CHAPTER 5

TONAL ANALYSIS

Tonal aspects play an important role in understanding and analyzing music, as can be seen from the vast number of pitch-related publications in music theory. Pitches are the basic building blocks of key, melody, and harmony of a piece of music.

5.1 Human Perception of Pitch

The human perception of *pitch* is directly related to the frequency of a signal in a way that higher frequencies will lead to the perception of a higher pitch. If the signal is a combination of sinusoidal components with the frequencies $f_0, 2f_0, 3f_0, \ldots$ (which is a reasonable approximation for tonal sounds produced by many musical instruments), then the *fundamental frequency* f_0 dominates the pitch perception. As a matter of fact, humans will usually even perceive the same pitch for this combination of harmonics if the fundamental frequency f_0 has low power or is missing.

5.1.1 Pitch Scales

The relation between the fundamental frequency and the perceived pitch is non-linear; at higher frequencies, two pitches with the same perceived pitch distance will have a larger frequency distance than at lower frequencies. Simply speaking, the non-linearity is tied to the frequency resolution of the human cochlea. There are different approaches to measure this cochlear frequency map and thus there exist different ways of describing it. The

An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics,
First Edition. Alexander Lerch.

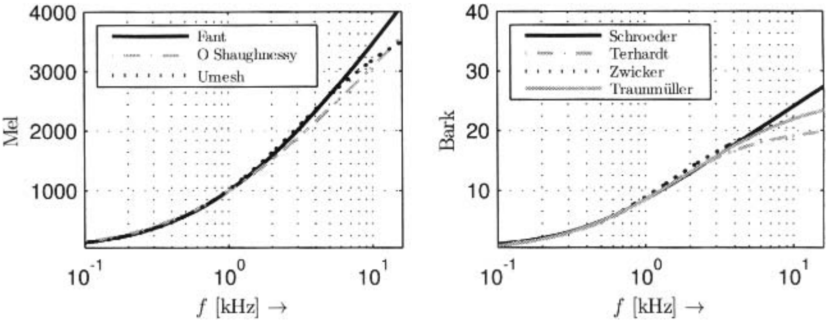

Figure 5.1 Different models for the non-linear mapping of frequency to mel (left) and bark (right)

most common models are the *mel scale* and the *critical band rate*, also called *bark scale* (Fig. 5.1). There has been a number of proposals for analytical functions to approximate the measurement data resulting from listening test for these scales. In most practical audio analysis applications, the use of one or another approximation is apparently circumstantial (compare [95, 96]). Nevertheless, several of these approximations will be presented below to illustrate the number of options.

5.1.1.1 Mel Scale

The term *mel* was introduced in 1937 by Stevens et al. as the name of a subjective pitch unit [82]. The *mel scale* is a measure of tone height. The empirical data used to build the numerous analytical models of the mel scale stems from only a limited number of psychological experiments: the most important test results have been presented by Stevens et al. [82], Stevens and Volkmann [97], and Siegel [98]. Three models will be presented here; the two older models by Fant [99]

$$\mathrm{m_F}(f) = 1000 \cdot \log_2\left(1 + \frac{f}{1000\,\mathrm{Hz}}\right) \tag{5.1}$$

and O'Shaughnessy [100]

$$\mathrm{m_S}(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700\,\mathrm{Hz}}\right) \tag{5.2}$$

are most commonly used. Note that the latter model is sometimes also referenced in the form

$$\mathrm{m_S}(f) = 1127 \cdot \log\left(1 + \frac{f}{700\,\mathrm{Hz}}\right). \tag{5.3}$$

The third model proposed by Umesh et al. [101] appears in this list mainly to demonstrate the variety of models and is not as widely known as the two other models:

$$\mathrm{m_U}(f) = \frac{f}{2.4 \cdot 10^{-4} f + 0.741}. \tag{5.4}$$

5.1.1.2 Bark Scale

The *bark scale* or *critical band rate* is constructed from the bandwidth of measured frequency groups, the critical bands. Zwicker and Fastl suggest that the bark scale is related to the mel scale by $1\,\mathrm{bark} = 100\,\mathrm{mel}$ [47].

The most prominent models of the bark scale have been proposed by Schroeder et al. [102]

$$\mathfrak{z}_{\mathrm{S}}(f) = 7 \cdot \operatorname{arcsinh}\left(\frac{f}{650\,\mathrm{Hz}}\right), \tag{5.5}$$

Terhardt [103]

$$\mathfrak{z}_{\mathrm{T}}(f) = 13.3 \cdot \arctan\left(0.75 \cdot \frac{f}{1000\,\mathrm{Hz}}\right), \tag{5.6}$$

and Zwicker and Terhardt [104]

$$\mathfrak{z}_{\mathrm{Z}}(f) = 13 \cdot \arctan\left(0.76 \cdot \frac{f}{1000\,\mathrm{Hz}}\right) + 3.5 \cdot \arctan\left(\frac{f}{7500\,\mathrm{Hz}}\right). \tag{5.7}$$

Traunmüller's model [105] is not as well known but is simple to calculate

$$\mathfrak{z}_{\mathrm{TM}}(f) = \frac{26.81}{1 + {}^{1960}/_{f}} - 0.53. \tag{5.8}$$

5.1.1.3 Other Models

Many more models have been proposed over the years for the non-linear transformation of frequency to perceptual frequency groups, pitch height, and position on the human *cochlea*.

Moore's model for the *Equivalent Rectangular Bandwidth (ERB)* can be seen as a model "competing" to the critical band rate [83]

$$\mathfrak{e}(f) = 9.26 \log\left(1 + \frac{f}{228.7}\right). \tag{5.9}$$

Terhardt introduced a function, which he named *SPINC*, as an alternative to the mel scale [106]:

$$\mathfrak{s}(f) = 1414 \arctan\left(\frac{f}{1414\,\mathrm{Hz}}\right), \tag{5.10}$$

and Greenwood proposed the following equation to compute the position (normed to the range of $[0; 1]$) of a specific frequency on the cochlea [107]:

$$\mathfrak{x}(f) = \frac{1}{2.1} \log_{10}\left(\frac{f}{165.4} + 1\right). \tag{5.11}$$

5.1.2 Chroma Perception

There is an additional facet to human pitch perception: not only do we perceive pitch height from low to high but we tend to group pitches with specific frequency ratios [108, 109]. More specifically, humans perceive frequencies with a frequency ratio of a power of 2 (such as $f_0, 2f_0, 4f_0, 8f_0, \ldots$) as very similar and closely related to each other. This phenomenon is usually called *chroma perception*.

Figure 5.2 visualizes this in a helix plot. On the one hand, the frequency is monotonically increasing on the z axis, modeling the tone or pitch height. On the other hand, points with the same (x, y) coordinates share a frequency ratio of a power of 2 — they share the same scale degree and are in the same pitch class (see Sect. 5.2.1), respectively. The circle which appears when looking directly on the (x, y) plane would thus encompass all pitch classes.

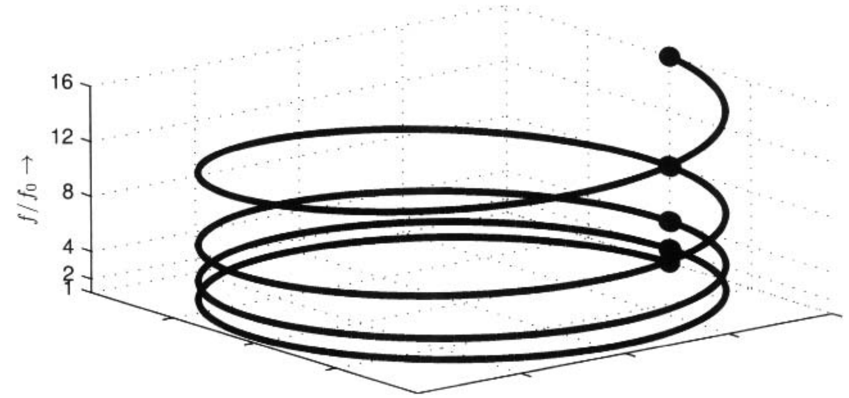

Figure 5.2 Helix visualizing the two facets of pitch perception: pitch height and chroma

5.2 Representation of Pitch in Music

Much can be said about pitch-related properties of music, covering not only the frequency mapping of specific pitches but also interaction of pitches in chords and melodies, harmony progression, and musical key. It is not the intention of this section to give a comprehensive overview on the (music) theory involved; instead, some basics will be covered in a simplified manner since the understanding of some theoretical background can be of help in the successful design of analysis algorithms.

5.2.1 Pitch Classes and Names

The concept of musical pitch in western music theory closely follows the pitch helix (see Fig. 5.2) in that each *octave*, i.e., each range with boundaries with a frequency ratio of 2 : 1, is divided into the same chunks: the 12 *pitch classes*.

The common labels of these octave-independent pitch classes are shown in Table 5.1.

Table 5.1 Pitch class indices and corresponding names of the chromatic pitch classes

0	1	2	3	4	5	6	7	8	9	10	11
C	$C\sharp/D\flat$	D	$D\sharp/E\flat$	E	F	$F\sharp/G\flat$	G	$G\sharp/A\flat$	A	$A\sharp/B\flat$	B

Seven of those pitch classes form the so-called *diatonic scale*.

Table 5.2 shows a diatonic scale, more specifically the major mode (see Sect. 5.2.3) based on a root note C with pitch class index, pitch class name, Solfège name and distance to the previous pitch class in semi-tones.

As can be seen from the table, the distance between two neighboring pitches is in most cases two semi-tones; twice, however, it is only one semi-tone. Any pitch between the presented diatonic pitches can be constructed be raising the pitch (e.g., $C \rightarrow C\sharp$) or lowering it (e.g., $E \rightarrow E\flat$). For now it will be assumed that, e.g., $C\sharp$ equals $D\flat$ (a relation that is

Table 5.2 Names and distance in semi-tones ΔST of diatonic pitch classes

Index	*Name*	*Solfège Name*	ΔST
0	*C*	Do	1
2	*D*	Re	2
4	*E*	Mi	2
5	*F*	Fa	1
7	*G*	Sol	2
9	*A*	La	2
11	*B*	Si	2

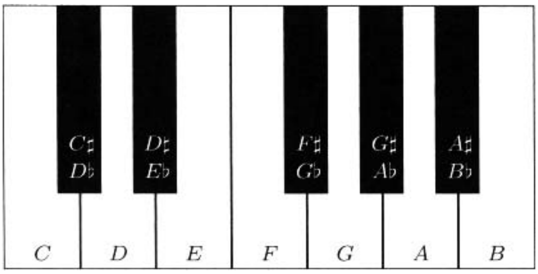

Figure 5.3 One octave on a piano keyboard with annotated pitch class names

referred to as *enharmonic equivalence*), resulting in 12 pitch classes per octave as shown in Table 5.1.

Figure 5.3 depicts these pitch classes on a piano keyboard; the white keys form the mode C $Major$ mode as shown in Table 5.2. Figure 5.4 displays the pitch classes in one octave in musical score notation.

A common convention for naming musical pitches used in the following is simply the pitch class name followed by an octave index, e.g., $C2$ or $A4$. Each new octave starts with a C by convention.

5.2.2 Intervals

The distance between two pitches is the musical *interval*. It is used for both the distance of simultaneously sounding pitches and pitches sounding one after another. Table 5.3 names commonly used intervals and their corresponding distance in semi-tones.

Figure 5.5 displays the most important intervals (rising from pitch $C4$) in musical score notation.

Humans will hear the same interval for different pairs of pitches if the ratio between their fundamental frequencies is the same.

5.2.3 Root Note, Mode, and Key

The musical *key* of a tonal piece of music is defined by both its *mode* and a *root note*.

The root note is the most important pitch class in a specific key. It is also referred to as the *first scale degree* and will usually appear most frequently in a piece of music.

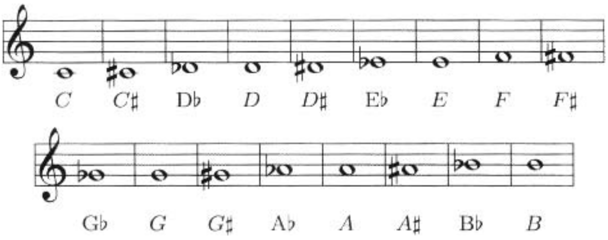

Figure 5.4 Musical pitches $C4 \ldots B4$ in musical score notation; enharmonically equivalent pitches are displayed twice

Table 5.3 Names of musical intervals, their enharmonic equivalents, and their pitch distance in semi-tones

Interval	*Enharmonic Equivalent*	ΔST
Unison	Diminished Second	0
Minor Second	Augmented Unison	1
(Major) Second	Diminished Third	2
Minor Third	Augmented Second	3
Major Third	Diminished Fourth	4
(Perfect) Fourth	Augmented Third	5
Augmented Fourth	Diminished Fifth/Tritone	6
(Perfect) Fifth	Diminished Sixth	7
Minor Sixth	Augmented Fifth	8
Major Sixth	Diminished Seventh	9
Minor Seventh	Augmented Sixth	10
Major Seventh	Diminished Octave	11
(Perfect) Octave	Augmented Seventh	12

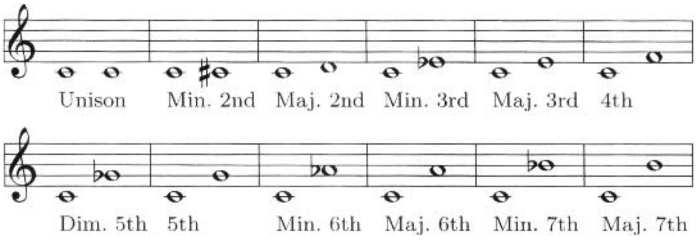

Figure 5.5 Musical intervals in musical score notation

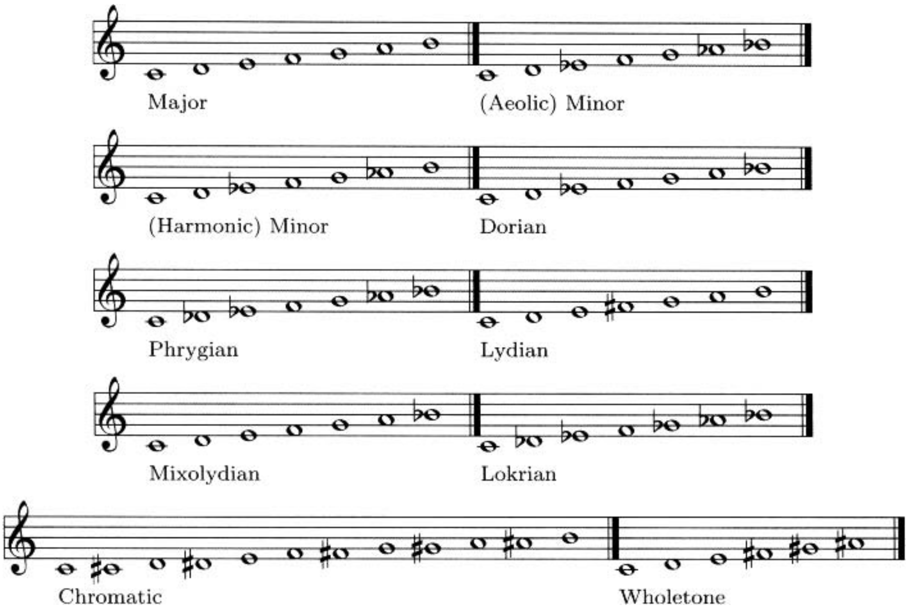

Figure 5.6 Different modes in musical score notation starting at the root note C

The *mode* defines a set of relative pitch relationships; an example would be: the distance between first and second scale degree is a major second, between first and third scale degree is a major third, etc. The most common modes are the *major mode* and the *minor mode*. Figure 5.6 displays an example set of different modes, all starting from the root note C. All modes except major mode and the two minor modes — aeolic mode and harmonic mode — are only of interest in specific musical styles and are usually not very important in the context of ACA.

The key thus defines the set of pitch classes which are used to construct the tonal aspects of a piece of music. In popular music it is common for a piece to have exactly one key. There are many exceptions to this rule: the key can change within a piece (when a so-called *modulation* occurs) and non-key pitches may be used for musical reasons.

Depending on the current root note and mode, up to six different accidentals have to be used to raise or lower the pitches in the musical score. The notational convention allows writing all key-inherent accidentals at the begin of a staff — the *key signature* — and to add only accidentals within the score where non-key-inherent pitches are used. Figure 5.7 shows major modes with their key signatures starting from all possible root notes.

As can be seen from Fig. 5.7, the root notes of keys that differ only in one accidental are always spaced by a fifth (F/C, C/G and G/D, etc.). This relationship can be visualized by the so-called *circle of fifths* (Fig. 5.8). Strictly speaking, this circle is only closed in the case of enharmonic equivalence when $F\sharp$ equals $G\flat$ (see Sect. 5.2.5.2).

Neighboring keys on the circle of fifths have all pitch classes but one in common; for example, $G\ Major$ has the same pitches as $C\ Major$ except for the F which is raised to an $F\sharp$ in $G\ Major$.

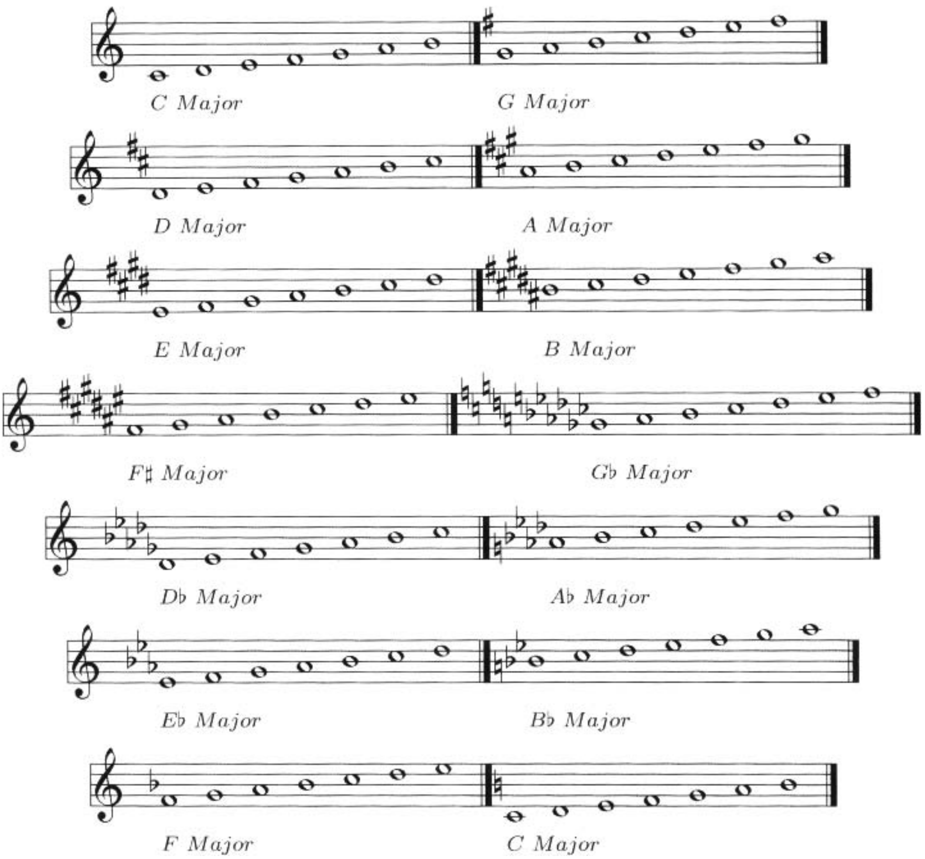

Figure 5.7 The twelve major keys in musical score notation, notated in the 4th octave

The circle of fifths also exists for the (aeolian) minor keys with *a minor* being the key without accidentals.

Since the circle of fifths shows the relation of different keys, it hints also at what modulations are more or less likely. To give an example, the most likely key changes from *F Major* would be either *C Major*, *B♭ Major*, or *d minor*. The circle of fifths can thus be understood as a model for a distance map between keys.

Keys construed from the same set of pitch classes such as *C Major* and *a minor* are called parallel keys; in the circle of fifths their distance is 0. In order to build an analytical model for key distances with a non-zero distance between parallel keys, the circle of fifths can be enhanced to a three-dimensional model. This model would feature two parallel planes, one containing the circle for major keys and the other for minor keys. Parallel keys thus share the same (x, y) coordinates but have a different z coordinate (which is then the distance between the two parallel keys).

Approaches to automatic key detection from audio signals can be found in Sect. 5.5.

5.2.4 Chords and Harmony

The simultaneous use of (usually no less than three) different pitches creates a chord. Chords built of three pitches are referred to as triads. The most common chord types are

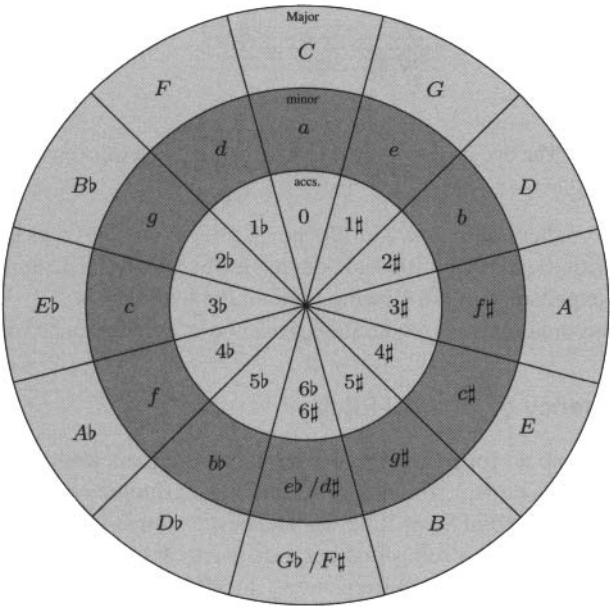

Figure 5.8 Circle of fifths for both major keys and minor keys, plus the number of accidentals of the key signature per key

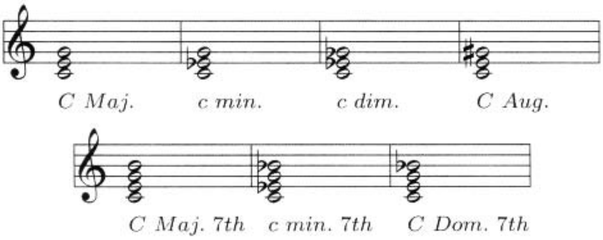

Figure 5.9 Common chords in musical score notation on a root note of C

built of third intervals — a few examples are displayed in Fig. 5.9 with respect to a root note of C. Note that the same naming convention is used to label both the key and some common chords (e.g., $C\ Major$), so one has to derive from the context which of the two is meant.

Chord-inherent pitches can be doubled in different octaves without changing the chord type. The chord's root note is the most commonly doubled pitch. If the lowest note of a sounding chord is not its root note it is called *inverted*. The first chord inversion of a $C\ Major$ triad would therefore have the pitch classes (bottom-up) $E - G - C$. Figure 5.10 shows the two possible inversions of a $D\ Major$ triad. The second inversion usually appears less frequent than the first.

Each chord can have one or more musical (harmonic) functions in a key and dependent on the musical context. The chord sharing the root note with the current key is referred to as the *tonic* and will most likely represent the tonal center. Other important harmonic

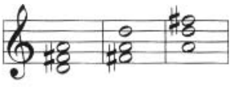

Figure 5.10 The two inversions of a *D Major* triad in musical score notation

functions are the so-called *dominant* with the key's fifth scale degree as the root note and the *subdominant* with the key's fourth scale degree as the root note. Dominant chords will usually induce an expectancy of a following tonic in the listener.

Approaches to automatic chord recognition from audio signals can be found in Sect. 5.6.

5.2.5 The Frequency of Musical Pitch

The most systematic model for relating musical pitch to frequency and vice versa is using the so-called equal temperament which also results in enharmonic equivalence (other temperaments will be mentioned in Sect. 5.2.5.2). The best example for such a transformation is the MIDI scale in which each semi-tone has a distance of 1 to its nearest neighbor. The equations for the transformation from frequency f to MIDI pitch $\mathfrak{p}$ and vice versa are

$$\mathfrak{p}(f) = 69 + 12 \cdot \log_2\left(\frac{f}{f_{A4}}\right), \quad (5.12)$$

$$f(\mathfrak{p}) = f_{A4} \cdot 2^{\frac{\mathfrak{p}-69}{12}}. \quad (5.13)$$

The reference frequency f_{A4} is the tuning frequency (see Sect. 5.2.5.1); using 12 times the logarithm to the base 2 ensures that each octave is divided into 12 parts of equal "length" and the constant 69 results in the pitch $A4$ having the index 69, a convention of the MIDI standard [3]. The MIDI pitch can be mapped easily to the pitch class index PC as introduced above by using a modulo operation:

$$\mathrm{PC}(\mathfrak{p}) = \mod(\mathfrak{p}, 12). \quad (5.14)$$

The unit cent $\Delta C(f_1, f_2)$ is a distance measure between two pitches or frequencies f_1 and f_2. It can be computed by

$$\begin{aligned}\Delta C(f_1, f_2) &= 100 \cdot (\mathfrak{p}(f_1) - \mathfrak{p}(f_2)) \\ &= 100 \cdot \left(\left(69 + 12 \cdot \log_2\left(\frac{f_1}{f_{A4}}\right)\right) - \left(69 + 12 \cdot \log_2\left(\frac{f_2}{f_{A4}}\right)\right)\right) \\ &= 1200 \cdot \log_2\left(\frac{f_1}{f_2}\right). \quad (5.15)\end{aligned}$$

A semi-tone interval has thus a distance of 100 cents and an octave has a distance of 1200 cents.

5.2.5.1 Tuning Frequency

The *tuning frequency* f_{A4} is the frequency of the *concert pitch* $A4$ (also: *standard pitch*) which is used for tuning one or more musical instruments. The tuning frequency is standardized internationally to 440 Hz [110], but the exact frequency used by musicians can

Table 5.4 Typical range of deviation of the tuning frequency from 440 Hz over three centuries

Year	*Lower Deviation*	*Upper Deviation*
1750	−50 Hz	+30 Hz
1850	−20 Hz	+20 Hz
1950	−5 Hz	+10 Hz

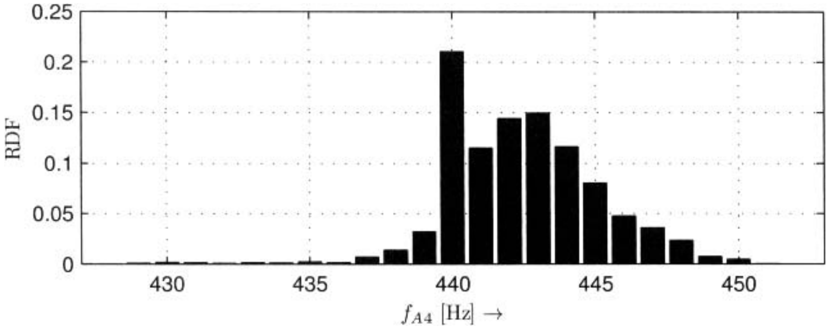

Figure 5.11 Distribution of tuning frequencies in Lerch's data set

vary due to various reasons such as the use of historic instruments or timbre preferences. Two performances of the same piece of music using different tuning frequencies will differ in their average pitch height.

The range of typical tuning frequencies decreased over the centuries. Table 5.4 shows this range for the past three centuries as deviation from 440 Hz [111].

Nowadays, while for many electronic music productions the "default" tuning frequency of 440 Hz is used, the tuning frequencies of orchestras still deviate from this standard tuning frequency. For example, the Chicago Symphony Orchestra and the New York Philharmonic tune at 442 Hz, while the Berliner Philharmoniker and the Wiener Philharmoniker have a tuning frequency of 443 Hz.[1] At least in the case of both European orchestras, the tuning frequency was higher in previous decades. The frequencies 442 and 443 Hz correspond to deviations of 7.85 and 11.76 cents from the standard tuning frequency, respectively.

There exist two studies analyzing the tuning frequency of a data set of recordings. Zhu et al. processed a database of 60 popular and 12 classical pieces[2] and found only three pieces of this database with a deviation of approximately 2–4 cents from the standard tuning frequency [112]. Lerch presented the results of a study with a large database of classical music, consisting of more than 3000 tracks and an overall playing time of approximately 291 hours [113]. A histogram of the extracted tuning frequencies as displayed in Fig. 5.11 shows a maximum at a tuning frequency of 440 Hz; the maximum itself consists of about 21% of the test database. The distribution has an arithmetic mean value of 442.38 Hz and a standard deviation of 2.75 Hz. The majority of the results (95%) is in the range from 439 to 448 Hz and only 50% of the results have a tuning frequency in the range of 440–443 Hz. The percentage of files below 439 Hz is about 3.3%.

[1] According to the orchestra's archivists, March and April 2006.

[2] The term *classical music* is in this context understood as "non-popular" music, as opposed to the epoch itself.

Table 5.5 Deviations of the Pythagorean, meantone, and two diatonic temperaments from the equally tempered scale in cents at a reference pitch of C (after Briner [111])

Pitch Class	*Equally*	*Pythagorean*	*Meantone*	*Diatonic Major*	*Diatonic Minor*
C	0	0	0	0	0
$C^{\#}$	0	–	–	–	–
D	0	+3.9	−6.9	+3.9	+3.9
E^b	0	–	–	–	+15.6
E	0	+7.8	−13.7	−13.7	–
F	0	−2.0	+3.4	−2.0	−2.0
$F^{\#}$	0	–	–	–	–
G	0	+2.0	−3.5	+2.0	+2.0
A^b	0	–	–	–	+13.7
A	0	+5.9	−10.2	−15.6	–
B^b	0	–	–	–	+17.6
B	0	+9.8	−17.1	−11.7	–

The tuning frequency is not necessarily static once the instruments have been tuned; it may change during a concert or a recording session. On the one hand the tuning frequency could be slowly decreasing as it sometimes happens at *a cappella* performances, on the other hand the tuning frequency may slightly increase, for example, due to a rising involvement of the musicians during the concert. The maximum range of this deviation can be assumed to be small in the case of professional musicians (about 3–5 cents).

Approaches to estimate the tuning frequency of a music signal can be found in Sect. 5.4.

5.2.5.2 Temperament

The temperament defines a system of frequency ratios for intervals.

In the equally tempered case assumed above the frequency ratio between the mid-frequencies of two pitches f_1 and f_2 spaced by N semi-tones (with N being a negative or positive integer) is always

$$\frac{f_1}{f_2} = 2^{N/12}. \tag{5.16}$$

Therefore, the distance between two pitches is constant and independent of tonal and harmonic context; it stays also constant if the enharmonic equivalents of the two pitches are used: the interval $B, F\sharp$ is the same as the intervals $B, G\flat$ or $C\flat, G\flat$. This, however, is only true for the equal temperament.

The Pythagorean, meantone and diatonic temperaments are examples of other temperaments in which the pitches are tuned depending on the key. Basically, the Pythagorean temperament is constructed from perfect fifths (frequency ratio $3 : 2$), the meantone temperament is constructed with nearly perfect thirds, and the diatonic temperament intends to use as small frequency ratios as possible toward the tonic for every scale degree.

Table 5.5 shows the deviations in cents of different temperaments from the equal temperament.

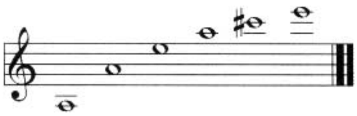

Figure 5.12 Six harmonics of the fundamental pitch $A3$ in musical score notation

5.2.5.3 Intonation

In contrast to temperament, the frequency of a certain pitch may vary over time depending on the musical context. This deviation from the frequency grid set by the temperament is called (expressive) intonation and is part of the musical performance.

Obviously, only musicians who are not forced to a pre-defined temperament can use expressive intonation. Examples are vocalists as well as players of string, brass, or woodwind instruments, while, for example, piano players have no means of changing the pitch frequency during a performance. It is generally assumed that musicians tend to produce pure frequency relationships such as $f_1/f_2 = 3/2$ for a fifth (seven semi-tones) because it sounds more "natural" [114]. Furthermore, if a specific note *leads* musically to the following, the frequency will in many cases be adjusted toward the following pitch. A good example are leading tones where the seventh scale degree leads to the first scale degree (in $C\ Major$: $B \rightarrow C$).

A special performance phenomenon related to expressive intonation is *vibrato*. Vibrato, a musical (performance) ornament, is a periodic frequency modulation of the pitch around its mean frequency. The extent and frequency of a vibrato is somewhat instrument dependent; typical frequencies are in the range of 5–9 Hz and the amplitude may be as large as two semi-tones [45, 115, 116].

5.3 Fundamental Frequency Detection

The basic assumption for all approaches to the estimation of the fundamental frequency (also referred to as *pitch detection* or *pitch tracking*) is that the signal is periodic or quasi-periodic. The periodic state of an acoustic tone can be represented in a Fourier series as introduced in Sect. 2.1.1, meaning that it is a superposition of weighted sinusoidals. The frequency of these sinusoidals is an integer multiple of the lowest — the *fundamental* — frequency. The different frequency components of a tone are called *harmonics* or *partials*, with the first harmonic being the *fundamental frequency*. Higher harmonics are also called *overtones*. The first six harmonics of the musical pitch $A3$ are displayed in Fig. 5.12 in traditional musical score notation.

This frequency structure can be found for most signals generated by acoustic or electronic instruments which are perceived as pitched sounds. However, there are certain deviations from this rule. For example, humans will hear the fundamental frequency of a harmonic series even without this frequency actually being present in the signal. Although an absent fundamental frequency occurs only rarely, it happens frequently that the first harmonic has lower energy than one or more of the higher harmonics.

The piano and string instruments are examples of acoustic signals in which the harmonics are not placed precisely at integer frequency multiples of the fundamental frequency but slightly off. A model of this deviation or inharmonicity is

$$f_k = kf_0\sqrt{1 + \lambda(k^2 - 1)} \qquad (5.17)$$

with k being the index of the harmonic (in this case better called partial as it is not entirely harmonic) and λ being the inharmonicity factor with typical values in the range $[10^{-3}; 10^{-4}]$ [117]. There are also instruments which are perceived as being pitched but show no clean harmonic pattern. Examples for this class of instruments are the xylophone, the vibraphone, and timpani. Nevertheless, the assumption of quasi-periodic states of a music signal has in many cases been proven to be valid and successful for *fundamental frequency detection*.

The common range of fundamental frequencies for musical instruments roughly starts between 20 and 50 Hz (e.g., on the double bass) and ends between 3–5 kHz (e.g., on the piccolo). Depending on the instrument, a number of at least three to seven harmonics should be considered to be important components of the sound (if components other than the fundamental frequency are of interest).

5.3.1 Detection Accuracy

Algorithms for fundamental frequency detection work either in the time domain by estimating the period length of the fundamental or in the frequency domain by finding the frequency of the fundamental. Both are discrete value domains and have thus a maximum accuracy determined by the distance of two time domain samples and two frequency domain bins, respectively. There are work-arounds for virtually enhancing the resolution such as frequency reassignment (see Sect. 2.2.3.1), but for now the effects of this discretization on the accuracy of fundamental frequency detection will be investigated.

5.3.1.1 Time Domain

As already mentioned above, the estimated period length of the fundamental frequency is quantized to samples in the time domain, resulting in the estimated period length being a multiple of the distance between two samples

$$T_\mathrm{Q} = j \cdot T_\mathrm{S} \tag{5.18}$$

with the integer multiplier j. This quantization leads to a certain amount of error depending on the sample rate and the period length. Figure 5.13 shows the minimum detection error in cent in dependency of the fundamental frequency for two different sample rates. The absolute worst-case error is small for low frequencies and increases with the frequency. Higher sample rates will result in smaller errors.

5.3.1.2 Frequency Domain

The trade-off between time resolution and frequency resolution is one of the big issues in STFT-based frequency analysis. While long analysis blocks increase the frequency resolution, they require a periodic and stationary signal during the analysis block in order to be useful. As can be easily seen from Eq. (2.44), the frequency resolution will stay constant if the ratio of sample rate f_S and block length $\mathcal{K}$ stays constant:

$$f_\mathrm{Q} = k \cdot \frac{f_\mathrm{S}}{\mathcal{K}}. \tag{5.19}$$

Table 5.6 displays the frequency resolution for different STFT sizes.

Figure 5.14 visualizes the error in cents for an analysis block length of 2048 samples at the sample rates 44.1 and 96 kHz. As the error is measured in cents (logarithmic scale) it decreases with increasing frequency (linear scale).

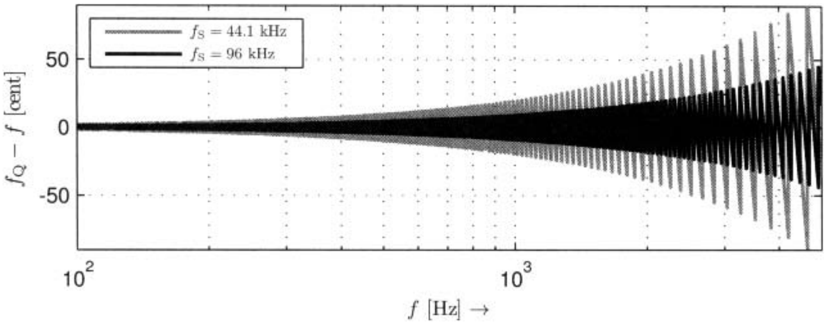

Figure 5.13 Detection error in cents resulting from quantization of the period length to a length in samples for two different sample rates

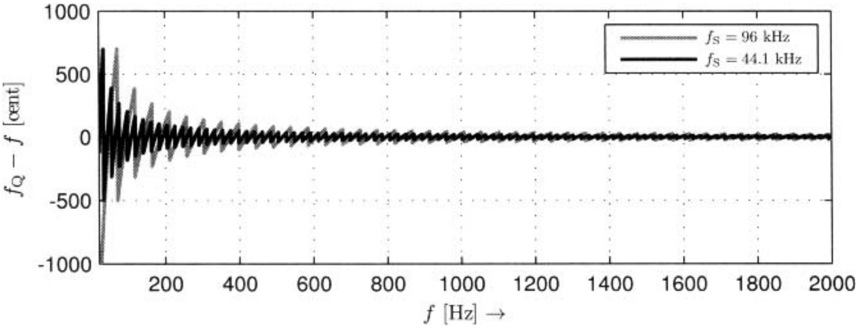

Figure 5.14 Detection error in cents resulting from quantization of the fundamental frequency to the STFT bin at an analysis block length of 2048 samples for two different sample rates

Table 5.6 Frequency resolution of the STFT for different block lengths at a sample rate of 48 kHz with bin index and frequency of the first bin k_{ST} with a distance to the following bin smaller than half a semi-tone

$\mathcal{K}$	Δf [Hz]	k_{ST}	$f(k_{\mathrm{ST}})$ [Hz]
256	187.5	35	6562.5
512	93.75	35	3281.25
1024	46.875	35	1640.625
2048	23.4375	35	820.3125
4096	11.7188	35	410.1563
8192	5.8594	35	205.0781
16384	2.9297	35	102.5391

5.3.2 Pre-Processing

As with nearly all signal processing algorithms, the results of fundamental frequency detection might be improved by applying appropriate pre-processing steps such as down-mixing, filtering, and sample rate conversion. Furthermore, it might be helpful in certain cases to remove noisy and non-tonal components before the actual detection.

5.3.2.1 Filtering and Down-Sampling

Since the range of fundamental frequencies is quite restricted compared to the sample rates used nowadays in audio signal processing, higher frequency components are frequently removed by both low-pass filtering and down-sampling the input signal. The cut-off frequency of the low-pass filter depends on the fundamental frequency range of the input signal as well as on the usefulness of higher harmonics for the subsequent detection algorithm. The target frequency of the resampling process might also depend on the required resolution for the period length estimation.

In addition, a high-pass filter removing components near DC may be helpful to clean up the input signal — typical cut-off frequencies would be in the range of 30–300 Hz.

5.3.2.2 Identification of Tonal Components

The differentiation of tonal and non-tonal (sinusoidal and noisy) signal components is a crucial pre-processing step for many audio signal processing systems and remained an active research topic through the last decades. The range of applications which can benefit from a reliable tonalness detector is wide: psycho-acoustic models in perceptual audio encoders can be optimized with a more accurate estimation of the signal-to-mask ratio, source separation algorithms may be improved by avoiding noise-like components, the quality of analysis/synthesis systems such as phase vocoders and audio restoration algorithms may increase by treating tonal and noisy parts differently, and the accuracy of pitch-based analysis systems such as key detection and chord recognition and ultimately music transcription systems can be enhanced.

As the variety and number of applications benefiting from the detection of tonalness suggest, there has been a plethora of publications dealing with the identification of tonal components.

Research on this topic began in the 1970s when spectral processing of digital audio became more and more common. In contrast to non-spectral approaches targeting a single monophonic source signal distorted by noise (see, e.g., [118]), the analysis of individual spectral bins allows the processing of multi-voiced input signals.

In the context of speech separation, Parsons identified peaks by finding local maxima in the magnitude spectrum — a first processing step that can be found in practically every publication dealing with the detection of tonal bins — and used the peak's symmetry, its proximity to the next peak as well as the continuity of the frequency bin's phase for detecting "peak overlaps," i.e., bins with supposedly two or more influencing sinusoidals [119]. Terhardt extended the concept of detecting the local maximum by expecting more distant bins (specifically the bins with a distance of 2 and 3) to be a certain level lower than the maximum itself [120]. Serra proposed a measure of "peakiness" of local maxima by comparing the bin magnitude with the surrounding local minima; he also discarded peaks outside of a pre-defined frequency and magnitude range [121].

An amplitude-based measure computing the correlation function between the magnitude spectrum and the shifted spectrum of the used window function has been presented by Peeters and Rodet [122] as well as Lagrange [123]. They also utilized a phase-derived measure comparing the bin frequency of a peak with its reassigned (instantaneous) frequency.

In addition to a local maximum feature similar to Terhardt's, Every proposed a threshold-based feature computed from the low-pass filtered magnitude spectrum, discarding peaks below the threshold [124].

Röbel et al. presented a set of features to classify spectral peaks into being sinusoidal or non-sinusoidal [125]. These features included the deviation of the bin frequency and its reassigned frequency, the peak's energy location according to its group delay, as well as the bandwidth of a spectral peak.

All of the publications presented above make a binary decision for a spectral bin being tonal or not; Kulesza and Czyzewski proposed an algorithm which aims at estimating the likelihood of a bin's tonalness [126]. They refer to this as a scoring classifier. This non-binary decision makes the algorithm probably most similar to the one outlined below; they use a so-called peakiness feature similar to Serra's, a frequency stability criterion for detected peaks, and a phase-based frequency coherence over subsequent blocks of the STFT. Their approach combines several features and uses a combination of heuristics and both binary and non-binary features to compute the resulting likelihood.

A feature-based approach to tonalness detection is a systematic and extensible way of computing the likelihood of an individual spectral bin being tonal. One example of such an approach is outlined in the following. It is based on the following assumptions for the input signal:

- it is a time-variant mixture of tonal and non-tonal signals,
- it has an undefined number of (tonal) voices (i.e., it is polyphonic), and
- the spectral envelope of both tonal and non-tonal components is unknown.

Furthermore, an individual tonal component is assumed to be

- salient, i.e., it is not masked by nearby components and has a certain intensity,
- deterministic, i.e., its phase cannot change erratically between the points of observation, and
- stationary for at least a minimum length of time.

These expected properties should be described by a set of features in the spectral domain. Each feature by itself should be simple to compute as well as simple to understand; it focuses on one individual property or aspect of a tonal component. Each feature $v_j(k,n)$ is the input of a Gaussian function $\varphi(x)$. Its output will be referred to as the *specific tonalness* $\Lambda_j(k,n) \in [0;1]$ which is a measure of likelihood of the bin k in frame n being a tonal component with respect to feature j

$$\Lambda_j(k,n) = \varphi\left(v_j(k,n)\right) = \exp\left(-\epsilon_j \cdot v_j(k,n)^2\right) \tag{5.20}$$

with ϵ_i being the normalization constant. The specific tonalness $\Lambda_j(k,n)$ is weighted and then combined to result in the *overall tonalness*.

Since the feature output $v_j(k,n)$ is in turn the input of the Gaussian function presented in Eq. (5.20), the features have to result in 0 output for tonalness and maximum output for non-tonalness. Examples of possible features are:

- *Local maximum*: Declaring the local maxima to be candidates for being tonal is a rather self-evident step in spectral analysis. In the presented variation the M surrounding bins are inspected for their magnitude being lower than the magnitude at the bin of interest:

$$v_1(k,n) = \sum_{m=1}^{M} \left(1 - \frac{m-1}{M}\right) \cdot \big(\Delta(k,m,n) + \Delta(k,-m,n)\big) \tag{5.21}$$

with

$$\Delta(k,m,n) = \begin{cases} 1, & \text{if } \big(|X(k,n)| - |X(k+m,n)|\big) \leq 0 \\ 0, & \text{if } \big(|X(k,n)| - |X(k+m,n)|\big) > 0. \end{cases} \tag{5.22}$$

The weighting increases the influence of nearby frequency bins on the likelihood.

- *Peakiness*: While the *local maximum* feature only takes into account whether the bin of interest has a higher magnitude than the neighboring bins, the *peakiness* evaluates their magnitude differences by relating the magnitude at the center bin with the mean of the neighboring magnitudes:

$$v_2(k,n) = \frac{|X(k-1,n)| + |X(k+1,n)|}{2 \cdot |X(k,n)|}. \tag{5.23}$$

The more pronounced a peak is, the lower will the feature value be and the higher will its tonalness be. The result $v_2(k,n)$ also depends on the used STFT windowing function and the spread of its main lobe. Because of these windowing effects, it could also be of advantage not to use the direct neighbors but bins with a bin distance of two or more for the averaging function, or to use the closest local minima as proposed by Serra [121].

- *Thresholding*: Since the tonal components are expected to be salient and to have more energy than noisy components, a magnitude threshold can be applied to increase the likelihood of bins with magnitudes above the threshold and decrease the likelihood of remaining bins correspondingly. This can be done by computing the ratio of a threshold $G(k,n)$ and the spectral magnitude $|X(k,n)|$:

$$v_3(k,n) = \frac{G(k,n)}{|X(k,n)|}. \tag{5.24}$$

The threshold can be determined by taking the maximum of an absolute threshold, a threshold relative to the highest magnitude in the current frame, and an adaptive threshold computed from the smoothed magnitude spectrum $X_{\mathrm{LP}}(k,n)$:

$$G(k,n) = \max\big(\lambda_1, \lambda_2 \cdot \max_{\forall k}(|X(k,n)|), \lambda_3 \cdot X_{\mathrm{LP}}(k,n)\big) \tag{5.25}$$

with λ_j being user-defined weighting parameters.

The smoothed magnitude spectrum X_{LP} may be computed with a single-pole filter:

$$X_{\mathrm{LP}}(k,n) = \alpha \cdot X_{\mathrm{LP}}(k-1,n) + (1-\alpha) \cdot |X(k,n)| \tag{5.26}$$

which is applied over the frequency in both the forward and the backward direction to ensure zero-phase response (see Sect. 2.2.1.2).

This thresholding process can be interpreted in different ways. On the one hand, one could see it as a pre-whitening process as used, for example, in pitch tracking systems with the goal of removing the spectral envelope from the signal [28]; on the other hand, it can be interpreted as a rudimentary model of a perceptual masking threshold applied to detect unmasked bins.

- *Frequency Coherence*: While the features presented above were based on the magnitude spectrum, the phase spectrum can also provide information on the tonalness of a frequency bin. More specifically, the instantaneous (or reassigned) frequency $f_I(k,n)$ (see Sect. 2.2.3.1) can be derived from the phase difference of overlapping spectra [17]. The instantaneous frequency has to be close to the bin frequency $f(k)$ in case of the main lobe of a stationary sinusoidal, otherwise, i.e., in the case of a noisy signal or a side lobe, the instantaneous frequency will probably deviate from the bin frequency. Furthermore, the instantaneous frequency should be comparably constant over consecutive blocks so that results can be averaged over two or more blocks. The final feature is then the difference of bin frequency and the (weighted) average of the instantaneous frequencies at this bin:

$$v_4(k,n) = f(k) - \frac{1}{\sum\limits_{\forall n_{\mathcal{O}}} b_{n_{\mathcal{O}}}} \sum_{n_{\mathcal{O}}=0}^{\mathcal{O}-1} b_{n_{\mathcal{O}}} \cdot f_I(k, n - n_{\mathcal{O}}). \qquad (5.27)$$

The *overall tonalness* $\Lambda(k,n)$ is the weighted combination of each specific tonalness. It can be computed by the weighted arithmetic mean of the specific tonalness, an approach that shows similarities to a simplified *Radial Basis Function (RBF)* network [127],

$$\Lambda_A(k,n) = \frac{1}{\lambda_s}\sum_{j=1}^{4} \lambda_j \cdot \Lambda_j(k,n)\,, \qquad \lambda_s = \sum_{\forall j} \lambda_j, \qquad (5.28)$$

or alternatively, when understood as a conditional probability, computed by the multiplication of the specific tonalness

$$\Lambda_G(k,n) = \prod_{j=1}^{4} \Lambda_j(k,n)^{\lambda_j/\lambda_s}\,, \qquad \lambda_s = \sum_{\forall j} \lambda_j. \qquad (5.29)$$

5.3.3 Monophonic Input Signals

A monophonic signal as opposed to a polyphonic signal[3] is single-voiced. There will never be more than one fundamental frequency present at a time. The problem of detecting the fundamental frequency in monophonic signals is basically limited to detecting the longest periodicity period as the frequencies of the harmonics will be integer multiples of the fundamental frequency. Since many algorithms for monophonic input signals make heavy use of this property, they are of no or only limited use in the context of polyphonic input signals which are mixtures of multiple voices with possibly different and time-variant timbre.

[3] The term *polyphonic* will be used for signals with multiple voices. It will not be used in the more music-theory-related restricted definition of quasi-independent voices as opposed to homophonic, where the multiple voices move together in harmony.

5.3.3.1 Zero Crossing Rate

The *zero crossing rate* has already been introduced in Sect. 3.4.3. There are two ways to estimate the fundamental period length with zero crossings; the first is to relate the number of zero crossings in an analysis block to its length, an approach that will only produce acceptable results for very large block sizes:

$$T_0(n) = \frac{2 \cdot \left(i_c(n) - i_s(n)\right)}{f_S \cdot \sum\limits_{i=i_s(n)}^{i_e(n)} \left|\text{sign}\left[x(i)\right] - \text{sign}\left[x(i-1)\right]\right|}, \tag{5.30}$$

and the second is to measure the interval $\Delta t_{ZC}(j)$ between neighboring zero crossings. This interval relates directly to the fundamental period length. If $\mathcal{Z}$ zero crossings have been detected in the analysis block, the fundamental period length can be estimated with

$$T_0(n) = \frac{2}{\mathcal{Z}-1} \sum_{z=0}^{\mathcal{Z}-2} \Delta t_{ZC}(z). \tag{5.31}$$

In case of large block lengths the time intervals can also be sorted into a histogram. The estimated period would then be the location of the histogram maximum multiplied by a factor of 2.

A related approach is to investigate the distance not only between zero crossing but between local extrema such as the maximum and minimum to increase robustness as explained by Rabiner and Schafer [128].

5.3.3.2 Autocorrelation Function

The use of the normalized ACF (see Sect.2.2.6) is very common in fundamental period length estimation. The lag of the maximum value is a direct estimation of the fundamental period length. Certain restrictions to maximum detection as exemplified in Sect. 3.4.1.4 can be applied to increase the algorithm's reliability.

Pre-Processing: Center Clipped Autocorrelation Function

The robustness of the detection can sometimes be improved by using so-called *center clipping* [128]. The non-linear function χ as shown in Fig. 5.15 (left) is applied to the input signal $x(i)$

$$x_c(i) = \chi\left(x(i)\right). \tag{5.32}$$

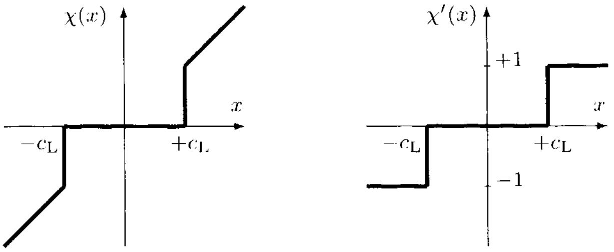

Figure 5.15 Non-linear pre-processing for ACF-based pitch period estimation: standard center clipping (left) and 3-level center clipping (right)

The idea of center clipping is to yield a result with more easily identifiable peaks. Typical thresholds c_{L} are chosen between 30% and 60% of the (instantaneous) maximum amplitude.

An alternative 3-level center clipping function χ' is shown in Fig. 5.15 (right). This function allows a very efficient computation of the ACF because the input signal then only has the three possible amplitudes $-1, 0, +1$. This performance optimization is usually not necessary anymore on modern hardware.

ACF Pre-Processing: Pre-Whitening

If the analyzed signal can be assumed to originate from a pulse-like excitation signal filtered with a transfer function, a common assumption in speech signal processing, a reasonable approach is to reverse the filtering process by estimating the smoothed spectral envelope of the current analysis block and apply the inverse filter to the signal — *pre-whitening*. The goal of the inverse filtering process is to convert the signal back to the initial pulse train in order to improve the results of the following correlation analysis.

There exist numerous ways to estimate the spectral envelope, including a low-order linear predictive filter (see Sect. 2.2.7), a smoothed power spectrum, or cepstrum-based iterative methods [129].

5.3.3.3 Average Magnitude Difference Function

The *Average Magnitude Difference Function (AMDF)* is similar to the ACF but avoids the use of multiplications and is therefore computationally efficient. The AMDF is computed by [130]

$$\mathrm{AMDF}_{xx}(\eta, n) = \frac{1}{i_{\mathrm{e}}(n) - i_{\mathrm{s}}(n) + 1} \sum_{i=i_{\mathrm{s}}(n)}^{i_{\mathrm{e}}(n)-\eta} \left|x(i) - x(i+\eta)\right|. \tag{5.33}$$

The estimated fundamental period length is then chosen to be the lag of the overall minimum if its value is also smaller than a certain (signal adaptive) threshold. The popular pitch tracking algorithm YIN utilizes the AMDF [131].

5.3.3.4 AMDF-Weighted Autocorrelation Function

The AMDF can also be used to weight the ACF. There are indications that this weighting leads to quite robust results [132]:

$$r'_{xx}(\eta, n) = \frac{r_{xx}(\eta, n)}{\mathrm{AMDF}_{xx}(\eta, n) + 1}. \tag{5.34}$$

5.3.3.5 Harmonic Product Spectrum and Harmonic Sum Spectrum

The *Harmonic Product Spectrum (HPS)* is an efficient method for finding the periodic harmonic pattern of an acoustic tone in its stationary state [133, 134]. It is defined by

$$X_{\mathrm{HPS}}(k, n) = \prod_{j=1}^{\mathcal{O}} |X(j \cdot k, n)|^2. \tag{5.35}$$

The parameter $\mathcal{O}$ is the order of the HPS.

The idea of the HPS is that the compression of the frequency axis by integer factors j causes higher harmonics at multiples of the fundamental frequency to coincide at the bin

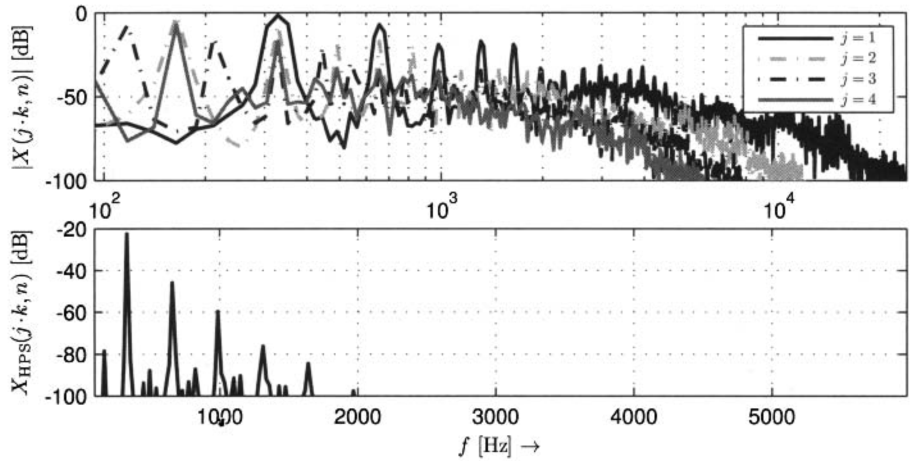

Figure 5.16 Compressed spectra with $j = 1, 2, 3, 4$ (top) and resulting HPS (bottom)

of the fundamental frequency. Thus, the first $\mathcal{O}$ harmonics will be mapped to their fundamental frequency. Since the harmonics can be assumed to have significantly higher power than any other signal components, the resulting (harmonic product) spectrum $X_{\mathrm{HPS}}(k, n)$ should have a clearly identifiable peak at the fundamental frequency. Figure 5.16 exemplifies this: the resulting peak in the HPS has a significantly higher distance to the second highest peak than in the original spectrum.

Alternative implementations of the HPS use the magnitude spectrum.

Typical Problems

If the fundamental frequency of the input signal is not located exactly on a frequency bin but between two bins, then the maxima of higher order harmonics will not be taken into account for the decimated spectra. The likelihood of missing the locations of the maxima increases with j. Two possible work-arounds can be used, but both will result in less efficient and more complicated implementations:

- increase the frequency resolution by using longer STFT block sizes or by interpolating (up-sampling) the spectrum, or
- take the maximum within a bin range for the multiplication. The bin range will have to increase by ± 1 bin with every increment of j.

If one of the harmonics is zero or near zero, the detection of the fundamental frequency will probably fail as the HPS at the fundamental frequency bin will be scaled with (near) zero. One approach to avoid this is to compute the *Harmonic Sum Spectrum (HSS)* with a definition similar to the HPS [134]:

$$X_{\mathrm{HSS}}(k, n) = \sum_{j=1}^{\mathcal{O}} |X(j \cdot k, n)|^2. \tag{5.36}$$

While the HSS is more robust against missing harmonics, the resulting maximum is usually not as pronounced as in the HPS.

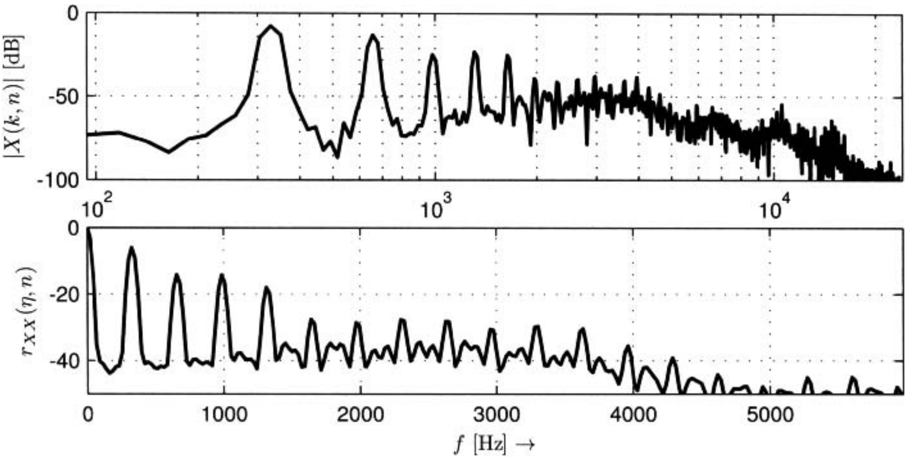

Figure 5.17 Magnitude spectrum (top) and ACF of this spectrum (bottom)

5.3.3.6 Autocorrelation Function of the Magnitude Spectrum

The usage of the ACF is a very intuitive approach to finding periodicities. Since the harmonics are equally spaced in the magnitude spectrum and the distance of neighboring harmonics equals the fundamental frequency, finding this periodicity is equivalent to finding the fundamental frequency. The lag of the maximum of the ACF is an estimate of the fundamental frequency in spectral bins. Figure 5.17 shows the result of the ACF of a magnitude spectrum.

5.3.3.7 Cepstral Pitch Detection

Cepstral pitch detection is based on the assumption that the analysis signal $x(i)$ is the result of a convolution of an excitation signal $e(i)$ with a transfer function $h(i)$

$$x(i) = e(i) * h(i). \tag{5.37}$$

This is a common assumption in speech signal processing; the excitation signal originates from the air streaming through the *glottis*. This excitation signal $e(i)$ consists of quasi-periodic pulses in the case of voiced (tonal) sounds [128]. The vocal tract and the nasal tract act as tubes that shape the frequency spectrum with the transfer function $h(i)$.

Equation (5.37) can be rephrased in the frequency domain (see Sect. B.1.3) as

$$X(\mathrm{j}\omega) = E(\mathrm{j}\omega) \cdot H(\mathrm{j}\omega). \tag{5.38}$$

If a (complex) logarithm is applied to this equation, the result is

$$\begin{aligned}\log\left(X(\mathrm{j}\omega)\right) &= \log\left(E(\mathrm{j}\omega)\cdot H(\mathrm{j}\omega)\right)\\ &= \log\left(E(\mathrm{j}\omega)\right) + \log\left(H(\mathrm{j}\omega)\right).\end{aligned} \tag{5.39}$$

Applying the logarithm allowed us to replace the multiplication with an addition. We define the *cepstrum* $c_x(i)$ to be the inverted logarithmic spectrum

$$\begin{aligned}c_x(i) &= \mathfrak{F}^{-1}\left\{\log\left(X(\mathrm{j}\omega)\right)\right\}\\ &= \mathfrak{F}^{-1}\left\{\log\left(E(\mathrm{j}\omega)\right) + \log\left(H(\mathrm{j}\omega)\right)\right\}\\ &= \mathfrak{F}^{-1}\left\{\log\left(E(\mathrm{j}\omega)\right)\right\} + \mathfrak{F}^{-1}\left\{\log\left(H(\mathrm{j}\omega)\right)\right\}.\end{aligned} \tag{5.40}$$

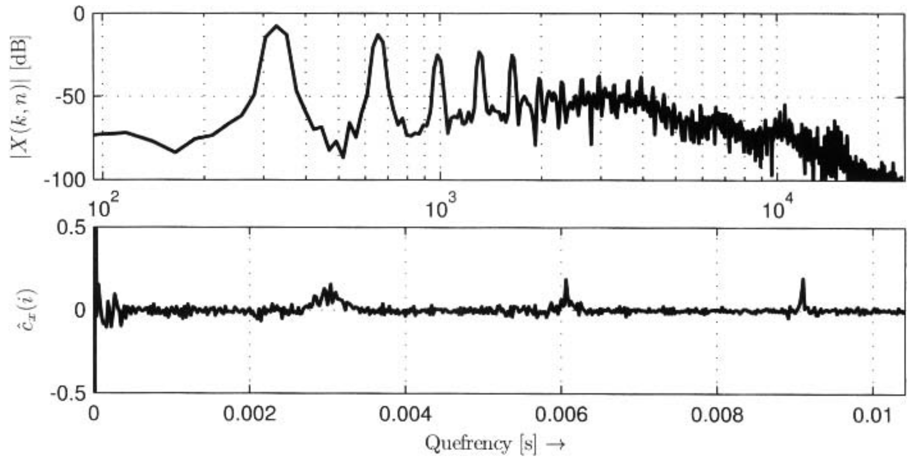

Figure 5.18 Magnitude spectrum (top) and cepstrum of this magnitude spectrum (bottom)

The inverse-transformed spectrum thus consists of signals $e(i)$ and $h(i)$ being *added* (although logarithmically) instead of being convolved. In order to emphasize the difference between the original time domain and the cepstral domain, the term *quefrency* is frequently used as axis label.

The cepstrum can be approximated by only using the magnitude spectrum

$$\hat{c}_x(i_s(n)\dots i_e(n)) = \sum_{k=0}^{\mathcal{K}/2-1} \log\left(|X(k,n)|\right) e^{jki\Delta\Omega} \tag{5.41}$$

and avoiding the use of the complex logarithm.

The cepstrum has two properties that are of particular interest in the context of pitch detection [133, 135]. First, a pulse-like excitation signal will also lead to a pulse in the cepstrum, and, second, the cepstrum will decay rapidly for large i. Therefore, the detection of a peak (or more accurately a pulse train) in the cepstrum should give the period length of the fundamental frequency. Figure 5.18 shows the cepstrum of an exemplary magnitude spectrum.

5.3.3.8 Auditory Motivated Pitch Tracking

A rather important class of pitch detection algorithms use models of human pitch perception to determine the pitch of a signal. Meddis and O'Mard describe the processing stages of such algorithms as [136]:

1. band-pass filtering,
2. HWR [see Eq. (3.33)] and band processing,
3. within-band periodicity extraction, and
4. across-band aggregation of periodicity estimates.

The filterbank used for band-pass filtering is frequently a gammatone filterbank as the "standard" filterbank for auditory processing. Gammtone filters have been introduced in Sect. 2.2.5.1.

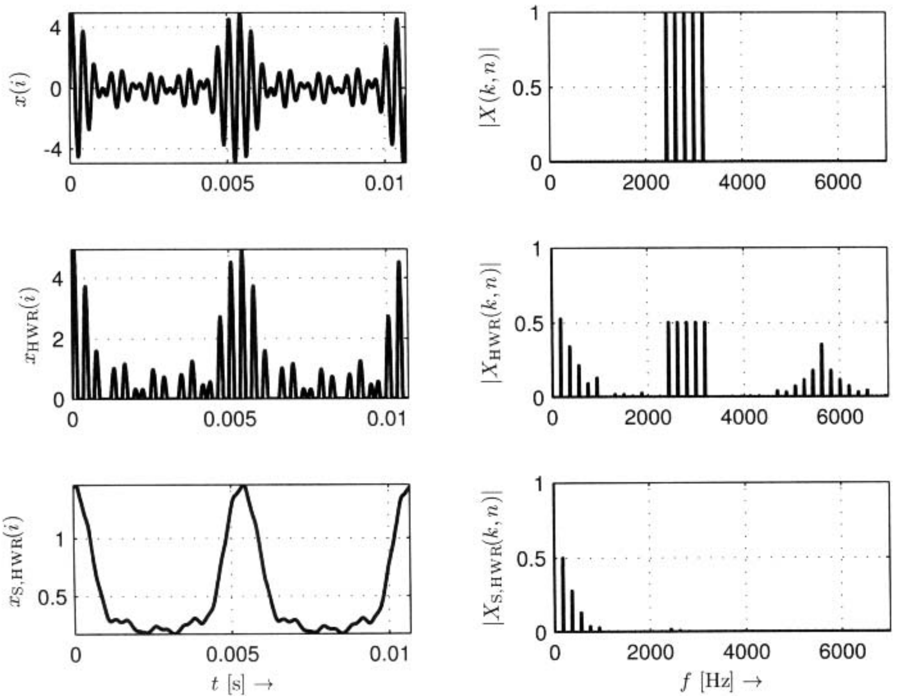

Figure 5.19 Time domain (left) and frequency domain (right) for a signal consisting of the 13th to 17th partial of a sound with a fundamental frequency of 187.5 Hz (top), the corresponding signal subjected to HWR (mid), and the low-pass filtered signal subjected to HWR (bottom)

Klapuri pointed out the importance of HWR on the filter channel outputs for fundamental pitch estimation [137]. Figure 5.19 shows this effect for one band containing the 13th–17th harmonics of a periodic sound: new frequency components are being generated at the distance of frequency components in the signal and thus around the fundamental frequency. Further band processing might include gain compression by, for instance, scaling the variance to unity [137].

A periodicity analysis of the filterbank outputs $z_c(i)$ may then be used for the detection of the fundamental period length by, for example, applying an ACF to each channel

$$r_{zz}(c,n,\eta) = \sum_{\eta=0}^{\mathcal{K}-1} z_c(i) \cdot z_c(i+\eta) \tag{5.42}$$

and summing the resulting ACFs

$$r_{\mathrm{A}}(n,\eta) = \sum_{c=0}^{\mathcal{C}-1} r_{zz}(c,n,\eta). \tag{5.43}$$

5.3.4 Polyphonic Input Signals

Most of the algorithms outlined above will not work well in the case of polyphonic signals with multiple simultaneous fundamental frequencies. The number of simultaneous pitches in polyphonic signals depends on genre, epoch, and musical context. In most cases the number of independent voices will be between one and eight.

One of the first *multi-pitch detection* systems was presented by Chafe et al. in 1985 [138]. They pick spectral magnitude peaks in a multi-resolution FT and derive candidates for fundamental frequencies by grouping the peaks with respect to their frequency ratio. The detected candidates are then tracked in the spectrogram to discard spurious detections. Nowadays, multi-pitch detection is a lively research field with numerous methods and approaches of which only a few basic ones will be described below.

5.3.4.1 Iterative Subtraction

The principle of iterative subtraction is to apply a fundamental frequency detection algorithm for *monophonic* input signals to the signal to extract the predominant fundamental frequency, then find a way to subtract this and related (mostly harmonic) frequency components from the original signal and repeat the process on the residual until the criterion for termination has been reached.

An early reference to such an algorithm in the spectral domain has been published by Parsons [119] who — inspired by the work of Schroeder [133] — constructed a histogram of spectral peaks and their integer submultiples, chose the largest peak for the first fundamental frequency estimate, and removed this estimate and its multiples from the histogram to detect the second fundamental frequency. Klapuri et al. published two adaptations on the iterative subtraction procedure more recently. They use an auditory-motivated monophonic fundamental frequency detection algorithm to find the predominant, most salient fundamental frequency and estimate the spectrum of the voice with this frequency to subtract it from the original spectrum for the detection of additional voices [139, 140].

Cheveigné proposed a system for the tracking of multiple pitches in the time domain [141, 142]. First, the squared AMDF (compare Sect. 5.3.3.3) is computed with

$$\mathrm{ASMDF}_{xx}(\eta, n) = \frac{1}{i_\mathrm{e}(n) - i_\mathrm{s}(n) + 1} \sum_{i=i_\mathrm{s}(n)}^{i_\mathrm{e}(n)} \left(x(i) - x(i+\eta)\right)^2 . \tag{5.44}$$

Then, the most salient period length is found at the lag of the ASMDF minimum:

$$\eta_\mathrm{min} = \eta \left|_{\mathrm{min}(\mathrm{ASMDF}_{xx}(\eta, n))} \right. . \tag{5.45}$$

In order to remove the detected frequency and its harmonics from the signal, a (FIR) comb cancellation filter is applied to the signal with a delay corresponding to the lag of the detected minimum. The comb filter has the impulse response

$$h(i) = \delta(i) - \delta(i - \eta_\mathrm{min}) \tag{5.46}$$

and not only attenuates the detected fundamental frequency but also its harmonics at integer multiples. The process can then be repeated for the detection of a second fundamental frequency. It is also possible to implement this approach non-iteratively by using an exhaustive search. In this case, two cascaded cancellation filters can be applied to the signal in all possible combinations of η_1 and η_2. The most likely pair of fundamental period lengths is the combination of η_1 and η_2 which minimizes the overall output power of the output signals.

Meddis and Hewitt use an auditory approach similar to the one described in Sect. 5.3.3.8 for detecting one fundamental frequency and then use all remaining filter channels, i.e., filter channels not showing a peak at the detected frequency, to detect more fundamental frequencies [143]. The iteration process is terminated if more than 80% of the channels have been removed.

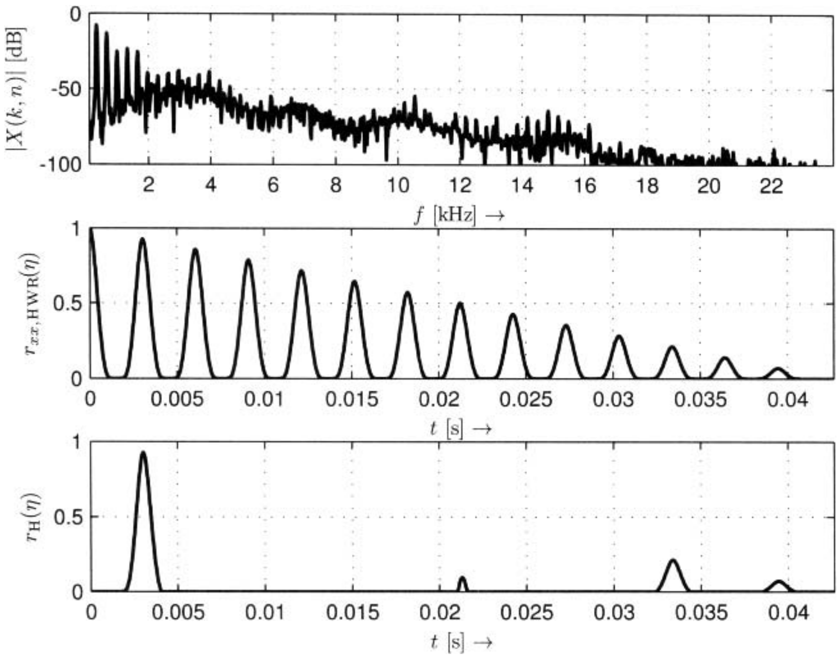

Figure 5.20 Magnitude spectrum (top), ACF of the signal subjected to HWR and harmonic ACF processing according to Karjalainen and Tolonen

5.3.4.2 Karjalainen and Tolonen

Another auditory-inspired multi-pitch detection algorithm focusing on computational efficiency has been published by Karjalainen and Tolonen [28, 144]. In a pre-processing step, they apply pre-whitening to flatten the spectral envelope. The spectral envelope is estimated by frequency-warped linear prediction. The next processing step splits the signal into a low-pass band and a high-pass band with a cut-off frequency of 1 kHz. The filter outputs are then subjected to HWR and smoothed with a low-pass filter. The periodicity within each band is estimated with an ACF. In an attempt to model auditory loudness scaling, the generalized ACF as introduced in Eq. (2.74) is used with an exponent of $\beta = {}^2\!/_3$. The two resulting functions are then added to result in an overall summary ACF.

This summary ACF is then harmonically processed with an idea similar to HPS. The harmonic post-processing is an iterative process in which an interpolated, time-scaled and half-wave rectified version is subtracted from the half-wave rectified ACF $r(\eta, n) = r(j, \eta, n)$:

$$r(j,\eta,n) = \mathrm{HWR}\left(\mathrm{HWR}\left(r(j-1,\eta,n)\right) - \mathrm{HWR}\left(r\left(j, {}^{\eta}\!/_{j}, n\right)\right)\right) \tag{5.47}$$

with $r\left(j, {}^{\eta}\!/_{j}, n\right)$ generated with linear interpolation.

The repeated application of HWR in combination with scaling aims at discarding frequencies other the fundamental frequencies. Figure 5.20 shows the result of a "normal" ACF with HWR and an harmonic ACF of the same block.

5.3.4.3 Klapuri

Klapuri makes use of an auditory approach by computing the normalized filterbank outputs after HWR [145]. He then computes the STFT of each filter band, sums their magnitudes

$$Z(k,n) = \sum_{c=0}^{C-1} |Z_c(k,n)|, \tag{5.48}$$

and weights the resulting overall spectrum $Z(k,n)$ with a low-pass filter transfer function. To identify possible fundamental frequency candidates, a set of delta pulse templates is used representing every detectable fundamental frequency with its harmonics. Similar to the calculation of the HSS, these templates are multiplied with the spectrum, and the result can be used as an estimate for the salience of each fundamental frequency. The components of the most salient frequency are then being removed from the spectrum to find the next fundamental frequency in an iterative process.

5.3.4.4 Other Methods

Probabilistic approaches aim at modeling the pitch tracking problem by means of a statistical framework. To give only one example, Kameoka et al. model a tone in the frequency domain as a superposition of weighted Gaussian distributions at integer multiples of the fundamental frequency and its power envelope function in the time domain by overlapping Gaussian distributions [146]. The spectrogram is then composed into clusters which model individual notes.

Non-negative matrix factorization, introduced by Lee and Seung [147], has attracted noteworthy attention in the context of multi-pitch detection during the last decade. It decomposes a time-frequency representation into a matrix containing the spectra of the individual sounds and another matrix containing the information on when each of the individual spectra is active. Smaragdis and Brown applied the technique to the magnitude spectrogram in order to detect pitches with promising results [148].

5.4 Tuning Frequency Estimation

The computation of the tuning frequency (see Sect. 5.2.5.1) is a prerequisite for every mapping from frequency to musical pitch given in Eq. (5.12). Examples of applications utilizing this mapping are key detection, chord recognition, automatic transcription and melody finding. Many of the published algorithms are based on the assumption that a tuning frequency of 440 Hz has been used for the recording. This assumption works reasonably well in many cases, but in general the tuning frequency should be detected from the audio in order to improve detection accuracy [113]. This is particularly true when the tuning frequency can be expected to change over time.

To estimate the tuning frequency from an audio signal, Scheirer used a set of narrow band-pass filters with their mid-frequencies at particular bands that had been handpicked to match pitches from the previously analyzed score [149]. These filters are swept over a small frequency range such as a semi-tone. The estimated tuning frequency is then determined by the mid-frequency of the maximum of the sum of the energy of all filterbank outputs.

Dixon proposed to use a peak detection algorithm in the frequency domain and to calculate the instantaneous frequency of the detected peaks [150]. The equally tempered reference frequencies are then modified iteratively until the distance between detected and reference frequencies is minimized. The amount of adaptation is the low-pass filtered geometric mean of previous and current reference frequency estimate.

Zhu et al. computed a CQT with the frequency spacing of 10 cents over a range of 7 octaves [112]. The detected peaks in the CQT spectrum are grouped based on the modulus distance against the concert pitch, resulting in a 10-dimensional histogram spanning 100 cents. If the maximum cumulative energy of the histogram is below a certain energy threshold, it is discarded. For the results of all processing blocks, a 10-dimensional average tuning pitch histogram is computed and the overall tuning frequency is chosen corresponding to the position of the histogram maximum.

In the context of single-voiced input signals, Ryynänen added the modulus distance of detected fundamental frequencies to a 10-dimensional histogram that is low-pass filtered over time [151]. Then, a "histogram mass center" is computed and the tuning frequency is adjusted according to this mass center.

Dressler and Streich modeled the tuning frequency deviation in cents with a circular model

$$
\begin{aligned}
z(n) &= r_n \cdot \exp\left(\mathrm{j}\frac{2\pi}{100}\Delta C(f_n, 440\,\mathrm{Hz})\right), &(5.49)\\
\mu_z &= \frac{1}{\mathcal{N}}\sum_{n=0}^{\mathcal{N}-1} z(n), &(5.50)\\
\hat{f}_{A4} &= 2^{\frac{\arg(\mu_z)}{2\pi\cdot 12}} \cdot 440\,[\mathrm{Hz}], &(5.51)
\end{aligned}
$$

and used different measures for r_n: the magnitude of salient spectral peaks, the magnitude of the estimated fundamental frequencies of the melody, and the magnitude of the average melody pitch, disregarding deviations such as vibrato [152]. As an alternative to computing the arithmetic mean, they propose constant adaption by low-pass filtering r_n with a single-pole filter.

Lerch proposed using a bank of steep resonance filters for detecting the tuning frequency, allowing both real-time processing and adaptation to a time-variant tuning frequency [113, 153]. In the range of two octaves, there are 24 groups of filters in (equally tempered) semi-tone distance, with each group consisting of 3 filters. The mid-frequencies of each group are narrowly spaced, and the mid-frequency of the centered filter is selected based on the most recent tuning frequency estimate. The filter output energy over a time window is then grouped based on the modulus distance against the concert pitch, resulting in a three-dimensional vector E. The symmetry of the distribution of the three accumulated energies gives an estimate on the deviation from the current tuning frequency compared to the assumption. If the distribution is symmetric, i.e., $E(0)$ equals $E(2)$, the assumption was correct. In the other case, all filter mid-frequencies are adjusted with the objective to symmetrize the energy distribution for the following processing blocks. The adaption rule used is the RPROP algorithm which allows fast and robust adaptation without the requirement of a signal-dependent adaption step size [154]. More specifically, the adaption rule for the adjustment of the estimated tuning frequency $\hat{f}_{A4}$ of the next processing block $n+1$ is

$$\hat{f}_{A4}(n+1) = \big(1 + \mathrm{sign}\left(E(2) - E(0)\right) \cdot \eta\big) \cdot \hat{f}_{A4}(n) \qquad (5.52)$$

with η being scaled up if the direction of the previous adaptation was the same and scaled down otherwise. The advantage of the scaling is an increasing step size as long as the sign does not change and a decreasing step size otherwise. Figure 5.21 shows the adaptation

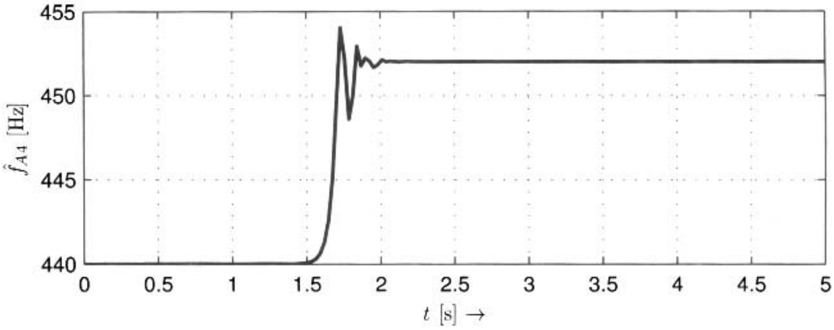

Figure 5.21 Adaptation of the tuning frequency estimate from an initial setting of 440 Hz to the target frequency of 452 Hz with the RPROP algorithm

from the initial tuning frequency of 440 Hz to the real frequency of 452 Hz. Adaptation can be tuned either for speed or for accuracy.

5.5 Key Detection

The musical key is an important tonal property of a piece of tonal music as it signifies the tonal center and restricts the pitch classes used within this key context. In classical music, the key is one descriptor used to identify a specific piece of music, complementing name, genre, and opus. In modern software applications for DJs the key is used to display the tonal compatibility between two tracks, i.e., to visualize how much "tonal overlap" they have to aid so-called harmonic mixing.

Since the key can be seen as a set of "allowed" pitch classes, the usual approach is to make use of an octave-independent representation of pitch for its automatic detection. The most common representation is the so-called pitch chroma.

5.5.1 Pitch Chroma

The *pitch chroma* (sometimes also referred to as *pitch chromagram* or *pitch class profile*) is a histogram-like 12-dimensional vector with each dimension representing one pitch class (C, $C\sharp$, D, ..., B; compare Sect. 5.2.1). It can be seen as a *pitch class distribution* for which the value of each dimension may represent both the number of occurrences of the specific pitch class in a time frame and its energy or velocity throughout the analysis block.

There are several advantages of using pitch chroma-based analysis. It is less dependent on timbre fluctuations and noise than many other feature. The pitch chroma is also robust against loudness fluctuations as well as against octave errors — a typical problem of pitch detection algorithms — with the self-evident disadvantage that the octave information is lost. It is, for example, not possible to distinguish between a note repetition an an octave interval with the pitch chroma representation.

While a representation of pitch similar to the pitch chroma has been frequently used in the past (see, e.g., Krumhansl's tonal distributions [155]), Bartsch and Wakefield were probably the first to propose its use in the context of audio signal processing [156]. Nowadays, this representation can be frequently found in publications in the context of ACA [157].

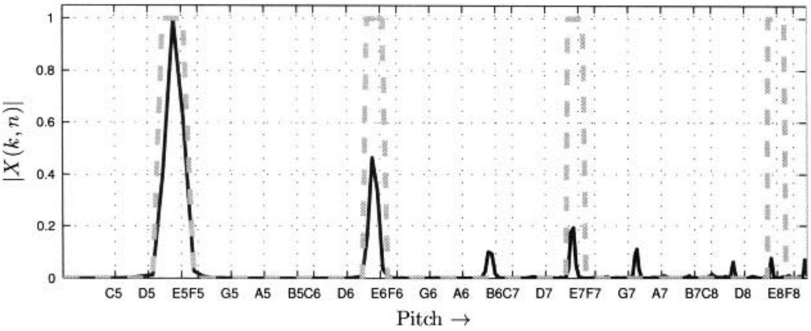

Figure 5.22 Window function for pitch class E for pitch chroma computation (octaves $5 \ldots 8$)

The exact algorithmic description of the pitch chroma computation varies from publication to publication; in all cases

- a frequency representation of the audio signal block is grouped into semi-tone bands,
- a measure of salience is computed in each band, and finally
- the sum of all bands (over all octaves) corresponding to a specific pitch class is calculated.

The simplest way to extract the pitch chroma sums the STFT magnitudes in each semitone band with the boundary indices $k_\mathrm{l}, k_\mathrm{u}$, and the result in every octave o is added to the corresponding pitch chroma entry with pitch class index j:

$$\nu(j,n) = \sum_{o=o_\mathrm{l}}^{o_\mathrm{u}} \left(\frac{1}{k_\mathrm{u}(o,j) - k_\mathrm{l}(o,j) + 1} \sum_{k=k_\mathrm{l}(o,j)}^{k_\mathrm{u}(o,j)} |X(k,n)| \right), \tag{5.53}$$

$$\boldsymbol{\nu}(n) = [\nu(0,n),\, \nu(1,n),\, \nu(2,n),\, \ldots,\, \nu(10,n),\, \nu(11,n)]^\mathrm{T}. \tag{5.54}$$

The indices $k_\mathrm{l}, k_\mathrm{u}$ are located at a distance of 50 cents from the mid-frequency of each equally tempered pitch.

Figure 5.22 shows the semi-tone bands for the pitch class E (pitch class index 4) in the octaves 5–8. Note that the window bandwidth is constant on the pitch axis; on the linear frequency axis it would increase with increasing frequency.

Frequently the pitch chroma is normed so that the sum of all possible pitch classes equals 1:

$$\boldsymbol{\nu}_\mathrm{N}(n) = \boldsymbol{\nu}(n) \cdot \frac{1}{\sum\limits_{j=0}^{11} \nu(j,n)}. \tag{5.55}$$

Occasionally the pitch chroma is interpreted as a vector and is normed to a length of 1 instead:

$$\boldsymbol{\nu}_\mathrm{N}(n) = \boldsymbol{\nu}(n) \cdot \sqrt{\frac{1}{\sum\limits_{j=0}^{11} \nu(j,n)^2}}. \tag{5.56}$$

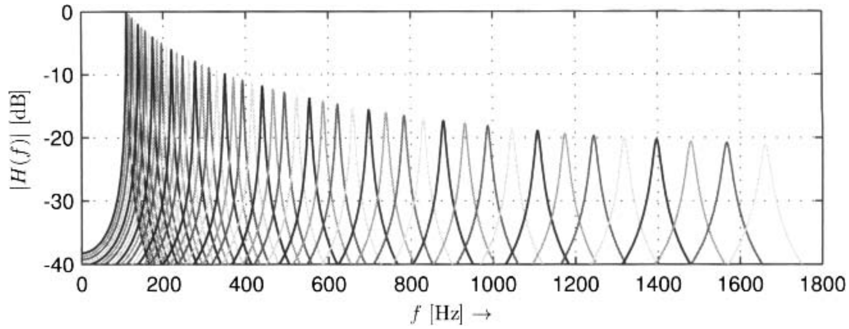

Figure 5.23 Frequency response of the resonance filterbank with one filter per semi-tone

5.5.1.1 Common Variants

The two main stages in the pitch chroma computation allowing alternative implementations are the type of the frequency transform and the selection of spectral content to sum.

When using an STFT as frequency transform, different windows with triangular, trapezoid, or sinusoidal shapes can be applied to each semi-tone band to weight bins in the center of the band higher than bins at the boundaries, as opposed to the rectangular window implicitly used in Eq. (5.53) and plotted in Fig. 5.22. Furthermore, the pitch chroma can be computed from a peak-picked or tonalness-weighted magnitude spectrum since only the tonal components are of interest for this feature.

A modification of the STFT has been used by Cremer and Derboven [158] for the computation of the pitch chroma: they utilize a so-called frequency-warped STFT as introduced by Oppenheim et al. which uses a chain of first-order all-pass filters to achieve non-equidistant spacing of the frequency bins [159]. The CQT as introduced in Sect. 2.2.4 can also be used for pitch chroma computation [160]. In this case the number of bands per semi-tone will equal 1 or be constant. It is also possible to use a filterbank with one filter for each semi-tone. Lerch used a bank of resonance filters as shown in Fig. 5.23 [153].

The analysis window length for the computation of one pitch chroma is another parameter allowing variations between different algorithms. While in the simplest case the STFT length equals the extraction window length [161, 162], other algorithms combine a fixed number of short analysis blocks [163] or adapt the extraction window length to the period between two neighboring beats [164] or to a measure of harmonic change in the audio [165]. The extraction window length obviously also depends on the task at hand; key detection requires comparably long windows as opposed to chord recognition which usually requires window lengths between a beat and a bar length.

5.5.1.2 Properties

In most audio applications, the pitch chroma is used as if only fundamental frequencies (and no overtones) have been mapped to it. In reality, all frequency content in the predefined frequency range of interest is mapped to the pitch chroma, regardless of it being a fundamental frequency or a higher harmonic. This leads to two possible problems:

- High non-power-of-two harmonics lead to distortions in the pitch chroma by adding undesired components. Figure 5.24 displays a series of harmonics with the fundamental pitch $A3$ and the resulting pitch chroma which shows spurious components at

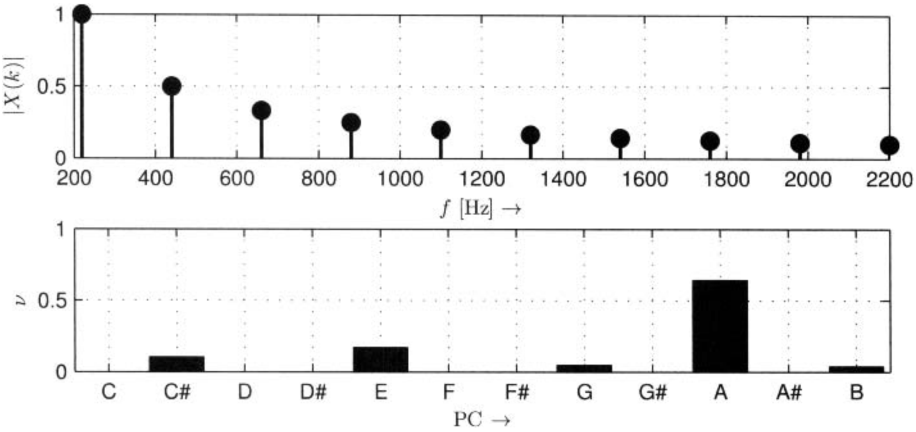

Figure 5.24 Pitch chroma of the pitch $A3$ with harmonics

Table 5.7 Deviation (in cents) of seven harmonics from the nearest equally tempered semi-tone mid-frequency

Harmonic	$\lvert\Delta C(f, f_T)\rvert$
$f = f_0$	0
$f = 2 \cdot f_0$	0
$f = 3 \cdot f_0$	1.955
$f = 4 \cdot f_0$	0
$f = 5 \cdot f_0$	13.6863
$f = 6 \cdot f_0$	1.955
$f = 7 \cdot f_0$	31.1741
$\mu_{\lvert\Delta C\rvert}$	6.9672

E and $C\sharp$ and to a lesser degree at G and B. It is possible to partly compensate for that effect by using an amplitude model for the harmonics and modifying the pitch chroma accordingly [160]. The effect can also be attenuated by applying a weighting function to de-emphasize higher frequencies (see e.g. [166]).

- High harmonics may even be able to distort the pitch chroma in a different way: since the harmonics will deviate from the equal temperament — the temperament the pitch chroma computation is based on — they may map to pitches not really part of the tonal context. This effect is, however, of limited influence since the deviation of the lower harmonics is relatively small as shown in Table 5.7.

The only possibility to avoid these artifacts is to use a multi-pitch detection system as a pre-processing step which ensures that only fundamental frequencies are mapped to the pitch chroma (compare [163, 167, 168]).

The pitch chroma entries may be ordered differently to reflect tonal relationships between the pitch classes. One way to do so is to replace the standard modulo operation for

Table 5.8 Pitch class order in the original and the rearranged pitch chroma

PC	C	$C\sharp$	D	$D\sharp$	E	F	$F\sharp$	G	$G\sharp$	A	$A\sharp$	B
$\mathrm{PC_S}$	C	G	D	A	E	B	$F\sharp$	$C\sharp$	$G\sharp$	$D\sharp$	$A\sharp$	F

computing the pitch chroma index as given in Eq. (5.14) by a shifted version:

$$\mathrm{PC_S}(\mathrm{p}) = \mod(7 \cdot \mathrm{p}, 12). \qquad (5.57)$$

This results in related keys with a distance of seven semi-tones to be close to each other. Table 5.8 shows the pitch class order of the rearranged chroma. Certain similarity measures such as the CCF will behave more gracefully or musically meaningful when using this resorted pitch chroma [112, 157].

5.5.1.3 Pitch Chroma Features

Similar to extracting instantaneous features from a spectrum, instantaneous features can be extracted from the 12-dimensional pitch chroma as well. The computation of statistical features such as the standard deviation is only of limited use due to the small number of elements. It should also be kept in mind that the pitch chroma is neither a series of observations nor a position in a space with 12 unrelated dimensions; it is a *distribution*.

There is no established set of features to be extracted from the pitch chroma, although the set of three features presented by Tzanetakis et al. proved to be simple yet effective [157]. They use both the index and the amplitude of the pitch chroma maximum as well as interval between its two highest bins. For the latter feature, the pitch chroma is rearranged in order to place related keys as neighbors (see above).

There exists a multitude of other features to extract from a pitch chroma that might be useful in certain applications. Possibilities include a pitch chroma crest factor or the centroid of the resorted pitch chroma.

5.5.2 Key Recognition

Using the assumptions that (a) the occurrence of key inherent pitch classes is more likely than the occurrence of non-key pitch classes and that (b) the number and energy of occurrences is an indication of the key, the two basic processing steps for automatic *key detection* are

- the extraction of a pitch chroma to estimate a pitch class distribution, and
- the computation of the likelihood of the extracted pitch chroma being a specific key and selecting the most likely one.

The various approaches of extracting the pitch chroma have been discussed in Sect. 5.5.1. The time frame over which the pitch chroma is extracted can vary from the whole file to a common texture window length. It is also possible to extract the pitch chroma only at the beginning and end of the audio file because the key is usually less ambivalent in these regions [153].

5.5.2.1 Key Profiles

The likelihood of a specific key is usually computed by comparing the *extracted* pitch chroma with a *template* pitch chroma, in the following referred to as key profile. The template minimizing the distance to the extracted pitch chroma will be the most likely key.

Various templates can be used (see also the summary in Table 5.9 for $C\ Major$):

- *Orthogonal* ν_{o}: The orthogonal template assumes that the root note is the most salient component of the pitch chroma and that all other keys have the same distance (and are thus unrelated). The template allows no distinction between major mode and minor mode.

- *Smoothed Orthogonal* ν_{s}: The smoothed orthogonal template is, as the name says, a low-pass filtered version of the simple orthogonal template. It results in an increasing distance with increasing index distance with a distance maximum for the key at the tritone.

- *Diatonic* ν_{d}: The diatonic template is 1 for every key inherent pitch class and 0 otherwise. The distance to other keys increases linearly like an "unwrapped" circle of fifths. It allows no distinction between major mode and minor mode.

- *Circle of Fifths* ν_5: The distance of two keys can be modeled by its distance in the circle of fifths with respect to its radius r. The coordinates of each key are given by the angle of the key in the circle of fifths. Each key has the same distance r to the point of origin. The model can be expanded into a third dimension to also include a distance between major mode and minor mode.

- *Probe Tone Ratings* ν_{p}: Krumhansl's probe tone ratings are not directly a key description but are the result of a listening test evaluating how a pitch fits into a given tonal context [155]. However, Krumhansl showed in experiments that these probe tone ratings correlate well with the number of occurrences of the pitch classes in real pieces of music.

- *Extracted Key Profiles* ν_{t}: Key profile vectors can also be derived from real-world data. The template given in Table 5.9 has been published by Temperley and has been extracted from symbolic data [169].

Table 5.9 Various key profile templates, normalized to a sum of 1

	ν_{o}	ν_{s}	ν_{d}	ν_5	ν_{p}	ν_{t}
$\nu(0)$	1	0.44721	0.37796	$r \cdot e^{j2\pi\frac{0}{12}}$	0.49483	0.49355
$\nu(1)$	0	0.44721	0	$r \cdot e^{j2\pi\frac{-5}{12}}$	0.17377	0.039589
$\nu(2)$	0	0.44721	0.37796	$r \cdot e^{j2\pi\frac{2}{12}}$	0.27118	0.32199
$\nu(3)$	0	0	0	$r \cdot e^{j2\pi\frac{-3}{12}}$	0.18157	0.054105
$\nu(4)$	0	0	0.37796	$r \cdot e^{j2\pi\frac{4}{12}}$	0.34131	0.44208
$\nu(5)$	0	0	0.37796	$r \cdot e^{j2\pi\frac{-1}{12}}$	0.31872	0.30352
$\nu(6)$	0	0	0	$r \cdot e^{j2\pi\frac{6}{12}}$	0.19637	0.063343
$\nu(7)$	0	0	0.37796	$r \cdot e^{j2\pi\frac{1}{12}}$	0.40443	0.47177
$\nu(8)$	0	0	0	$r \cdot e^{j2\pi\frac{-4}{12}}$	0.18624	0.068621
$\nu(9)$	0	0	0.37796	$r \cdot e^{j2\pi\frac{3}{12}}$	0.28521	0.24149
$\nu(10)$	0	0.44721	0	$r \cdot e^{j2\pi\frac{-2}{12}}$	0.17845	0.03761
$\nu(11)$	0	0.44721	0.37796	$r \cdot e^{j2\pi\frac{5}{12}}$	0.22443	0.26393

The top left plot of Fig. 5.25 shows the listed key profiles in a bar graph. In this specific plot they are normed to a vector length of 1 with Eq. (5.56) as opposed to a sum of 1 as the entries in Table 5.9.

5.5.2.2 Similarity Measure between Template and Extracted Vector

The most likely key can be estimated by finding the minimum distance between the extracted pitch chroma ν_e and the key profile template. Under the reasonable assumption that the template key profiles for different keys are identical but shifted versions of the $C\ Major$ profile, only 1 template needs to be stored per mode. Detecting modes other than major mode and minor mode is not of relevance in practice; thus, only 2 template vectors have to be stored to generate an overall set of 24 shifted templates.

The index m of the most likely key can then be computed by finding the minimum distance d between the extracted pitch chroma ν_e and the set of 24 template key profiles ν_t:

$$m = \min_{0 \leq s \leq 24} \left(d(s)\right). \tag{5.58}$$

Typical distance measures are

- *Euclidean Distance*:

$$d_\mathrm{E}(s) = \sqrt{\sum_{j=0}^{11} \left(\nu_\mathrm{e}(j) - \nu_{\mathrm{t,s}}(j)\right)^2}. \tag{5.59}$$

- *Manhattan Distance*:

$$d_\mathrm{M}(s) = \sum_{j=0}^{11} \left|\nu_\mathrm{e}(j) - \nu_{\mathrm{t,s}}(j)\right|. \tag{5.60}$$

- *Cosine Distance*:

$$d_\mathrm{C}(s) = 1 - \left(\frac{\sum\limits_{j=0}^{11} \nu_\mathrm{e}(j) \cdot \nu_{\mathrm{t,s}}(j)}{\sqrt{\sum\limits_{j=0}^{11} \nu_\mathrm{e}(j)^2} \cdot \sqrt{\sum\limits_{j=0}^{11} \nu_{\mathrm{t,s}}(j)^2}} \right). \tag{5.61}$$

- *Kullback-Leibler Divergence*:

$$d_\mathrm{KL}(s) = \sum_{j=0}^{11} \nu_\mathrm{e}(j) \cdot \log\left(\frac{\nu_\mathrm{e}(j)}{\nu_{\mathrm{t,s}}(j)}\right). \tag{5.62}$$

Figure 5.25 plots the inter-key distances between the shifted key profile templates given above with respect to $C\ Major$; the distances are shown both in a line plot (bottom left) and within the circle of fifths (right). The Euclidean distance has been used to compute these diagrams; therefore it is only logical that the input key profiles have been normed to a vector length of 1 (as opposed to a sum of 1). The distances basically behave as anticipated: the orthogonal template has the same high distance to all other keys while for the smoothed orthogonal template the distance increases to a maximum at the tritone — this would be musically reasonable were it not for the fact that the keys with root notes

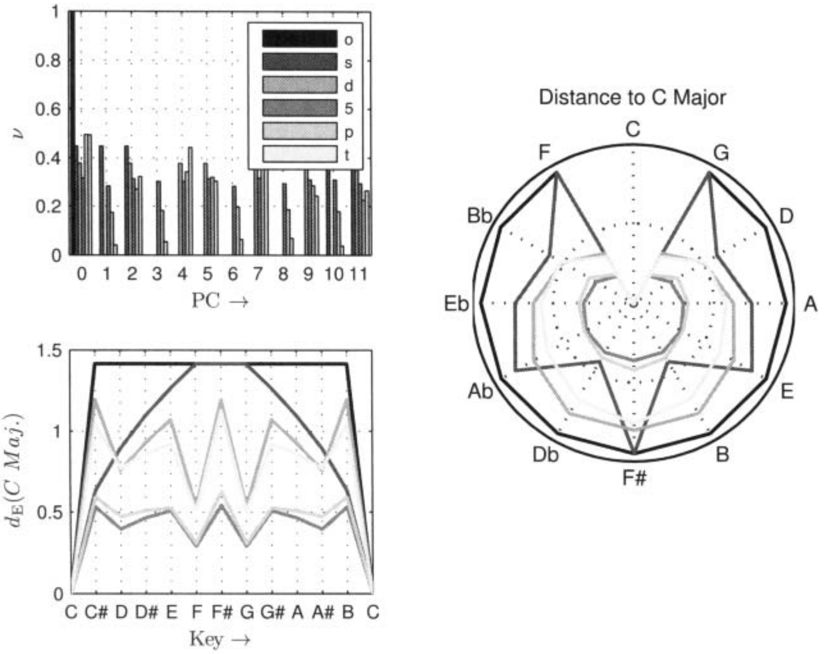

Figure 5.25 Key profile vectors (top left) and the resulting inter-key distances (Major) to $C\ Major$ (bottom left and right)

surrounding the tritone are the most closely related but have the same maximum distance. The remaining profiles basically show all the tendency of greater distances for keys more distant according to the circle of fifths. This is demonstrated by the cardioid shape of these distances in the polar diagram. It should be pointed out that the graph only visualizes the distances for keys in major mode and that none of the templates but the probe tone ratings and the extracted key profiles are able to separate between major mode and minor mode keys (although the model based on the circle of fifths can be adapted to a 3D model).

Theoretically, these inter-key distance measures are systematically wrong as the pitch chroma is a distribution and no vector. However, vector distances have been used and appear to work reasonably well in the absence of more fitting distance measures.

It is possible to smooth the pitch chroma to avoid maximum distances for nearby pitches [170] which in the case of tonality perception is probably most effective if the pitch chroma is reordered in a way that related keys have similar pitch chroma indices as described in Sect. 5.5.1.2.

There are other more complex mathematical models for estimating a distance between two keys or pitch class distributions; one example is Chew's *spiral array model* [171, 172]. Another model targeted specifically at automatic tonalness detection is the multidimensional *tonal centroid* proposed by Harte et al. [165] for which the pitch chroma is converted into a six-dimensional representation called tonal centroid based on the so-called harmonic network or *Tonnetz*. If enharmonic equivalence can be assumed, the Tonnetz can be transformed into three two-dimensional planes representing the circle of fifths, major thirds, and minor thirds. Another model is Gatzsche's *circular pitch space* [173].

5.5.2.3 Typical Key Detection Errors

The most frequent error of automatic key detection systems is — unsurprisingly — the confusion with closely related keys, namely the keys with a dominant and a subdominant relationship as well as the parallel key. Another typical error surfacing mainly with popular music is the confusion of major modes with minor modes and vice versa while correctly estimating the root note.

When detecting the key of pieces with modulations, i.e., of pieces with a changing key, analyzing only small sections at the start and end of the piece of music will improve results as it is more than common that the piece will start and end in the same (main) key [153, 166, 174].

5.6 Chord Recognition

Similar to key detection systems, the automatic recognition of chords utilizes the pitch chroma, with the difference that the pitch chroma is extracted from a shorter segment in the piece of music as the focus is on the local tonal context. Most modern systems extract one pitch chroma for the period between each pair of neighboring beats.

The pitch chroma is commonly mapped to a chord probability vector; the chord estimation itself is often based on heuristic or statistical models for the progression of chords over time. The transformation of the pitch chroma into the chord space is in the simplest case done with a linear transformation by the chord transformation matrix $\mathbf{\Gamma}$. Since the pitch chroma has the dimensions 12×1, the transformation matrix would be of dimension $\mathcal{T} \times 12$ with $\mathcal{T}$ being the number of chord templates. For example, the number of chord templates will be $\mathcal{T} = 24$ if only all major and minor triads are allowed. The transformation can be formulated as

$$\psi(n) = \mathbf{\Gamma} \cdot \nu(n) \qquad (5.63)$$

and can also be interpreted as the correlation between pitch chroma and each template (each row).

The resulting chord vector $\psi(n)$ then has the dimension $\mathcal{T} \times 1$ and is a measure of the salience or likelihood of a specific chord given the pitch chroma $\nu(n)$.

Each row of the transformation matrix represents one chord template; the simplest template would be to weight each pitch which is part of the chord by 1 and the remaining pitches with 0. Each row could be normalized to the sum of all entries in this row in order to compute the arithmetic mean. In this case, the template for a $C\ Major$ triad would be $[1/3, 0, 0, 0, 1/3, 0, 0, 1/3, 0, 0, 0]$.

Other transformations than a linear transformation are possible; the matrix multiplication (which can be seen as the calculation of the arithmetic mean of the result of multiplying each chord template with the pitch chroma):

$$\psi(0, n) = \sum_{j=0}^{11} \Gamma(0, j) \cdot \nu(j, n) \qquad (5.64)$$

could, for example, be replaced by computing the geometric mean

$$\psi_{\mathrm{G}}(0, n) = \prod_{j=0}^{11} \nu(j, n)^{\Gamma(0,j)}. \qquad (5.65)$$

It is also possible to use each row of the matrix directly as a template and compute a distance measure between extracted pitch chroma and the template.

Simply calculating the instantaneous probability of a chord while neglecting the likelihood of certain chord progressions would mean to ignore typical and well-known musical "standards" and to dismiss valuable information allowing us to improve the algorithm's reliability and robustness. Thus, nearly every system for automatic chord detection utilizes a model for chord progressions. This model is either analytically derived or trained from a set of data. There are three basic properties that determine the musicological validity of the model:[4]

- *Key Dependence*: a specific chord will have different preferred progressions dependent on the tonal context. Let us consider a typical cadence progression with the chords on the scale degrees I → IV → V → I. In the key $A\ Major$ ($A \rightarrow D \rightarrow E \rightarrow A$) this would imply that the dominant's ($E\ Major$) preferred transition would usually be to the tonic ($A\ Major$). However, were we in the key $B\ Major$, $E\ Major$'s harmonic function would be the subdominant with a relatively high likelihood of the following chord being $F\sharp\ Major$ (which would not even be part of the key $A\ Major$). Thus, chord progression models (theoretically) have to use the key of the piece of music.

- *Model Order*: musical context spans usually a larger area than just the neighboring chords. Therefore, the likelihood of a specific chord may depend not only on the directly preceding chord but on several preceding chords and possibly on some following chords.

- *Style Dependence*: different musical styles are based on different rule sets and different musical expectations. The transition probabilities between different chords will depend on the style of the piece of music being analyzed.

A typical approach to model the transition probability between chords is to use a *Hidden Markov Model (HMM)*. The states of the HMM represent the possible chords through the (possibly transformed) pitch chromas. The observation vectors can be either set a priori by applying "musical" knowledge or can be trained from a training data set. Examples for a priori settings are the simple chord template mentioned above [164] or the same templates weighted to take into account the influence of harmonics [161].

A very simple (and key independent) model for the chord progression likelihood is to use a circle of fifths [164] with the model errors mentioned above. A related approach is to utilize the correlation between Krumhansl's key profiles as chord distances [161].

The training data set can either be annotated audio [164], MIDI-synthesized audio [175], or a symbolic score format [161].

An example for a system acknowledging the key dependence has been published by Lee and Slaney by using one HMM for each of the 24 keys [176].

[4] It should be noted that these properties do not necessarily directly reflect on the algorithm's accuracy.