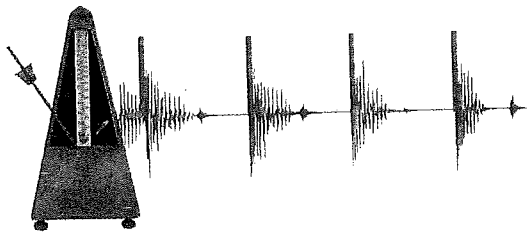
Chapter 6
Tempo and Beat Tracking



Temporal and structural regularities are perhaps the most important incentives for people to get involved and to interact with music. It is the **beat** that drives music forward and provides the temporal framework of a piece of music. Intuitively, the beat corresponds to the pulse a human taps along when listening to music. The beat is often described as a sequence of perceived pulse positions, which are typically equally spaced in time and specified by two parameters: the **phase** and the **period** (see Figure 6.1b). The term **tempo** refers to the rate of the pulse and is given by the reciprocal of the beat period. Tempo and beat are fundamental aspects of music, and the automated extraction of such information from audio recordings constitutes one of the central and well-studied research areas in music processing. In this chapter, we introduce some key techniques used in tempo estimation and beat tracking. Furthermore, we discuss some of the challenges one has to face when dealing with music where certain model assumptions are not fulfilled.

When listening to a piece of music, we as humans are often able to tap along with the musical beat without difficulty—sometimes, we even do this unconsciously. In the case that we lose track at some point in time, maybe because of a tempo change or rhythmic displacement, we are able to recover quickly and resume tapping. However, simulating this cognitive process with an automated beat tracking system is much harder than one may think. Recent beat tracking systems can cope well with modern pop and rock music that has a strong and steady beat. In deriving this information, most systems are based on the assumptions that beats correspond to note onsets (typically percussive in nature) and that beats are periodically spaced in time. However, there are many types of music where these assumptions are violated. For example, in string music a note may be played softly with a barely noticeable onset, or a musician may slightly lengthen certain notes to shape musical phrases. In general, musicians do not play mechanically at a fixed tempo, but slow down or accelerate at certain positions to create tension and release. As a consequence, the

M. Müller, *Fundamentals of Music Processing*,
DOI 10.1007/978-3-319-21945-5_6

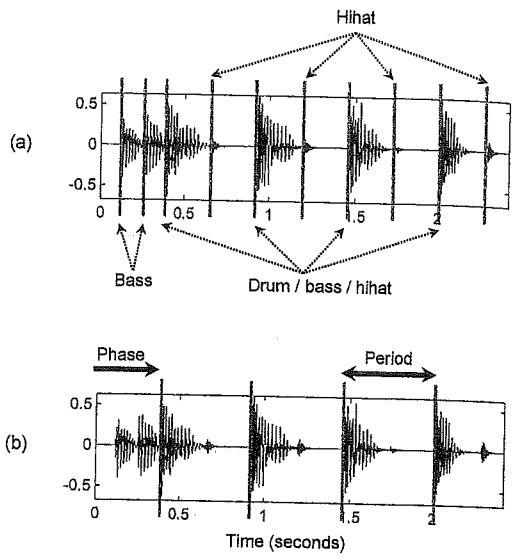


Fig. 6.1 Waveform representation of an excerpt of "Another one bites the dust" by Queen. **(a)** Note onsets. **(b)** Beat positions.

presence of such local tempo changes makes the extraction of beat positions a very challenging task. Still, at least when familiar with the type of music, humans are capable of anticipating local tempo changes and tracking the beats even for highly complex music.

In most approaches to automated tempo and beat tracking, the first step is to estimate the positions of note onsets within the music signal (see Figure 6.1a). This task, which is also referred to as **onset detection**, is discussed in Section 6.1. In particular, we show how to transform a given music signal into a novelty representation that captures certain changes in the signal's energy or spectrum. The peaks of such a representation yield good indicators for note onset candidates. We have seen a similar concept when applying novelty detection to music structure analysis (see Section 4.4). In Section 6.2, we introduce the notion of a tempogram, which represents local tempo information on different pulse levels. Such a time–tempo representation is obtained by analyzing a novelty representation with regard to reoccurring patterns and quasiperiodic pulse trains. In this context, we study two important methods for periodicity analysis, one using Fourier and the other using autocorrelation analysis techniques. We then continue in Section 6.3 with the topic of beat tracking. First, we introduce a mid-level representation that captures meaningful local pulse information even in the presence of significant tempo changes. Then, based on a dynamic programming approach, we discuss a robust beat tracking procedure, which assumes a roughly constant tempo throughout the recording.

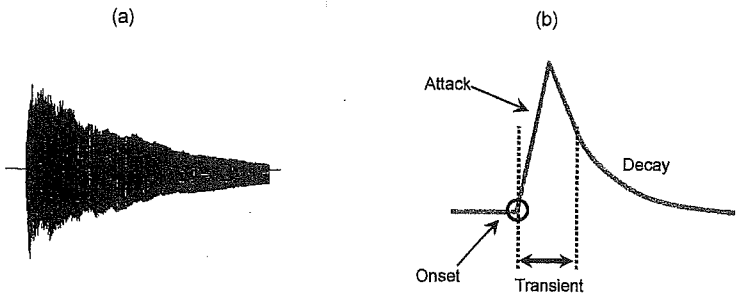


Fig. 6.2 Illustration of attack, transient, onset, and decay of a single note (based on [1]). **(a)** Note played on a piano. **(b)** Idealized amplitude envelope.

6.1 Onset Detection

Generally speaking, **onset detection** is the task of determining the starting times of notes or other musical events as they occur in a music recording. In practice, however, the notion of an onset can be rather vague and is related to other concepts such as attacks or transients. As discussed in Section 1.3.4, there is often a sudden increase of energy at the beginning of a musical tone (see Figure 6.2a). The **attack** of a note refers to the phase where the sound builds up, which typically goes along with a sharply increasing amplitude envelope. This is also reflected by the initial phase of the ADSR model shown in Figure 1.22. The concept of a **transient** is more difficult to grasp. As noted in Section 1.3.4, a transient may be described as a noise-like sound component of short duration and high amplitude typically occurring at the beginning of a musical tone or a more general sound event. However, the release or offset of a sustained note may also contain a transient-like component. In transient regions, the signal evolves quickly in an unpredictable and rather chaotic way. For example, in the case of a piano, the transient corresponds to the initial phase where a key is hit, the damper is raised, the hammer strikes the strings, the strings start to vibrate, and the vibrations are transmitted to the large soundboard that starts resonating to finally yield a steady and sustained sound. As opposed to the attack and transient, the **onset** of a note refers to the single instant (rather than a period) that marks the beginning of the transient, or the earliest time point at which the transient can be reliably detected (see Figure 6.2b).

To detect note onsets in the signal, the general idea is to capture sudden changes that often mark the beginning of transient regions. For notes that have a pronounced attack phase, onset candidates may be determined by locating time positions where the signal's amplitude envelope starts increasing. When this is not the case, such as for nonpercussive music with soft onsets and blurred note transitions, the detection of onsets is much more challenging. For example, the waveform of a violin sound, as shown in Figure 1.23b, may exhibit a slow energy increase rather than an abrupt change as in a piano sound. For soft sounds, it is hard to determine the exact onset

position. The detection of individual note onsets becomes even harder when dealing with complex polyphonic music. Simultaneously occurring sound events may result in masking effects, where no significant changes in the signal's energy are measurable. In these cases, more refined onset detection methods are needed, e.g., by looking at changes in the signal's short-time spectrum or other statistical properties.

In this section, we study four different approaches for onset detection: an energy-based approach (Section 6.1.1), a spectral-based approach (Section 6.1.2), a phase-based approach (Section 6.1.3), and a complex-domain approach (Section 6.1.4). All approaches follow the same algorithmic pipeline, but differ in the signal properties that are exploited to derive onset candidates. In this pipeline, the signal is first converted into a suitable feature representation that better reflects the properties of interest. Then, a type of derivative operator is applied to the feature sequence and a novelty function is derived. Finally, a peak-picking algorithm is employed to locate the onset candidates. Note that this general procedure is exactly the same as for novelty detection in the context of music structure analysis (see Section 4.4.1). However, the features and, in particular, the temporal levels that are relevant in structure analysis and onset detection are quite different. While a tolerance window of 500 ms up to a couple of seconds may be used in the case of structural boundaries, the accuracy needed in onset detection is usually far below 100 ms, sometimes even on the order of 10 ms.[1]

6.1.1 Energy-Based Novelty

We have seen that playing a note on an instrument often coincides with a sudden increase of the signal's energy. For example, this holds when striking a key on a piano, plucking a string on a guitar, or hitting a drum with a stick. Based on this observation, a straightforward way to detect note onsets is to transform the signal into a local energy function that indicates the local energy of the signal for each time instance and then to look for sudden changes in this function. Mathematically, this procedure can be realized as follows: Let x be a DT-signal. As in the case of a discrete STFT (see Section 2.5.3), we fix a discrete window function $w : \mathbb{Z} \to \mathbb{R}$, which is shifted over the signal x to determine local sections. In particular, we assume that w is a bell-shaped function centered at time zero[2] and that $w(m)$ for $m \in [-M : M]$ comprises the nonzero samples of w for some $M \in \mathbb{N}$. The **local energy** of x with regard to w is defined to be the function $\mathrm{E}_w^x : \mathbb{Z} \to \mathbb{R}$ given by

$$\mathrm{E}_w^x(n) := \sum_{m=-M}^{M} |x(n+m)w(m)|^2 = \sum_{m\in\mathbb{Z}} |x(m)w(m-n)|^2 \tag{6.1}$$

[1] This is the range where the human ear is no longer capable of distinguishing between two subsequent transients [52].

[2] In Section 2.5.3, to simplify notation, we considered the noncentered case assuming that the nonzero window coefficients are $w(n)$ for $n \in [0 : N-1]$.

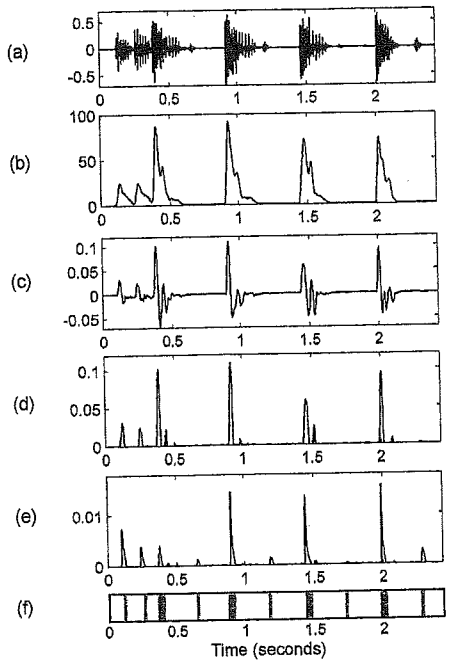


Fig. 6.3 Computation of an energy-based novelty function of the signal from Figure 6.1. **(a)** Waveform. **(b)** Local energy function. **(c)** Discrete derivative. **(d)** Novelty function Δ_{Energy} obtained after half-wave rectification. **(e)** Novelty function $\Delta^{\text{Log}}_{\text{Energy}}$ based on a logarithmic energy function. **(f)** Annotated note onsets (the four beat positions are marked by thick lines).

for $n \in \mathbb{Z}$. In other words, $\mathrm{E}^x_w(n)$ contains the energy (as defined in (2.41)) of the signal x multiplied with a window shifted by n samples. Let us have a look at the example shown in Figure 6.3b, which shows a local energy function for the beginning of "Another one bites the dust" by Queen. Starting with an offbeat consisting of two sixteenth notes played only by bass, four percussive beats (played by kick drum, snare drum, hihat, and bass) follow (see Figure 6.1). Furthermore, between each two subsequent beats, there is an additional hihat stroke. As the energy function shows, the percussive beats contain a lot of energy, whereas the low-energy hihat strokes are not as strongly captured.

Intuitively, to measure energy changes, we take a derivative of the local energy function. In the discrete case, the easiest way to realize such a derivative is to take the difference between two subsequent energy values (see Figure 6.3c). Furthermore, since we are interested in energy increases (and not decreases), we keep only the positive differences while setting the negative differences to zero. The latter step is known as **half-wave rectification** and is notated as:

$$|r|_{\geq 0} := \frac{r+|r|}{2} = \begin{cases} r, & \text{if } r \geq 0, \\ 0, & \text{if } r < 0 \end{cases} \tag{6.2}$$

for $r \in \mathbb{R}$. Altogether, we obtain an **energy-based novelty function** $\Delta_{\text{Energy}} : \mathbb{Z} \to \mathbb{R}$ given by

$$\Delta_{\text{Energy}}(n) := |\mathrm{E}^x_w(n+1) - \mathrm{E}^x_w(n)|_{\geq 0} \tag{6.3}$$

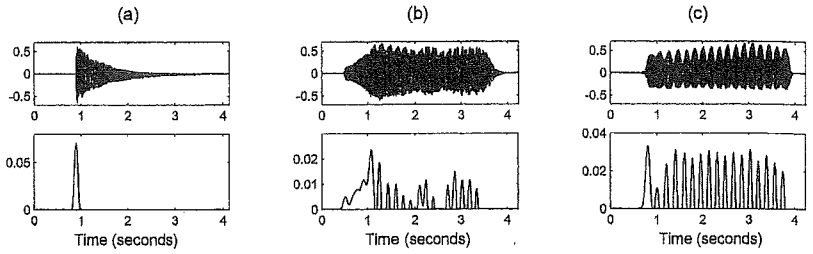


Fig. 6.4 Waveform and energy-based novelty function of the note C4 (261.6 Hz) played by different instruments (see Figure 1.23). **(a)** Piano. **(b)** Violin. **(c)** Flute.

for $n \in \mathbb{Z}$. The resulting function is shown in Figure 6.3d for our example "Another one bites the dust." The four quarter-note drum beats correspond to the four highest peaks. Therefore, these beats can be correctly detected by a simple peak-picking procedure. Also, the two beginning offbeats played by the bass are correctly identified by the first two peaks. However, the four hihat strokes between the beat positions do not show up in Δ_{Energy} (see Figure 6.3d). As mentioned before, these four hihat events contain relatively little energy and, when compared with the high-energy drum events, become invisible in the energy-based novelty function.

As we discussed in Section 1.3.3, the human perception of sound intensity is logarithmic in nature. Therefore, even musical events of rather low energy may still be perceptually relevant. For example, the hihat is clearly audible even at the beat positions where it is overlaid with the strong drum hits. To account for such phenomena, one often applies a logarithm to the energy values, for example, by switching to the logarithmic decibel scale (1.6) or by applying logarithmic compression (3.7). Note that, in the logarithmic case, the resulting novelty function corresponds to (the logarithm of) energy ratios rather than differences as shown by the following equation:

$$\Delta^{\text{Log}}_{\text{Energy}}(n) := |\log(\mathrm{E}^x_w(n+1)) - \log(\mathrm{E}^x_w(n))|_{\geq 0} = \left|\log\left(\frac{\mathrm{E}^x_w(n+1)}{\mathrm{E}^x_w(n)}\right)\right|_{\geq 0}. \quad (6.4)$$

As can be seen in Figure 6.3e, even the weak hihat onsets become visible in the logarithmic novelty function. On the downside, however, the logarithm may also amplify noise-like sound components, possibly leading to spurious peaks.

Another general problem in onset detection is energy fluctuation in nonsteady sounds as a result of vibrato or tremolo (see Section 1.3.4). Especially for purely energy-based procedures, amplitude modulations often lead to spurious peaks in the resulting novelty function. This is demonstrated by Figure 6.4, which shows the energy-based novelty function for the note C4 played by different instruments. While the novelty function shows a single clear peak in the case of a piano sound, there are many additional peaks in the case of a violin or flute sound. Furthermore, the relatively slow energy increase at the beginning of the violin sound leads to a smeared and temporally inaccurate onset peak.

To increase the robustness of onset detection, a typical approach is to first decompose the signal into several subbands that contain complementary frequency information. Then one computes a novelty function for each subband separately and suitably combines the individual functions to derive the onset information. For example, the subbands may correspond to musical pitches as discussed in Section 3.1.1, which results in pitch-based novelty functions. To exploit prior knowledge, one may use broader frequency bands that correspond to typical ranges of musical instruments (see Exercise 1.11). In the next section, we study an approach that decomposes a signal into subbands that correspond to the spectral coefficients. In this case, the resulting novelty function measures spectral changes, which yields more refined information than purely energy-based approaches.

6.1.2 Spectral-Based Novelty

Onset detection becomes a much harder problem for polyphonic music with simultaneously occurring sound events. A musical event of low intensity may be masked by an event of high intensity. Energy fluctuations (e.g., coming from vibrato) in the sustain phase of one instrument may be stronger than energy increases in the attack phase of other instruments. Therefore, in the case of multiple instruments playing at the same time, it is generally hard to detect all onsets when using purely energy-based methods. However, the characteristics of note onsets may strongly depend on the respective type of instrument. For example, for percussive instruments with an impulse-like onset, one can observe a sudden increase in energy that is spread across the entire spectrum of frequencies (see Figure 2.21a). Such noise-like broadband transients may be observable in certain frequency bands even in polyphonic mixtures. In particular, since the energy of harmonic sources is concentrated more in the lower part of the spectrum, transients are often well detectable in the higher-frequency region.

Motivated by such observations, the idea of spectral-based novelty detection is to first convert the signal into a time–frequency representation and then to capture changes in the frequency content. In the following, let $\mathcal{X}$ be the discrete STFT of the DT-signal x as defined in (2.26) or (2.148). For a discussion of the various parameters, including the sampling rate $F_s = 1/T$, the window length N of the discrete window w, and the hop size H, we refer to Section 2.5.3 or Section 3.1.1. For the moment, we only need to keep in mind that $\mathcal{X}(n,k) \in \mathbb{C}$ denotes the k^{th} Fourier coefficient for frequency index $k \in [0:K]$ and time frame $n \in \mathbb{Z}$, where $K = N/2$ is the frequency index corresponding to the Nyquist frequency.

To detect spectral changes in the signal, one basically computes the difference between subsequent spectral vectors using a suitable distance measure. This results in a **spectral-based novelty function**, which is also known as the **spectral flux**. There are many different ways of computing such a novelty function, which depend not only on the parameters of the STFT and the distance measure, but also on pre- and postprocessing steps that are often applied.

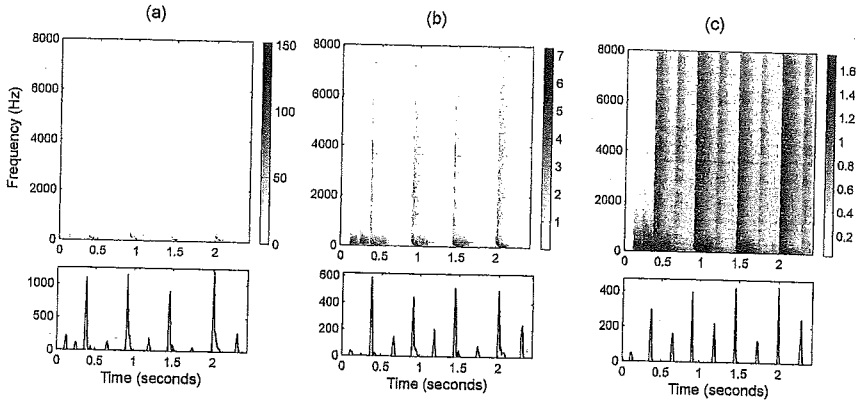


Fig. 6.5 Logarithmic compression (using the same audio excerpt as in Figure 6.1). The figure shows the respective magnitude spectrogram (top) and the resulting novelty function Δ_{Spectral} (bottom). **(a)** Magnitude spectrogram. **(b)** Compressed spectrogram using $\gamma = 1$. **(c)** Compressed spectrogram using $\gamma = 1000$.

In the following, we describe a typical procedure. First, to enhance weak spectral components, we apply a **logarithmic compression** to the spectral coefficients. Such a step, as we have already encountered in the context of chroma features (Figure 3.7), is often applied to account for the logarithmic sensation of sound intensity and to balance out the dynamic range of the signal. To obtain the compressed spectrogram, we apply the function Γ_γ of (3.7) to the magnitude spectrogram $|\mathcal{X}|$. This yields

$$\mathcal{Y} := \Gamma_\gamma(|\mathcal{X}|) = \log(1 + \gamma \cdot |\mathcal{X}|) \tag{6.5}$$

for a suitable constant $\gamma \geq 1$. In onset detection, logarithmic compression is particularly helpful for enhancing the comparatively weak high-frequency information. This is also illustrated by Figure 6.5, which continues our example "Another one bites the dust" from Figure 6.1. In the visualization of the original spectrogram $|\mathcal{X}|$ (Figure 6.5a), the harmonic components of the bass are visible in the low-frequency part. However, the transients at the beat positions can hardly be recognized. Using a compressed spectrogram with $\gamma = 1$ (Figure 6.5b), the vertical structures of the transients become more prominent—even the weak transients of the hihat between subsequent beats become visible. By increasing γ, the low-intensity values are further enhanced. On the downside, a large compression factor γ may also amplify nonrelevant noise-like components.

In the next step, we compute the discrete temporal derivative of the compressed spectrum $\mathcal{Y}$. Similarly to the energy-based novelty function, we only consider the positive differences (increase in intensity) and discard negative ones. This yields the spectral-based novelty function $\Delta_{\text{Spectral}} : \mathbb{Z} \to \mathbb{R}$ defined by

$$\Delta_{\text{Spectral}}(n) := \sum_{k=0}^{K} |\mathcal{Y}(n+1,k) - \mathcal{Y}(n,k)|_{\geq 0} \tag{6.6}$$

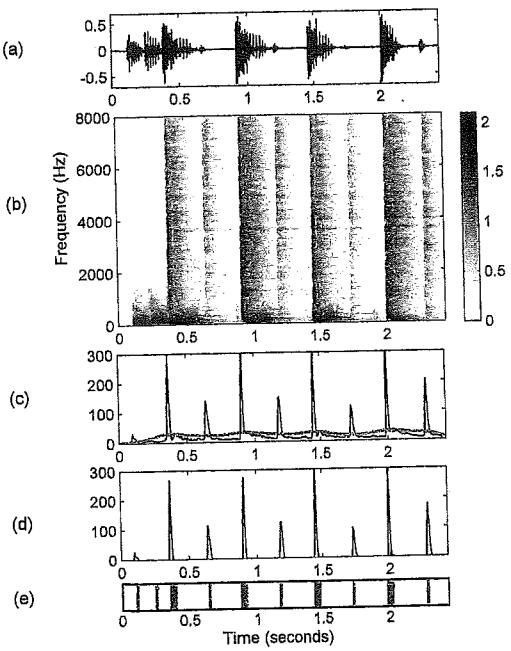


Fig. 6.6 Computation of a spectral-based novelty function for the signal from Figure 6.1. **(a)** Waveform. **(b)** Compressed spectrogram using $\gamma = 100$. **(c)** Novelty function Δ_{Spectral} and local average function μ (in thick/red). **(d)** Novelty function $\bar{\Delta}_{\text{Spectral}}$. **(e)** Annotated note onsets (the four beat positions are marked by thick lines).

for $n \in \mathbb{Z}$, where we use the half-wave rectification as introduced in (6.2). One can further enhance the properties of the novelty function by applying suitable postprocessing steps. For example, in view of a subsequent peak-picking step, one objective may be to enhance the peak structure of the novelty function, while suppressing small fluctuations. To this end, we introduce a local average function $\mu : \mathbb{Z} \to \mathbb{R}$ by setting

$$\mu(n) := \frac{1}{2M+1} \sum_{m=-M}^{M} \Delta_{\text{Spectral}}(n+m), \tag{6.7}$$

$n \in \mathbb{Z}$, where the parameter $M \in \mathbb{N}$ determines the size of an averaging window. The enhanced novelty function $\bar{\Delta}_{\text{Spectral}}$ is obtained by subtracting the local average from Δ_{Spectral} and by only keeping the positive part (half-wave rectification):

$$\bar{\Delta}_{\text{Spectral}}(n) := \big|\Delta_{\text{Spectral}}(n) - \mu(n)\big|_{\geq 0} \tag{6.8}$$

for $n \in \mathbb{Z}$. Figure 6.6 illustrates the computational pipeline by means of our running example. As opposed to the energy-based novelty functions (Figure 6.3), the enhanced spectral-based novelty function $\bar{\Delta}_{\text{Spectral}}$ (Figure 6.6d) not only indicates the onsets at the four beat positions, but also has significant peaks at the four weak hihat onsets between the beats. Even though the hihat sounds have a comparatively low intensity, they produce sharp transients, which are captured well by the compressed magnitude spectrogram (see also Figure 6.5c).

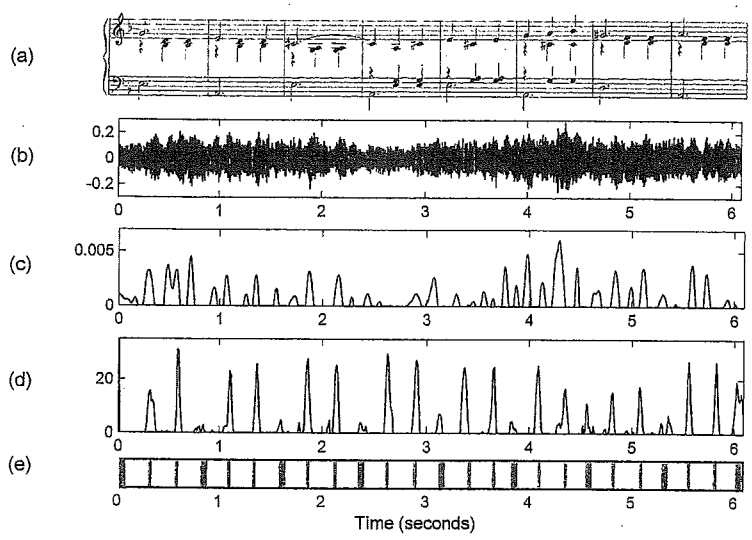


Fig. 6.7 Different novelty functions for an audio excerpt of Shostakovich's Waltz No. 2 from the "Suite for Variety Orchestra No. 1." **(a)** Score representation (in a piano reduced version). **(b)** Waveform. **(c)** Energy-based novelty function. **(d)** Spectral-based novelty function. **(e)** Annotated note onsets (downbeat positions are marked by thicker lines).

As a second example, let us have a look at an excerpt of an orchestra recording of the Waltz No. 2 from Dimitri Shostakovich's Suite for Variety Orchestra No. 1, an example we have already used in Figure 4.11. The first beats (downbeats) of the 3/4 meter are played softly by nonpercussive instruments, leading to relatively weak and blurred onsets. In contrast, the second and third beats are played sharply ("staccato"), supported by percussive instruments. These properties are also reflected by the spectral-based novelty function shown in Figure 6.7d. The peaks that correspond to downbeats are hardly visible or even missing, whereas the peaks that correspond to the percussive beats are much more pronounced. The figure also shows the improvements one obtains for this example when using spectral-based methods (Figure 6.7d) compared with purely energy-based methods (Figure 6.7c).

As said before, there are many more approaches for computing spectral-based novelty functions. For example, as with the energy-based case, it may be beneficial to first split up the spectrum into several frequency bands (often five to eight logarithmically spaced bands are used). The resulting bandwise novelty functions are then weighted and summed up to yield the single overall novelty function (see Exercise 6.4).

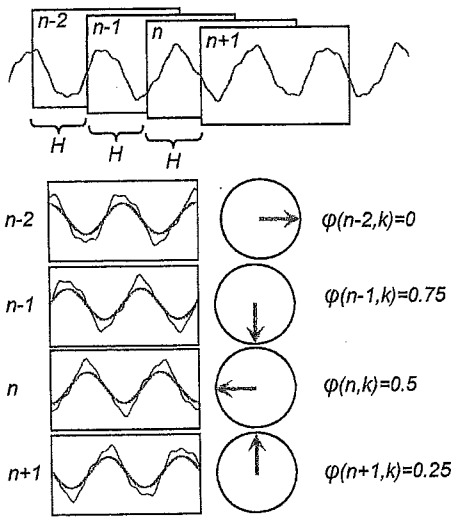


Fig. 6.8 Locally stationary signal and its correlation to a sinusoid corresponding to frequency index k for the frames $n-2$, $n-1$, n, and $n+1$. The angular representation of the phases is indicated by the circles.

6.1.3 Phase-Based Novelty

In the definition of the spectral-based novelty function, we have only used the magnitude of the spectral coefficients. However, the phases of the complex coefficients are also an important source of information for various audio analysis and synthesis tasks. In the following, we show how the phase information can be used for onset detection. In particular, we exploit the fact that stationary tones have a stable phase, while transients have an unstable phase. For another application of the phase information, we refer to Section 8.2.1.

As before, let $\mathcal{X}(n,k) \in \mathbb{C}$ be the complex-valued Fourier coefficient for frequency index $k \in [0:K]$ and time frame $n \in \mathbb{Z}$. Using the polar coordinate representation (2.9), this complex coefficient can be written as

$$\mathcal{X}(n,k) = |\mathcal{X}(n,k)| \exp(2\pi i \varphi(n,k)) \tag{6.9}$$

with the phase $\varphi(n,k) \in [0,1)$ (see also Section 2.3.2.2). Intuitively, as we explained in Section 2.1.1.1, the phase $\varphi(n,k)$ determines how the sinusoid of frequency $F_{\text{coef}}(k) = F_s \cdot k/N$ (see (2.28)) has to be shifted to best correlate with the windowed signal corresponding to the n^{th} frame. Let us assume that the signal x has a high correlation with this sinusoid (i.e., $|\mathcal{X}(n,k)|$ is large) and shows a steady behavior in a region of a number of subsequent frames $\ldots, n-2, n-1, n, n+1, \ldots$ (i.e., x is locally stationary). Then the phases $\ldots, \varphi(n-2,k)$, $\varphi(n-1,k)$, $\varphi(n,k)$, $\varphi(n+1,k), \ldots$ increase from frame to frame in a fashion that is linear in the hop size H of the STFT (see Figure 6.8). Therefore, the frame-wise phase difference in this region remains approximately constant (possibly up to some integer, as we discuss shortly in this section):

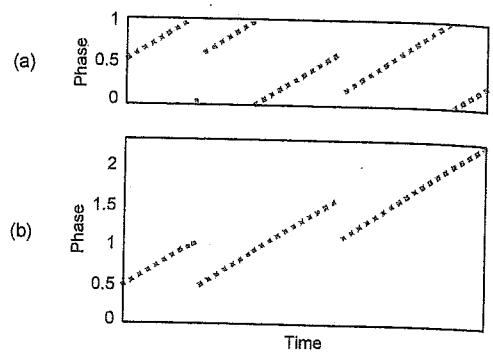


Fig. 6.9 Illustration of phase unwrapping. **(a)** Wrapped phase. **(b)** Unwrapped phase.

$$\varphi(n,k) - \varphi(n-1,k) \approx \varphi(n-1,k) - \varphi(n-2,k). \tag{6.10}$$

Let us define the first-order difference by

$$\varphi'(n,k) := \varphi(n,k) - \varphi(n-1,k) \tag{6.11}$$

and the second-order difference by

$$\varphi''(n,k) := \varphi'(n,k) - \varphi'(n-1,k). \tag{6.12}$$

Note that one obtains $\varphi''(n,k) \approx 0$ in steady regions of x. However, in transient regions, the phase behaves quite unpredictably across the entire frequency range. As a result, a simultaneous disturbance of the values $\varphi''(n,k)$ for $k \in [0:K]$ is a good indicator for note onsets. Motivated by this observation, we define the **phase-based novelty function** Δ_{Phase} by

$$\Delta_{\text{Phase}}(n) = \sum_{k=0}^{K} |\varphi''(n,k)| \tag{6.13}$$

for $n \in \mathbb{Z}$.

At this point, we need to discuss a technical issue. Recall that the phase γ (in radians) of a complex number $c \in \mathbb{C}$ is defined only up to integer multiples of 2π (see (2.9)). Therefore, the phase is often constrained to the interval $[0, 2\pi)$ and the number $\gamma \in [0, 2\pi)$ is called the **principal value** of the phase. In the scenario of Fourier analysis, we are using the normalized phases $\varphi = \gamma/(2\pi)$. In this case, the interval $[0, 1)$ represents the principal values. When considering a function or a time series of phase values (e.g., the phase values over the frames of an STFT as above), the choice of principal values may introduce unwanted discontinuities. These artificial phase jumps are the results of **phase wrapping**, where a phase value just below one is followed by a value just above zero (or vice versa). To avoid such discontinuities, one often applies a procedure called **phase unwrapping**, where the objective is to recover a possibly continuous sequence of (unwrapped) phase values (see Figure 6.9). Such a procedure, however, is in general not well defined since the

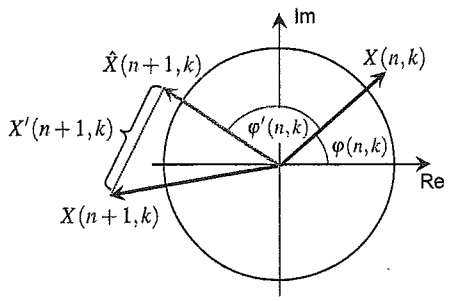


Fig. 6.10 Illustration of the complex-domain difference $\mathcal{X}'(n,k)$ between an estimated spectral coefficient $\hat{\mathcal{X}}(n+1,k)$ and the actual coefficient $\mathcal{X}(n+1,k)$.

original time series may possess "real" discontinuities that are hard to distinguish from "artificial" phase jumps. In the onset detection context, phase jumps due to wrapping may occur when computing the differences in (6.11) and (6.12). In these cases, one needs to use an unwrapped version of the phase. As an alternative, we introduce a **principal argument function**

$$\Psi : \mathbb{R} \to [-0.5, 0.5] \tag{6.14}$$

which maps phase differences into the range $[-0.5, 0.5]$. To this end, a suitable integer value is added to or subtracted from the original phase difference to yield a value in $[-0.5, 0.5]$. The differences as defined in (6.11) and (6.12) are then replaced by

$$\varphi'(n,k) := \Psi\big(\varphi(n,k) - \varphi(n-1,k)\big), \tag{6.15}$$
$$\varphi''(n,k) := \Psi\big(\varphi'(n,k) - \varphi'(n-1,k)\big). \tag{6.16}$$

Even though the principal argument function may cancel out large discontinuities in the phase differences, this effect is attenuated since we consider in (6.13) the sum of differences over all frequency indices.

6.1.4 Complex-Domain Novelty

We have seen that steady regions within a signal may be characterized by a phase-based criterion in the case that the sinusoid correlates well with the signal. However, if the magnitude of the Fourier coefficient $\mathcal{X}(n,k)$ is very small, the phase $\varphi(n,k)$ may exhibit a rather chaotic behavior due to small noise-like fluctuations that may occur even within a steady region of the signal. To obtain a more robust detector, one idea is to weight the phase information with the magnitude of the spectral coefficient. This leads to a complex-domain variant of the novelty function, which jointly considers phase and magnitude. The assumption of this variant is that the phase differences as well as the magnitude stay more or less constant in steady regions. Therefore, given the Fourier coefficient $\mathcal{X}(n,k)$, one obtains a steady-state estimate

$\hat{\mathcal{X}}(n+1,k)$ for the next frame by setting

$$\hat{\mathcal{X}}(n+1,k) = |\mathcal{X}(n,k)| \exp(2\pi i(\varphi(n,k)+\varphi'(n,k))) \tag{6.17}$$

(see Figure 6.10). Then, we can use the magnitude between the estimate $\hat{\mathcal{X}}(n+1,k)$ and the actual coefficient $\mathcal{X}(n+1,k)$ to obtain a measure of novelty:

$$\mathcal{X}'(n+1,k) = |\hat{\mathcal{X}}(n+1,k) - \mathcal{X}(n+1,k)|. \tag{6.18}$$

The complex-domain difference $\mathcal{X}'(n,k)$ quantifies the degree of nonstationarity for frame n and coefficient k. Note that this number does not discriminate between note onsets (energy increase) and note offsets (energy decrease). Therefore, we decompose $\mathcal{X}'(n,k)$ into a component $\mathcal{X}^+(n,k)$ of increasing magnitude and a component $\mathcal{X}^-(n,k)$ of decreasing magnitude:

$$\mathcal{X}^+(n,k) = \begin{cases} \mathcal{X}'(n,k) & \text{for } |\mathcal{X}(n,k)| > |\mathcal{X}(n-1,k)| \\ 0 & \text{otherwise,} \end{cases} \tag{6.19}$$

$$\mathcal{X}^-(n,k) = \begin{cases} \mathcal{X}'(n,k) & \text{for } |\mathcal{X}(n,k)| \leq |\mathcal{X}(n-1,k)| \\ 0 & \text{otherwise.} \end{cases} \tag{6.20}$$

A **complex-domain novelty function** Δ_{Complex} for detecting note onsets can then be defined by summing the values $\mathcal{X}^+(n,k)$ over all frequency coefficients:

$$\Delta_{\text{Complex}}(n,k) = \sum_{k=0}^{K} \mathcal{X}^+(n,k). \tag{6.21}$$

Similarly, for detecting general transients or note offsets, one may compute a novelty function using $\mathcal{X}'(n,k)$ or $\mathcal{X}^-(n,k)$, respectively.

6.2 Tempo Analysis

The extraction of tempo and beat information from audio recordings is a challenging problem in particular for music with weak note onsets and local tempo changes. For example, in the case of romantic piano music, the pianist often takes the freedom of speeding up and slowing down the tempo—an artistic means also referred to as **tempo rubato**. There is a wide range of music where the notions of tempo and beat remain rather vague or are even nonexistent. Sometimes, the rhythmic flow of music is deliberately interrupted or disturbed by **syncopation**, where certain notes outside the regular grid of beat positions are stressed. To make the problem of tempo and beat tracking feasible, most automated approaches rely on two basic assumptions. The first assumption is that beat positions occur at note onset positions, and the second assumption is that beat positions are more or less equally spaced—at least for a certain period of time. Even though both assumptions may be violated and

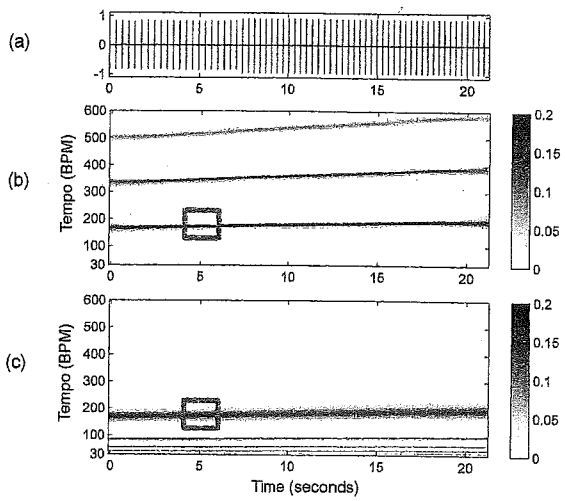


Fig. 6.11 Illustration of two different tempogram representations $\mathcal{T}$ of a click track with increasing tempo (170 to 200 BPM). The large values $\mathcal{T}(t,\tau)$ around $t = 5$ sec and $\tau = 180$ BPM are highlighted by the rectangular frames. **(a)** Novelty function of click track. **(b)** Tempogram with harmonics. **(c)** Tempogram with subharmonics.

inappropriate for certain types of music, they are convenient and reasonable for a wide range of music including most rock and popular songs.

Based on these two assumptions, we discuss in this section various time–tempo or tempogram representations, which capture local tempo characteristics of music signals (Section 6.2.1). To derive such representations, we study two methods for analyzing novelty functions with regard to reoccurring or quasiperiodic patterns. Using Fourier analysis, we show how to derive a tempogram by comparing the novelty function with templates that consist of windowed sinusoids, each representing a specific frequency or tempo (Section 6.2.2). For the second method, we discuss an autocorrelation approach where a tempogram is obtained by comparing a novelty function with localized time-shifted copies of itself (Section 6.2.3). Finally, we introduce robust mid-level representations referred to as cyclic tempograms (Section 6.2.4), which are the tempo-related counterpart of the harmony-related chroma representations. The properties of the tempogram representations are illustrated in the context of music segmentation.

6.2.1 Tempogram Representations

In Section 2.5.2, we studied the concept of a (magnitude) spectrogram, which represents the time–frequency content of a given signal. A large value $\mathrm{Spec}(t,\omega)$ of a spectrogram indicates that the signal contains at time instance t a periodic component that corresponds to the frequency ω (see (2.141)). We now introduce a similar concept referred to as a **tempogram**, which indicates for each time instance the local relevance of a specific tempo for a given music recording. Mathematically, we model a tempogram as a function

$$\mathcal{T} : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0} \tag{6.22}$$

depending on a time parameter $t \in \mathbb{R}$ measured in seconds and a tempo parameter $\tau \in \mathbb{R}_{>0}$ measured in beats per minute (BPM). Intuitively, the value $\mathcal{T}(t,\tau)$ indicates the extent to which the signal contains a locally periodic pulse of a given tempo τ in a neighborhood of time instance t. For example, the tempogram of Figure 6.11b has a large value $\mathcal{T}(5,180)$, thus indicating that the music signal has a dominant tempo of $\tau = 180$ BPM around time position $t = 5$ sec. Just as with spectrograms (Section 2.5.3), one computes a tempogram in practice only on a discrete time–tempo grid. As before, we assume that the sampled time axis is given by $[1:N]$. To avoid boundary cases and to simplify the notation in the subsequent considerations, we extend this axis to $\mathbb{Z}$. (The respective representations are then extended by, e.g., zero-padding.) Furthermore, let $\Theta \subset \mathbb{R}_{>0}$ be a finite set of tempi specified in BPM. Then, a **discrete tempogram** is a function

$$\mathcal{T} : \mathbb{Z} \times \Theta \to \mathbb{R}_{\geq 0}. \tag{6.23}$$

Most approaches for deriving a tempogram representation from a given audio recording proceed in two steps. Based on the assumption that pulse positions usually go along with note onsets, the music signal is first converted into a novelty function (see Section 6.1). This function typically consists of impulse-like spikes, each indicating a note onset position. In the second step, the locally periodic behavior of the novelty function is analyzed. To obtain a tempogram, one quantifies the periodic behavior for various periods $T > 0$ (given in seconds) in a neighborhood of a given time instance. The rate $\omega = 1/T$ (measured in Hz) and the tempo τ (measured in BPM) are related by

$$\tau = 60 \cdot \omega. \tag{6.24}$$

For example, a sequence of impulse-like spikes that are regularly spaced with period $T = 0.5$ sec corresponds to a rate of $\omega = 1/T = 2$ Hz or a tempo of $\tau = 120$ BPM.

One major problem in determining the tempo of a music recording arises from the fact that pulses in music are often organized in complex hierarchies that represent the rhythm. In particular, there are various levels that are presumed to contribute to the human perception of tempo and beat. For example, as illustrated by Figure 6.12, one may consider the tempo on the **tactus** level, which typically corresponds to the quarter note level and often matches the foot tapping rate. Thinking at a larger musical scale, one may also perceive the tempo at the **measure** level, in particular when listening to fast music or to highly expressive music with strong rubato. Finally, one may also consider the **tatum** (temporal atom) level, which refers to the fastest repetition rate of musically meaningful accents occurring in the signal.

Often the tempo ambiguity that arises from the existence of different pulse levels is also reflected in a tempogram $\mathcal{T}$. Higher pulse levels often correspond to integer multiples $\tau, 2\tau, 3\tau, \ldots$ of a given tempo τ. As with pitch (Section 1.3.2), we call such integer multiples **(tempo) harmonics** of τ. Furthermore, integer fractions $\tau, \tau/2, \tau/3, \ldots$ are referred to as **(tempo) subharmonics** of τ. Analogous to the

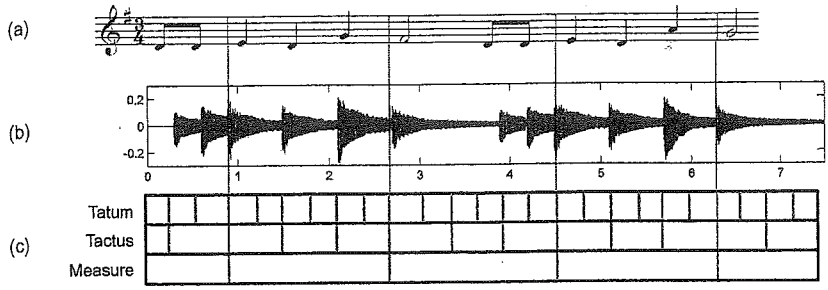


Fig. 6.12 Illustration of various pulse levels. In this example, the tactus level corresponds to the quarter note and the tatum level to the eighth note level. **(a)** Score representation. **(b)** Waveform of an audio excerpt of "Happy Birthday to you." **(c)** Annotation of pulse levels.

notion of an octave for musical pitches (see Section 1.1.1), the difference between two tempi with half or double the value is called a **tempo octave**. For an illustration, we refer to Figure 6.11, which shows two different types of tempograms for a click track of increasing tempo (raising from 170 to 200 BPM over the course of 20 sec). The tempogram of Figure 6.11b emphasizes tempo harmonics, whereas the tempogram of Figure 6.11c emphasizes tempo subharmonics. In the following, we will study two conceptually different methods that are used to derive these two tempograms.

6.2.2 *Fourier Tempogram*

As a first periodicity estimation method, we show how a short-time Fourier transform can be used to derive a tempogram from a given novelty function $\Delta : \mathbb{Z} \to \mathbb{R}$. Dealing with a discrete-time signal Δ, we consider the discrete version of the STFT as discussed in Section 2.5.3. To this end, we fix a window function $w : \mathbb{Z} \to \mathbb{R}$ of finite length centered at $n = 0$ (e.g., a sampled Hann window as defined in (2.140)). Then, for a frequency parameter $\omega \in \mathbb{R}_{\geq 0}$ and time parameter $n \in \mathbb{Z}$, the complex Fourier coefficient $\mathcal{F}(n, \omega)$ is defined by

$$\mathcal{F}(n, \omega) := \widetilde{\Delta^w}(n, \omega) = \sum_{m \in \mathbb{Z}} \Delta(m) \overline{w}(m - n) \exp(-2\pi i \omega m). \tag{6.25}$$

This definition corresponds to (2.143) when using a hop size $H = 1$. Converting frequency to tempo values based on (6.24), we define the (discrete) **Fourier tempogram** $\mathcal{T}^{\mathrm{F}} : \mathbb{Z} \times \Theta \to \mathbb{R}_{\geq 0}$ by

$$\mathcal{T}^{\mathrm{F}}(n, \tau) := |\mathcal{F}(n, \tau/60)|. \tag{6.26}$$

The Fourier-based analysis of the novelty function is also illustrated by Figure 6.13, which continues our Shostakovich example from Figure 6.7. As the Fourier tempogram $\mathcal{T}^{\mathrm{F}}$ (Figure 6.13b) reveals, the dominant tempo of this excerpt is between 200 and 300 BPM. Starting with roughly $\tau = 225$ BPM, the tempo slightly increases over time. An entry $\mathcal{T}^{\mathrm{F}}(n, \tau)$ of the tempogram is obtained by locally comparing the novelty function Δ in a neighborhood of n with a windowed sinusoid that represents the tempo τ. This kind of analysis is shown in Figure 6.13c for a time index n that corresponds to the physical time $t = 2$ sec and a frequency parameter ω that corresponds to the tempo $\tau = 230$ BPM. In this case, the positive parts of the windowed sinusoid nicely align with the impulse-like peaks of the novelty function Δ, whereas the negative parts of the sinusoid fall into the zero-regions of Δ. As a result, there is a high correlation between the windowed sinusoid and Δ, which leads to a large coefficient $\mathcal{T}^{\mathrm{F}}(n, \tau)$. In contrast, using a sinusoid that represents only half this tempo leads to a small coefficient, as illustrated by Figure 6.13d. In this case, every second peak of Δ falls into the positive parts of the sinusoid, whereas the remaining peaks of Δ fall into the negative parts of the sinusoid. Because of the resulting cancellations, the correlation between Δ and the sinusoid becomes small. Finally, Figure 6.13e illustrates that one obtains a high correlation when using a sinusoid that represents twice the main tempo. In this case, the peaks of Δ are aligned with every second positive part of the sinusoid, whereas all other parts of the sinusoid fall into the zero-regions of Δ. Our discussion shows that a Fourier tempogram generally indicates tempo harmonics, but suppresses tempo subharmonics. This fact is illustrated by Figure 6.11b, which shows the Fourier tempogram of a synthetic click track. Also, in our Shostakovich example, the second tempo harmonic starting at $\tau = 450$ BPM is clearly visible in $\mathcal{T}^{\mathrm{F}}$ (Figure 6.13b). Interestingly, because of the weak downbeats every third beat within the 3/4 meter (see our discussion of Figure 6.7), the tempogram $\mathcal{T}^{\mathrm{F}}$ also shows some larger coefficients that correspond to $1/3$ and $2/3$ of the main tempo (see Exercise 6.5)

For practical applications, $\mathcal{T}^{\mathrm{F}}$ is computed only for a small number of tempo parameters. For example, one may choose the set $\Theta = [30 : 600]$ covering the (integer) musical tempi between 30 and 600 BPM. The bounds are motivated by the assumption that only musical events showing a temporal separation between roughly 100 ms (600 BPM) and 2 sec (30 BPM) contribute to the perception of tempo. This tempo range requires a spectral analysis of high resolution in the lower frequency range. Therefore, a straightforward FFT as discussed in Section 2.4.3 is not suitable. However, since only relatively few frequency bands (tempo values) are needed for the tempogram, computing the required Fourier coefficients individually according to (6.25) still has a reasonable computational complexity. As for the temporal resolution, one can set w to be a sampled Hann window as defined in (2.140) of size $2N + 1$ for some $N \in \mathbb{N}$. Depending on the respective application and the nature of the music recording, a window size corresponding to 4–12 sec of audio is a reasonable range. Finally, note that the feature rate of the resulting tempogram can be adjusted by introducing a hop size parameter H in (6.25) as used in (2.143).

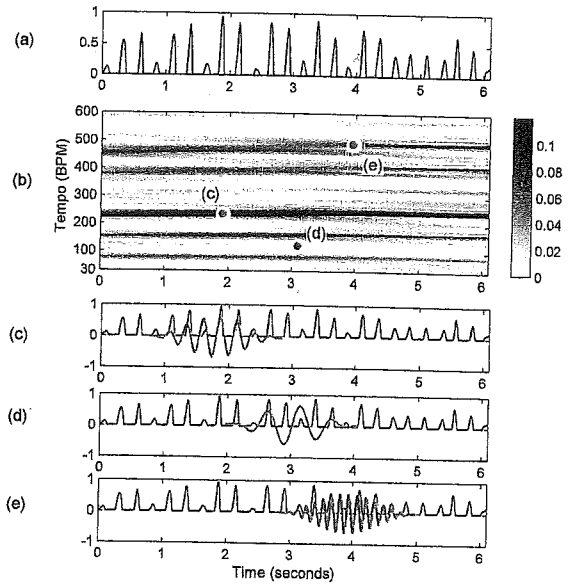


Fig. 6.13 Fourier-based tempo analysis for the Shostakovich example from Figure 6.7. **(a)** Novelty function Δ. **(b)** Fourier tempogram $\mathcal{T}^{\mathrm{F}}$. **(c–e)** Correlation of Δ and various analyzing windowed sinusoids.

6.2.3 Autocorrelation Tempogram

As a second periodicity estimation method, we now discuss an autocorrelation-based approach. Generally speaking, the **autocorrelation** is a mathematical tool for measuring the similarity of a signal with a time-shifted version of itself. Since the inner product as defined in (2.43) is used for this measurement, this technique is also known as the **sliding inner product**. In the following, we only consider the case of discrete-time and real-valued signals. Let $x \in \ell^2(\mathbb{Z})$ be such a signal having finite energy (see (2.41)). The autocorrelation $R_{xx} : \mathbb{Z} \to \mathbb{R}$ of the real-valued signal x is defined by

$$R_{xx}(\ell) = \sum_{m \in \mathbb{Z}} x(m)x(m-\ell), \tag{6.27}$$

which yields a function that depends on the time-shift or **lag** parameter $\ell \in \mathbb{Z}$. As shown in Exercise 6.6, the autocorrelation is well defined for signals in the space $\ell^2(\mathbb{Z})$. Furthermore, $R_{xx}(\ell)$ is maximal for $\ell = 0$ and symmetric in ℓ. Intuitively, if the autocorrelation is large for a given lag, then the signal contains repeating patterns that are separated by a time period as specified by the lag parameter.

We now apply the autocorrelation in a local fashion for analyzing a given novelty function $\Delta : \mathbb{Z} \to \mathbb{R}$ in the neighborhood of a given time parameter n. As in the case of the Fourier tempogram discussed in the last section, we fix a window function $w : \mathbb{Z} \to \mathbb{R}$ of finite length centered at $n = 0$. The windowed version $\Delta_{w,n} : \mathbb{Z} \to \mathbb{R}$ localized at point $n \in \mathbb{Z}$ is defined by

$$\Delta_{w,n}(m) := \Delta(m)w(m-n), \tag{6.28}$$

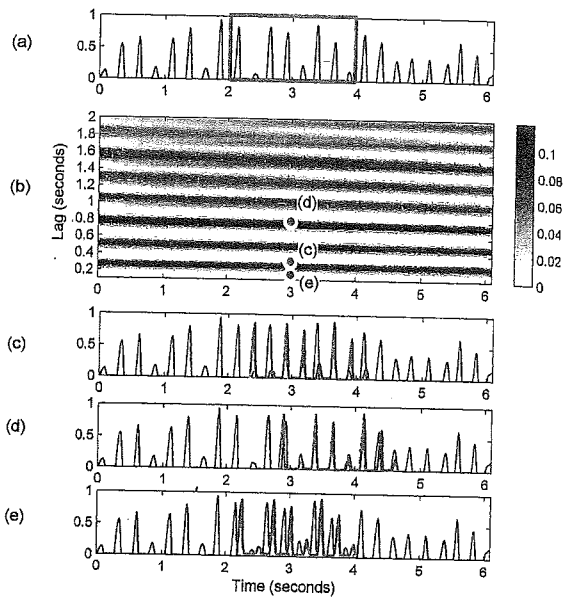


Fig. 6.14 Autocorrelation-based tempo analysis for the Shostakovich example from Figure 6.7. **(a)** Novelty function Δ. **(b)** Time-lag representation $\mathcal{A}$. **(c–e)** Correlation of Δ and various time-shifted windowed sections.

$m \in \mathbb{Z}$. Recall that we have used a similar definition when introducing the STFT (see (2.133)). To obtain the **short-time autocorrelation** $\mathcal{A} : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$, we apply (6.27) to $\Delta_{w,n}$ and define

$$\mathcal{A}(n,\ell) := \sum_{m\in\mathbb{Z}} \Delta(m)w(m-n)\Delta(m-\ell)w(m-n-\ell). \tag{6.29}$$

When assuming that the window function w is of finite length, the autocorrelation of the localized novelty function is zero for all but a finite number of time lag parameters. In the following, let us assume that the support of the window function w lies in the interval $[-L : L]$ for some $L \in \mathbb{N}$. Then one has $\mathcal{A}(n,\ell) = 0$ for $|\ell| \geq 2L+1$ (see Exercise 6.7). Because of this property and the symmetry of the autocorrelation, one only needs to consider the time lag parameters $\ell \in [0 : 2L]$. Furthermore, because of the windowing, at most $2L+1-\ell$ of the summands in (6.29) are nonzero. To balance out the effect of the windowing, the value $\mathcal{A}(n,\ell)$ may be divided by a factor that depends on the window properties and the overlap $2L+1-\ell$ of the window and its time-shifted version.

Visualizing the short-time autocorrelation $\mathcal{A}$ leads to a **time-lag representation**. Before we discuss how this representation can be converted into a time–tempo representation, let us first have a look at Figure 6.14, which continues our Shostakovich example. The window w used in this example is a rectangular window that has a length corresponding to 2 sec of the original audio recording. Let us consider the time index n corresponding to the time instance $t = 3$ sec. To compute $\mathcal{A}(n,\ell)$, one only considers the section of the novelty function Δ between 2 sec and 4 sec (Figure 6.14a). We have seen that the tempo of our Shostakovich recording is

roughly 230 BPM in this section. In other words, the duration of the interval between two subsequent beats is roughly $s = 0.26$ sec. Let us consider the lag parameter ℓ that corresponds to a time shift of $s = 0.26$ sec. Then, as illustrated by Figure 6.14c, the novelty function in this section nicely correlates with its time-shifted version: the peaks of the section fall onto peaks of the section shifted by one beat period. The same holds when shifting the section by two, three or more beat periods. For example, Figure 6.14d shows the case $s = 0.78$ sec (three beat periods). This period corresponds to a tempo of 77 BPM, which is the tempo on the measure level. In contrast, when using a lag ℓ that corresponds to half a beat period $s = 0.13$ sec (double tempo 461 BPM), the peaks of the section and the peaks of the shifted section miss each other, thus resulting in a coefficient $\mathcal{A}(n, \ell)$ close to zero. This case is illustrated by Figure 6.14e.

To obtain a time–tempo representation from the time-lag representation, one needs to convert the lag parameter into a tempo parameter. To this end, one requires the frame rate or time resolution of the novelty function. Suppose that each time frame corresponds to r seconds, then a time lag of ℓ (given in frames) corresponds to $\ell \cdot r$ seconds. Since a shift of $\ell \cdot r$ seconds corresponds to a rate of $1/(\ell \cdot r)$ Hz, one obtains from (6.24) the tempo

$$\tau = \frac{60}{r \cdot \ell} \text{ BPM}. \tag{6.30}$$

Based on this conversion, the lag axis can be interpreted as a tempo axis as illustrated by Figure 6.15b. This allows us to define the **autocorrelation tempogram** $\mathcal{T}^{\mathrm{A}}$ by setting

$$\mathcal{T}^{\mathrm{A}}(n, \tau) := \mathcal{A}(n, \ell) \tag{6.31}$$

for each tempo $\tau = 60/(r \cdot \ell)$, $\ell \in [1 : L]$. Note that in this case, since the tempo values are reciprocal to the linearly sampled lag values, the tempo axis is sampled in a nonlinear fashion. To obtain a tempogram $\mathcal{T}^{\mathrm{A}} : \mathbb{Z} \times \Theta \to \mathbb{R}_{\geq 0}$ that is defined on the same tempo set Θ as the Fourier tempogram $\mathcal{T}^{\mathrm{F}}$, one can use standard resampling and interpolation techniques applied to the tempo domain. The result of such an interpolation step is shown in Figure 6.15c.

As another example, Figure 6.11c shows the autocorrelation tempogram of a click track. This figure illustrates that, as opposed to the Fourier tempogram, an autocorrelation tempogram exhibits tempo subharmonics, but suppresses tempo harmonics. We have already given the argument for this behavior when discussing Figure 6.14: a high correlation of a local section of the novelty function with the section shifted by ℓ samples also implies a high correlation with a section shifted by $k \cdot \ell$ lags for integers $k \in \mathbb{N}$. Assuming that ℓ corresponds to tempo τ, the lag $k \cdot \ell$ corresponds to the subharmonic τ/k.

This property is also evident in our Shostakovich example. Similar to the Fourier tempogram $\mathcal{T}^{\mathrm{F}}$ (Figure 6.13b), the autocorrelation tempogram $\mathcal{T}^{\mathrm{A}}$ (Figure 6.15c) reveals the dominant tempo at $\tau = 225$ BPM, which corresponds to the quarter note level. However, as opposed to $\mathcal{T}^{\mathrm{F}}$, the dominant tempo revealed by $\mathcal{T}^{\mathrm{A}}$ is at $\tau = 75$ BPM, which corresponds to the tempo on the measure level and is the third

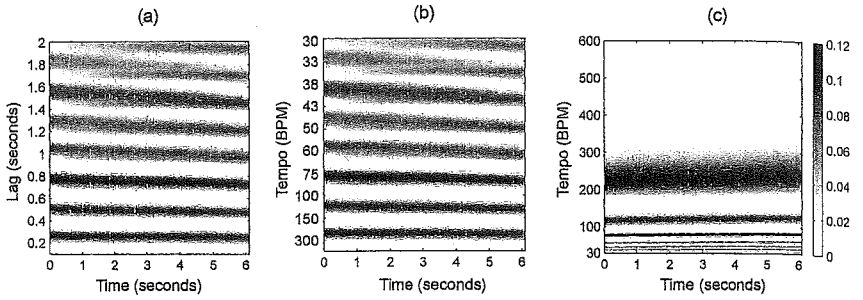


Fig. 6.15 Conversion from lag to tempo. **(a)** Time-lag representation with linear lag axis. **(b)** Representation from (a) with tempo axis. **(c)** Time–tempo representation with linear tempo axis.

subharmonic of $\tau = 225$ BPM. Reflecting the 3/4 meter of the waltz, the dominance of the tempi $\tau = 225$ BPM and $\tau = 75$ BPM is also of musical relevance. In conclusion, one may say that the Fourier tempogram and autocorrelation tempogram yield different types of tempo information and ideally complement each other.

Assuming a more or less steady tempo, it suffices to determine one **global** tempo value for the entire recording. Such a value may be obtained by averaging the tempo values obtained from a frame-wise periodicity analysis. For example, based on a tempogram representation, one can average the tempo values over all time frames to obtain a function $\mathcal{T}_{\text{Average}} : \Theta \to \mathbb{R}_{\geq 0}$ that only depends on $\tau \in \Theta$. Assuming that the relevant time positions lie in the interval $[1 : N]$, one may define $\mathcal{T}_{\text{Average}}$ by

$$\mathcal{T}_{\text{Average}}(\tau) := \frac{1}{N} \sum_{n \in [1:N]} \mathcal{T}(n, \tau). \tag{6.32}$$

The maximum

$$\hat{\tau} := \max\{\mathcal{T}_{\text{Average}}(\tau) \mid \tau \in \Theta\} \tag{6.33}$$

of this function then yields an estimate for the global tempo of the recording. Of course, more refined methods for estimating a single tempo value may be applied. For example, instead of using a simple average in (6.32), we may apply median filtering, which is more robust to outliers and noise. Also, to alleviate the problem of tempo octave confusion, one may improve the result by a combined usage of the Fourier and autocorrelation tempograms.

When dealing with music that exhibits significant tempo changes, one needs to estimate the **local** tempo in the neighborhood of each time instance, which is a much harder problem than global tempo estimation. Having computed a tempogram, the frame-wise maximum yields a good indicator of the locally dominating tempo. In the case that the tempo is relatively steady over longer periods of time, one may increase the window size to obtain more robust and smoother tempo estimates. However, it then becomes harder to detect sudden tempo changes and local tempo fluctuations—the same trade-off we have already encountered in the case of the

STFT (see Section 2.5.2). Furthermore, instead of simply taking the frame-wise maximum—a strategy that is prone to local inconsistencies and outliers—global optimization techniques based on dynamic programming may be used to obtain smooth tempo trajectories. Such strategies will be discussed in Section 6.3 in the context of beat tracking. In both global and local tempo estimation, one often has to struggle with confusions of tempo harmonics and subharmonics, which are the result of the existence of various pulse levels such as measure, tactus, and tatum. In the following section, we introduce a robust mid-level representation that is impervious to tempo octave confusions while still capturing local tempo information.

6.2.4 Cyclic Tempogram

The various pulse levels mentioned above can be seen in analogy to the existence of harmonics in the pitch context (see Section 1.3.2). To reduce the effects of harmonics, we introduced in Section 3.1.2 the concept of chroma-based audio features. By identifying pitches that differ by one or several octaves, we obtained a cyclic mid-level representation that captures harmonic information while being robust to changes in timbre. Inspired by the concept of chroma features, we now introduce the concept of cyclic tempograms. The idea is to form tempo equivalence classes by identifying tempi that differ by a power of two. More precisely, we say that two tempi τ_1 and τ_2 are **octave equivalent**, if they are related by $\tau_1 = 2^k \tau_2$ for some $k \in \mathbb{Z}$. For a tempo parameter τ, we denote the resulting tempo equivalence class by $[\tau]$. For example, for $\tau = 120$ one obtains $[\tau] = \{\ldots, 30, 60, 120, 240, 480 \ldots\}$. Given a tempogram representation $\mathcal{T} : \mathbb{Z} \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$, we define the **cyclic tempogram** by

$$C(n, [\tau]) := \sum_{\lambda \in [\tau]} \mathcal{T}(n, \lambda). \tag{6.34}$$

Note that the tempo equivalence classes topologically correspond to a circle. Fixing a reference tempo τ_0, the cyclic tempogram can be represented by a mapping $\mathcal{C}_{\tau_0} : \mathbb{Z} \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ defined by

$$\mathcal{C}_{\tau_0}(n, s) := C(n, [s \cdot \tau_0]) \tag{6.35}$$

for $n \in \mathbb{Z}$ and a **scaling parameter** $s \in \mathbb{R}_{>0}$. Note that $\mathcal{C}_{\tau_0}(n, s) = \mathcal{C}_{\tau_0}(n, 2^k s)$ for $k \in \mathbb{Z}$. In particular, $\mathcal{C}_{\tau_0}$ is completely determined by its values $s \in [1, 2)$.

These definitions are illustrated by Figure 6.16, which shows various tempograms for a click track of increasing tempo (110 to 130 BPM), similar to the one used in Figure 6.11. As demonstrated by Figure 6.16a, the Fourier tempogram $\mathcal{T}^{\mathrm{F}}$ indicates the tempo as well as its tempo harmonics. Using a reference tempo $\tau_0 = 60$ BPM, the resulting **cyclic Fourier tempogram**, which we denote by $\mathcal{C}^{\mathrm{F}}_{\tau_0}$, is shown in Figure 6.16c. In the pitch context, given a reference frequency ω, the frequency 3ω is an octave plus a fifth higher, and 3ω can be regarded as the **dominant** to the **tonic** ω. In analogy to the pitch context, we call the tempo class $[3\tau]$, which