CHAPTER 6

TEMPORAL ANALYSIS

The temporal aspects of music signals such as the tempo and the rhythm are important musical properties. A fundamental building block of these aspects is the onset: the beginning of a musical sound event such as a tone or the stroke on a percussive instrument. The start time of an event is usually considered to be more important than the time of the end of that event, as listeners apparently perceive musical events more in terms of onset-to-onset intervals [177].

6.1 Human Perception of Temporal Events

During the process of human perception, the audio stream will be segmented into a series of events; speaking of segmentation is a simplification because musical meaning and even rhythm can be conveyed by audio streams with no such clear division into distinct events [178]. However, this simplification can be assumed to be sufficiently valid in the context of western music for the majority of possible input signals — other incarnations of temporal information will be ignored for the sake of simplicity.

6.1.1 Onsets

As stated above, an *onset* is the start of a (musical) sound event. The term *onset* is frequently used as a synonym to onset time, but it should be more correct to state that its time

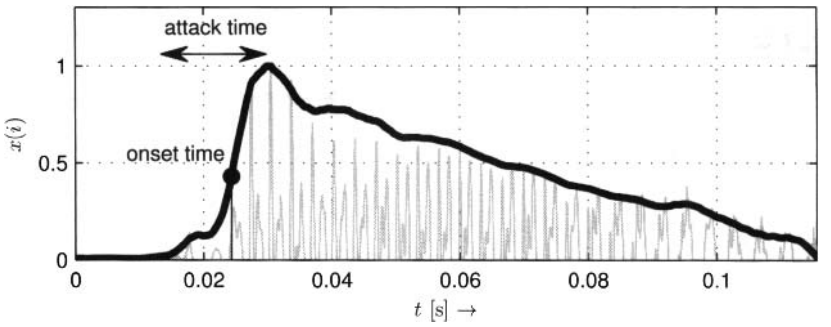

Figure 6.1 Visualization of an envelope and attack time and one possible location for an approximation of the perceptual onset time

position (i.e., the onset time) is one — most definitely the main — property of the onset while it can have other properties such as its strength.

In reality, the start of a musical sound usually is not an exact point in time, but a time span, the *attack time* or *rise time*. It is the time from the first instrument-induced oscillation until a maximum amplitude is reached. An example of an attack phase is shown in Fig. 6.1.

Sometimes the attack time is differentiated from the *initial transient time* which ends when the note reaches its quasi-periodic state. Obviously this differentiation works only for tonal events. The attack time can vary significantly between different musical instruments or groups of instruments. It ranges from about 5 ms for some percussive instruments to up to 200 ms for woodwind instruments (flute) under certain conditions [46].

The exact usage of the terms *onset*, *attack*, and *transient* is sometimes inconsistent and confusing. To give an example of a different naming convention than the one used here, Bello et al. propose to use the terms *attack* for our attack time, *transient* as a description of the initial phase of a musical event in which "the signal evolves quickly in some nontrivial or relatively unpredictable way" (the period covered by our attack time), and *onset* as a single instant chosen to mark the temporally extended transient (our onset time) [179].

Repp pointed out that three definitions of onset times can generally be distinguished [180]:

1. *Note Onset Time (NOT)*: the time when the instrument is triggered to make a sound. In the MIDI domain, the NOT is exactly the time of the Note-On command. Depending on the instrument or sample used for sound generation, this is not necessarily the time when the signal becomes detectable or audible.

2. *Acoustic Onset Time (AOT)*: the first time when a signal or an acoustic event is theoretically measurable. Sometimes the AOT is called *physical onset time*.

3. *Perceptual Onset Time (POT)*: the first time when the event can be perceived by the listener. The POT might also be distinguished from the *Perceptual Attack Time (PAT)* which is the instant of time that is relevant for the perception of rhythmic patterns [181]. While the PAT might occur later than the POT, they will be equal in many cases. For the sake of simplicity, there will be no distinction of POT and PAT in the following.

The POT can never occur before the AOT, which in turn never occurs before the NOT. Due to the "perceptual" definition of the POT, the exact location cannot be determined acoustically but has to be measured in a listening test. Both Gordon and Zwicker found strong location drifts of the PAT depending on the waveform properties during the rise time [47, 181]. There are indications of the POT to be correlated with the envelope slope [181].

Given the three definitions above, the following question arises: which of the three onset times is on the one hand detectable in the signal and on the other hand of the utmost interest in automatic onset detection and any rhythm-related task? Due to the symbolic nature of the NOT, it simply cannot be detected from the audio signal. The choice between AOT and POT might be application-dependent; assuming that musicians adapt their timing to their sound perception and that most ACA systems try to analyze the *perceptible* audio content, the POT is most likely the point in time desired as result. This reflection, however, is rather academic since in reality the accuracy of automatic onset detection systems is usually too poor to differentiate between the different onset times for all but a small class of signal types. The algorithm designer will probably strive to improve the detection performance of an onset detection system as opposed to its accuracy.

The human ability to locate onset times and to distinguish closely spaced onsets is of specific interest when estimating the required time accuracy of an onset detection system since most systems aim to be at least as accurate as the human perception.

Hirsh found that temporal discrimination of two onsets is possible for humans if the onset time difference is as little as 2 ms [182]. However, in order to determine the order of the stimuli, their distance had to be about 20 ms. The measurements were done with synthetic signals with short rise times.

Gordon reported a standard deviation of 12 ms for the accuracy of onset times specified by test listeners, using 16 real-world monophonic sounds of different instruments played in an infinitely long loop pattern with *Inter-Onset Intervals (IOIs)* of 600 ms [181]. Friberg and Sundberg undertook a similar experiment using tone stimuli [183]. For IOIs smaller than 240 ms, they reported a just noticeable difference of about 10 ms, and increasing values for larger IOIs.

Repp reported for the manual annotation of onset times by one listener in the context of piano recordings a mean absolute measurement error of about 4.3 ms and a maximum error of about 35 ms [184]. In a recent investigation, Leveau et al. had three test subjects annotating the onset times in audio files of various genres and instrumentations [185]. The results showed a mean absolute measurement error over all test data of about 10 ms; for one piece of classical music, the mean absolute measurement error nearly reached 30 ms.

Rasch evaluated the onset time differences between instruments in three ensemble performances [186]. He found synchronization deviations in a range between 30 and 50 ms between the (string and woodwind) instruments, while the mean onset time differences were in the range of ± 6 ms. However, since the measurement accuracy has not been evaluated in this case, it is unknown how much of the actual time differences can be attributed to the performance itself.

For piano duet performance, Shaffer reported standard deviations within the voices between 14 and 38 ms [187].

It may be concluded that the accuracy of human onset perception depends on the test data and that deviations evoked by motoric abilities seem to be in the same range. The presented results imply that an automatic onset detection system aiming at human detection accuracy (or being evaluated with test data annotated by humans) will have a minimum mean absolute error in the range of 5-10 ms; the error can be expected to be as high as 10 times more for specific instruments and pitches with long rise times. A real-world aspect

with negative impact on the onset detection "accuracy" is the occurrence of several quasi-simultaneous onsets in polyphonic music. In this case the deviation between the individual voices will virtually decrease the system's accuracy, although it may be argued that in this case there is no single reference onset.

6.1.2 Tempo and Meter

The *tempo* is the rate at which perceived pulses with equal duration units occur at a moderate and natural rate [188]. This perceived tempo is called the *tactus* [189] and is sometimes simply referred to as the *foot tapping rate* [190]. A typical natural rate would be located around 100 *Beats per Minute (BPM)* [191].

For segments of music with constant tempo, the tempo $\mathfrak{T}$ in BPM can be computed using the length of the segment Δt_s in seconds and the number of beats $\mathcal{B}$ in the segment:

$$\mathfrak{T} = \frac{\mathcal{B} \cdot 60\,\mathrm{s}}{\Delta t_\mathrm{s}} \ [\mathrm{BPM}]. \tag{6.1}$$

In the case of a dynamic tempo, the local tempo can be extracted by identifying the event time of every beat t_b and computing the distance between two neighboring beats with indices j and $j+1$:

$$\mathfrak{T}_\mathrm{local}(j) = \frac{60\,\mathrm{s}}{t_\mathrm{b}(j+1) - t_\mathrm{b}(j)} \ [\mathrm{BPM}]. \tag{6.2}$$

Deriving an *overall* tempo becomes increasingly difficult when the tempo is not constant; in this case the mean tempo given in Eq. (6.1) does not necessarily match the *perceived tempo* a listener would indicate. This led Gabrielsson to distinguishing between the mean tempo and the *main tempo*, the latter being a measure ignoring slow beginnings or final *ritardandi* [192]. Repp found good correlation of the perceived tempo with the mean value of a logarithmic IOI distribution [193]. Goebl proposed a *mode tempo* which is computed by sweeping a window over the histogram of *Inter-Beat Intervals (IBIs)* and selecting the maximum position as mode tempo [194].

McKinney and Moelants complicated matters further by arguing that a single tempo does not sufficiently describe the (listener) group response when presented with a piece of music. They propose a representation of the overall tempo with two BPM values instead of a single one [195].

The *meter* is a regular alternation of strong and weak musical elements which are grouped with a length of normally three to seven beats or a length of around 5 s.

6.1.3 Rhythm

The perception of *rhythm* can — similar to the meter — be described by its grouping properties. The grouping properties allow a hierarchical segmentation into smaller subsequences forming different grouping levels. The length of the groups can range from the length of a few notes up to whole parts of the work defining musical form [189].[1] Groups of a length between one beat and the length of the meter are most commonly referred to as rhythm. The rhythm is then defined by its accents and time intervals; if the durations of subsequent intervals relate to simple integer ratios, then the group usually has a closer binding than otherwise [196].

[1] However, we will use the term *rhythm* only for groups with a length of up to several beats.

The various hierarchical levels of temporal grouping are an important property of many (western) pieces of music. Humans perceive pulses at different levels and with different tempi; at all levels the grouping of strong and weaker events occurs. The basic building block on the lowest (and shortest level is commonly referred to as tatum [197], although other terms such as *atomic beat* have been used [198]. The tatum specifies the lowest period length or the period of the regular pulse train with the highest frequency represented in the music. Every rhythm is built of the tatums which can be interpreted as a rhythmic grid or a time quantization. The length of the highest level grouping depends on the definition of grouping and could go up to the level defining musical form such as the length of musical phrases or even longer structures which form groups.

6.1.4 Timing

The *timing* of individual notes or temporal events in a music performance does not necessarily exactly reflect the structural properties of the rhythm or meter but shows systematic temporal deviations from the underlying rhythmic structure [45]. A detailed overview of expressive timing will be given in Chap. 10.

6.2 Representation of Temporal Events in Music

The representation of musical (temporal) events is closely related to the perception of such events for both terms and the musical score.

6.2.1 Tempo and Time Signature

The *overall tempo* of a piece of music is usually chosen by the performing artists even if the composer indicates a preferred tempo. Tempo instructions for the performers became more and more explicit over the centuries. While many pieces from the Baroque period do not contain instructions due to the composer's assumption that the tempo was specified by performance conventions, it became more and more common in later epochs to indicate the tempo with Italian terms such as *Largo* (very slow), *Adagio* (slow), *Andante* (walking pace), *Moderato* (moderately), *Allegro* (fast), and *Presto* (very fast). During the last century it became more common to make specific tempo indications in BPM.

The *local tempo* varies throughout a piece of music for nearly all genres. The possibilities to include instructions for such variations in the score are limited besides adding tempo indicators; examples of tempo instructions for sliding tempo changes are *ritardando* (slowing down) and *accelerando* (speeding up).

The *bar* (also called a *measure*) is the score equivalent of the (perceptual) meter. A score marks the beginning of each bar by a vertical line. The first beat of a bar usually has the highest (perceptual) weight and is referred to as *downbeat*.

The *time signature* is a way to convey information on the properties of a bar, namely the number of beats grouped together in one bar as well as the note value constituting one beat. The time signatures in Fig. 6.2 group four, three, two, and two beats, respectively. The fourth example differs from the first three in grouping half notes instead of quarter notes. The denominator of the time signature thus indicates the note value of one beat while its numerator indicates the number of beats per bar.[2]

[2]There are exceptions from this rule such as a time signature 6/8 with a beat length of three eighth notes.

Figure 6.2 Frequently used time signatures

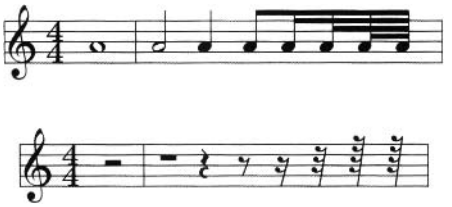

Figure 6.3 Note values (top) and corresponding rest values (bottom) with decreasing length in musical score notation

6.2.2 Note Value

The *note value* defines the relative length of a note with respect to time signature and tempo. Notational convention requires that the sum of note values and rest values per bar (except a few special cases) must result in the numerator of the time signature. Thus the absolute onset time of each note is specified by the bar index and the note's position in the bar, given a specific tempo.

The offset time (also the *note off time*) is determined by the note's onset time and its note value in the score but is not necessarily as clearly defined in a real-world performance. Sometimes a note value is shortened and a rest is appended to give the performing artists indications of the preferred articulation. In general, however, it is not unusual to place the responsibility for such articulation decisions on the performers rather than the musical score, but this depends also on epoch, style, and composer.

Figure 6.3 shows the most common note values (top) starting from a whole note and decreasing the value down to two sixty-fourth notes and the corresponding rests (bottom).

6.3 Onset Detection

Segmenting the audio stream into separate musical events can be an important processing step in applications such as tempo detection or the automatic transcription of music.

The flowchart of a typical *onset detection* system (also called *onset tracking* system) is shown in Fig. 6.4. First, a novelty function is computed extracting the amount of "new" information in the audio signal for each analysis block. The second processing step consists of identifying the locations of the significant maxima which can then be regarded as onset times. This second processing step is usually referred to as peak picking.

Overview articles for different approaches to detecting onsets have been published by Bello et al. and Dixon [179, 199].

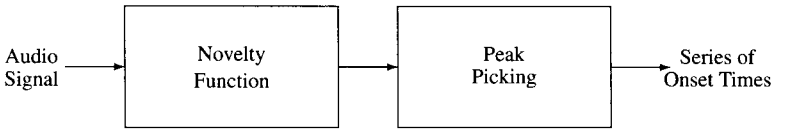

Figure 6.4 General flowchart of an onset detection system

6.3.1 Novelty Function

An important property of the beginning of musical sound events is that "something new happens." Thus, the first step toward automatic onset detection is the computation of a *novelty function* which indicates the amount of audio signal changes over time [200]. Other names of this function are *detection function* [179] or *difference function* [201].

The first step in the computation of the novelty function is usually the calculation of the difference between current and preceding feature values. The result is then smoothed and negative values are discarded by applying HWR. The latter processing step is usually helpful for onset detection as an (amplitude or energy) increase might be expected at onset times while a decrease should make an onset less likely.

In one of the first publications on onset detection in (percussive) music signals, Schloss presented an algorithm that makes direct use of the audio signal's envelope slope, extracted in the time domain [202]. He extracts the envelope of the audio signal by computing the maximum of the magnitude of the signal within a block of samples and recommends to adjust the block length to the length of the period of the lowest frequency present. The envelope slope is then computed by using linear regression over several points of the peak amplitude.

As pointed out in Chap. 4, there exist different possibilities to extract the envelope of a signal, including taking the block's maximum amplitude and low-pass filtering either the signal's magnitude or its RMS. This kind of envelope analysis is nowadays usually applied to specific frequency subbands as opposed to the time domain signal.

Most of the concurrent systems use STFT-based techniques for computing the novelty function, computed from the differences between subsequent (overlapping) STFT blocks. This can either be done with each individual spectral bin or with multiple bins grouped into frequency bands. The advantage of computing the novelty function in the frequency domain is that the onset detection is not only based on amplitude and envelope differences but might also take into account spectral differences such as a pitch change. The disadvantage of the frequency domain computation is the comparably poor time resolution which affects the algorithm's detection accuracy.

The number of frequency bands of onset detection systems varies. Scheirer uses a filterbank of 6 bands basically covering a one-octave range [203]. Klapuri uses 21 non-overlapping bands with their band width and mid-frequency inspired by Zwicker's critical bands [204]. Duxbury uses 5 bands up to 22 kHz with constant Q. The individual results per subband are either combined into one overall novelty function or they are processed per frequency band to be combined only for the final result.

While spectral domain onset detection systems differ in the number of frequency bands they analyze, their main difference is the distance measure $d(n)$ between consecutive STFTs.

Laroche used a distance similar to the spectral flux with an additional square root function to increase lower signal amplitudes [205]:

$$d_{\mathrm{lar}}(n) = \sum_{k=k(f_{\min})}^{k(f_{\max})} \sqrt{|X(k,n)|} - \sqrt{|X(k,n-1)|}, \tag{6.3}$$

Duxbury et al. proposed the distance between complex STFT bins [206]:

$$d_{\mathrm{dux}}(n) = \sum_{k=0}^{\mathcal{K}/2-1} |X(k,n) - X(k,n-1)|, \tag{6.4}$$

while Hainsworth and Macleod calculated a logarithmic distance [201]:

$$d_{\mathrm{hai}}(n) = \sum_{k=0}^{\mathcal{K}/2-1} \log_2 \left(\frac{|X(k,n)|}{|X(k,n-1)|} \right). \tag{6.5}$$

It is also possible to compute the distance with the cosine distance between two STFT frames as suggested by Foote [200]:

$$d_{\mathrm{foo}}(n) = 1 - \frac{\sum_{k=0}^{\mathcal{K}/2-1} |X(k,n)| \cdot |X(k,n-1)|}{\sqrt{\left(\sum_{k=0}^{\mathcal{K}/2-1} |X(k.n)|^2 \right) \cdot \left(\sum_{k=0}^{\mathcal{K}/2-1} |X(k,n-1)|^2 \right)}}. \tag{6.6}$$

Bello et al. pointed out that phase relations may be used for the detection of novelty in an audio stream as well [207]. Here, the principles of instantaneous frequency computation (see Sect. 2.2.3.1) are applied and the difference of the unwrapped phases $\tilde{\Phi}$ is used:

$$d_{\mathrm{bel}}(k,n) = \mathrm{princarg} \left[\tilde{\Phi}_X(k,n) - 2\tilde{\Phi}_X(k,n-1) + \tilde{\Phi}_X(k,n-2) \right]. \tag{6.7}$$

Goto and Muraoka proposed a distance which compensates for slow frequency variation over time [208]. They identify all bin indices k with

(a) higher power than the maximum of the four closest preceding bins:

$$\begin{aligned} A &= X(k,n-1)^2, \\ B &= X(k-1,n-1)^2, \\ C &= X(k+1,n-1)^2, \\ D &= X(k,n-2)^2, \\ E_{\max}(k,n) &= \max(A,B,C,D), \end{aligned} \tag{6.8}$$

and

(b) the same condition fulfilled for the maximum power of the three closest neighboring bins:

$$E_{k,n+1} = \max\left(X(k,n+1)^2, X(k-1,n+1)^2, X(k+1,n+1)^2\right). \tag{6.9}$$

The distance is then computed from the maximum of the current and following power value $E_t(k,n) = \max\left(X(k,n)^2, X(k,n+1)^2\right)$ by

$$d_{\mathrm{got}}(k,n) = \begin{cases} E_t(k,n) - E_{\max}(k,n), & \text{if } \left(X(k,n)^2 > E_{\max}(k,n)\right) \wedge \\ & \left(E_{k,n+1} > E_{\max}(k,n)\right) \\ 0, & \text{otherwise} \end{cases}, \tag{6.10}$$

$$d_{\mathrm{got}}(n) = \sum_{k=0}^{\mathcal{K}/2-1} d(k,n). \tag{6.11}$$

This distance strongly depends on the ratio of STFT size and sample rate as well as the block overlap ratio.

Röbel proposed a transient detection that utilizes the COG of the instantaneous energy [209]. He calculates this COG per arbitrary frequency band with

$$t_{\mathrm{cg}}(t_m) = \frac{\int_{\omega_l}^{\omega_h} -\frac{\partial\Phi(\omega,t_m)}{\partial\omega} |X(\omega,t_m|^2\, d\omega}{\int_{\omega_l}^{\omega_h} |X(\omega,t_m)|^2\, d\omega}. \tag{6.12}$$

The derivation of time reassignment is closely related to frequency reassignment and is a way of virtually improving the time resolution.

Zhou and Reiss presented an onset detector utilizing a filterbank of first-order complex resonators with an overall number of 960 frequency bands and 10 filters covering the range of a semi-tone, respectively [210]. They distinguish between hard and soft onsets and use the half-wave rectified energy difference per band for the detection of hard onsets and a pitch-based detection for so-called soft onsets by detecting steady-state tonal components and locating the corresponding onset position by searching for a salient energy increase.

6.3.2 Peak Picking

Although some systems for the automatic extraction of tempo and rhythm features utilize the extracted novelty function directly, other systems use it only as an intermediate result from which a series of onset times is derived. This is done by *peak picking* the novelty function. The final tempo and rhythm features are then derived from this series of onsets.

A flawless novelty function would indicate an onset at each local maximum. In reality, however, using the locations of local maxima directly as onset times will cause a large number of falsely detected onsets [*False Positives (FPs)*]. To suppress peaks of no interest, a threshold G is applied to the novelty function and only peaks above this threshold are considered as onset position candidates.

In the simplest case, the threshold is a fixed threshold:

$$G_{d,\mathrm{c}} = \lambda_1. \tag{6.13}$$

An alternative to this fixed threshold is using a signal-adaptive threshold. This adaptive threshold could be computed from the smoothed version of the novelty function. A typical smoothing filter is the MA filter:

$$G_{d,\mathrm{ma}} = \lambda_2 + \sum_{j=0}^{\mathcal{O}-1} b(j) \cdot d(i-j) \tag{6.14}$$

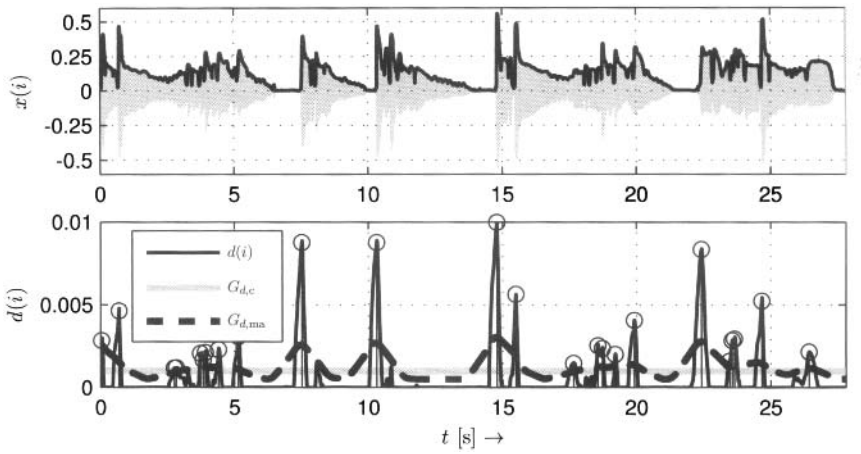

Figure 6.5 Audio signal and extracted envelope (top) and novelty function with exemplary thresholds for the peak picking process

with $b(j)$ representing a user-defined window function. The weight λ_2 shifts the threshold for adjusting the algorithm's sensitivity. Alternatively, a *median filter* may replace the MA filter; its output $\hat{Q}_d(0.5)$ is an estimate of the median in the block of an appropriate length $\mathcal{K}$:

$$G_{d,\mathrm{me}} = \lambda_2 + \hat{Q}_d(0.5). \tag{6.15}$$

Figure 6.5 exemplifies the process of peak picking by displaying a simple amplitude-based novelty function and two thresholds (static and adaptive) for the detection of local maxima. The input signal is a monophonic saxophone recording.

There are other ways of increasing the robustness of the onset detection system. Examples of additional criteria are the usage of the "amplitude" distance between the local maximum and the preceding local minimum as an indicator of onset strength and the extraction of the novelty function's slope before the local maximum as a cue for the likelihood of an onset.

An additional post-processing step is applied occasionally: the detected onsets may have very close proximity in time, and depending on the task at hand it might be beneficial to combine two or more closely neighbored onsets into one. The exact process of combining several onsets is, however, not as straightforward as one could wish. Possibilities are to choose the earliest onset, to choose the onset with the highest weight, or to compute some kind of average as the resulting (combined) onset time.

The final result of the onset detection system is a series of estimated onset times $\hat{t}_\mathrm{o}(j)$, possibly including an additional weight or intensity information as additional property. The markers in Fig. 6.5 indicate the estimated onset positions.

6.3.3 Evaluation

The *evaluation* of onset detection systems is a good example for the evaluation of ACA systems in general. Early publications on onset detection described their evaluation methodology and evaluation metrics as far less elaborate than the algorithm itself. To give an example, the number of correct detections was commonly reported with only a fuzzy de-

scription of what the definition of a correct detection actually is. The lack of information on both the test procedure and the test signals made it nearly impossible to estimate the algorithm's detection performance and to compare the results between different publications on onset detection. Only during the last decade has the problem of proper evaluation moved into the researcher's focus. In the context of the *Music Information Retrieval Evaluation eXchange (MIREX)*,[3] an annual evaluation campaign for MIR algorithms coupled to the *International Society for Music Information Retrieval (ISMIR)*, effort has been made in proposing a standardized test environment for audio onset detection systems.

The main problems in evaluating audio onset detection systems can be summarized as:

- *Lack of proper definition of the term onset*: Frequently it is neither entirely clear what the system aims to detect exactly (e.g., AOT vs. POT) nor what the required measurement accuracy is.
- *Lack of an adequate amount of test material*: The effort and error-proneness of the manual annotation of onset times in real-world signals makes it difficult to produce a sufficient amount of ground truth test data.
- *Lack of standardized and critical test material*: Comparison between different algorithms is hard without publicly available training and test databases.
- *Lack of simple yet meaningful evaluation metrics*: The evaluation results have to be computed and presented in a way that enables the estimation and comparison of the system's accuracy and robustness.

6.3.3.1 Procedure

The following parameters may be of interest when evaluating onset detection systems:

- detection performance,
- detection accuracy,
- robustness for noisy and band-limited input signals, and
- workload of the algorithm.

For each of these parameters, the definition of meaningful rating metrics with a predefined range is desirable. The type, the properties, and the amount of test signals and ground truth has to be specified to make results as comparable as possible.

Detection Performance

The *detection performance* is probably of the highest interest in the evaluation of onset detection systems. Obviously it should be a measure of how many onsets were correctly detected and how many onsets were incorrectly detected. The extracted onset times have to be compared with previously defined reference onset times as given by the ground truth. Two possible errors can occur: a *False Negative (FN)* indicating that no onset is detected at the time of a reference onset, and a *False Positive (FP)* which is an onset that is wrongly detected where no reference onset is found. Both of these measurements presume the definition of a time tolerance window around the reference onset time in which a detection

[3]MIREX Home. http://www.music-ir.org/mirex. Last retrieved on Nov. 25, 2011.

Table 6.1 Overview of different descriptions of the number of correct and incorrect detections and their relation

	Det. Positives	*Det. Negatives*	Σ
Ref. Positives	O_{TP}	O_{FP}	O_{DP}
Ref. Negatives	O_{FN}	O_{TN}	O_{DN}
Σ	O_{RP}	O_{RN}	

is counted as correct; usually, the length of this window is set to 50–100 ms. Similar to the definition of FPs and FNs, the *True Positives (TPs)* are the correctly detected onsets and the *True Negatives (TNs)* are the positions at which correctly no onset has been detected. In summary, the detection performance can depend on the following values:

- the number of TPs (correctly detected onsets) O_{TP},
- the number of FPs (falsely detected onsets) O_{FP},
- the number of FNs (missed onsets) O_{FN},
- the number of onsets in the reference data set O_{RP}, and
- the number of detected onsets $O_{\mathrm{DP}} = O_{\mathrm{TP}} + O_{\mathrm{FP}}$

Table 6.1 visualizes these numbers and their inter-relationship. It refers to onsets as *positives* and non-onsets as *negatives* to generalize the table to a two-class problem.

The internal parameters of the onset detection system should be adjusted for the desired "working point" before the evaluation itself can be carried out. The so-called *Receiver Operating Curve (ROC)* is an intuitive way to visualize the trade-off between TPs and false detections [211]. In the case of the evaluation of onset detection we would plot on one axis the TPs, on the other axis the sum of FPs and FNs. Each different parametrization of the algorithm results in exactly one point in the two-dimensional space of the ROC plot. The parameterizations of interest are the ones that maximize the TPs and at the same time minimize the false detections. The ratio of FPs and FNs can also be of interest for algorithm parameterization. If these two errors are considered to be equally bad, then the ratio should be near the value 1.

Several measurements of detection performance have been proposed in the past. Cemgil et al. proposed the relation of the total number of detections O_{DP}, the number of FNs O_{FN}, and the total number of reference onsets O_{RP} as a measure of detection performance [212]:

$$q_{\mathrm{cemgil},1} = \frac{O_{\mathrm{DP}} - O_{\mathrm{FN}}}{O_{\mathrm{RP}}}. \tag{6.16}$$

While this is a simple definition of the detection rate, it does not take into account the falsely detected additional onsets, and thus can result in misleading values in the case of many FPs.

Liu et al. proposed a similar value for the detection rate, additionally taking into account the number of FPs O_{FP} [213]:

$$q_{\mathrm{liu}} = \frac{\max(O_{\mathrm{DP}}, O_{\mathrm{RP}}) - (O_{\mathrm{FN}} + O_{\mathrm{FP}})}{\max(O_{\mathrm{DP}}, O_{\mathrm{RP}})}. \tag{6.17}$$

At least theoretically, the result can be negative; this is not desirable for a detection rate measure that should be in the range between 0 and 1.

A different reliability measurement (a measurement of relative error) has been proposed by Lerch [214]

$$: q_{\text{lerch}} = \frac{O_{\text{DP}} - (O_{\text{FN}} + O_{\text{FP}})}{O_{\text{RP}} + (O_{\text{FN}} + O_{\text{FP}})}. \tag{6.18}$$

The resulting value has the desired range between 0 and 1. The number of missing detections has the same weight as the number of false positives. In some contexts it might be desirable to weight O_{FN} and O_{FP} by different values when one should have a higher influence than the other. In these cases, a scaling factor λ in the range of $[0; 1]$ can be introduced which weights the sum of missing and falsely detected onsets: $\lambda \cdot O_{\text{FN}} + (1 - \lambda) \cdot O_{\text{FP}}$. Then, however, there is the possibility of negative results of Eq. (6.18).

The established statistical evaluation measures precision P and recall R allow a more systematic approach. Precision is the fraction of correctly detected onsets from all detected onsets:

$$P = \frac{O_{\text{TP}}}{O_{\text{TP}} + O_{\text{FP}}} = \frac{O_{\text{TP}}}{O_{\text{DP}}} \tag{6.19}$$

and recall is the fraction of correct detections from all reference onsets:

$$R = \frac{O_{\text{TP}}}{O_{\text{TP}} + O_{\text{FN}}} = \frac{O_{\text{TP}}}{O_{\text{RP}}}. \tag{6.20}$$

Precision and recall can be combined into the so-called *F-Measure* F. Mathematically it is the harmonic mean of precision and recall:

$$F = \frac{2PR}{P + R} = \frac{2 \cdot O_{\text{TP}}}{2 \cdot O_{\text{TP}} + O_{\text{FP}} + O_{\text{FN}}} = \frac{2 \cdot O_{\text{TP}}}{O_{\text{RP}} + O_{\text{DP}}}. \tag{6.21}$$

Note that for the F-Measure to produce reliable results, the number of positives has to roughly equal the number of negatives in the test set. If this is not the case (which may very well be true in the case of onset detection) the results will be biased.

Until now we only evaluated correct versus incorrect detection. In addition, a detection performance measurement could also include the time distance between the reference and the detected onset time and thus include a measure of *detection accuracy* (see below). A detected onset would be weighted with respect to its proximity to the reference. An intuitive way to do so could be to weight the distance $\Delta t_{\text{R,D}}$ between reference and detected time with a window function $w(\Delta t)$. Cemgil et al. proposed such a measure in the context of evaluation of beat tracking systems with a Gaussian window function [compare the RBF in Eq. (5.20)] [215]. Adapted to the onset detection evaluation problem the evaluation measure would be

$$q_{\text{cemgil},2} = \frac{\sum_{\forall r} \max_{\forall t} w(\Delta t_{\text{R,D}})}{(O_{\text{RP}} + O_{\text{DP}})/2}. \tag{6.22}$$

This measurement has again the limitation that it is not able to correctly handle FPs.

Detection Accuracy

The *detection accuracy* evaluates the timing accuracy of the algorithm, as opposed to the *detection performance* which evaluated the number of correct and false detection within a relatively large tolerance window. The accuracy can be measured by investigating the time difference $\Delta t_{\text{R,D}}$ between reference and detected onset times. The distribution of the resulting time differences contains the necessary data for timing evaluation; values of

interest are the arithmetic mean which indicates the tendency of the system detecting onsets systematically too early or too late:

$$d_{\text{mean}} = \sum_{\forall j} \Delta t_{\text{R,D}}(j), \tag{6.23}$$

the absolute mean value indicating the average time distance between detected and reference onset is

$$d_{\text{abs}} = \sum_{\forall j} |\Delta t_{\text{R,D}}(j)|, \tag{6.24}$$

the standard deviation or a confidence interval

$$\sigma_d = \sqrt{\frac{1}{O_{\text{DP}}} \sum_{\forall j} (\Delta t_{\text{R,D}}(j) - d_{\text{mean}})^2}, \tag{6.25}$$

and the absolute maximum value of the deviation

$$d_{\text{max}} = \max_{\forall j} |\Delta t_{\text{R,D}}(j)|. \tag{6.26}$$

Furthermore, a measure of statistical significance such as the p-value (see, e.g., [216]) can be given to attest the reliability of the results.

Robustness and Workload

Some target applications require robustness of the algorithm against noisy and distorted signals or have specific requirements on the workload of a system.

The robustness against noise and bandwidth limitations can be carried out using the test scenario and metrics described above but with modified test signals. Possible modifications depend on the target application but could include added noise, down-sampling, and encoding with lossy audio encoders or speech encoders.

The evaluation of the algorithmic *workload* gives an estimate of the complexity and real-time capabilities of the system. This might be of particular interest if the algorithm has to run on embedded devices or has to process vast amounts of data. The actual investigation of the workload is more complex than it might seem at first glance. Even estimating the theoretical algorithmic complexity in a number of operations gets complicated as soon as specific functions such as trigonometric or exponential functions are used. The measurable execution time itself may be influenced by many different conditions, for example:

- the *hardware*: processor clock speed, vectorization (SIMD) functionality, cache size as well as memory access speed,
- the *implementation*: optimization of the source code, optimization of the compiler, and
- the *software*: operating system efficiency, audio and file IO.

Thus workload measurements usually give only rough impressions of the algorithms processing performance even on comparable hardware. The result of workload measurements can be given as the ratio between the required computation time t_{C} and the overall length of the tested audio data of the test database t_{A} by calculating $t_{\text{C}}/t_{\text{A}}$ with respect to the used processor.

6.3.3.2 Test Signal Databases

The test signal database for the evaluation of onset detection performance should preferably contain real-world signals such as signals in CD quality with the onset times annotated per hand as the ground truth. However, as several researchers point out, the manual annotation is a tedious and time-consuming task [184, 185]. Therefore, two alternatives for the generation of test sequences may be considered: acoustic recordings with a symbolic trigger (such as recordings of the Yamaha Disklavier) and audio data synthesized from symbolic data. In both cases, the reference onsets are available in the MIDI format, allowing the easy automated extraction of NOTs. Given the range of the typical tolerance interval of 50 ms, the difference between NOT and POT can sometimes be neglected.

The test database should include the following signal types to make the evaluation as general as possible:

- *various genres* (pop, rock, symphonic, chamber music, electronic, etc.),
- *various instrumentations*,
- *different tempi and musical complexity*, and
- *critical signals* which are signals with very low detection performance. In the case of onset detection these might include noisy signals and signals containing various kinds of tremolo and vibrato.

As mentioned above, databases with manually annotated audio files are difficult to find. This is on the one hand due to intellectual property issues, on the other hand due to the time-consuming task of annotation. Two examples for publicly available databases are the data set published by Leveau[4] and the data set published by Glover.[5]

6.4 Beat Histogram

The *beat histogram* or *beat spectrum* is a way to visualize some rhythmic properties of the signal. Similar to the "normal" magnitude spectrum, the frequency (in this case with the unit BPM) is assigned to the abscissa and the magnitude (beat strength) is assigned to the ordinate. Peaks in the histogram should therefore correspond to the main tactus and its integer multipliers and divisors. The beat histogram can be interpreted as the frequency domain representation of the novelty function. There are multiple ways of computing such a beat histogram.

Scheirer used a closely spaced filterbank of comb resonance filters and used the filter's output energy as indication of the *beat strength* [203].

Foote and Uchihashi proposed to construct a similarity matrix (compare the distance matrix $\boldsymbol{D}$ in Sect. 7.1) from the cosine distance between all pairs of STFTs from the audio file and then derive the beat histogram by summing the similarity matrix along its diagonal [217].

Tzanetakis and Cook split the audio signal into four octave bands and extract the envelope per band by applying four processing steps [60]:

1. *Full-Wave Rectification (FWR)* by computing the absolute value,

[4]Leveau, Pierre. https://sites.google.com/site/pierreleveau/research. Last retrieved on Nov. 25, 2011.
[5]Glover, John. http://www.johnglover.net/audiosoftware.html. Last retrieved on Nov. 25, 2011.

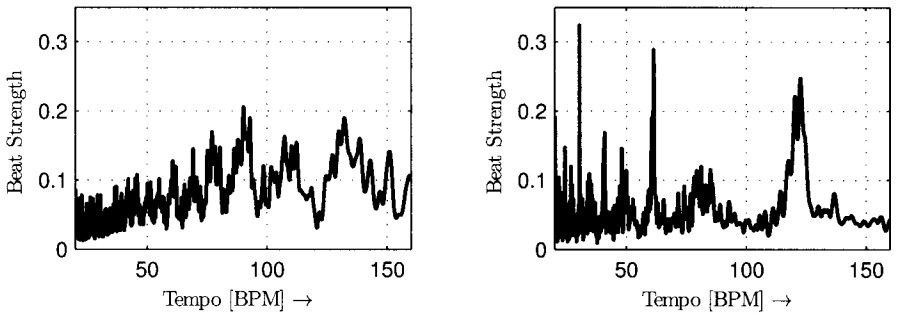

Figure 6.6 Beat histogram of a piece of popular music (left) and of a string quartet performance (right)

2. *envelope smoothing* by low-pass filtering,
3. *down-sampling* to reduce the complexity by reducing the sample rate, and
4. *DC removal* by subtracting the arithmetic mean.

An ACF (with harmonic processing as described in Sect. 5.3.4.2) is then computed in order to identify (rhythmic) envelope regularities. The beat histogram is construed by taking three peaks in the search range and adding their amplitude to the beat histogram. This is done for each texture window.

Figure 6.6 visualizes the beat histogram of an excerpt of popular music in comparison with the histogram extracted from a string quartet performance. Note that for this plot the beat histogram calculation is based on a very simple novelty function derived in the time domain from the signal's magnitude. After computing the ACF, the amplitude of each individual lag is mapped into the beat domain as the beat strength (this leads to a high resolution at low frequencies and a low resolution at high frequencies). The beat histogram computed from the popular music example has clearly defined peaks at multiples of a base frequency; such a pattern is not identifiable in the right beat histogram computed from a string quartet recording.

6.4.1 Beat Histogram Features

Similar to an audio magnitude spectrum, the beat histogram can be represented by (simple yet meaningful) features. Widely used is the set of features introduced by Tzanetakis and Cook consisting of [60]

- the overall sum of the histogram,
- the relative amplitude of the highest peak,
- the relative amplitude of the second highest peak,
- the amplitude ratio of second highest to highest peak, and
- the BPM frequencies of the highest and second highest peak.

Burred and Lerch evaluated a feature set including statistical features of the beat histogram such as its arithmetic mean, standard deviation, kurtosis, skewness, and entropy; they additionally used a measure for what they called *rhythmic regularity*. The rhythmic regularity is a measure of how much the computed ACF differs from the linear weighting function of a block-wise ACF [218].

6.5 Detection of Tempo and Beat Phase

Systems for *tempo detection* (also referred to as *tempo induction* or *tempo tracking*) usually compute some kind of novelty function in a first processing step. Regardless of whether discrete onset times are picked from the novelty function or the novelty function is used directly, it is far from trivial to derive the pulse which would be perceived as tactus since the main periodicity is frequently not clearly identifiable.

The detection of specific beat positions is not necessarily required for detecting the tempo itself as the tempo is based on the *distance* between the beats rather than their absolute position. Therefore, a periodicity analysis of the novelty function is sufficient for tempo estimation when knowledge of the absolute beat positions is not required.

Given a tempo, the absolute beat position is — due to the similarity to the relation of frequency and phase — referred to as the *beat phase*. This beat phase is, for example, required for the so-called beat matching technique for which two or more pieces of music with different musical content (but the same tempo) are synchronized at their beat positions to generate a so-called *mash-up*. Without knowing the exact beat positions it would not be possible to mix those pieces in a musically meaningful way.

The first beat tracking systems avoided the complexity of extracting a novelty function and focused exclusively on the beat tracking part; they used onset times from symbolic data such as MIDI files as input of their beat tracking system. Thus, they assumed a perfect computation of the novelty function and of the onset detection, respectively. Allen and Dannenberg used a beam search to consider multiple hypotheses of the beat phase, utilizing heuristics to select the most likely hypothesis [219]. Large presented a system utilizing an oscillator for generating pulses; the system was able to simultaneously adapt its current tempo estimate and beat phase estimate to the onset times [220]. The adaption speed is based on the distance between the estimated beat position and the actual onset position.

Goto and Muraoka used multiple agents to estimate tempo and beat phase. Each agent has its own hypothesis of the tempo and the beat phase and computes its own reliability by measuring the coincidence of the estimated beat positions with the (extracted) onset positions [208]. Each agent will have different parameter and initialization settings. The agent with the highest reliability is chosen as the one providing the most likely tempo and beat phase estimate. The system has been developed for several years; a predecessor of the system is described in [221]. Later versions of the system extend the computation of the novelty function by detecting changes in the tonal components to get additional information on the salience of onsets and estimated beats [222, 223].

A system which is not based on using discrete onset times has been presented by Scheirer [203]. He computes the envelope in six frequency bands and subjects it to HWR; the six output signals are then used as inputs to a filterbank of closely spaced comb resonance filters. A major difference between this approach and most others is that on the one hand it yields the strength of all detectable tempi and can therefore be used to calculate a beat histogram (see above), but on the other hand the tempo estimate is restricted to the

filterbank's resolution. A similar system based on comb resonance filters has also been used by Klapuri [224].

Dixon's tempo detection system analyzes clusters of all IOIs within a short time window, quantizes them and uses the IOI histogram to find the best tempo estimate [225]. In order to detect the beat phase, multiple competing agents work in parallel with different initializations; the evaluation function for choosing the agent with the most likely beat phase estimate throughout the piece is a measure of regularity of the IBIs and the salience of the chosen onsets [226]. In a related publication he argues that a beat tracking system can benefit from more sophisticated information such as a salience measure based on note duration (or to be more exact, the IOI), intensity, and pitch [227]. Meudic modified Dixon's beat tracking system to work in real-time [228].

Similar to Scheirer, Gouyon and Herrera presented a system utilizing a continuous novelty function as opposed to a series of discrete onset times. The overall novelty function is derived from a set of various features [229]. The tempo is then found by seeking periodicities in the ACF of the novelty function. The choice of the most likely tempo candidate is computed by using a "harmonic grid"; the ACF values of multiples of the current tempo hypothesis are used to estimate its likelihood. A related approach to harmonic processing of the ACF can be found in the fundamental frequency detection system of Karjalainen and Tolonen as described in Sect. 5.3.4.2.

Laroche used a spectral flux-based novelty function and computed the correlation function between a set of quantized template delta pulses for various tempi and beat positions in a window with the length of several seconds [205]. The most salient 10 to 15 maxima are used as initial tempo candidates. Finally, he applies *Dynamic Programming (DP)* techniques to find the most likely overall path through all the tempo candidates for all analysis windows over the whole audio file [for a related algorithm see the description of *Dynamic Time Warping (DTW)* in Sect. 7.1]. A similar approach has been published by Peeters [230]. The main difference to Laroche's approach is that he computes both an ACF and an STFT from the novelty function and combines the results of both to estimate the tempo candidates. The advantage of this combined representation is increased robustness against octave errors, one of the typical problems of periodicity analysis. Furthermore, he adds three different meter templates as possible states to the tempo candidates; his DP approach utilizes the *Viterbi algorithm* [231] to find the most likely tempo path through the "audio file." The advantage of the latter two approaches is that their DP techniques should enable them to deal with sudden changes in tempo and, in the case of Peeters, meter.

6.6 Detection of Meter and Downbeat

The relation of *meter* and *downbeat* is very similar to the relation of tempo and beat phase. Just as the tempo is derived from the distance between two neighboring beats, the meter is (usually) the length of a bar while the downbeat marks the beginning of a bar.

The hierarchical metrical structure of music makes the differentiation of detecting the beat and the downbeat basically a question of the hierarchical level to investigate. Now, we are just interested in long-term periodicities. The algorithms are therefore quite similar; the main differences can be found in the search range and in the computation of the novelty function.

Brown weighted the series of onsets with their IOI (in order to increase the impact of long notes) and computed the ACF of this series of weighted onsets to detect the meter

[232]. Toiviainen and Eerola used a similar approach and evaluated different weighting functions for the IOIs [233].

Uhle and Herre derived bar length candidates from integer multiples of a previously detected tatum and then computed the CCF of two snippets of the novelty function per frequency band to derive a measure of the likelihood of the individual bar length candidates [190].

Escalona-Espinosa argued that it is not only the onset pattern itself that is of interest for the estimation of meter and downbeat but other features should be used for computing the novelty function as well [234]. More specifically, he assumed that in the western tradition of music (and even more so in the case of popular music) the position of a downbeat increases the likelihood of both

- a note or harmony change and
- the occurrence of a new bass note

compared to positions between downbeats. Therefore he proposed the computation of two novelty functions, one based on the pitch chroma difference and the second on the bass energy increase. The time resolution for this computation is signal adaptive: it is the previously estimated tatum. Using the tatum has the two advantages of a signal-related segmentation and higher computational efficiency of the following processing steps. As an alternative to the ACF-based approaches he constructed two matrices which contain the (self-) similarity between all pairs of samples of the two novelty functions. The matrices are called self-similarity matrices. When averaging the diagonals of each similarity matrix the result is a measure of periodicity with respect to the distance from the main diagonal. Depending on the similarity (or distance) measure used for computing the similarity matrix, this function can be closely related to the ACF. The lag of the main peak within a pre-defined search range is then the detected bar length. In combination with a tempo estimate the result allows him to derive the time signature of the piece of music. The most likely downbeat position is estimated with the extracted bar length by computing the CCF of each novelty function with a delta pulse spaced with the bar length period. The lag of the CCF's maximum indicates the downbeat position.