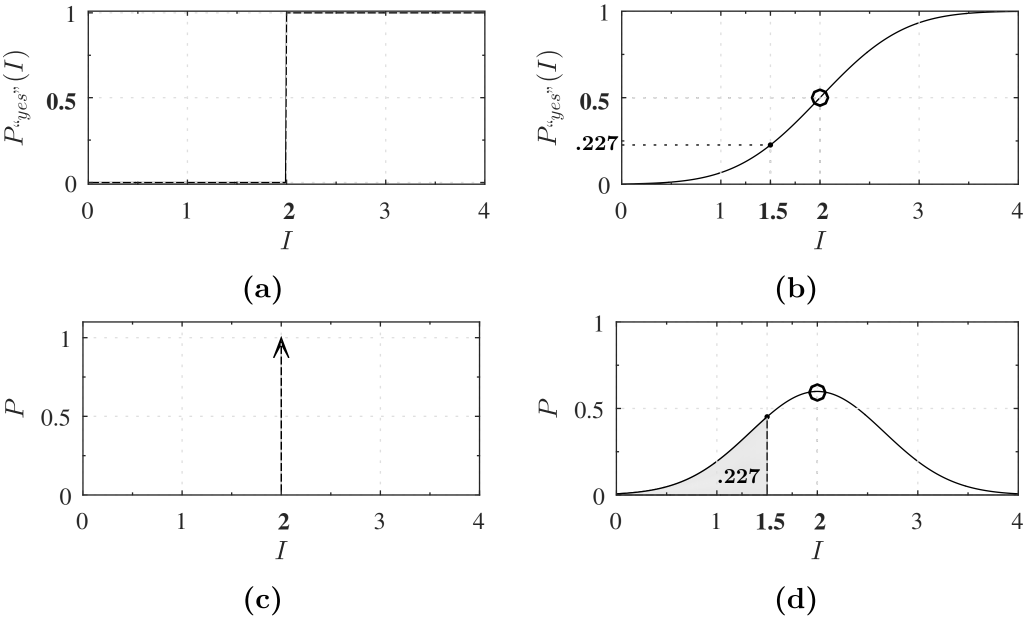
Detection of a Stimulus

Early threshold theories stated, that a stimulus is only detectable if its intensity (e.g. level, size, duration) outreaches a specific threshold. The question, whether or not a stimulus was perceived, would be answered “yes” 100 % of the time, if its intensity is higher than said threshold (see Fig. 4.4a).

The *Classical Threshold Theory* agrees with that, but adds a primary assumption, that this threshold is not constant but rather varies over time. This means that a stimulus only gets detected, when its intensity outreaches the *momentary threshold*, which is constantly changing. Therefore, a fixed threshold as in Fig. 4.4c can no longer be assumed. However, there exists a *probability density function* (PDF) for the momentary threshold $f(I)$. A typical distribution for sensory thresholds - the normal distribution $\mathcal{N}(\mu, \sigma)$ - with hypothetical parameters ($\mu = 2$, $\sigma = \frac{2}{3}$) is shown in Fig. 4.4d). For a stimulus intensity value I, the proportion of time for which the momentary threshold lies underneath (i.e. the stimulus is detectable) can be calculated:

$$F(I) = \int_{-\infty}^{I} f(I')\,\mathrm{d}I' \tag{4.1}$$

This is also known as the *psychometric function* $F(I)$, which reflects the probability of detecting a stimulus $P_{\text{“}yes\text{”}}(I)$ with a certain intensity I. The psychometric function is equivalent to the *cumulative density function* (CDF), in case of a normal distribution for $f(I)$, $F(I)$ is equal to the cumulative normal distribution $\Phi(I)$ also referred to as *normal ogive* [30], which is depicted in Fig. 4.4b. As can be seen in comparing the psychometric function with its underlying distribution, the area below the distribution (e.g. 0.227 for $I = 1.5$) equals the detection probability. The intensity of which the probability of detection reaches 50 % is referred to the *absolute threshold.*[12] A stimulus at that intensity level gets detected half of the trials.[11] It is also the expected value of the PDF (the mean in case of a normal distribution).



(a)

(b)

(c)

(d)

Figure 4.4 Two psychometrical functions $F(I)$ and their underlying probability density functions $f(I)$ of the momentary threshold. (a) and (c) early threshold theories; (b) and (d) classical threshold theory.

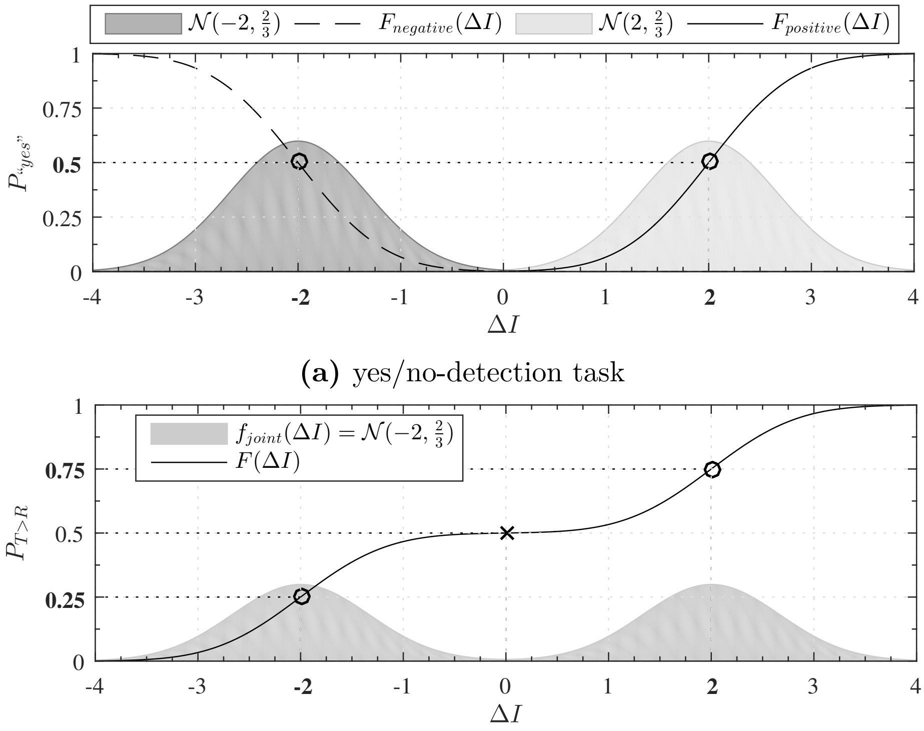
Discrimination of Two Different Stimuli

Not only a stimulus itself can be detected, but also a difference between two stimuli. This is called discrimination. Therefore, two stimuli are presented, a reference stimulus with a base intensity $I_{reference}$ and a target stimulus with a intensity difference ΔI. The threshold of discrimination is called *difference threshold* and is calculated depending on the type of discrimination task.

> **NOTE:** The following example is not fully proven by literature but gives a good understanding of the origins of different psychometric functions.

Similar to the detection of a stimulus, there is a threshold from which a intensity difference ΔI is more likely to be detected, i.e. a target stimulus with a intensity $I_{target} = I_{reference} + \Delta I$ can be discriminated more likely from a reference stimulus with intensity $I_{reference}$. This threshold is also known as the *just noticable difference* (JND). A **yes/no-detection task** can be used to measure this threshold: the subject listens to both stimuli (reference and target) and has to decide whether or not (YES or NO) they were different. In other words, they have to *detect* a difference. Hence, the psychometric function is the same as the one of a detection task (see Fig. 4.5a). The right half side of the figure contains the psychometric function $F_{+}(\Delta I)$ (solid line), its underlying distribution of the momentary threshold $f_{+}(\Delta I)$ (light grey area) and the JND for detecting a *positive* intensity difference ΔI, marked by a circle (50 %). It seems logical that there also has to be a difference threshold for detecting negative intensity differences (i.e. the target stimulus has a lower intensity than the reference). This case is represented in the left half of the figure: the psychometric function $F_{-}(\Delta I)$ (dashed line) and its underlying distribution $f_{-}(\Delta I)$ (dark grey area).[1] The JND is also marked with a circle. For simplicity, both distributions were chosen to be symmetrical, even though they don't have to be in general. In sum, for high absolute intensity differences (here: $\Delta I < -2$ and $\Delta I > 2$) the probability of detecting a difference between both stimuli is higher than 50 %. For intensity differences between both thresholds ($-2 < \Delta I < 2$) the probability is smaller and goes down to its minimum for no difference ($\Delta I = 0$). For a yes/no-detection task both negative and positive momentary threshold distributions have to be evaluated separately. This changes for a different type of discrimination task.

[1] The integral for the negative psychometric function in equation 4.3 has to be evaluated in the direction of increasing absolute intensity, i.e. with a lower integral limit of ∞ instead of $-\infty$ and an additional negative sign prior to the integral for the result to remain positive valued.



(a) yes/no-detection task

(b) two-alternative forced choice task

Figure 4.5 Illustration of the occurrence of different thresholds for yes/no-detection tasks (a) and 2AFC tasks (b) for assumed parameters of the psychometrical functions (grey areas). The lines represent the psychometric function $F(\Delta I)$.

The **two-alternative forced choice task** (2AFC) is another way to determine the JND. Here the subjects listen to the same stimuli as described above but have to decide, which one has a higher intensity (e.g. has a longer duration). To visualize the psychometric function, both momentary threshold distributions can be combined, resulting in a joint distribution $f_{joint}(\Delta I)$ (grey area in Fig. 4.5b). Applying Equation 4.3 yields the psychometric function $F(\Delta I)$ for a 2AFC task (solid line), for both negativ and positive intensity differences ΔI. For a low absolute intensity difference (e.g. $\Delta I = 0$), the subjects usually don't perceive any difference and due to being forced to choose one of both stimuli to be at a higher intensity, the subjects have to guess. This results in a probability of choosing the target stimulus for the higher intensity $P_{T>R} = 50\,\%$. This point (marked with an x) is also known as the point of subjective equality (PSE). This means that the 50 % level is not suitable for specifying a difference threshold. Usually for a 2AFC task the 75 % level is used for the positive JND and the 25 % level for the negative (circles), making it correspond to the 50 % level of a yes/no detection task [22].

Used Terms

In literature, different terms and symbols are used to describe the very same entity. To combine results of different studies, a common notation is necessary. The following section represents terms, which are used in this thesis, specifically regarding the audibility of rhythm perturbances:

Inter Onset Interval (IOI) is the time duration T of the reference interval. It is measured between the beginnings of two consecutive sound events. Smaller T values represent a higher tempo of the sequence and vice versa. T is used as reference/base duration for the discrimination tasks.

Distortion refers to altering the duration of a time interval, i.e. either lengthening or shortening, depending on the type of perturbation. I.e. the difference between the duration of the target interval and the duration of the reference interval. It can be denoted with absolute or relative values:

$\boldsymbol{\Delta T}$ Absolute amount of distortion in seconds.

$\boldsymbol{d}$ Relative amount of distortion in relation to the duration of the base interval T:

$$d = \frac{\Delta T}{T} \tag{4.2}$$

$\boldsymbol{D_{50}}$ (given for different IOI) is used as term for the JND of two time intervals. It represents the amount of distortion needed to achieve a detection rate of 50 %. The values are given in percent of the base duration (relative distortion). It is equivalent to the terms JND and difference threshold in literature.

$\boldsymbol{\Delta T_{50}}$ is the absolute equivalent to D_{50}. The values are given in seconds and can be calculated with:

$$\Delta T_{50} = D_{50} * T \tag{4.3}$$

4.2.2 Prior Research

Different researchers [17, 15, 25, 14, 31, 10] have studied the ability of subjects to discriminate durations in rhythm sequences. For this purpose, different methods and stimuli were used making the results only suitable for specific applications. A small selection of applicable studies is presented below.

Duration Discrimination in a Series of Rhythmic Events

Halpern and Darwin [14] investigated the duration discrimination of the last interval of a series of four clicks. Eight subjects with a variety of musical backgrounds listened to series of four clicks, the first three equally spaced in time with IOI reaching from 400 ms to 1450 ms. The duration of the last interval was either shortened or lengthened by an amount of 0 %, 2 %, 4 %, 6 %, 8 % or 10 % of the base IOI. The subjects had to state whether the last click came early or late i.e. whether the last interval was shortened or lengthened, respectively.

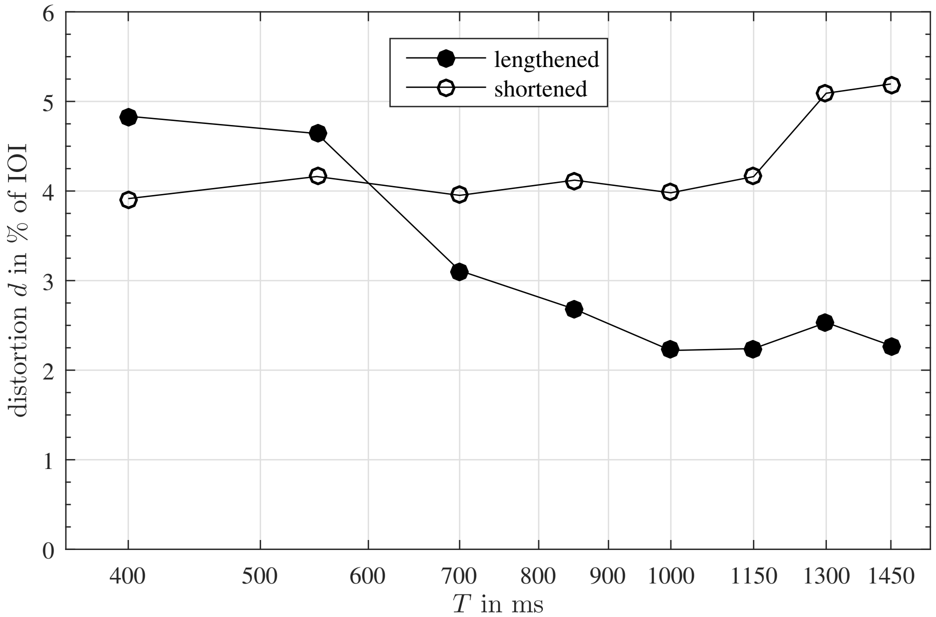


Figure 4.6 JND of duration of the last of three intervals found by Halpern and Darwin. Data points represent D_{50} values as a function of the base IOI (abscissa). Filled circles (●) indicate the condition "lengthened", open circles (○) the condition "shortened" for the type of perturbation. Created from values given in [14].

The gathered data was fitted to a cumulative normal distribution to derive the psychometric function for each subject. The means and standard deviations of each normal distribution were then used to describe the results. To make the results comparable to the other found results and those of the listening tests conducted for the purpose of this thesis, the data given in [14] was used to calculate the absolute level of discrimination. The results are shown in Fig. 4.6. The filled (●) and open circles (○) represent the D_{50} values for the two types of perturbation "lengthened" and "shortened". The JND for the shortened intervals seems to be constant ($D_{50} \approx 4\,\%$) for base durations between 400 and 1500 ms. For $T = 1300$ ms and $T = 1450$ ms the JND is about 5 %. The "lengthened" JND values show a higher dependence on the base duration. For lower IOI ($T < 600$ ms) the threshold is at about 5 %, with decreasing values towards higher IOI ($T > 600$ ms) with a minimum of approx. 2 %. These results also show a high dependence on type of perturbation especially towards higher IOI.

Rhythm Perception in Repetitive Sound Sequence

Hibi [17] investigated the ability of listeners to tell whether or not a sequence of isochronous tones has been distorted. Therefore, he used the method of constant stimuli within a yes/no-detection task to sample the psychometric function and gain the just noticeable difference (JND) for small disturbances of a monotonic, isochronous sequence.

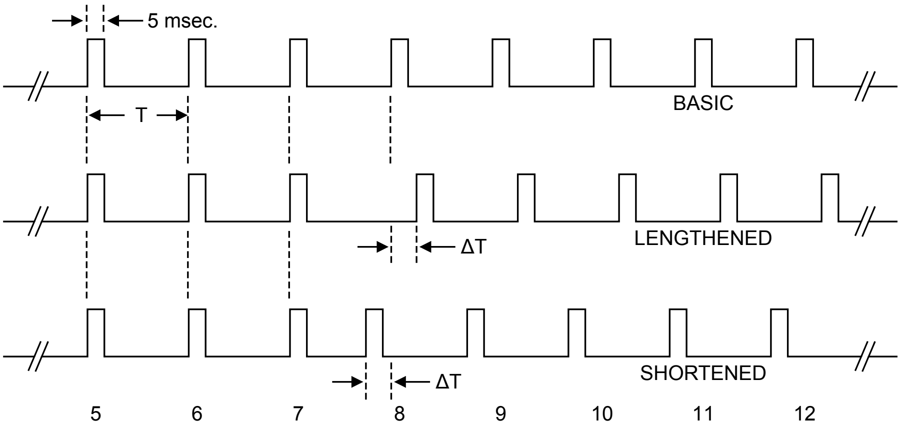


Figure 4.7 Examples of distorted sequences (basic, lengthened and shortened). Adapted from [17].

There were three types of stimuli, which are depicted in Fig. 4.7. The basic (or undistorted) sequences consisted of 15 uniformly spaced tone bursts with IOIs reaching from 143 ms to 1500 ms. The two distorted types emerged from lengthening (resp. shortening) one IOI of the basic sequence by the relative duration d (2, 4, …, 14 and

16 % of the basic IOI). The altered IOI appeared either between the 7th and 8th, 8th and 9th or 9th and 10th position (the former in Fig. 4.7), shifting all subsequent tone bursts backwards (resp. forwards).[2]

Three subjects with normal hearing participated at the experiment. Their task was to listen to the played sequences and tell whether or not there was a distortion of the metrical regularity. Out of the obtained psychometric function the author calculated the D_{50} values for either the lengthened and shortened sequence of each rate of succession.

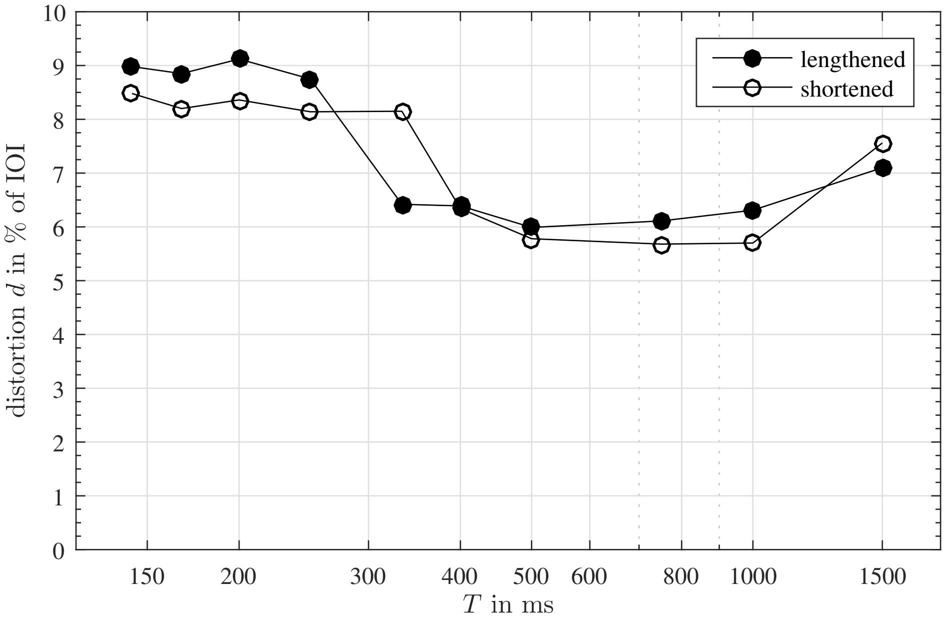


Figure 4.8 JND of duration found by Hibi. D_{50} values as a function of T. Filled circles (•) indicate the condition "lengthened", open circles (∘) the condition "shortened" for the type of perturbation. (adapted from [17])

The results are shown in Fig. 4.8. The ordinate contains the amount of distortion of a sequence, the abscissa the base interval T. The found D_{50} values are indicated depending on the type of perturbation: filled circles (•) for the condition "lengthened" and open circles (∘) for the condition "shortened".

There were found two areas of constant D_{50} values, one reaching from T=143 ms to 250 ms and the other from T=400 ms to 1000 ms. For the first area, the D_{50} value varied between 8.9 % (lengthened) and 8.2 % (shortened) depending on the type of perturbation. In contrast to the first area, there was no significant difference between "lengthened" and "shortened" found in the second area (mean value of D_{50} was 6.1 %).

[2] This creates a scenario which would occur if a loop is not perfectly enclosed. The changed interval embodies the transition from the end of a phrase to the beginning of the repetition.

The dependency of the type of perturbation on the thresholds in the transition area (T=250 ms to 400 ms) was even higher. [10]

Compared to the results of Halpern and Darwin (Fig. 4.6) the thresholds found were much higher in Hibi's study. Beside the different methods of gathering and processing the data, the wide range of musical education of Halpern's participants could also be an explanation for their better performance.

Out of doubts, that such a small group of subjects ($n = 3$) can produce reliable data, the author carried out a small selftest on himself. The resulting just noticeable differences were smaller than the ones Hibi found. The authors musical training also raised suspicion of a strong dependency on musicality.

Is there a Difference Between Musicians and Non-Musicians?

The question arises of whether there are diverging thresholds for musicians and non-musicians. In Hibi's listening test (described above) this was not taken into account. And for this applications it seems to be important since the addressed group for a guitar looper is in fact musicians.

In different studies focusing on human timing abilities, a significant difference was found between musicians and non-musicians:

- Franěk et al. [9] conducted a finger tapping task, in which the subjects had to follow repetitive rhythmic tonal patterns. It revealed that motor timing among musicians was more accurate than among non-musicians.
- The study of Drake and Botte [7] highlights significant differences between subjects with and without musical training in a tempo sensitivity task. Musicians were able to detect much smaller changes in tempo than non-musicians.
- Rammsayer and Altenmüller [25] investigated the temporal information processing of musicians and non-musicians in a variety of auditory tasks, including temporal discrimination and rhythm perception tasks. For both of them, musicians performed significantly better in terms of discrimination of smaller changes.

To verify whether the different time discrimination abilities of musicians and non-musicians also apply for the perception of rhythm perturbances, two listening tests were conducted: the first replicating the listening test of Hibi [17], but including the additional factor *musicality*; the second investigating the influence of the degree of presence of the tatum within a musical context.

4.2.3 Listening Test: Rhythm Perception Revisited

The first experiment investigated the audibility of rhythm perturbances in a monotonic, isochronous sequence. The aim of the study was to find the just noticeable difference of duration for sequences with different tempos ($150\,\text{ms} \leq T \leq 600\,\text{ms}$). Of further interest was whether musicians are more sensitive to rhythm disturbances than non-musicians.

Therefore, the same experiment as Hibi conducted was repeated, yet with two modifications: Firstly, there were two groups of subjects, musicians (M) and non-musicians (NM). Secondly, in order to reduce the duration of the test, an adaptive stair-case model was used instead of the method of constant stimuli for measuring the JND.

Subjects

There was a total of 34 participants. 18 of them were members of the expert listening panel [28] of the Institute of Electronic Music and Acoustics. A questionary prior to the experiment was used to divide them into the groups “musicians” (M) and “non-musicians” (NM). This questionary can be found in Appendix A. As a consequence of a poor reliability value (see section “data analysis”), four members of the “musicians” group were excluded from the analysis, resulting in a total number of 19 members in group M and 11 in group NM, respectively.

Stimuli

The basic, non distorted sequences consisted of 15 uniformly spaced tone bursts. Each tone burst was 1 kHz in frequency with a duration of 5 ms. To avoid clicks during playback it was filtered with a second order Butterworth lowpass filter with a cutoff frequency of 2 kHz. The time intervals T between each onset (IOI) were 150, 200, 250, 300, 350, 400, 450, 500, 550 and 600 ms. To indicate the start of a new stimulus, the first tone burst was shifted one octave higher.

The two distorted types were created by lengthening (resp. shortening) one interval of the basic sequence by the duration ΔT. The distorted interval appeared randomly either between the 7th and 8th, 8th and 9th or 9th and 10th position. This prevents a subject’s anticipation to the position of the distortion. The amount of distortion depended on the adaptive staircase procedure with values between 0 % and 15 % of the base interval T. The subsequent tone bursts were shifted backwards (resp. forwards), as depicted in Fig. 4.7 on page 39.

Procedure

The experiment was divided into 11 blocks, one for each of the ten IOI plus an additional 350 ms run at the beginning of the experiment for familiarization with the procedure. This layout was chosen, so subjects did not have to adapt to a new tempo for each trial. Except for the first test run, the subsequent blocks were presented in a randomized order to avoid position effects. Each block consisted of five trial sets: a basic set of six undistorted sequences, and four adaptive 1-up/1-down staircase procedures: two for each perturbation type (“lengthened” and “shortened”) with initial distortion values 0 % and 15 % of the base interval T, representing start values below and above the perception threshold. For a “detected” response, the distortion value decreased by the step size and increased for a “not-detected” response, respectively. The step size of each staircase was initially 5 % and decreased to 3 %, 2 % and 1 % after each reversal. The staircases were set to terminate after six reversals. Due to the 1-up/1-down method, the staircases converged to the 50 % level, the D_{50} threshold. As in an interleaved staircase procedure, for each trial, one of the five trial sets was chosen randomly and its next trial was presented to the subject.

The experiment was conducted in a sound proof room, where the subjects listened to the stimuli through a loudspeaker. The tone bursts were presented at 86 dB, measured with a continuous 1 kHz sinus tone at the listening position. After a stimulus was presented, the subjects had to state whether or not there was a distortion. To do so, they had to press ‘s’ for “Störung” (german for “distortion”) or ‘k’ for “keine Störung” (german for “no distortion”) on a keyboard. After the forced choice, the next trial was presented. Once a block was completed the subjects were able to have a break before continuing with the next block. The experiment lasted about 45 min.

Data analysis

The basic sets without distortion of the sequences were used as a measure for the subject’s reliability. A subject was considered less reliable the more “detected” responses he or she gave, even though there was no perturbation. A value of 0.78 was chosen as threshold as a compromise between reliable data and a sufficient number of subjects. Subjects with a lower reliability were discarded from the analysis. A profound insight into the effect of that threshold is presented in Appendix B.

A Lilliefors-test was used to test the data for normal distribution. Four out of the 40 combinations of IOI, musicality and type of perturbation were significant. That means that the data does not arise from a normal distribution. Nevertheless, due to its robustness against violations of the normal distribution [4] a three-way ANOVA with repeated measures was performed. The ANOVA had a 2*2*10 design with one between-subjects factor *musicality* with its conditions musician (M) and non-musician (NM), and two within-subjects factors *TOP* (type of perturbation) and *IOI* with their conditions lengthened and shortened, and the 10 inter onset intervals 150, 200, 250, 300, 350, 400, 450, 500, 550 and 600 ms, respectively. Mauchly's sphericity test revealed that for the factor *IOI* sphericity was not given ($p < .001$ with $\epsilon = .598$). Therefore, the Greenhouse-Geisser correction was used for further analysis.

Results

The results are depicted in Figure 4.9. As can easily be seen, the found thresholds for musicians and non-musicians are remarkably different. Members of the group *musicians* perceived way smaller perturbations than non-musicians. This is confirmed by the significance test of the between-subjects effect *musicality* ($F(1) = 86.1$, $p < 0.001$). The average D_{50} value is 5.4 % for musicians (M) and 9.0 % for non-musicians (NM), respectively. The results of the significance tests for within-subjects effects are given in Table 4.1.

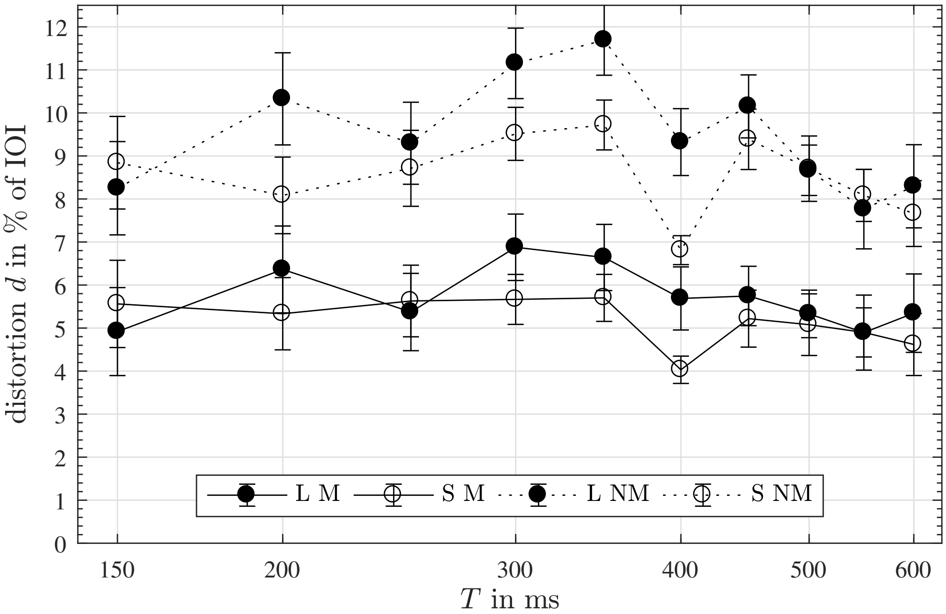


Figure 4.9 Results of the first listening test. D_{50} values as a function of T. The dotted line represents the results of the non-musicians (NM), the solid line those of the musicians (M). Filled circles (●) indicate the condition "lengthened", open circles (○) the condition "shortened" for the type of perturbation.

Table 4.1 Significance tests of within-subjects effects.

Source	*df*	*F*	*p*
IOI	**5.38**	**9.95**	**0.000**
IOI*musicality	9	1.31	0.23
TOP	**1**	**12.54**	**0.001**
TOP*musicality	1	0.89	0.35
IOI*TOP	**9**	**8.61**	**0.000**
IOI*TOP*musicality	9	0.87	0.55

The D_{50} values vary significantly for different IOI values ($F(5.38) = 9.95$, $p < 0.001$), with higher values for medium rates of succession and a decrease towards higher and lower rates. The drop at $T = 400\,\mathrm{ms}$ is an exception for that trend. It appears consistently in all four combinations of the factors *musicality* and *type of perturbation*. A systematic error due to excitation of the room stands to reason, although a measurement yielded not high enough reverberation times for one tone burst to interfere with a successive one. Therefore, the values for this particular IOI should be considered cautiously.

Also the type of perturbation (TOP) showed significant differences. The “shortened” thresholds (average D_{50} value of 6.88 %) were lower than those of the “lengthened” condition (average D_{50} value of 7.6 %). Thus, all main effects showed significant differences.

The IOI*TOP interaction showed significant differences. For the IOIs 200, 300, 350, 400 and 600 ms the difference between the conditions “lengthened” and “shortened” was significant with positive differences, i.e. the thresholds for the “lengthened” condition are higher.

All other interactions (IOI*musicality, TOP*musicality, IOI*TOP*musicality) showed no significant differences.

Discussion

The findings show that the perception of rhythm perturbances indeed depends on the subject’s musical background. Musical training reduces the threshold of perturbance perception from 9.0 % to 5.4 % of the base IOI, on average. Hence the hypothesis, that musicians are more sensitive to rhythm disturbances, is confirmed.

For lower IOIs ($T < 300\,\mathrm{ms}$) the results of the non-musicians are consistent with Hibi’s findings, as Fig. 4.10 shows. For higher IOIs ($T > 300\,\mathrm{ms}$) Hibi’s thresholds diverge from the non-musicians and agree more with the musicians results. Also the results of Halpern's musically trained participants assort well with the latter ones in the overlapping area between 400 and 600 ms.

The absolute amount of distortion ΔT, which is barely perceivable, lies between about 7.4 ms and 32.1 ms for musicians, and 12.3 ms and 49.8 ms for non-musicians, depending on the type of perturbation and the rate of succession. For the algorithm, this means that the aligned start and stop cues should not be set wider apart. Otherwise, a users timing error would not be compensated. However, these values are only valid for isochronous tone bursts and not for musical conditions, which might bring different outcomes.

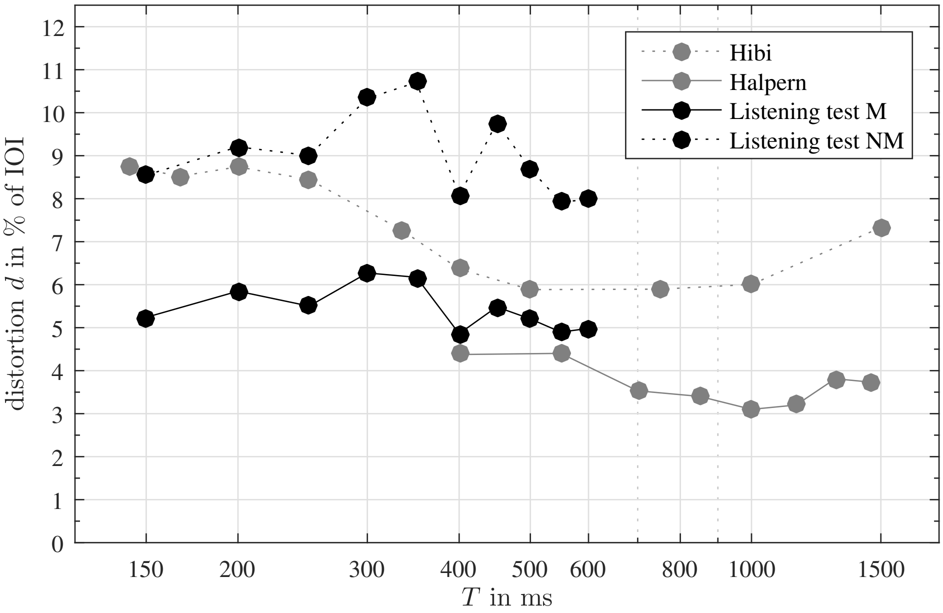


Figure 4.10 Comparison of results of the listening test with those of Hibi and Halpern. D_{50} values as a function of T. The grey lines represent the results of the literature (solid: Halpern, dotted: Hibi). The black lines represent the results of the first conducted listening test (dotted for the group non-musicians, solid for the musicians). For reasons of clarity, the type of perturbation was disregarded by computing the mean.

4.2.4 Listening Test: Impact of the Tatum's Presence

The tone burst stimuli, used in the first listening test, acted as sterile instruments, without musical aspects, used to investigate the human ability of detecting rhythm perturbances. For the validation, it is of interest, whether these findings would be different within a musical context, in which two successive onsets usually vary in intensity, shape and timbre. Due to different rhythms, syncopation and emphasis on certain beats (on-beat, off-beat), the underlying tatum can be less apparent.

The effect of the tatum presence on the detection threshold was investigated in the second listening test. To create a musical context, a synthetic drum beat was chosen as a stimulus archetype due to its advantages of being both repeatable and manipulable. The level of the hi-hat semiquaver notes L_{16th} was used to make the tatum more and less present. In contrast to the first listening test, a fixed tempo was used (100 bpm). Nevertheless, due to different L_{16th} levels, the hi-hat tatum is either 300 ms (without semiquavers) or 150 ms (with semiquavers) with a higher or lower presence depending on the level. A more precise stimulus description is given below (section "stimuli").

The listening test was divided into two methods. One, creating a scenario which could occur while using a guitar looper, in which the position of a possible distortion would appear after every iteration, and one with a random position, providing results which are comparable to those of the first listening test.

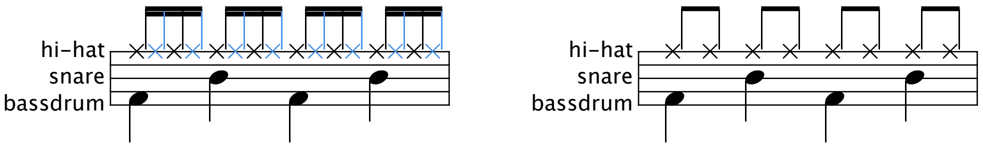
Subjects

There was a total of 28 participants. 17 of them were members of the expert listening panel [28] of the Institute of Electronic Music and Acoustics. The same questionary as in the first listening test was used to divide them into the groups "musicians" (M) and "non-musicians" (NM). As only the musicians group is of interest, the data of the five non-musicians was discarded. Also six subjects with a poor reliability value (see section "data analysis") were excluded from the analysis. Thus, the number of the remaining participants was 17. Eight of them also participated in the first listening test.

Stimuli

A synthetic drumbeat in four-four time served as the basic sequence. The bass drum was played on the counts 1 and 3, the snare drum on 2 and 4. The hi-hat played semiquaver notes. The score of the drumbeat is depicted in Fig. 4.11a. The level L_{16th} of every other semiquaver note (beginning with the second one, indicated blue) was used as an independent variable with the values $-\infty$, -20, -10 and 0 dB. With $L_{16th} = 0\,\mathrm{dB}$ the hi-hat figure consists of even loud semiquavers. Lowering the level results in a more naturally accentuated drumbeat, and with $L_{16th} = -\infty\,\mathrm{dB}$ the semiquavers turn to quavers, as can be seen in Fig. 4.11b. This means that the hi-hat IOI doubles. The played tempo was constant with 100 bpm, ergo the IOI for semiquavers was 150 ms, for quavers 300 ms and for crotchets 600 ms.

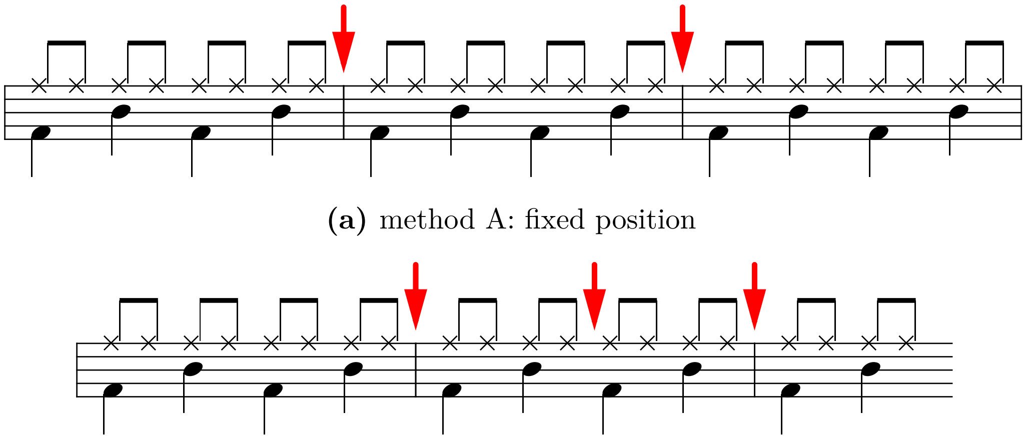
The distorted stimuli were created by lengthening (resp. shortening) one interval of the basic stimulus by the duration ΔT. The amount of distortion depended on



(a) basic drumbeat **(b)** drumbeat with $L_{16th} = -\infty\,\mathrm{dB}$

Figure 4.11 Scores of the drumbeats used in the second listening test. a) shows the basic hi-hat figure: the level of every other note (blue) was adjusted (L_{16th}) b) example with $L_{16th} = -\infty\,\mathrm{dB}$.

the adaptive staircase procedure with values between 0 ms and 50 ms (corresponding to relative values of 0 % and 16.7 % for $T = 300\,\mathrm{ms}$). The subsequent sequence was shifted backwards (resp. forwards). The position of the distortion depended on the method. For method A, the distortion occurred after each bar, meaning with a stimulus length of three bars, the distortion occurred two times, as depicted in Fig. 4.12a. This represents a loop scenario assuming the looped phrase had the length of one bar. For method B, the distortion occurred randomly either at the first or third count in bar 2 or the first count in bar 3 (Fig. 4.12b), preventing the subjects anticipation to the position of the distortion. The length of the sequences for method B was 2.5 bars.



(a) method A: fixed position

(b) method B: random position

Figure 4.12 Positions of the distorted interval of the second listening test.

The stimuli were created by convolving impulse sequences with real drum samples. The distortion took place prior to the convolution, to prevent audio glitches in the drum samples.

Procedure

The experiment was divided into the two methods A and B, which chronological order changed for every subject to avoid position effects. Each method was divided into four blocks, one for each L_{16th} value. Also the blocks were presented in a randomized order to avoid position effects. Each block consisted of five trial sets: a basic set of eight undistorted sequences, and four adaptive 1-up/1-down staircase procedures: two for each perturbation type ("lengthened" and "shortened") with initial distortion values 0 ms and 50 ms, representing start values below and above the perception threshold.. For a "detected" response, the distortion value decreased by the step size and increased for a "not-detected" response, respectively. The step size of each staircase was initially 20 ms and decreased to 10 ms, 8 ms and 5 ms after each reversal. The staircases were set to terminate after five reversals. Due to the 1-up/1-down method, the staircases converged to the 50 % level, the ΔT_{50} threshold. As in an interleaved staircase procedure, for each trial, one of the five trial sets was chosen randomly and its next trial was presented to the subject.

The experiment was conducted in the *Produktionsstudio* of the IEM, where the subjects listened to the stimuli through headphones. The stimuli were presented at $L_{Aeq} = 70\,\text{dB}$ (measured with a dummy head). After a stimulus was presented, the subjects had to response whether or not there was a distortion. To do so, they had to press 's' for "Störung" (german for "distortion") or 'k' for "keine Störung" (german for "no distortion") on a keyboard. After the forced choice, the next trial was presented. In between the two methods the subjects were able to have a break before continuing with the next one. The experiment lasted about 45 min.

Data analysis

The basic sets without distortion of the sequences were used as a measure for the subject's reliability. A subject was considered less reliable the more "detected" responses he or she gave, even though there was no perturbation. A value of 0.82 was chosen as threshold as a compromise between reliable data and a sufficient number of subjects. Subjects with a lower reliability were discarded from the analysis. A profound insight into the effect of that threshold is presented in Appendix B.

A Lilliefors-test was used to test the data for normal distribution. Three out of the 16 combinations of L_{16th}, *method* and *type of perturbation* were significant. This means that the data did not arise from a normal distribution. Nevertheless, due to its robustness against violations of the normal distribution [4] a three-way ANOVA with repeated measures was performed.

The ANOVA had a 2*2*4 design with three within-subjects factors *method* (A and B), *TOP* (type of perturbation) and L_{16th} with their conditions lengthened and shortened, and the 4 semiquaver levels $-\infty$, -20, -10 and 0 dB, respectively. Mauchly's sphericity test revealed that for the factor L_{16th} sphericity was not given ($p < .05$ with $\epsilon = .644$). Therefore, the Greenhouse-Geisser correction was used for further analysis.

Results

Figure 4.13 shows the results of the second listening test. The outcomes of the significance tests are listed in table 4.2. In general, the found ΔT_{50} thresholds for method A (average: 14.87 ms) were lower than those of method B (average: 16.23 ms). This is confirmed by the significance test of the within-subjects effect *method* ($F(1) = 4.84$, $p = 0.043$). The data has the tendency to decrease towards higher L_{16th} values. The associated effect L_{16th} is also significant ($F(1.93) = 18.82$, $p < 0.001$) and shows significant differences for all pairwise comparisons except for the combinations -20 and 0 dB, and -20 and -10 dB.

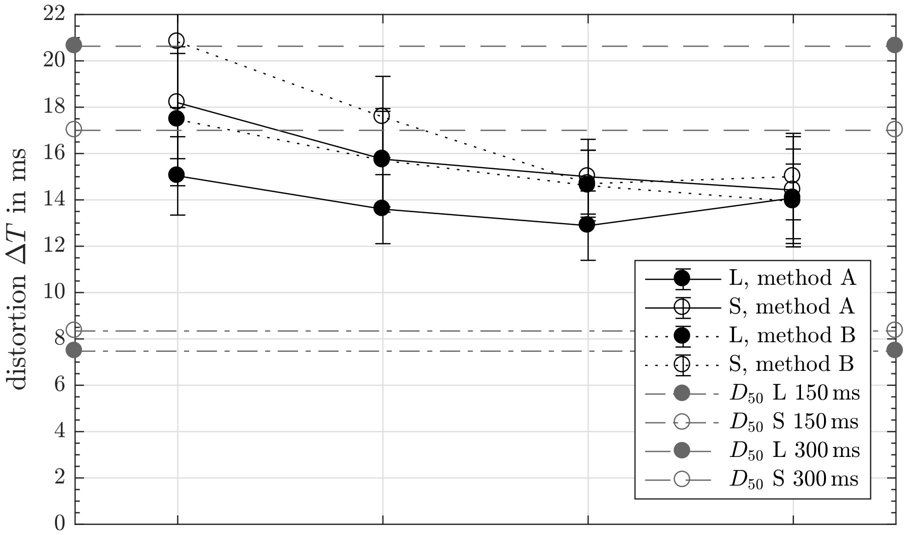


Figure 4.13 Results of the second listening test. ΔT_{50} values in as a function of L_{16th}. The solid line represents the results for method A, the dotted line these for method B. Filled circles (●) indicate the condition "lengthened" (L), open circles (○) the condition "shortened" (S) for the type of perturbation. The horizontal lines represent the ΔT_{50} values of the first listening test for $T = 150$ ms (dashed line) and $T = 300$ ms (dash-dot line).

There was also one significant interaction *method**L_{16th} ($F(3) = 4.49$, $p = 0.007$). A posttest revealed, that the measures for two methods differ significantly for $-\infty$ and -20 dB, whereas for -20 and -10 dB no significant differences were found. Looking at both methods separately, pairwise comparisons of different L_{16th} values yield the following results:

- **Method A** The levels -20, -10 and 0 dB can be grouped together and do not differ significantly within the group but to the level $-\infty$ dB.
- **Method B** The levels -10 and 0 dB can be grouped together and do not differ significantly within the group but to the levels $-\infty$ and -20 dB, which also differ significantly from each other.

Like in the first listening test, the factor *type of perturbation* is also significant in the second one. In contrast to the first listening test, the *lengthened* condition with a total mean of 14.67 ms lies below the *shortened* condition (average: 16.44 ms).

Except for the above mentioned *method**L_{16th} interaction, no interaction showed significant differences.

Table 4.2 Significance tests of within-subjects effects.

Source	*df*	*F*	*p*
method	**1**	**4.84**	**0.043**
L$_{16th}$	**1.93**	**18.82**	**0.000**
TOP	**1**	**8.12**	**0.012**
method*L$_{16th}$	**3**	**4.49**	**0.007**
method*TOP	1	0.13	0.727
L_{16th}*TOP	3	2.22	0.098
method*L_{16th}*TOP	3	1.11	0.356

Discussion

The two horizontal lines in Fig. 4.13 indicate the results of the first listening test for $T = 300$ ms. Despite a switched *type of perturbation* behavior, these thresholds seem to agree with the found thresholds for method B with a semiquaver level of $L_{16th} = -\infty$ dB (i.e. hi-hat IOI of 300 ms). Then again, with $L_{16th} = 0$ dB the hi-hat IOI halves to $T = 150$ ms, resulting in lower ΔT_{50} values. However, the findings of the first listening test for the 150 ms IOI are substantially lower with 7.5 ms for the condition *lengthened* and 8.3 ms for *shortened*.

This could be explained by the difference in the used stimuli. Whereas in the first listening test the tone bursts had very pronounced onsets, those of the hi-hat in the second listening test were less pronounced and interfered by the bass-drum and snare-drum samples, which occurred simultaneously. The resulting blurred onsets might be responsible for the distortions to be less perceptible and also for the switched *type of perturbation* behavior.

The matter of fact that the thresholds of method A are lower than those of method B is not surprising due to expectancy and the possibility to listen to the distortion again, as it occurred two times at fixed positions in the stimuli. As well as with method B, the findings of method A do not fully agree with the results of the first listening test.

The presence of the tatum decreases towards lower semiquaver levels, starting with its maximum at $L_{16th} = 0\,\mathrm{dB}$. At $L_{16th} = -\infty\,\mathrm{dB}$, the presence jumps to its maximum again with a doubled tatum value. The question, whether or not the presence on the tatum has an impact on the perception of rhythm perturbances, can not be fully answered. The significant dependency on L_{16th} could imply, that the presence of the tatum has indeed an impact on the perception of rhythm perturbances. However, looking at the translation between L_{16th} and the presence of the tatum shows different results for the two methods A and B. As described in the section "results", the found thresholds for **method A** show no significant differences for the levels -20, -10 and $0\,\mathrm{dB}$. However, this is the range where the presence changes. Hence, it can be concluded that there is no dependency on the tatum's presence. For **method B** there is a significant difference between the $-20\,\mathrm{dB}$ level and both the -10 and $0\,\mathrm{dB}$ levels, but not between the latter ones. Therefore, there is a dependency on the tatum's presence, even if a small decrease ($-10\,\mathrm{dB}$) does not show significant changes in the thresholds, whereas a bigger one ($-20\,\mathrm{dB}$) does.

For the purpose of this thesis, the gathered results are sufficient and can be used for the validation of the algorithm. However, to prove the explanation attempts above, more listening tests are needed. It would be of interest, to investigate further with crossings of both listening tests. The following list of possible tests comes to mind:

- A repetition of the first listening test with a different level for every other tone burst as a better way to investigate the impact of the tatum's presence.

- Replacing the tone bursts with hi-hat samples in the first and the just described listening test, to investigate the influence of less pronounced onsets.

- Adding bass-drum and snare or even more instruments (like electrical bass) to look into the effect of a musical context.

4.2.5 Summary of Results

Both listening tests yielded results, which give an insight into the perception of rhythm perturbances. A summary of the results, especially of those which have an influence on the algorithm's requirements for validation are listed below:

- The threshold of detecting a rhythm distortion depends on
  - the musical training.
  - the IOI of the tatum.
  - the type of perturbation.
- The dependency on the tatum's presence could neither be fully confirmed nor ruled out. Assuming it being given yields a generally applicable validation.
- Musical context increases the detection threshold for specific IOI, making rhythmic distortions less perceivable.
- The repetition-based nature of a looper lowers the detection threshold of rhythmic distortions in a sequence.

4.3 Validation of the Algorithm

The above gathered information was used to validate the algorithm's performance regarding the reduction of perceivable gaps induced due bad timing of the start and stop cues. Actually, for a seamless performance these cues do not have to sit perfectly on the same musical measure. It is sufficient when both cues have the same offset to an arbitrary measure. As a consequence, only the difference of both offsets is used as a measure for the gap. The results of the two listening tests hold information about the perceptibility of this gap, whether positive (lengthened) or negativ (shortened).

4.3.1 Method

To validate the algorithm in an analytical and ecological valid way, the used signal has to be both reproducible and within a musical context. Consequently, a live played phrase would be unsuitable as parameters like rhythm fluctuation cannot be controlled and the right start and stop cues are difficult to determine, even though the signal would be perfect for the purpose of musical context.

As a compromise, the drum beats of the second listening test were used. The advantages are:

- easy reproducibility
- control of parameters (rhythm fluctuation and presence of the tatum)
- availability of suitable thresholds (listening test)
- start and stop cues are easy to determine
- musical context.

Rhythm Fluctuation is realized with a 1/f noise jitter as described in Section 3.4.4, with standard derivation values between 0 and 10 ms. The *tatum presence* was adjusted with the L_{16th} level analogous to the second listening test (Section 4.2.4).

For different combinations of the above mentioned parameters, the algorithm processes audio files with pre-defined and perfectly timed start and stop cues. The alteration of these cues leads to a gap introduced by the algorithm, which serves as a measure for performance as it shows how well the beats can be aligned at best.

4.3.2 Results

The validation's results are depicted in Table 4.3. The data shows the gap introduced by the algorithm. The smaller the values, the better the algorithm's performance. Negative values indicate a negative gap, meaning the interval was shortened (otherwise lengthened). For $L_{16th} = -30\,\mathrm{dB}$ and $L_{16th} = -\infty\,\mathrm{dB}$ the tatum found was 300 ms corresponding to the quaver notes in the drumbeat. As can be seen for no introduced jitter ($\sigma = 0$) the algorithm creates a gap, especially when the tatum doubles. Nevertheless, when comparing to the results gathered in the second listening test, all gaps are smaller then the found thresholds of perception. Even the more strict criteria of the first listening test is satisfied.

Table 4.3 Results of validation of the algorithm. Values are given in ms and show the resulting gaps introduced by the algorithm. Marked (*) L_{16th} levels indicate the algorithm finding the tatum for quaver notes (300 ms) instead of semiquavers (150 ms).

σ 1/f noise STD in samples / ms	L_{16th} in dB				
	0	-10	-20	-30*	-∞*
0 / 0	0.23	0.29	0.44	1.92	1.88
100 / 2.08	0.17	0.15	0.42	0.10	1.54
200 / 4.17	-0.56	1.02	2.96	-0.75	1.23
300 / 6.25	-3.40	-1.73	-2.06	2.65	0.23
400 / 8.33	-6.31	1.21	0.85	-1.56	-4.35
500 / 10.42	5.02	4.25	1.00	0.73	-2.44

Chapter 5

Conclusio

The beat-tracking algorithm described in Chapter 3 embodies an extension to guitar loop pedals, which supports the musician's performance. Possible audible gaps due to bad timing get corrected by quantizing the start and stop cues of the recorded phrase to the estimated beats. This lowers the temporal extent of the gaps below the threshold of audibility, which was found with the listening tests described in Chapter 4. Including phase information within the beat estimation process leads to an increased precision and enables real-time capability.

Side Benefit of Beat Estimation

With knowing the locations of not only the beats at the beginning and the end but for the whole recorded phrase, the looper can be extended with a rhythm section accompanying the musician with exactly the tempo recorded. This can be done in a simple way by treating every beat as a semiquaver note and adding bass drum, snare drum and hi-hats to corresponding musical measures (for example as in Fig. 4.11b on page 48). A more sophisticated method could be realized by including an analysis of the characteristics of the recorded phrase in terms of syncopation, swing feel or emphasis on specific musical measures. With this information an individual drum track could be created which adopts to the musicians style. Also with an estimation for the time signature, the musician would not have to set it beforehand but can start recording and gets accompanied within seconds after pressing the stop button.

This can be further augmented by analyzing the tonality of the phrase in order to create a bass track perfectly tailored to rhythm and harmony, thus developing a whole accompaniment for practice or live-performance purposes.

Bibliography

[1] Baumgärtel, T. (2016). *Schleifen: Zur Geschichte und Ästhetik des Loops.* Kulturverlag Kadmos.

[2] Bilmes, J. (1993). *Timing is of the Essence: Perceptual and Computational Techniques for Representing, Learning, and Reproducing Expressive Timing in Percussive Rhythm.* PhD thesis, Massachusetts Institute of Technology, Program in Media Arts & Sciences.

[3] Böck, S., Krebs, F., and Schedl, M. (2012). Evaluating the Online Capabilities of Onset Detection Methods. *ISMIR.*

[4] Bühner, M. and Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler.* Pearson Deutschland GmbH.

[5] Davies, M. E. and Plumbley, M. D. (2007). Context-dependent beat tracking of musical audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1009–1020.

[6] Deleuze, G. (1994). *Difference and Repetition.* Columbia University Press.

[7] Drake, C. and Botte, M.-C. (1993). Tempo sensitivity in auditory sequences: Evidence for a multiple-look model. *Perception & Psychophysics*, 54(3):277–286.

[8] Fechner, G. T. (1860). *Elemente der Psychophysik.* Breitkopf und Härtel.

[9] Franěk, M., Mates, J., Radil, T., Beck, K., and Pöppel, E. (1991). Finger tapping in musicians and nonmusicians. *International Journal of Psychophysiology*, 11(3):277–279.

[10] Friberg, A. and Sundberg, J. (1995). Time discrimination in a monotonic, isochronous sequence. *The Journal of the Acoustical Society of America*, 98(5):2524–2531.

[11] Gescheider, G. A. (1985). *Psychophysics; method, theory, and application.* Erlbaum, second edition.

[12] Goldstein, E. (2009). *Sensation and Perception.* Cengage Learning.

[13] Grob, M. (2009). Live Looping - Growth due to limitations. http://www.livelooping.org/history_concepts/theory/growth-along-the-limitations-of-the-tools/. Last accessed on Jan 11, 2017.

[14] Halpern, A. R. and Darwin, C. J. (1982). Duration discrimination in a series of rhythmic events. *Perception & Psychophysics*, 31(1):86–89.

[15] Hennig, H., Fleischmann, R., Fredebohm, A., Hagmayer, Y., Nagler, J., Witt, A., Theis, F. J., and Geisel, T. (2011). The Nature and Perception of Fluctuations in Human Musical Rhythms. *PloS one*, 6(10).

[16] Hennig, H., Fleischmann, R., and Geisel, T. (2012). Musical rhythms: The science of being slightly off. *Physics Today*, 65(7):64.

[17] Hibi, S. (1983). Rhythm perception in repetitive sound sequence. *Journal of the Acoustical Society of Japan (E)*, 4(2):83–95.

[18] Margulis, E. H. (2013). Aesthetic Responses to Repetition in Unfamiliar Music. *Empirical Studies of the Arts*, 31(1):45–57.

[19] Margulis, E. H. (2014). *On repeat: How music plays the mind.* Oxford University Press.

[20] Masri, P. (1996). *Computer modelling of sound for transformation and synthesis of musical signals.* PhD thesis, University of Bristol.

[21] Meyer, M. (1903). Experimental Studies in the Psychology of Music. *The American Journal of Psychology*, 14(3/4):192.

[22] Montag, E. D. (n.d.). Forced choice. http://cis.rit.edu/people/faculty/montag/vandplite/pages/chap_4/ch4p5.html. Last accessed on Dec 21, 2016.

[23] Müller, M., Jiang, N., and Grohganz, H. G. (2014). Sm toolbox: Matlab implementations for computing and enhancing similarity matrices. In *Proceedings of 53rd Audio Engineering Society (AES)*, London, UK.

[24] Peretz, I., Gaudreau, D., and Bonnel, A.-M. (1998). Exposure effects on music preference and recognition. *Memory & Cognition*, 26(5):884–902.

[25] Rammsayer, T. and Altenmüller, E. (2006). Temporal Information Processing in Musicians and Nonmusicians. *Music Perception: An Interdisciplinary Journal*, 24(1):37–48.

[26] Räsänen, E., Pulkkinen, O., Virtanen, T., Zollner, M., and Hennig, H. (2015). Fluctuations of Hi-Hat Timing and Dynamics in a Virtuoso Drum Track of a Popular Music Recording. *PloS one*, 10(6).

[27] Rickard, S. (2011). The beautiful math behind the world's ugliest music. https://www.ted.com/talks/1331. Last accessed on Dez 21, 2016.

[28] Sontacchi, A. and Pomberger, H. (2009). Recruiting and evaluation process of an expert listening panel. *Fortschritte der Akustik, NAG/DAGA, Rotterdam.*

[29] Stowell, D. and Plumbley, M. (2007). Adaptive whitening for improved real-time audio onset detection. In *Proceedings of the International Computer Music Conference (ICMC 2007).*

[30] Weiner, I. B. (2003). *Handbook of Psychology, Experimental Psychology.* John Wiley & Sons.

[31] Westheimer, G. (1999). Discrimination of short time intervals by the human observer. *Experimental Brain Research*, 129(1):121–126.

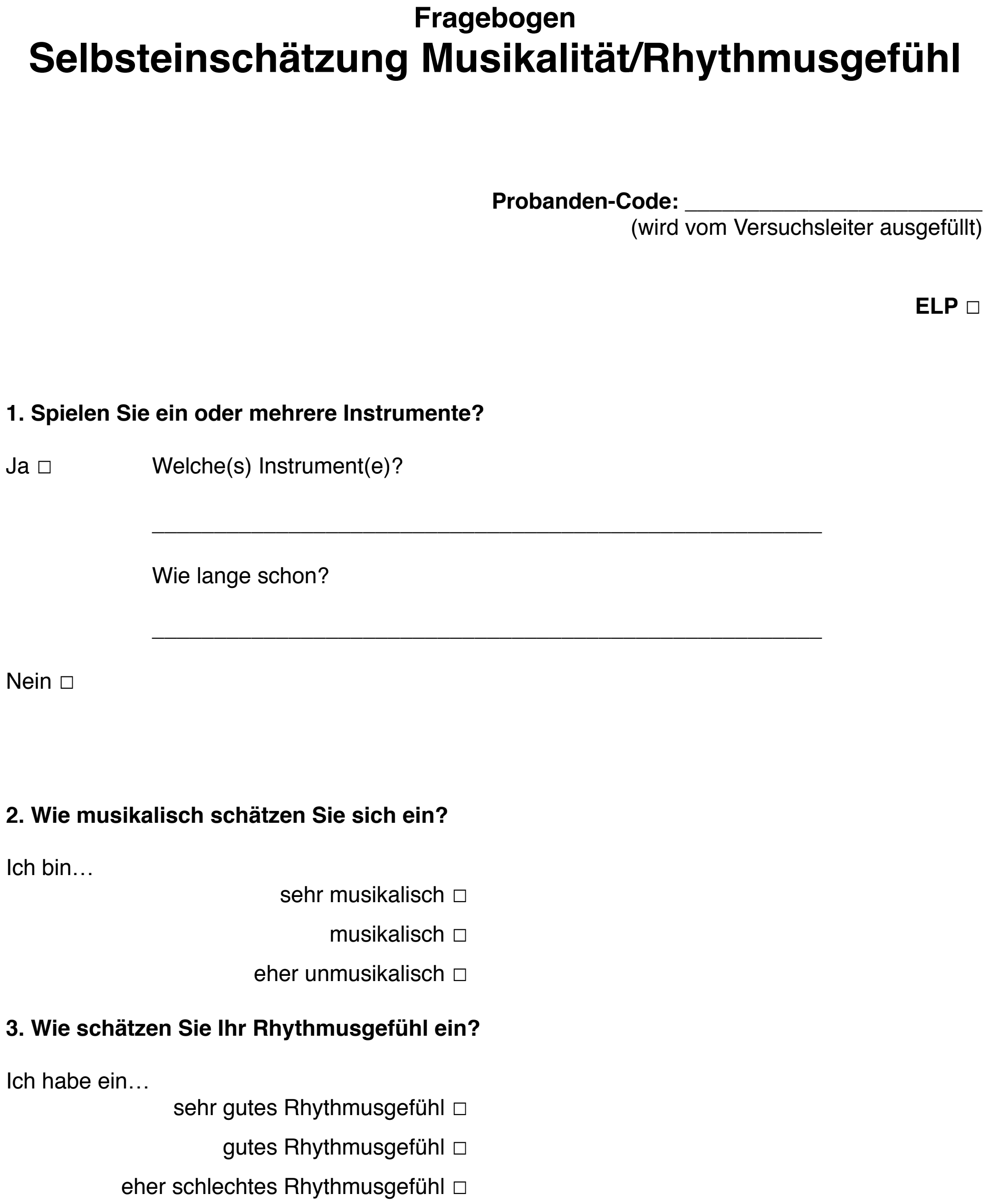
[32] Wu, F., Lee, T. C., Jang, J., Chang, K. K., and Lu, C. H. (2011). A Two-Fold Dynamic Programming Approach to Beat Tracking for Audio Music with Time-Varying Tempo. *ISMIR.*

[33] Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2, Pt.2):1–27.

Appendix A

Questionary

The following questionary was used to divide the subjects into the groups *musicians* and *non-musicians* for the two listenings tests.



Fragebogen

Selbsteinschätzung Musikalität/Rhythmusgefühl

Probanden-Code: ________________________
(wird vom Versuchsleiter ausgefüllt)

ELP □

1. Spielen Sie ein oder mehrere Instrumente?

Ja □ Welche(s) Instrument(e)?

____

Wie lange schon?

____

Nein □

2. Wie musikalisch schätzen Sie sich ein?

Ich bin…

sehr musikalisch □
musikalisch □
eher unmusikalisch □

3. Wie schätzen Sie Ihr Rhythmusgefühl ein?

Ich habe ein…

sehr gutes Rhythmusgefühl □
gutes Rhythmusgefühl □
eher schlechtes Rhythmusgefühl □

Appendix B

Effect of the Data Discard Threshold

Let's assume that every subject holds an individual probability p, with which he or she answers correctly. With a value of $p = 1$, the subject gives only correct and reliable answers. Accordingly, getting presented a stimulus without any distortions, the subject would answer that he or she did not perceive any distortions. On the other hand, with $p = 0.5$ the subject would guess all the time.

Using the binomial distribution $B(n,p)$, the probability of responding correctly for exactly k out of n trials can be calculated:

$$P(X = k) = f(k,n,p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{B.1}$$

With defining a relative threshold r of correctly answered reliability trials, below which the data of a subject gets discarded, the probability of remaining in the analysis can be calculated as follows:

$$P(X > k) = 1 - P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \tag{B.2}$$

with $k = \lfloor rn \rfloor$.

Figure B.1 shows this probability evaluated for different p values and with n and r values corresponding to those of the two listening tests. As can be seen, subjects with a low probability p get discarded more likely from the analysis than those with higher p values, i.e. more reliable subjects.

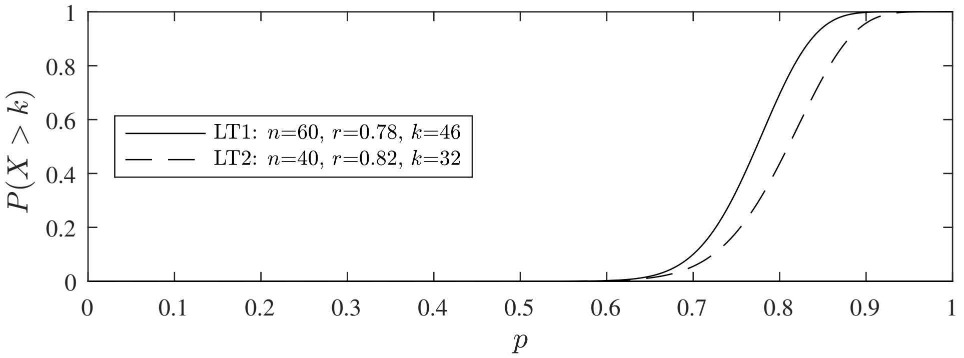


Figure B.1 Probability of remaining in the analysis P(X>k) as a function of p. The solid line represents the results with the parameters n and r of the first listening test, the dashed line the results with those of the second listening test.

The expected value $\mathbf{E}(p)$ can be used as a measure of the mean reliability of the remaining subjects and can be obtained with

$$\mathbf{E}(p) = \frac{\int_0^1 P(X > k) p \, dp}{\int_0^1 P(X > k) \, dp} \tag{B.3}$$

and yields $\mathbf{E}(p) = 0.88$ for the first listening test and $\mathbf{E}(p) = 0.89$ for the second one, respectively. Fig. B.2 depicts the expected value as a function of the threshold r.

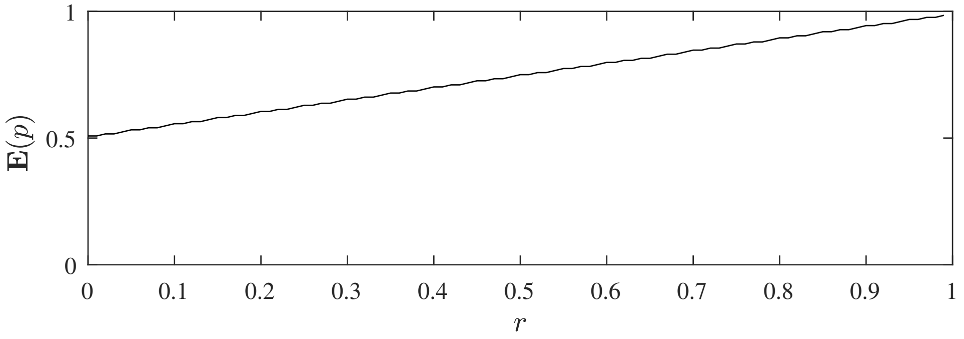


Figure B.2 Expected value $\mathbf{E}(p)$ as a function of the relative threshold r for the parameters of the first listening test. The graph with parameters of the second listening test only differs barely and is not depicted here.

Nevertheless, these expected values are only valid, if the assumed p values occur equally often in the random sample, which - in general - is not the case. The p value depends on factors like concentration, but also of the standard derivation of the momentary threshold, as higher standard derivations lead to a wider psychometrical function and a more random response behavior due to the lower gradient around the guessing probability $P_{``yes"} = 0.5$.