10. Pitch and Voicing Determination of Speech with an Extension Toward Music Signals

W. J. Hess

This chapter reviews selected methods for pitch determination of speech and music signals. As both these signals are time variant we first define what is subsumed under the term *pitch*. Then we subdivide pitch determination algorithms (PDAs) into short-term analysis algorithms, which apply some spectral transform and derive pitch from a frequency or lag domain representation, and time-domain algorithms, which analyze the signal directly and apply structural analysis or determine individual periods from the first partial or compute the instant of glottal closure in speech. In the 1970s, when many of these algorithms were developed, the main application in speech technology was the vocoder, whereas nowadays prosody recognition in speech understanding systems and high-accuracy pitch period determination for speech synthesis corpora are emphasized. In musical acoustics, pitch determination is applied in melody recognition or automatic musical transcription, where we also have the problem that several pitches can exist simultaneously.

Pitch and voicing determination of speech signals are the two subproblems of voice source analysis. In voiced speech, the vocal cords vibrate in a quasiperiodic way. Speech segments with voiceless excitation are generated by turbulent air flow at a constriction or by the release of a closure in the vocal tract. The parameters we have to determine are the manner of excitation, i. e., the presence of a voiced excitation and the presence of a voiceless excitation, a problem we will refer to as *voicing determination* and, for the segments of the speech signal in which a voiced excitation is present, the rate of vocal cord vibration, which is usually referred to as *pitch determination* or *fundamental frequency determination* in the literature.

Unlike the analysis of vocal-tract parameters, where a number of independent and equivalent representations are possible, there is no alternative to the parameters pitch and voicing, and the quality of a synthesized signal critically depends on their reliable and accurate determination. This chapter presents a selection of the methods applied in pitch and voicing determination. The emphasis, however, is on pitch determination.

Over the last two decades, the task of fundamental frequency determination has become increasingly popular in musical acoustics as well. In the beginning the methodology was largely imported from the speech community, but then the musical acoustics community developed algorithms and applications of their own, which in turn became increasingly interesting to the speech communication area. Hence it appears justified to include the aspect of fundamental frequency determination of music signals and to present some of the methods and specific problems of this area. One specific problem is multipitch determination from polyphonic signals, a problem that might also occur in speech when we have to separate two or more simultaneously speaking voices.

Pitch determination has a rather long history which goes back even beyond the times of vocoding. Literally hundreds of pitch determination algorithms (PDAs) have been developed. The most important developments leading to today's state of the art were made in the 1960s and 1970s; most of the methods that are briefly reviewed in this chapter were extensively discussed during this period [10.1]. Since then, least-squares and other statistical methods, particularly in connection with sinusoidal models [10.2], entered the domain. A number of known methods were improved and refined, whereas other solutions that required an amount of computational effort that appeared prohibitive at the time the algorithm was first developed were revived. With the widespread use of databases containing many labeled and processed speech data, it has nowadays also become possible to thoroughly evaluate the performance of the algorithms.

The bibliography in [10.1], dating from 1983, includes about 2000 entries. To give a complete overview of the more-recent developments, at least another 1000 bibliographic entries would have to be added. It goes without saying that this is not possible here given the space limitations. So we will necessarily have to present a selection, and many important contributions cannot be described.

The remainder of this chapter is organized as follows. In Sect. 10.1 the problems of pitch and voicing determination are described, definitions of what is subsumed under the term *pitch* are given, and the various PDAs are grossly categorized. Sections 10.2 and 10.3 give a more-detailed description of selected PDAs. Section 10.4 shortly reviews a selection of voicing determination algorithms (VDAs); Sect. 10.5 deals with questions of error analysis and evaluation. Selected applications are discussed in Sect. 10.6, and Sect. 10.7 finally presents a couple of new developments, such as determining the instant of glottal closure or processing signals that contain more than one pitch, such as polyphonic music.

10.1 Pitch in Time-Variant Quasiperiodic Acoustic Signals

10.1.1 Basic Definitions

Pitch, i. e., the fundamental frequency F_0 and fundamental period T_0 of a (quasi)periodic signal, can be measured in many ways. If a signal is completely stationary and periodic, all these strategies – provided they operate correctly – lead to identical results. Since both speech and musical signals, however, are nonstationary and time variant, aspects of each strategy such as the starting point of the measurement, the length of the measuring interval, the way of averaging (if any), or the operating domain (time, frequency, lag etc.) start to influence the results and may lead to estimates that differ from algorithm to algorithm even if all these results are *correct* and *accurate*. Before entering a discussion on individual methods and applications, we must therefore have a look at the parameter pitch and provide a clear definition of what should be measured and what is actually measured.

A word on terminology first. There are three points of view for looking at such a problem of acoustic signal processing [10.3]: the *production*, the *signal-processing*, and the *perception* points of view. For pitch determination of speech, the production point of view is obviously oriented toward phonation in the human larynx; we will thus have to start from a time-domain representation of the waveform as a train of laryngeal pulses. If a pitch determination algorithm (PDA) works in a speech-production oriented way, it measures individual *laryngeal excitation cycles* or, if some averaging is performed, the *rate of vocal-fold vibration*. The signal-processing point of view, which can be applied

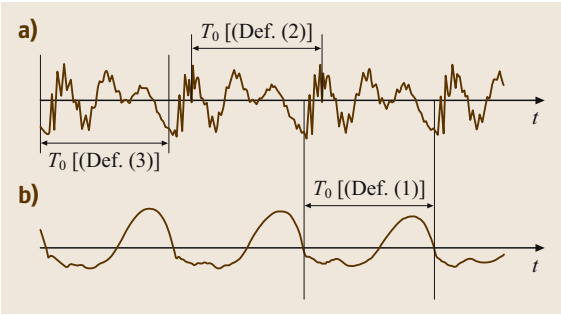

Fig. 10.1a,b Time-domain definitions of T_0. **(a)** Speech signal (a couple of periods), **(b)** glottal waveform (reconstructed). For further detail, see the text

to any acoustic signal, means that (quasi)periodicity or at least cyclic behavior is observed, and that the task is to extract those features that best represent this periodicity. The pertinent terms are *fundamental frequency* and *fundamental period*. If individual cycles are determined, we may (somewhat inconsistently) speak of *pitch periods* or simply of *periods*. The perception point of view leads to a frequency-domain representation since pitch sensation primarily corresponds to a frequency [10.4, 5] even if a time-domain mechanism is involved [10.6]. This point of view is associated with the original meaning of the term *pitch*. Yet the term *pitch* has consistently been used as some kind of common denominator and a general name for all those terms, at least in the technical literature on speech [10.7]. In the following, we will therefore use the term *pitch* in this wider sense, even for musical signals.

When we proceed from production to perception, we arrive at five basic definitions of pitch that apply to speech signals and could read as follows ([10.1, 8, 9]; Fig. 10.1):

1. *T_0 is defined as the elapsed time between two successive laryngeal pulses. Measurement starts at a well-specified point within the glottal cycle, preferably at the instant of glottal closure.*
   PDAs that obey this definition will be able to locate the points of glottal closure and to delimit individual laryngeal cycles. This goes beyond the scope of ordinary pitch determination in speech; if only the signal is available for the analysis, it must be totally undistorted if reliable results are to be expected. For music signals we can apply this definition if we analyze a human voice or an instrument that operates in a way similar to the human voice.
2. *T_0 is defined as the elapsed time between two successive laryngeal pulses. Measurement starts at an arbitrary point within an excitation cycle. The choice of this point depends on the individual method, but for a given PDA it is always located at the same position within the cycle.*
   Time-domain PDAs usually follow this definition. The reference point can be a significant extreme, a certain zero crossing, etc. The signal is tracked period by period in a synchronous way yielding *individual pitch periods*. This principle can be applied to both speech and music signals. In speech it may even be possible to derive the point of glottal closure from the reference point when the signal is undistorted.
3. *T_0 is defined as the elapsed time between two successive excitation cycles. Measurement starts at an arbitrary instant which is fixed according to external conditions, and ends when a complete cycle has elapsed.*
   This is an *incremental* definition. T_0 still equals the length of an individual period, but no longer from the production point of view, since the definition has nothing to do with an individual excitation cycle. The synchronous method of processing is maintained, but the phase relations between the laryngeal waveform and the markers, i. e., the pitch period delimiters at the output of the algorithm, are lost. Once a reference point in time has been established, it is kept as long as the measurement is correct and the signal remains cyclic, for instance as long as voicing continues. If this synchronization is interrupted, the reference point is lost, and the next reference point may be completely different with respect to its position within an excitation cycle.

4a. *T_0 is defined as the average length of several periods. The way in which averaging is performed, and how many periods are involved, is a matter of the individual algorithm.*
   This is the standard definition of T_0 for any PDA that applies stationary short-term analysis, including the implementations of frequency-domain PDAs. Well-known methods, such as cepstrum [10.10] or autocorrelation [10.11] approaches follow this definition. The pertinent frequency-domain definition reads as follows.

4b. *F_0 is defined as the fundamental frequency of an (approximately) harmonic pattern in the (short-term) spectral representation of the signal. It depends on the particular method in which way F_0 is calculated from this pattern.*

The perception point of view of the problem leads to a different definition of pitch [10.5]:

5. *F_0 is defined as the frequency of the sinusoid that evokes the same perceived pitch (residue pitch, virtual pitch, etc.) as the complex sound that represents the input speech signal.*
   Above all, this definition is a long-term definition [10.12]. Pitch perception theories were first developed for stationary complex sounds with constant F_0. The question of the behavior of the human ear with respect to short-term perception of time-variant pitch is not yet fully understood. The difference limen for F_0 changes, for instance, goes up by at least an order of magnitude when time-variant stimuli are involved [10.13, 14]. In practice even such PDAs that claim to be perception oriented [10.15, 16] enter the frequency domain in a similar way as in definition 4b, i. e., by some discrete Fourier transform (DFT) with previous time-domain signal windowing.

Since the results of individual algorithms differ according to the definition they follow, and since these five definitions are partly given in the time (or lag) domain and partly in the frequency domain, it is necessary to reestablish the relation between the time- and frequency-domain representations of pitch,

$$F_0 = 1/T_0 \tag{10.1}$$

in such a way that, whenever a measurement is carried out in one of these domains, however T_0 or F_0 is defined there, the representation in the other domain will always be established by this relation.

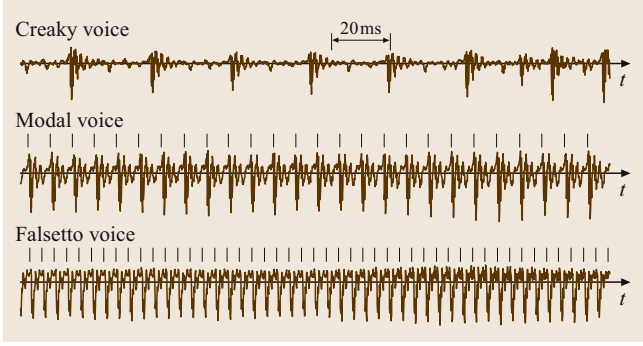

Fig. 10.2 Speech signal excitation with different voice registers (male speaker, vowel [ɛ])

10.1.2 Why is the Problem Difficult?

Literally hundreds of methods for pitch determination have been developed. None of them has been reported to be error free for any signal, application, or environmental condition.

At first glance the task appears simple: one just has to detect the fundamental frequency of a quasiperiodic signal. When dealing with speech signals, however, the assumption of (quasi)periodicity is often far from reality. For a number of reasons, the task of pitch determination must be counted among the most difficult problems in speech analysis.

- In principle, speech is a nonstationary process; the momentary position of the vocal tract may change abruptly at any time. This leads to drastic variations in the temporal structure of the signal, even between subsequent pitch periods.
- In fluent speech there are voiced segments that last only a few pitch periods [10.17].
- Due to the flexibility of articulatory gestures and the wide variety of voices, there exist a multitude of possible temporal structures. Narrowband formants at low harmonics (especially at the second or third harmonic) are a particular source of trouble.
- For an arbitrary speech signal uttered by an unknown speaker, the fundamental frequency can vary over a range of almost four octaves (50–800 Hz). Especially for female voices, F_0 thus often coincides with the first formant (the latter being about 200–1400 Hz). This causes problems when inverse-filtering techniques are applied.
- The excitation signal itself is not always regular. Even under normal conditions, i. e., when the voice is neither hoarse nor pathologic, the glottal waveform exhibits occasional irregularities. In addition, the voice may temporarily fall into vocal fry or creak ([10.18, 19]; Fig. 10.2).
- Additional problems arise in speech communication systems, where the signal is often distorted or band limited (for instance, in telephone or even mobile-phone channels).

For music signals, the situation is comparable. The range of F_0 can be even wider than for speech. However, structural changes of the signal usually occur more slowly for music. The maximum speed at which a musical instrument can be played is about 10 notes per second so that a single note usually lasts at least 100 ms. For speech, on the other hand, 100 ms is already a lot of time which

can consist of three or more segments. An additional problem in music is that we may have to analyze polyphonic signals with several pitches present at the same time.

10.1.3 Categorizing the Methods

A PDA is defined as consisting of three processing steps: (a) the preprocessor, (b) the basic extractor, and (c) the postprocessor [10.1, 20]. The basic extractor performs the main task of converting the input signal into a series of pitch estimates. The task of the preprocessor is data reduction and enhancement in order to facilitate the operation of the basic extractor. The postprocessor (Sect. 10.5.4) is more application oriented. Typical tasks are error correction, pitch tracking, and contour smoothing, or visualization.

The existing PDA principles can be split into two gross categories when the input signal of the basic extractor is taken as a criterion. If this signal has the same time base as the original input signal, the PDA operates in the time domain. It will thus measure T_0 according to one of the definitions 1–3 above. In all other cases, somewhere in the preprocessor the time domain is left. Since the input signal is time variant, this is done by a short-term transform; and we will usually determine T_0 or F_0 according to definitions 4a,b or 5; in some cases definition 3 may apply as well. Accordingly, we have the two categories: time-domain PDAs, and short-term analysis PDAs. These will be discussed in the next two sections.

10.2 Short-Term Analysis PDAs

In any short-term analysis PDA a short-term (or short-time) transformation is performed in the preprocessor. The speech signal is split into a series of frames; an individual frame is obtained by taking a limited number of consecutive samples of the signal $s(n)$ from a starting point, $n = q$, to the ending point, $n = q + K$. The frame length K (or $K + 1$) is chosen short enough so that the parameter(s) to be measured can be assumed approximately constant within the frame. On the other hand, K must be large enough to guarantee that the parameter remains measurable. For most short-term analysis PDAs a frame thus requires two or three complete periods at least. In extreme cases, when F_0 changes abruptly, or when the signal is irregular, these two conditions are in conflict with each other and may become a source of error [10.21]. The frame interval Q, i. e., the distance between consecutive frames (or its reciprocal, the frame rate), is determined in such a way that any significant parameter change is documented in the measurements. 100 frames/s, i. e., $Q = 10$ ms, is a usual value.

The short-term transform can be thought of as behaving like a concave mirror that focuses all the information on pitch scattered across the frame into one single peak in the spectral domain. This peak is then determined by a peak detector (the usual implementation of the basic extractor in this type of PDAs). Hence this algorithm yields a sequence of average pitch estimates. The short-term transform causes the phase relations between the spectral domain and the original signal to be lost. At the same time, however, the algorithm loses much of its sensitivity to phase distortions and signal degradation.

Not all the known spectral transforms show the desired focusing effect. Those that do are in some way related to the (power) spectrum: correlation techniques, frequency-domain analysis, active modeling, and statistical approaches (Fig. 10.3). These methods will be discussed in more detail in the following.

10.2.1 Correlation and Distance Function

Among the correlation techniques we find the well-known short-term autocorrelation function (ACF)

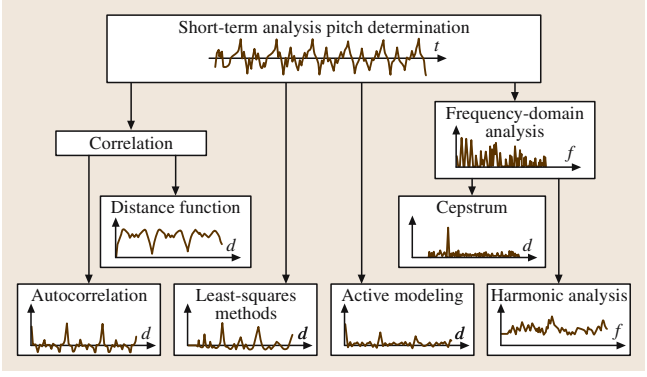

Fig. 10.3 Methods of short-term analysis (short-time analysis) pitch determination. (Time t and lag d scales are identical; the frequency f scale in the box 'Harmonic analysis' has been magnified)

usually given by

$$r(d,q) = \sum_{n=q}^{q+K-d} s(n)s(n+d)\,. \tag{10.2}$$

The autocorrelation function of a periodic signal exhibits a strong peak when the lag d equals the period T_0/T of the signal, T being the time-domain sampling interval of the signal.

The autocorrelation PDA is among the oldest principles for short-term analysis PDAs. However, it tends to fail when the signal has a strong formant at the second or third harmonic. Therefore this technique became successful in pitch determination of band-limited speech signals when it was combined with time-domain linear or nonlinear preprocessing, such as center clipping or inverse filtering [10.22, 23].

The counterpart to autocorrelation is given by applying a distance function, for instance the average magnitude difference function (AMDF) [10.24, 25]:

$$\mathrm{AMDF}(d,q) = \sum_{n=q}^{q+K} |s(n) - s(n+d)|\,. \tag{10.3}$$

If the signal were strictly periodic, the distance function would vanish at the lag (delay time) $d = T_0/T$. For quasiperiodic signals there will be at least a strong minimum at this value of d. So, in contrast to other short-term PDAs where the estimate of T_0 or F_0 is indicated by a maximum whose position and value have to be determined, the minimum has an ideal target value of zero so that we only need to determine its position. For this reason, distance functions do not require (quasi)-stationarity within the measuring interval; they can cope with very short frames of one pitch period or even less. This principle is thus able to follow definition 3.

Shimamura and *Kobayashi* [10.26] combine ACF and AMDF in that they weight the short-term ACF with the reciprocal of the AMDF, thus enhancing the principal peak of the ACF at $d = T_0/T$. For the PDA they named YIN, *De Cheveigné* and *Kawahara* [10.27] start from a squared distance function,

$$D(d,q) = \sum_{n=q}^{q+K} [s(n) - s(n+d)]^2 \tag{10.4}$$

and normalize it to increase its values at low lags,

$$D'(d,q) = \frac{D(d,q)}{\frac{1}{d}\sum_{\delta=1}^{d} D(\delta,q)}\;;\; d>0 \tag{10.5}$$

with $D'(0) = 1$. In doing so, the authors were able to drop the high-frequency limit of the measuring range and to apply their PDA to high-pitched music signals as well. The normalized distance function is locally interpolated around its minima to increase the accuracy of the value of D' at the minima and the pitch estimate at the same time.

Knowing that many errors arise from a mismatch during short-term analysis (which results in too few or too many pitch periods within a given frame), *Fujisaki* et al. [10.21] investigated the influence of the relations between the error rate, the frame length, and the actual value of T_0 for an autocorrelation PDA that operates on the linear prediction residual. The optimum occurs when the frame contains about three pitch periods. Since this value is different for every individual voice, a fixed-frame PDA runs nonoptimally for most situations. For an exponential window, however, this optimum converges to a time constant of about 10 ms for all voices. For a number of PDAs, especially for the autocorrelation PDA, such a window permits recursive updating of the autocorrelation function, i. e., sample-by-sample pitch estimation without excessive computational effort.

Hirose et al. [10.28] and *Talkin* [10.17] showed that the autocorrelation function can also be computed in a nonstationary way using a suitable normalization,

$$r(d,q) = \frac{\sum_{n=q}^{q+K} s(n)s(n+d)}{\sqrt{\left[\sum_{n=q}^{q+K} s^2(n)\right]\left[\sum_{n=q}^{q+K} s^2(n+d)\right]}}\,. \tag{10.6}$$

In Talkin's PDA a 7.5 ms frame is used; the effective frame length is of course 7.5 ms plus the longest pitch period in the measuring range.

Terez [10.29] applies a multidimensional embedding function and a scatter plot procedure derived from chaos theory. The underlying idea, however, is quite straightforward and leads to a distance function in a multidimensional state space. The problem is how to convert the one-dimensional speech signal into a multidimensional representation. In Terez's algorithm a vector is formed from several equally spaced samples of the signal,

$$\boldsymbol{s}(n) = \left[s(n)\; s(n+d)\; \cdots\; s(n+Nd)\right]^{\mathrm{T}}\,, \tag{10.7}$$

(where the frame reference point q has been omitted here and in the following for sake of simplicity) whose components create an N-dimensional space, the state space. In Terez's algorithm, $N = 3$ and $d = 12$ samples gave the best results. If the signal is voiced, i. e., cyclic, the vector $\boldsymbol{s}$ will describe a closed curve in the state space as time proceeds, and after one pitch period it is

expected to come back near the starting point. We can thus expect the (Euclidian) distance

$$D(n,\,p) = \|\boldsymbol{s}(n) - \boldsymbol{s}(n+p)\| \tag{10.8}$$

generally to become a minimum when the trial period p equals the true period T_0/T. If we compute $D(n,\,p)$ for all samples $s(n)$ within the frame and all values of p within the measuring range and count the number of events, depending on p, where D lies below a predetermined threshold, we arrive at a periodicity histogram that shows a sharp maximum at $p = T_0/T$.

As it develops the distance function D for all samples of a frame, this PDA follows the short-term analysis principle. Yet one can think of running it with a comparatively short window, thus following definition 3.

The idea of using a multidimensional representation of the signal for a PDA (and VDA) dates back to the 1950s [10.1]. In 1964 *Rader* [10.30] published the *vector* PDA where he used the output signals from a filterbank (cf. *Yaggi* [10.31], Sect. 10.3.3) and their Hilbert transforms to form a multidimensional vector $\boldsymbol{s}(n, q)$. Rader then used the Euclidian distance between the vector at the starting point $n = q$ of the measurement and the points $q + p$ to set up a distance function which shows a strong minimum when p equals the true period T_0/T. This PDA follows definition 3 as well.

Medan et al. [10.32] present a PDA (they called the super-resolution PDA) that explicitly addresses the problem of granularity due to signal sampling and applies a short-term window whose length depends on the trial pitch period p in that it takes on a length of exactly $2p$. A similarity function is formulated that expresses the relation between the two periods in the window,

$$\begin{aligned} \boldsymbol{s}(n, q) &= a \cdot \boldsymbol{s}(n + p, q) + e(n, q)\,; \\ & n = q, \cdots, q + p - 1\,. \end{aligned} \tag{10.9}$$

Here, a is a positive amplitude factor that takes into account possible intensity changes between adjacent periods. Equation (10.9) is optimized with respect to a and the unknown period p applying a least-squares criterion to minimize the error e,

$$\hat{p} = \operatorname{argmin}_{a,e} \left(\frac{\sum_{n=q}^{q+p-1} [s(n) - as(n+p)]^2}{\sum_{n=q}^{q+p-1} s^2(n)} \right)\,. \tag{10.10}$$

This optimization finally results in maximization of the correlation term

$$\hat{p} = \operatorname{argmax} \left(\frac{\left[\sum_{n=q}^{q+p-1} s(n)s(n+p) \right]^2}{\left[\sum_{n=q}^{q+p-1} s^2(n) \right] \left[\sum_{n=q}^{q+p-1} s^2(n+p) \right]} \right)\,. \tag{10.11}$$

This resembles Talkin's ACF approach [10.17] except that here the trial period p determines the length of the window as well.

From (10.11) a pitch period estimate can only be derived as an integer number of samples. In a second pass, this estimate is refined (to yield the super-resolution) by expanding (10.11) for a fraction of a sample using linear interpolation.

10.2.2 Cepstrum and Other Double-Transform Methods

The sensitivity against strong first formants, especially when they coincide with the second or third harmonic, is one of the big problems in pitch determination. This problem is suitably met by some procedure for spectral flattening.

Spectral flattening can be achieved in several ways. One of them is time-domain nonlinear distortion, such as center clipping ([10.11, 22]; see previous section). A second way is linear spectral distortion by inverse filtering (e.g., [10.23]). A third way is frequency-domain amplitude compression by nonlinear distortion of the spectrum. This algorithm operates as follows: (1) short-term analysis and transformation into the frequency domain via a suitable discrete Fourier transform (DFT), (2) nonlinear distortion in the frequency domain, and (3) inverse DFT back into the time domain (which we will call the lag domain to avoid ambiguity).

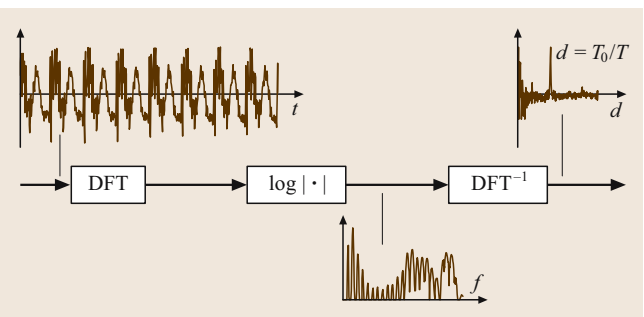

Fig. 10.4 Cepstrum pitch determination. Signal: vowel [i], 48 ms

If we take the logarithm of the power spectrum as the frequency-domain nonlinear distortion, we arrive at the well-known cepstrum PDA ([10.10]; Fig. 10.4). Instead of the logarithmic spectrum, however, one can also compute the amplitude spectrum or its square root and transform it back [10.33, 34]. The inverse Fourier transform of the power spectrum gives the autocorrelation function. All these so-called double-transform techniques [10.33] lead to a lag-domain representation that exhibits a strong maximum at the lag $d = T_0/T$. The independent variable, lag (or *quefrency*, as it is called with respect to the cepstrum [10.10]), has the physical dimension of time, but as long as the phase relations are discarded by the nonlinear distortion, all values of d refer to a virtual point $d = 0$ where we will always find a pitch pulse, and then the next one necessarily shows up at $d = T_0/T$.

Two members of this group were already mentioned: the autocorrelation PDA [10.11] and the cepstrum PDA [10.10]. It is well known that the autocorrelation function can be computed as the inverse Fourier transform of the power spectrum. Here, the distortion consists of taking the squared magnitude of the complex spectrum. The cepstrum, on the other hand, uses the logarithm of the spectrum. The two methods therefore differ only in the characteristics of the respective nonlinear distortions applied in the spectral domain. The cepstrum PDA is known to be rather insensitive to strong formants at higher harmonics but to develop a certain sensitivity to additive noise. The autocorrelation PDA, on the other hand, is insensitive to noise but rather sensitive to strong formants. Regarding the slope of the distortion characteristic, we observe the dynamic range of the spectrum being expanded by squaring the spectrum for the autocorrelation PDA, whereas the spectrum is substantially flattened by taking the logarithm. The two requirements – robustness against strong formants and robustness against additive (white) noise – are in some way contradictory. Expanding the dynamic range of the spectrum emphasizes strong individual components, such as formants, and suppresses wideband noise, whereas spectral flattening equalizes strong components and at the same time raises the level of low-energy regions in the spectrum, thus raising the level of the noise as well. Thus it is worth looking for other characteristics that perform spectral amplitude compression. *Sreenivas* [10.36] proposed the fourth root of the power spectrum instead of the logarithm. For larger amplitudes this characteristic behaves very much like the logarithm; for small amplitudes, however, it has the advantage of going to zero and not to $-\infty$. *Weiss* et al. [10.33] used the amplitude spectrum, i. e., the magnitude of the complex spectrum.

10.2.3 Frequency-Domain Methods: Harmonic Analysis

Direct determination of F_0 as the location of the lowest peak in the power or amplitude spectrum is unreliable and inaccurate; it is preferable to investigate the harmonic structure of the signal so that all harmonics contribute to the estimate of F_0. One way to do this is spectral compression combined with harmonic pattern matching, which computes the fundamental frequency as the greatest common divider of all harmonics. The power spectrum is compressed along the frequency axis by a factor of two, three etc. and then added to the original power spectrum. This operation gives a peak at F_0 resulting from the coherent additive contribution of the higher harmonics [10.35, 37]. Some of these PDAs stem from theories and functional models of pitch perception in the human ear [10.12, 15, 16].

The PDA described by *Martin* [10.35] (Fig. 10.5) modifies the harmonic pattern-matching principle in such a way that the computational effort for the spectral transform is minimized. The signal is first decimated to 4 kHz and then Fourier transformed by a 128 point fast Fourier transform (FFT). This yields a spectral resolution of about 30 Hz, which is sufficient to represent

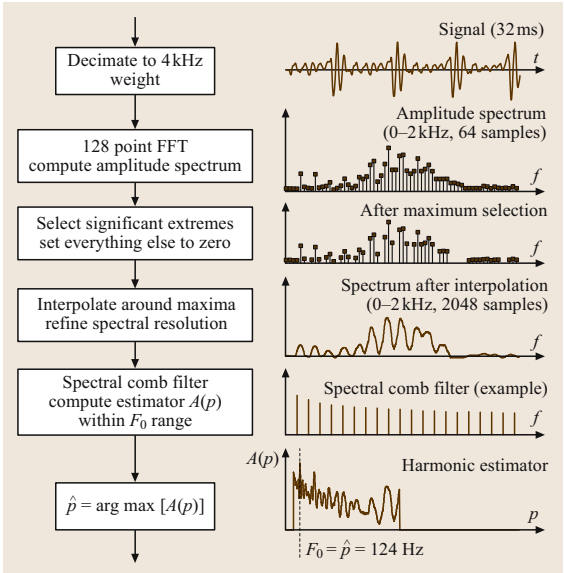

Fig. 10.5 Frequency-domain PDA by harmonic compression and pattern matching (after *Martin* [10.35])

values of F_0 down to 60 Hz. The algorithm then enhances any spectral information that may pertain to a harmonic structure by preserving only those spectral samples that represent local maxima or their immediate neighbors and setting everything else to zero. To measure F_0 with sufficient accuracy, the spectral resolution is then increased to 1 Hz. A spectral comb filter, which is applied over the whole range of F_0, yields the harmonic estimation function $A(p)$; the value of p where this function reaches its maximum is taken as the estimate for F_0. In a more-recent version of this PDA, *Martin* [10.38, 39] applies a logarithmic frequency scale for the computation of $A(p)$, which results in another substantial reduction of the computational effort and has the additional advantage that the relative accuracy of the PDA is now constant over the whole range of F_0.

Similar to *Martin*'s [10.38] PDA for speech, *Brown* [10.41] developed a frequency-domain PDA for music which uses a logarithmic frequency scale. In Brown's PDA the spacing on the frequency axis equals a quarter tone (about 3%), i. e., 1/24 of an octave. In such a scale that corresponds to a musical interval scale, a harmonic structure, if the timbre is unchanged, always takes on the same pattern regardless of the value of its fundamental frequency (Fig. 10.6). Consequently, a pattern-recognition algorithm is applied to detect such patterns in the spectrum and to locate their starting point corresponding to F_0. The patterns themselves depend on the kind of instrument analyzed and can be adjusted according to the respective instruments.

Special attention is given to the frequency resolution of the PDA. To apply a pattern-recognition method, patterns are expected to align with the semitone scale. This requires the frequency scale spacing of a quarter tone. To make sure that a harmonic shows up in one and only one frequency bin, we need a window length of 34 fundamental periods to satisfy the sampling theorem in the frequency domain. For $F_0 = 100$ Hz this would give a window of 340 ms, which is far beyond reality for speech and even excessive for music. However, if we adapt the window-length requirement to the fundamental frequency to be determined, we would need 34 periods at any F_0 to be measured, which results in much shorter windows for higher-frequency bins. For the DFT, this leads to a window whose length is inversely proportional to the bin's frequency. If the spectral values are computed individually, both an individual time-domain window for each frequency bin and unequal spacing of the frequency bins are possible. *Brown* and *Puckette* [10.40] showed that a fast Fourier transform can be applied if its kernel is slightly modified. The PDA by *Medan* et al. [10.32] is a time-domain counterpart to this approach.

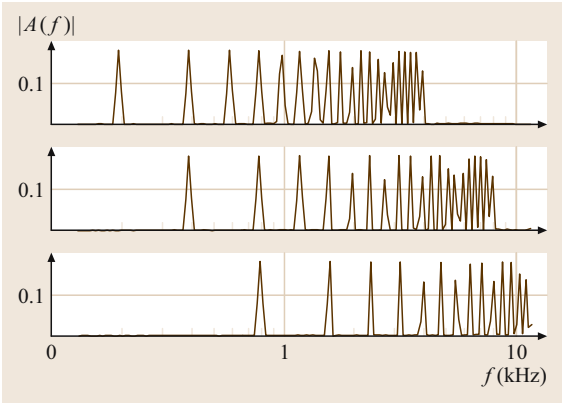

Fig. 10.6 Harmonic patterns in log frequency scale (after *Brown* and *Puckette* [10.40])

As the accuracy of this PDA was not sufficient to determine F_0 for instruments that can vary their frequencies continuously (such as string or wind instruments or a human voice), and as the required window length was excessive even for music, a PDA with a 25 ms window was developed [10.40] whose frequency resolution was refined using a phase-change technique. This technique is based on the instantaneous-frequency approach by *Charpentier* [10.42] (see below) who used the short-term phase spectrum at two successive samples in time to determine the frequencies of the individual harmonics without needing spectral interpolation. When a Hann window is used to weight the time signal, the time shift of one sample can be recursively computed in the frequency domain without needing another DFT.

Lahat et al. [10.43] transfer the autocorrelation method into the frequency domain. The amplitude spectrum is passed through a bank of 16 spectral filters (*lifters*), which cover the measuring range of F_0. At the output of each lifter a frequency domain autocorrelation function is calculated covering the respective range of each lifter. The estimate for F_0 is then determined as the location of the maximum of that function and refined by interpolation.

For harmonic analysis it is often convenient to estimate the number of harmonics, i. e., the order of a harmonic model, simultaneously with the fundamental frequency. For instance, *Doval* and *Rodet* [10.44] apply such a procedure for a PDA with an extremely wide measuring range (40–4000 Hz) for music signals. The algorithm is based on a harmonic-matching procedure using a maximum-likelihood criterion. To obtain

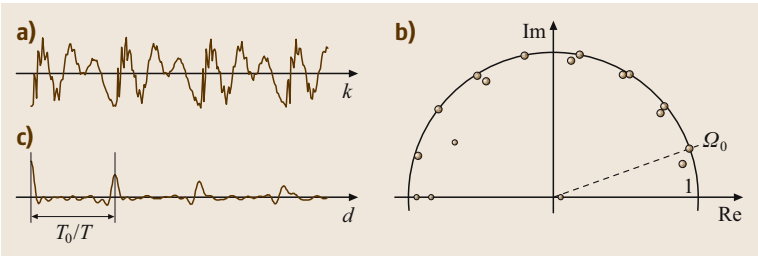

Fig. 10.7a–c PDA with active modeling. **(a)** Signal: 32 ms, vowel [e], male voice; **(b)** zeros of the 41-st-order LP polynomial in the z plane (upper half; sampling frequency reduced to 2 kHz); **(c)** reconstructed impulse response with zero phase and equal amplitude of all partials

a rapid initial estimate, the measuring range is split into subranges that have an equal number of partials in the frequency range for which the spectrum is analyzed. The initial estimate is obtained by selecting the subrange that is most likely to contain the optimal estimate for F_0. For the final step of refining the estimate, only this subrange is further processed.

The principle of instantaneous frequency (IF) was introduced into pitch determination by *Charpentier* [10.42]. Instantaneous frequency is defined as the time derivative of the instantaneous phase,

$$\dot{\varphi}(t) := \frac{\mathrm{d}\varphi}{\mathrm{d}t} \quad \text{for } s(t) = a(t)\exp[\mathrm{i}\omega(t)t]\,, \tag{10.12}$$

where $a(t)$ is the instantaneous amplitude. The short-term Fourier transform can be viewed as a set of bandpass filters as follows,

$$S(f,t) = \int_{-\infty}^{\infty} s(\tau)w(t-\tau)\mathrm{e}^{-2\pi\mathrm{i}f\tau}\,\mathrm{d}\tau \tag{10.13}$$

$$= \mathrm{e}^{-2\pi\mathrm{i}ft}\int_{-\infty}^{\infty} s(\tau)w(t-\tau)\mathrm{e}^{-2\pi\mathrm{i}f(t-\tau)}\,\mathrm{d}\tau$$

$$F(f,t) = \mathrm{e}^{2\pi\mathrm{i}ft}S(f,t)\,. \tag{10.14}$$

Here, the signal F is the output of the *bandpass* centered around the frequency f. The IF for this signal becomes

$$\dot{\varphi}(f,t) = \frac{\partial}{\partial t}\arg[F(f,t)]\,. \tag{10.15}$$

There are different ways to effectively compute the IF from the discrete short-term spectrum [10.42,45,46]. The bandpass filters have a certain bandwidth that depends on the time-domain window applied [10.42] and extends over more than one DFT coefficient. If we now compute the IF for each frequency bin of the DFT spectrum of a voiced speech signal, the IFs of bins adjacent to a strong harmonic tend to cluster around the true frequency of this harmonic, and so it is possible to enhance the harmonics in the spectrum if the bins are regrouped according to their respective IFs, thus forming the so-called IF amplitude spectrum. *Abe* et al. [10.45] transform the IF amplitude spectrum back into the time domain, thus obtaining a representation similar to that of a double-spectral-transform PDA. *Nakatani* and *Irino* [10.46] define a spectral dominance function that suppresses insignificant information in the IF amplitude spectrum and derive F_0 by harmonic matching of this dominance function.

10.2.4 Active Modeling

Linear prediction (LP) is usually applied to estimating the transfer function of the vocal tract. If a high-order LP filter is applied to a voiced speech signal, however, its poles will match the individual harmonics of the signal rather than the resonances of the vocal tract. A PDA based on this principle was designed by *Atal* (unpublished; see [10.47], or [10.1]). The algorithm operates as follows (Fig. 10.7):

- After decimation to 2 kHz, the signal is analyzed with a 41-st-order LP filter using the covariance method of linear prediction. The high order guarantees that even at the low end of the F_0 range, i. e., at $F_0 = 50$ Hz, two complex poles are available for each harmonic. Each complex pole pair represents an exponentially decaying (or rising) oscillation.
- To eliminate phase information, all residues at the pole locations in the z plane are set to unity. The pertinent computation can be avoided when the locations of the poles are explicitly determined.
- The impulse response of the filter now supplies a waveform for the estimation of T_0 (Fig. 10.7c). When the poles are explicitly available, it is sufficient to determine and to sum up the impulse responses of the individual second-order partial filters. This method has the advantage that the sampling frequency of the impulse response – and with it the measurement accuracy – can easily be increased. In addition, poles that are situated outside or far inside the unit circle can be modified or excluded from further processing.

Arévalo [10.48] showed that this PDA is extremely robust to noise and that one can also use it with short frame lengths so that it matches definition 3.

10.2.5 Least Squares and Other Statistical Methods

The first statistical approach in pitch determination is based on a least-squares principle. Originally this approach was based on a mathematical procedure to separate a periodic signal of unknown period T_0 from Gaussian noise within a finite signal window [10.49]. It computes the energy of the periodic component at a trial period τ and varies τ over the whole range of T_0. The value of τ that maximizes the energy of the periodic component for the given signal is then taken as the estimate of T_0. *Friedman* [10.50] showed that this PDA has a trivial maximum when τ equals the window length K, and developed a work-around.

With respect to robustness, the least-squares PDA behaves like the autocorrelation principle: it is extremely robust against noise but somewhat sensitive to strong formants. However, there is no algorithmic shortcut so that an order of K^2 operations are needed to compute the estimate for a frame. So this PDA was slower than its counterparts that can make use of the FFT; hence this principle was not further pursued until more powerful computers became available.

The method was revived with the upcoming of the sinusoidal model of speech [10.2]. The continuous speech signal $s(n)$ is modeled as a sum of sinusoids with time-varying amplitudes, frequencies, and phases. Within a short-term frame, these parameters can be assumed constant,

$$s(n) = \sum_{m=1}^{M} S_m \ \exp\left(\mathrm{i}\Omega_m n + \varphi_m\right) \ . \tag{10.16}$$

The parameters of this model are estimated from the peaks within the short-term Fourier spectrum of the frame. This can be converted into a PDA [10.51] when the sinusoidal representation in (10.16), whose frequencies are generally *not* harmonics of a fundamental, is matched against a harmonic model,

$$u(n) = \sum_{k=1}^{K} U_k \ \exp\left(\mathrm{i}k\Omega_0 n + \psi_k\right) \ . \tag{10.17}$$

Starting from the difference between $s(n)$ and $u(n)$, the match is done using a modified least-squares criterion, which finally results in maximizing the expression with respect to the trial (angular) fundamental frequency p,

$$\varrho(p) = \sum_{k=1}^{K(p)} U(kp) \times \left\{ \max_{\Omega_m \in L(kp)} \left[S_m D(\Omega_m - kp)\right] - \frac{1}{2} U(kp) \right\} \ . \tag{10.18}$$

Like the model of virtual-pitch perception [10.4], this criterion takes into account near-coincidences between a harmonic kp and the (angular) frequency Ω_m of the respective component of the sinusoidal model, and it defines a lobe L of width p around each harmonic and the corresponding weighting function

$$D(\Omega - kp) = \frac{\sin[2\pi(\Omega - kp)/p\,]}{(\Omega - kp)/p} \tag{10.19}$$

within the lobe, and zero outside. The lobe becomes narrow for low values of p and broader for higher values. If there are several components Ω_m within a lobe, only the largest is taken. The amplitude estimates $U(kp)$ are derived from a smoothed Fourier or LP spectrum of the frame. The measurement may be confined to a subband of the signal, e.g., to 2 kHz.

Both the sinusoidal model and the PDA have been applied to speech and music signals.

For pitch detection in noisy speech, third-order statistics are occasionally applied. One such PDA was developed by *Moreno* and *Fonollosa* [10.52]. Their PDA applies a special third-order cumulant,

$$C(0, d) := \sum_{n} s(n)\, s(n)\, s(n+d) \ , \tag{10.20}$$

which tends to vanish for noises with symmetrical distribution, such as Gaussian noise. It also tends to vanish for voiceless fricatives, as *Wells* [10.53] discovered for his VDA. If the signal is periodic, the cumulant C is also periodic, but one cannot expect a maximum to occur at $d = 0$. This PDA thus treats $C(0, d)$ like an ordinary signal, takes the autocorrelation function, and derives T_0 therefrom in the same way as it is done in an ordinary autocorrelation PDA. The algorithm was tested with speech in various additive noises at various signal-to-noise ratios (SNRs) against an autocorrelation PDA (without any pre- or postprocessing) and found superior, especially when noise levels were high.

The PDA by *Tabrikian* et al. [10.54] determines the parameters of a harmonic-plus-noise model by maximizing a log-likelihood measure, i. e., the unknown fundamental frequency, the spectral amplitudes of the

harmonics, and the variance of the noise, which is modeled to be Gaussian. It then performs pitch tracking over consecutive frames using a method based on the maximum a posteriori probability. The authors tested their PDA under extreme noise conditions and found that it worked even at SNRs of −5 dB and worse. A similar algorithm was developed by *Godsill* and *Davy* [10.55] for music signals.

10.2.6 Concluding Remarks

Short-term analysis PDAs provide a sequence of average pitch estimates rather than a measurement of individual periods. They are not very sensitive to phase distortions or to the absence of the first partial. They collect information about pitch from many features and (mostly) from several periods of the signal. They are thus robust against additive noise. Some of them still work at SNRs of 0 dB or worse. On the other hand, they are sensitive when the signal does not fulfil their basic requirement, i. e., periodicity. Rapid within-frame changes of F_0 of irregularly excited signals (e.g., laryngealizations) lead to failure.

One advantage of this principle that is not always explicitly mentioned is the ability to give rather accurate estimates and to overcome measurement granularity due to signal sampling. To decrease computational complexity, many of these PDAs perform some moderate low-pass filtering and/or decrease the sampling frequency in the first step and thus increase the granularity. Once a crude estimate is available, it can be refined via a local interpolation routine, which is frequently implemented. This is most evident in active modeling (Sect. 10.2.4) where the impulse response of the model filter can be generated with an arbitrarily high sampling frequency independently from the sampling frequency of the signal. However, any other representation from which pitch is derived – ACF, AMDF, cepstrum, etc. – can be treated like a signal and can be locally upsampled, e.g., via a standard multirate finite impulse response (FIR) filter, to increase measurement accuracy. The evaluation by *McGonegal* et al. [10.56] showed that an increased accuracy is honored by the human ear when listening to synthetic speech generated with such a pitch contour.

10.3 Selected Time-Domain Methods

This category of PDAs is less homogenous than that of the short-term analysis methods. One possibility to group them is according to how the data reduction is distributed between the preprocessor and the basic extractor, and we find most of these PDAs between two extremes.

- Data reduction is done in the preprocessor. In the extreme case, only the waveform of the first harmonic is offered to the basic extractor. The basic extractor processes this harmonic and derives pitch from it.
- Data reduction is done in the basic extractor, which then has to cope with the whole complexity of the temporal signal structure. In the extreme case, the preprocessor is totally omitted. The basic extractor investigates the temporal structure of the signal, extracts some key features, and derives the information on pitch therefrom.

A third principle is situated somewhere in the middle of these extremes. Temporal structure simplification performs a moderate data reduction in the preprocessor but preserves the harmonic structure of the signal.

Time-domain PDAs are principally able to track the signal period by period. At the output of the basic extractor we usually obtain a sequence of period boundaries (pitch *markers*). Since the local information on pitch is taken from each period individually, time-domain PDAs are sensitive to local signal degradations and are thus less reliable than most of their short-term analysis counterparts. On the other hand, time-domain PDAs may still operate correctly even when the signal itself is aperiodic (but still cyclic), in speech for instance due to temporary voice perturbation or laryngealization.

Most time-domain PDAs, especially those which follow definitions 2 and 3, were developed before the 1990s. With the introduction of time-domain pitch modification methods [10.57], research in this area concentrated on high-precision algorithms for determination of the instant of glottal closure. This issue will be discussed in Sect. 10.7.1.

10.3.1 Temporal Structure Investigation

A pitch period in speech is the truncated response of the vocal tract to an individual glottal impulse. Since the vocal tract behaves like a lossy linear system, its impulse response consists of a sum of exponentially damped oscillations. It is therefore to be expected that

the magnitude of the significant peaks in the signal is greater at the beginning of the period than versus the end. Appropriate investigation of the signal peaks (maxima and/or minima) leads to an indication of periodicity.

There are some problems with this approach, however. First, the frequencies of the dominant damped waveforms are determined by the local formant pattern and may change abruptly. Second, damping of the formants, particularly of a low first formant, is often quite weak and can be hidden by temporary changes of the signal level due to articulation. Third, if the signal is phase distorted, different formants may be excited at different points in time. These problems are solvable but lead to complex algorithmic solutions investigating a great variety of temporal structures.

The usual way to carry out the analysis is the following [10.58].

- Do a moderate low-pass filtering to remove the influence of higher formants.
- Determine all the local maxima and minima in the signal.
- Exclude those extremes that are found to be insignificant until one significant point per period is left; this point will become the local pitch marker.
- Reject obviously wrong markers by local correction.

Many individual (and heuristic) solutions have been developed, but they cannot be reviewed here for lack of space. For more details, the reader is referred to the literature [10.1].

10.3.2 Fundamental Harmonic Processing

F_0 can be detected in the signal via the waveform of the fundamental harmonic. If present in the signal, this harmonic is extracted from the signal by extensive low-pass filtering in the preprocessor. The basic extractor can then be relatively simple. Figure 10.8 shows the principle of three basic extractors: zero-crossings analysis as the simplest one, nonzero threshold analysis, and finally threshold analysis with hysteresis. The zero-crossings analysis basic extractor sets a marker whenever the zero axis is crossed with a defined polarity. This requires that the input waveform has two and only two zero crossings per period. The threshold analysis basic extractor sets a marker whenever a given nonzero threshold is exceeded. When operating with hysteresis, the marker is only set when a second (lower) threshold is crossed in the opposite direction. This more-elaborate device requires less low-pass filtering in the preprocessor.

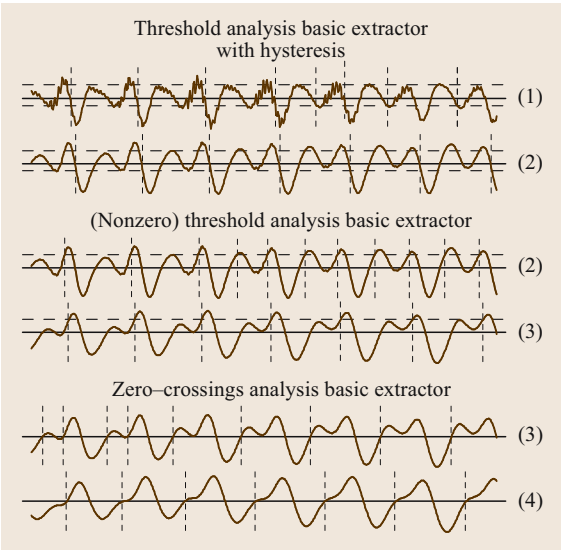

Fig. 10.8 Example of the performance of basic extractors for fundamental harmonic extraction in speech. Signals: (1) original (vowel [i], 32 ms), (2) low-pass filtered at 6 dB/oct, (3) low-pass filtered at 12 dB/oct, and (4) low-pass filtered at 18 dB/oct. The signal is such that success and failure are displayed at the same time

The requirement for extensive low-pass filtering is a severe weak point of this otherwise fast and simple principle when applied to speech signals. In a number of applications, however, such as voice quality measurement or the preparation of reference elements for time-domain speech synthesis, where the signals are expected to be clean, the use of a PDA applying first-partial processing may be advantageous. *Dologlou* and *Carayannis* [10.59] proposed a PDA that overcomes a great deal of the problems associated with the low-pass filter. An adaptive linear-phase low-pass filter that consists of a variable-length cascade of second-order filters with a double zero in the z plane at $z = -1$ is applied. These filters are consecutively applied to the input signal; after each iteration the algorithm tests whether the higher harmonics are sufficiently attenuated; if they are, the filter stops. T_0 is then derived from the remaining first partial by a simple maximum detector. Very low-frequency noise is tolerable since it barely influences the positions of the maxima.

10.3.3 Temporal Structure Simplification

Algorithms of this type take some intermediate position between the principles of structural analysis and

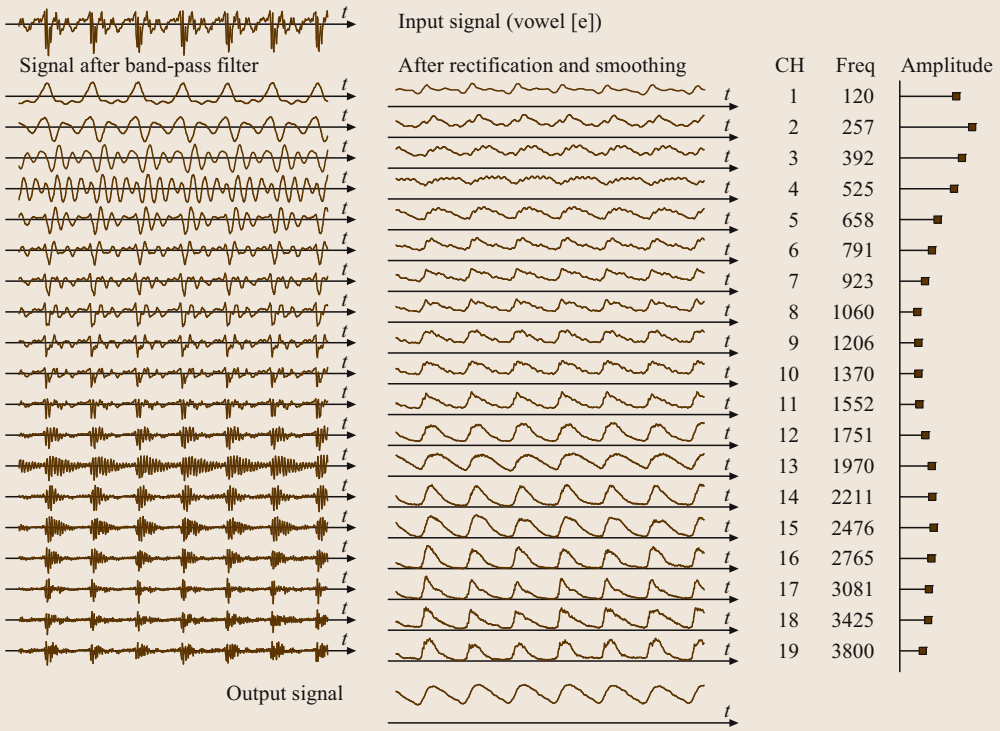

Fig. 10.9 Example PDA by *Yaggi* (1962 [10.31]). The signal is split into 19 subbands. In each channel (CH) the filtered signal is rectified and smoothed; the weighted outputs of the channels are added, and pitch markers are derived from the resulting signal via maximum detection

fundamental harmonic extraction. The majority of these algorithms follow one of two principles: (1) inverse filtering, and (2) epoch detection. Both of these principles deal with the fact that the laryngeal excitation function has a temporal structure that is much simpler and more regular than the temporal structure of the speech signal itself, and both methods when they work properly are able to follow definition 1 if the signal is not grossly phase distorted.

The inverse filter approach cancels the transfer function of the vocal tract and thus reconstructs the laryngeal excitation function. If one is interested in pitch only and not in the excitation function itself, a crude approximation of the inverse filter is sufficient. For instance, we can confine the analysis to the first formant [10.60].

The second principle, epoch extraction [10.61], is based on the fact that at the beginning of each laryngeal pulse there is a discontinuity in the second derivative of the excitation function. Usually this discontinuity cannot be reliably detected in the speech signal because of phase distortions that occur when the waveform passes the vocal tract. The signal is thus first phase shifted by 90° (by applying a Hilbert transform). The squares of the original and the phase-shifted signals are then added to yield a new signal that shows a distinct peak at the time when the discontinuity in the excitation function occurs. In principle this yields the instantaneous amplitude of the complex analytic signal constructed from the original signal as its real part and the phase-shifted signal as its imaginary part.

The original method [10.61] works best when the spectrum of the investigated signal is flat to some extent. To enforce spectral flatness, the analyzed signal can be band limited to high frequencies well above the narrowband lower formants. Another way is to analyze the LP residual or to filter the signal into subbands.

One prototype of these algorithms, which never became widely known, was developed by *Yaggi* [10.31]. It splits the signal into 19 subbands and subsequently rectifies and smoothes the signal in each channel so that the envelope is extracted. The individual channels are then added, and the individual periods are derived from

the resulting signal (Fig. 10.9). Another prototype is the PDA by *Dolanský* [10.62] that models the envelope of the pitch period by a decaying exponential function (in analog technology) and sets a marker whenever the signal exceeds the modeled envelope, resetting the envelope at the same time.

Both the inverse filter approach and the epoch detection principle have one weak point, which frequently arises with female voices. When F_0 is high, it may coincide with the first formant $F1$. In this case the signal becomes nearly sinusoidal since we have something like a double pole *glottal formant* and $F1$ at the same frequency) in the overall transfer function. If an inverse filter is not blocked in this case, it removes the fundamental harmonic from the signal and brings the PDA to failure. For epoch detection, we know that the envelope of a sinusoid is a constant ($\cos^2 x + \sin^2 x = 1$) and does not show any discontinuity. Hence these algorithms need a work-around for low values of $F1$.

This drawback was overcome by the finding that the global statistical properties of the waveform change with glottal opening and closing as well. We will come back to this issue in Sect. 10.7.1.

Structural analysis of the signal itself or of some simplified representation, especially when many possible structures have to be reviewed, is a good candidate for self-organizing, nonlinear pattern-recognition methods, i. e., for artificial neural networks. Such a PDA for speech was introduced by *Howard* and *Walliker* [10.63]. The speech signal is divided into nine subbands with a subsequent half-wave rectification and second-order linear smoothing in each channel. The underlying idea is to obtain a representation similar to that in a wide-band spectrogram. The basic extractor consists of a four-layer perceptron structure, the input layer consisting of 41 successive samples in each subband. Two hidden layers with 10 units each and a fully connected network are followed by a one-unit output layer, which is intended to yield an impulse when the network encounters a signal structure associated to the instant of glottal closure. The network is trained using (differentiated) output signals of a laryngograph as reference data. Such a structure has the advantage that it can be based upon several features occurring at different instants during a pitch period.

10.3.4 Cascaded Solutions

Among the many possibilities of such solutions, one is of particular interest here: the cascade of a robust short-term PDA and an accurate but less-robust time-domain PDA. Such an algorithm is further described in Sect. 10.7.1.

10.4 A Short Look into Voicing Determination

The task of voicing determination of speech signals may be split up into two subtasks: (1) a decision of whether or not a voiced excitation is present and (2) a decision of whether or not a voiceless excitation is present. If neither of these excitations is active, the current segment represents pause or silence; if both excitations are simultaneously active, we speak of mixed excitation. The two features *voiced* and *voiceless* are binary unless they occur simultaneously. In segments with mixed excitation the degree of voicing – for instance, the energy ratio of the voiced and voiceless excitations – may play a role, although this feature is rarely exploited.

Most voicing determination algorithms (VDAs) thus apply decisions. VDAs exploit almost any elementary speech signal parameter that may be computed independently of the type of input signal: energy, amplitude, short-term autocorrelation coefficients, zero-crossings count, ratio of signal amplitudes in different subbands or after different types of filtering, linear prediction error, or the salience of a pitch estimate. According to the method applied, VDAs can be grouped into three major categories: (1) simple threshold analysis algorithms, which exploit only a few basic parameters [10.64]; (2) more-complex algorithms based on pattern recognition methods; and (3) integrated algorithms for both voicing and pitch determination.

In this section, we distinguish between *voiceless* and *unvoiced*. Unvoiced means that a frame can be anything but voiced, i. e., it can be voiceless or silence. Voiceless means that voiceless excitation is present so that the frame is neither voiced nor silence. We will not review such algorithms that distinguish between speech and nonspeech – many of such algorithms have been developed for other applications, such as voice over Internet protocol (IP) or bandwidth reduction in mobile telephony (see, for instance, *Davis* et al. [10.65], for a survey). The basic task of the VDA in the context of pitch determination is to decide whether a frame or signal segment is voiced (and thus subject to pitch determination) or unvoiced.

10.4.1 Simultaneous Pitch and Voicing Determination

A number of PDAs – usually pertaining to the short-term analysis category – permit estimation the salience of their results without having to know the actual value of the pitch estimate. This is always possible when the amplitude of the significant maximum or minimum at T_0 or F_0 in the basic extractor of the PDA can be compared to a reference value. As an example, the ratio of the values of the autocorrelation function at $d = T_0$ and at $d = 0$ (the latter equalling the signal energy) gives a direct measure of the degree of periodicity of the signal. It is dangerous, however, to rely on this feature alone. Of course, it is correct (and almost trivial) to state that pitch can exist only when the signal is voiced. However, this statement cannot be simply reversed; i. e., we cannot say that a segment is unvoiced because pitch is not existent (or not measurable). The corresponding PDA may momentarily fail, or the signal may be voiced but irregular ([10.47, 66]; see also Fig. 10.2 or Sect. 10.1.2). Amplitude changes and especially small intraframe changes of F_0 severely degrade the *salience* of the pitch estimate [10.36]. It is thus at least necessary to check adjacent frames before making a decision [10.10]. In this respect, such a VDA behaves very much like a median smoother in pitch determination.

In principle, these VDAs do not make a voiced–unvoiced discrimination; rather they check for the presence of a (sufficient but not absolutely necessary) condition for a voiced signal. An improvement is to be expected when such criteria are only used for declaring a frame as voiced, and when the decision to declare it as unvoiced is based on additional criteria [10.67].

The VDA by *Lobanov* [10.68] avoids this problem, although it is based on a similar principle. A voiceless segment of speech represents a stochastic signal which is continuously excited. In contrast, the excitation of a voiced signal is confined to a few instants per period; major parts of the pitch period are composed of exponentially decaying oscillations, and adjacent samples of the signal are highly correlated. This contrast of a highly stochastic versus a highly deterministic signal is preserved even when a voiced signal becomes irregular or aperiodic. Lobanov's VDA exploits this feature by computing the Hilbert transform of the speech signal, combining the original signal and its Hilbert transform to yield the complex analytic signal, and plotting the momentary amplitude and phase of the analytic signal in the so-called phase plane. For voiced frames the analytic signal will describe a closed curve. During unvoiced segments, where the signal and its Hilbert transform are much less correlated, the curve will touch almost any point in the phase plane within a short interval. In Lobanov's algorithm the phase plane is crudely quantized, and the algorithm simply counts the number of points which have been touched within a given frame.

Talkin's PDA [10.17] integrates the VDA into the postprocessor that applies dynamic programming. Among the various estimates for T_0 to be tracked, there is always a candidate *unvoiced*, which is selected when it lies on the optimal path (Sect. 10.5.4).

Ahmadi and *Spanias* [10.67] present an improved VDA module within an implementation of the cepstrum PDA [10.10] for telephone-bandwidth speech. An utterance is processed in two passes. The first pass, covering the whole utterance, is to derive gross initial thresholds for a rough voiced–unvoiced decision. Distributions are taken for the relative amplitude of the main cepstral peak, the relative zero-crossings rate, and normalized signal energy. The medians of these distributions serve as initial thresholds for the decisions to be made in the second pass. A frame is roughly declared unvoiced if its energy and cepstral peak amplitudes are below and its zero crossings rate is above the respective threshold. Frames are declared voiced according to their cepstral peak amplitudes and a continuity criterion. The algorithm was evaluated on data from the TIMIT corpus; reference values were obtained using the PDA by *McAulay* and *Quatieri* [10.51] with visual inspection of uncertain frames. For clean speech, voiced-to-unvoiced and unvoiced-to-voiced errors together were about 1.5%.

McAuley and *Quatieri* [10.51] use their harmonic-model PDA (Sect. 10.2.5) to incorporate a VDA. It is based on the energy ratio between the harmonic energy and the energy of the nonharmonic part of the signal (the *noise*) which consists of everything not captured by the harmonic structure. Frames for which this ratio is above 10 dB are certainly voiced, while those for which the ratio is below 4 dB are certainly unvoiced.

Fisher et al. [10.69] start from a generalized log likelihood measure that is separately and independently evaluated for the two hypotheses that the frame is (1) voiced, or that it is (2) unvoiced. The measure for the frame being voiced is based on the aforementioned ratio between harmonic and nonharmonic energy, whereas the measure for unvoiced is based on a model of colored Gaussian noise. The hypothesis with the higher likelihood value wins for each frame; a dynamic-programming postprocessor (Sect. 10.5.4) integrates the VDA into the PDA which is also based on the harmonic-plus-noise model.

10.4.2 Pattern-Recognition VDAs

One of the motivations for applying a pattern-recognition algorithm in a VDA was the wish to get away from the conjunction of voicing determination and pitch determination [10.70]. The VDA by *Atal* and *Rabiner* [10.70] (the first of a series of VDAs developed at Bell Laboratories in the late 1970s) uses a statistical classifier and is based on five parameters: the signal energy, the zero-crossing rate, the autocorrelation coefficient at a delay of one sample, the first predictor coefficient of a 12 pole LP analysis, and the energy of the normalized prediction error. For a given environmental condition the algorithm works well, but it is rather sensitive to environmental changes, e.g., from high-quality speech to a telephone channel [10.47].

The usual classification problems in speech recognition, where we have to cope with a large number of different classes, require that the input parameters form specific clusters in the parameter space, which are then separated by the classifier. In contrast, the voicing determination problem has at most four categories (silence, voiced, voiceless, and mixed) and the distribution of the patterns in the parameter space is rather diffuse. It is thus appropriate to concentrate the VDA on patterns that are situated at or near the boundaries between the different categories in the parameter space. Such a VDA was developed by *Siegel* and *Bessey* [10.66]. For some applications, such as high-quality analysis–synthesis systems, incorporation of a mixed excitation source is desirable: (1) for voiced fricatives, and (2) for some high vowels, which tend to become partly devoiced in connected speech [10.71]. Siegel and Bessey further found that for the voiced–voiceless–mixed classification, the number of features used for a voiced-unvoiced classifier is insufficient. Their VDA is realized in two steps using a binary decision tree structure. The first step is a classifier which separates frames that are predominantly voiced from those that are predominantly unvoiced. It uses a minimum-distance statistical classifier exploiting seven features: normalized autocorrelation coefficient at unit sample delay, minimum normalized LP error, zero-crossings rate, signal energy, overall degree of periodicity (via AMDF), and degree of periodicity measured via the cepstrum in two subbands. In both categories the mixed frames are split off during the second step. The voiced–mixed decision uses another six features, mostly cepstral and LP measures, whereas the voiceless–mixed decision is based on two features alone. The VDA is reported to work with 94% overall accuracy and 77% correct identification of the mixed frames.

10.5 Evaluation and Postprocessing

To evaluate the performance of a measuring device, one should have another instrument with at least the same accuracy. If this is not available, at least objective criteria – or data – are required to check and adjust the behavior of the new device. In pitch and voicing determination both these bases of comparison are tedious to generate. There is no VDA or PDA that operates without errors [10.47]. There is no reference algorithm, even with instrumental support, that operates completely without manual inspection or control [10.8, 72]. Yet nowadays speech databases with reference pitch contours and voicing information have become widely available so that at least reliable reference data are there and are being used for evaluation.

In this section, we first deal with the question of how to generate reference data (Sect. 10.5.1). Then we consider the question of error analysis (Sect. 10.5.2) and present the results of some comparative evaluations (Sect. 10.5.3). Finally, we describe the problem of pitch tracking (Sect. 10.5.4), which is the foremost task of the postprocessor.

10.5.1 Developing Reference PDAs with Instrumental Help

A number of evaluations compared the algorithm(s) to be tested to the results of a well-known algorithm such as the cepstrum PDA, whose performance was known to be good. *Rabiner* et al. [10.47] used an interactive PDA to generate reference data. This procedure proved reliable and accurate but required a great deal of human work. Other evaluations [10.1] used the output signal of a vocoder for which the pitch contour was exactly known or the output signal of a mechanic accelerometer which derives the information on pitch from the vibrations of the neck tissue at the larynx. The latter device [10.73] was used by *Viswanathan* and *Russell* [10.74] for their evaluation of five PDAs. *Indefrey* et al. [10.34] used a laryngograph to obtain the signal for generating a reference contour.

Among the algorithms used for determining a reference pitch contour, methods that make use of an instrument (such as a mechanic accelerometer or a laryn-

gograph) that derives pitch directly from the laryngeal waveform have been shown to be most effective. This type of algorithm avoids most errors pertinent to the problem of pitch determination from the speech signal, and permits using natural speech for the evaluation of the performance of PDAs. Among the many instruments available, the laryngograph [10.72,75] is especially well suited for this kind of application. It is robust and reliable, does not prevent the speaker from natural articulation, and gives a good estimate for the instant of glottal closure. A number of PDAs have been designed for this device [10.8,72]. In addition, *Childers* et al. [10.76] propose a four-category VDA that exploits the speech signal and the laryngogram.

The principle of the laryngograph [10.75] is well known. A small high-frequency electric current is passed through the larynx by a pair of electrodes that are pressed against the neck at the position of the larynx from both sides. The opening and closing of the glottis during each pitch period cause the laryngeal conductance to vary; thus the high-frequency current is amplitude modulated. In the receiver the current is demodulated and amplified. Finally, the resulting signal is high-pass filtered to remove unwanted low-frequency components due to vertical movement of the larynx.

Figure 10.10 shows an example of the laryngogram (the output signal of the laryngograph) together with the corresponding speech signal. In contrast to the speech signal, the laryngogram is barely affected by the instantaneous configuration of the vocal tract, and the changes in shape or amplitude are comparatively small. Since every glottal cycle is represented by a single pulse, the use of the laryngograph reliably suppresses gross period-determination errors. In addition, it supplies the basis for a good voiced–unvoiced discrimination since the laryngogram is almost zero during unvoiced segments, when the glottis is always open. Nonetheless, the laryngograph is not free from problems: it may fail temporarily or permanently for some individual speakers, or it may miss the beginning or end of a voiced segment by a short interval – for instance, when the vocal folds, during the silent phase of a plosive, continue to oscillate without producing a signal, or when voicing is resumed after a plosive and the glottis does not completely close during the first periods [10.72]. For such reasons, visual inspection of the reference contour is necessary even with this configuration; these checks, however, can be confined to limited segments of the signal.

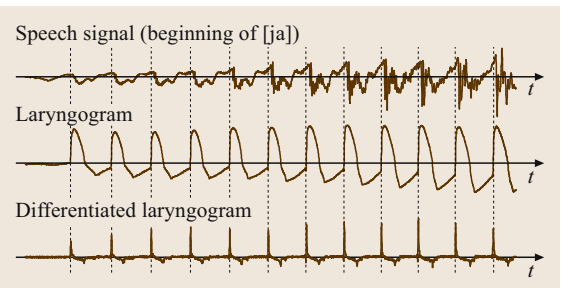

Fig. 10.10 Speech signal, laryngogram, differentiated laryngogram, and instants of glottal closure

The instant of glottal closure is the point of maximum vocal-tract excitation, and it is justifiable to define this instant as the beginning of a pitch period. In the laryngogram this feature is well documented. As long as the glottis is open, the conductance of the larynx is at a minimum and the laryngogram is low and almost flat. When the glottis closes, the laryngeal conductance goes up and the laryngogram shows a steep upward slope. The point of inflection during the steep rise of the laryngogram, i. e., the instant of maximum change of the laryngeal conductance, was found best suited to serve as the reference point for this event.

10.5.2 Error Analysis

According to the classic study by *Rabiner* et al. [10.47], which established the guidelines for the performance evaluation of these algorithms for speech, PDAs and VDAs commit four types of errors:

1. Gross pitch-determination errors
2. Fine pitch-determination errors, i. e., measurement inaccuracies
3. Voiced-to-unvoiced errors
4. Unvoiced-to-voiced errors

The latter two types represent errors of voicing determination, whereas the former two refer to pitch determination.

Gross pitch-determination errors are *drastic failures of a particular method or algorithm to determine pitch* [10.47]. Usually an error is considered to be gross when the deviation between the correct value of T_0 or F_0 and the estimate of the PDA exceeds the maximum rate of change a voice can produce without becoming irregular (*Rabiner* et al. [10.47]: 1 ms; *Hess* and *Indefrey* [10.8], *Mousset* et al. [10.77]: 10%; *Krubsack* and *Niederjohn* [10.78]: 0.25 octave). Typical gross errors are confusions between F_0 and the formant $F1$, which usually falls into the measuring range. Other typical errors are the so-called octave errors, i. e., taking $F_0/2$ or $2F_0$ as the pitch estimate.

On the other hand, error situations may also arise from *drastic failures of the voice to produce a regular excitation pattern*, which is not very frequent in well-behaved speech but is nearly always the case when the voice temporarily falls into creak or vocal fry. *Hedelin* and *Huber* [10.18] distinguish between four main types of irregularity of phonation that occur frequently in running speech: (1) laryngealization (a temporal near-closure of the glottis resulting in single, irregular glottal pulses); (2) creak or vocal fry as a temporal voice register (Fig. 10.2); (3) creaky voice, which is even less structured than creak; and (4) diplophonic excitation, which shows an irregular pattern between adjacent periods but a more-regular one between every second pitch period. A further problem, which may sometimes become severe, is the rate of change of fundamental frequency. *Xu* and *Sun* [10.79], also referring to earlier studies, give data for the maximum rate of change of F_0 that a human voice is able to produce without becoming irregular. They found that a human voice can change its F_0 at a speed up to 100 semitones per second, and that this limit is frequently reached during running speech. One hundred semitones per second means one semitone (6%) per 10 ms or two semitones (12%) per 20 ms. According to *Sreenivas* [10.36], a 4% within-frame F_0 change already affects the salience of the estimate in the same way as additive noise with 15 dB SNR. As we see from these data, a 10% change of F_0 within a frame can easily occur. If we interpret these data with respect to individual pitch periods, we see that deep male voices with long periods (10 ms and more) are more strongly affected than female voices. Nonetheless, a deviation of 10% for F_0 estimates between adjacent frames seems reasonable as a lower bound for gross errors because a larger change is beyond the capabilities of a human voice.

Hence, gross errors arise mainly from three standard situations.

- Adverse signal conditions: strong first formants, rapid change of the vocal-tract position, band-limited or noisy recordings. Good algorithms reduce these errors to a great extent but cannot completely avoid them [10.47].
- Insufficient algorithm performance, e.g., mismatch of F_0 and frame length [10.21]; temporary absence of the key feature in some algorithms.
- Errors that arise from irregular excitation of voiced signals. Since most algorithms perform some averaging or regularity check, they can do nothing but fail when the source becomes irregular.

When a PDA is equipped with an error-detecting routine (the majority of cases, even if no postprocessor is used), and when it detects that an individual estimate may be wrong, it is usually not able to decide reliably whether this situation is a true measurement error, which should be corrected or at least indicated, or a signal irregularity, where the estimate may be correct and should be preserved as it is. This inability of most PDAs to distinguish between the different sources of error situations is one of the great unsolved problems in pitch determination.

In the study by *Rabiner* et al. [10.47] gross errors are simply counted, and the percentage of frames with gross errors compared to the total number of (correctly recognized) voiced frames is given as the gross error rate. However, the perceptual importance of gross errors depends on the deviation between the estimate and the correct value as well as the energy of the frame [10.74, 80], from which a weighted gross error measure was derived [10.67],

$$\mathrm{GPE} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{E(k)}{E_{\max}}\right)^2 \left|\frac{p(k)-F_0(k)}{F_0(k)}\right| , \quad (10.21)$$

where $p(k)$ is the incorrect estimate, $E(k)$ is the energy of the frame, and $E_{\max}$ is the maximum energy in the utterance. It appears useful to include both these measures in an evaluation. The gross error count evaluates the PDA performance from a signal-processing point of view, whereas GPE says something about their perceptual relevance.

Measurement inaccuracies cause noisiness of the obtained T_0 or F_0 contour. They are small deviations from the correct value but can nevertheless be annoying to the listener. Again there are three main causes.

- Inaccurate determination of the key feature. This applies especially to algorithms that exploit the temporal structure of the signal, for instance, when the key feature is a principal maximum whose position within a pitch period depends on the formant $F1$.
- Intrinsic measurement inaccuracies, such as the ones introduced by sampling in digital systems.
- Errors from small fluctuations (*jitter*) of the voice, which contribute to the perception of *naturalness* and should thus be preserved.

Voicing errors are misclassifications of the VDA. We have to distinguish between voiced-to-unvoiced errors, in which a frame is classified as unvoiced although it is in fact voiced, and unvoiced-to-voiced errors, with the opposite misclassification. This scheme, as established

in [10.47], does not take into account mixed excitation. Voiced-to-unvoiced and unvoiced-to-voiced errors must be regarded separately because they are perceptually inequivalent [10.74], and the reasons for such errors in an actual implementation may be different and even contradictory. A number of VDAs even allow a tradeoff between these two errors by adjusting their parameters.

10.5.3 Evaluation of PDAs and VDAs– Some Results

Due to the absence of reliable criteria and systematic guidelines, few publications on early PDAs included a quantitative evaluation. As this situation has thoroughly changed, publications presenting new PDAs increasingly also evaluate them. Mostly the newly developed PDA is evaluated against some well-known PDA(s) to show that the new approach is in some way or for some kind of signals and conditions better or at least equivalent to the known algorithms [10.46,52]. The *Keele* database [10.81] has played a major role in this respect. We will not discuss these evaluations here due to lack of space; we rather deal with a couple of studies that did not aim at developing a new PDA but were done to establish guidelines and show the state of the art.

The classic studies by *Rabiner* et al. [10.47] and *McGonegal* et al. [10.56] investigated seven PDAs (two time domain, five short-term analysis) with respect to pitch and voicing determination. The main results were:

- None of the PDAs investigated were error free, even under good recording conditions. Each PDA had its own favorite error; nevertheless, all error conditions occurred for all of the PDAs.
- Almost any gross error was perceptible; in addition, unnatural noisiness of a pitch contour was well perceived.
- The subjective evaluation did not match the preference of the objective evaluation. In fact, none of the objective criteria (number of gross errors, noisiness of the pitch contour, or voicing errors) correlated well with the subjective scale of preference.

Hence the question of which errors in pitch and voicing determination are really annoying for the human ear remained open. This issue was further pursued by *Viswanathan* and *Russell* [10.74], who developed objective evaluation methods that are more closely correlated to the subjective judgments. The individual error categories are weighted according to the consistency of the error, i. e., the number of consecutive erroneous frames, the momentary signal energy, the magnitude of the error, and the special context.

Indefrey et al. [10.34], concentrating on the evaluation of pitch determination errors only, investigated several short-term PDAs in various configurations. They showed that in many situations different short-term analysis PDAs behave in a complementary way so that combining them in a multichannel PDA could lead to better overall performance.

Indefrey et al. [10.34] also investigated the performance of double-transform PDAs (cf. Sect. 10.2.2) with additive Gaussian noise. Under this condition these algorithms tend to break down at SNRs between 0 and −6 dB. It does not make a big difference whether the SNR is defined globally (i. e., with a constant noise level over a whole utterance) or segmentally (i. e., with the same SNR for each frame), except that the slope of the error curve at the breakdown point is larger for segmental SNR. These results were confirmed in a number of other studies [10.26, 46, 52]. *Moreno* and *Fonollosa* [10.52] evaluated several autocorrelation PDAs (among them their own) with several kinds of noise signals and found that for the low-frequency-biased car noise the breakdown starts at an SNR of about 6 dB. The same holds for babble noise [10.46].

De Cheveigné and *Kawahara* [10.27] investigated eight PDAs whose software was available via the Internet together with two of their own developments ([10.27], cf. Sect. 10.2.1; [10.82], based on an IF principle). Only gross errors were considered. The evaluation was based upon an extensive database (almost two hours of speech) with samples from three languages (English, French, and Japanese), including falsetto speech from male speakers, and laryngograms as reference signals. Obviously aperiodic voiced signals were excluded. Postprocessors and VDAs were disabled in the algorithms as far as possible. The evaluation showed great differences between algorithms and partly rather bad performance (more than 10% gross errors for some of them); the best one produced about 0.6% on average. The evaluation also showed considerable dependency of the error rate on the data so that the authors claim the need for large databases when performing such evaluations.

All these evaluations show that there is still no PDA that works without errors, although they work better now than 20 years ago. A gross error count of 0.6% is regarded as excellent; nonetheless we must not forget that, with the usual frame rate of 100 frames per second, such an algorithm still produces a grossly wrong estimate every two seconds of speech on average.

10.5.4 Postprocessing and Pitch Tracking

A standard procedure for the reduction of pitch determination errors is smoothing. Smoothing is possible when a pitch contour is given as a sequence of T_0 or F_0 estimates and not as delimiters of individual periods. The two standard smoothing methods are linear smoothing using some kind of low-pass filter and (nonlinear) median smoothing [10.83]. Linear smoothing reduces measurement inaccuracies but is unable to cope with the effect of gross pitch determination errors, which are reduced in size and distributed over a larger amount of time but are not really removed. Median smoothing, on the other hand, replaces each pitch estimate with the middle value of an ordered sequence of three, five, or seven adjacent estimates; gross outliers are removed, but measurement inaccuracies remain unchanged. *Rabiner* et al. [10.83] combine these methods and propose a two-step smoothing procedure with median smoothing coming first, followed by a linear smoother. Linear smoothing, however, can be dangerous since it may replace a gross error that has been left in the median-smoothed contour by some estimates lying between the correct value and the error and so cause an inflection in the contour that is not due to the signal.

Applying such a smoothing algorithm was shown to substantially improve the (objective and subjective) performance of any PDA to which it was applied [10.47, 56]. *Specker* [10.84] showed that postprocessing is able to reduce the number of gross errors in a time-domain PDA by almost an order of magnitude.

Secrest and *Doddington* [10.80] used dynamic programming methods to find an optimal path through a list of pitch estimate candidates with the smoothness of the contour as the major criterion. They showed that this technique performed better than any linear, nonlinear, or median smoothing. This approach was further developed by *Talkin* [10.17]. Dynamic programming is well suited to pitch tracking since it allows the basic extractor to give several pitch candidates so that we can deal with more than only the best choice in each frame. Each candidate is accompanied by a salience measure (usually the relative amplitude of the corresponding peak in the representation from where the estimate is derived, with respect to the reference point, e.g., the value of the ACF at zero lag). In addition, Talkin's PDA supplies one candidate *unvoiced* per frame. Pitch tracking is done by searching for an optimal path through the candidates from consecutive frames by minimizing a global cost function. This global cost function is formed as the sum of weighted local per-frame cost functions of two types: (1) candidate costs, and (2) transition costs between consecutive frames.

Candidate costs distinguish between pitch candidates and the unvoiced candidate. The cost of a pitch candidate equals one minus the salience measure of this candidate. The cost of the unvoiced candidate is a constant penalty plus the maximum salience measure within the current frame.

The transition cost between consecutive frames also depends on voicing. Between two unvoiced candidates it is zero. Between two pitch candidates it depends on the difference in frequency between the two estimates, and special attention is given to octave jumps, which are made costly but not totally impossible. The costs for voiced-to-unvoiced transitions and vice versa include a term with the reciprocal *Itakura* distance [10.85], an energy term, and an extra penalty for this transition. The rationale is that (1) these transitions are not too frequent, (2) there are usually large spectral changes between a voiced and an unvoiced frame, and (3) energy usually decreases at the end of a voiced part and increases at its beginning.

There is no latency limit for the algorithm to find the optimal path; in principle the search can extend over a whole utterance. *Talkin* [10.17], however, reports that it rarely takes more than 100 ms for all possible paths to converge to a single point. The algorithm is part of the well-known ESPS software package for speech analysis. A comparable postprocessor operating on a probabilistic approach is described in [10.86].

10.6 Applications in Speech and Music

Applications for PDAs in speech can be grouped into four areas [10.1]:

1. Speech technology
2. Basic linguistic research, e.g., investigation of intonation
3. Education, such as teaching intonation to foreign-language learners or to hard-of-hearing persons

4. Clinical applications, such as the investigation of voice quality

Some of these application areas have changed greatly over the last two decades. The vocoder, which was the main application for earlier speech technology, has almost disappeared. Instead, investigating prosodic events such as intonation, particularly in spontaneous speech, has become an important issue in speech understanding systems [10.87], and many of these systems now contain a prosody module. As it is a long-term goal in speech technology to make such devices operable from almost anywhere, a PDA may even have to cope with signals from mobile phones, which can be extremely bad and inconsistent. Another new application area is data-driven speech synthesis. Algorithms for time-domain pitch modification, such as the well-known *pitch-synchronous overlap add* (PSOLA) algorithm [10.57], require precise pitch period determination to work properly, and with the recent technology of nonuniform unit selection synthesis large speech corpora have to be analyzed, yet usually with excellent-quality signals free from phase distortions.

What are the implications of this application shift for the development of algorithms? PDAs for precise pitch period determination of good-quality speech signals have been known for a long time; nonetheless the main problem is exact synchronization with laryngeal cycles, such as the instant of glottal closure. Such algorithms, which originally come from clinical applications where they were applied to isolated vowels, have been extended to work for running speech as well.

In prosody recognition, intonation research and speech technology now go together to a certain extent. Prosody recognition needs intonation contours, not individual periods, and a certain lag between the running time of the signal and the time of release of an estimate is tolerable, in contrast to a vocoder where the result must be available without delay. On the other hand, prosody recognition must rely on automatic estimates and cope with adverse conditions; above all, this requires robustness.

The number of devices available for computer-aided intonation teaching has been small [10.88]. However, with the increased use of high-quality PDAs for intonation research, this will change. In the clinical area, digital hearing prostheses have created a new application area [10.63, 89]. We cannot discuss these applications here for reasons of space; the reader is referred to [10.88, 89] for surveys.

Pitch determination of musical signals has three main application areas, two of which are closely related:

1. Automatic notation of melodies and tunes, and automatic scoring
2. Information retrieval from audio data
3. Real-time capturing of tone frequencies of musical instruments for musical instrument digital interface (MIDI) applications

Automatic notation of melodies is a long-standing problem. As early as 1979 *Askenfelt* [10.90] reported on a project to automatically transcribe folk melodies that had been collected in a large corpus. In this case most of the melodies were one-voiced so that PDAs developed for speech signals could be used.

This is not always the case for music, however, and we must therefore count as a particular problem in music that more than one note can be played at the same time, and that PDAs for this application may have to cope with multiple pitches and have to determine them all, if possible.

Information retrieval from audio data (audio data mining) also involves pitch determination of musical signals. This application area is closely related to the preceding one and also involves the problem of multiple pitches. An evolving application is the so-called query by humming. Consider a person who listens to an unknown tune in the radio, likes it, calls a call center with a musical database, hums or sings the melody, and wants the title of the song retrieved. This is a difficult task with pitch determination being a part thereof.

Transcription of a musical melody into musical notes is much more than mere pitch determination. One has to recover the rhythm from the timing of the melody, and one has to take care of the timing, i. e., when one note ends and another one starts. These questions, which are partly still open, go beyond the scope of pitch determination and will not be further discussed here.

Real-time capturing for MIDI is closely related to the old vocoder application in speech. A PDA is always desirable if a MIDI synthesizer is to be driven from an instrument other than a keyboard, e.g., from an electric guitar. Here the main problem is that the response must be almost instantaneous; a delay of even 50 ms could be detrimental for a live performance.

10.7 Some New Challenges and Developments

This section will concentrate on three problems:

1. Detecting the instant of glottal closure for applications in speech synthesis
2. Multiple pitch detection, particularly in music [10.93]
3. Instantaneousness versus reliability

10.7.1 Detecting the Instant of Glottal Closure

If a PDA in speech is required to be very accurate, it is optimal to synchronize it with the instants of glottal closure. A cascaded PDA for this purpose was developed by *Hess* [10.92] based on a previous algorithm by *Cheng* and *O'Shaughnessy* [10.91]. It is intended for work in time-domain speech synthesis and determines the glottal closure instant (GCI) from undistorted signals. The first part of the cascade is a short-term PDA applying a double-spectral-transform principle [10.34]. The second part uses the estimate of the first PDA to restrict its momentary F_0 range to an octave around this value.

According to *Mousset* et al. [10.77] a GCI detector consists of four steps:

1. Acoustic speech signal pre-emphasis (optional)
2. A transformation aiming to produce peaks at GCIs
3. Postprocessing aiming to increase contrast in the resulting signal (optional)
4. A peak picking operation

The algorithm described here follows this scheme.

In the source-filter approach the speech signal is the response of the supraglottal system to the pulse train generated by the source, and a pitch period can be regarded as the beginning of the impulse response of the vocal tract. If we now compute the correlation function $c(n)$ between the speech signal $s(n)$ and the beginning of the pertinent impulse response $h(n)$ of the vocal tract,

$$c(n) = \sum_k s(n+k)h(k) \qquad (10.22)$$

there will be maxima at those instants n that correspond to an GCI. The impulse response $h(n)$ is estimated via linear prediction. The main peaks of $c(n)$ synchronize well with the individual GCIs. Strong formants, however, also show up in $c(n)$, and further processing is required. Cheng and O'Shaughnessy suggested that one should calculate the envelope of $c(n)$, send it through a high-pass filter and a half-way rectifier to enhance the leading amplitudes at the beginning of each pitch period, and multiply $c(n)$ by the resulting waveform $d(n)$. The envelope

$$e(n) = \sqrt{[c(n)]^2 + [c_H(n)]^2}\,, \qquad (10.23)$$

where $c_H(n)$ is the Hilbert transform of $c(n)$, is calculated using a digital Hilbert filter. Figure 10.11 shows an example.

This approach has the problem known from all PDAs that apply some sort of inverse filtering (cf. Sect. 10.3.3) that it fails systematically when the formant $F1$ coincides with F_0 and the signal becomes almost sinusoidal. Then $e(n)$ becomes a constant, and $d(n)$ is either zero or fluctuates at random around zero instead of displaying the envelope of a damped formant oscillation. A way out of this problem is to partly bridge the high-pass filter so that the constant component of the envelope is not completely removed. If the short-term analysis enters a reliable estimate of T_0, rather stringent correction routines can remove the unwanted peaks in $c(n)$ and at the same time preserve correct processing. The exact position of the GCIs is derived from $c(n)$ anyway.

Group delay methods have also been proposed for GCI determination [10.94, 95]. Group delay is defined as the derivative of the phase spectrum with respect to

Fig. 10.11a–e Determining the instant of glottal closure via a maximum-likelihood criterion [10.91, 92]: example of performance. **(a)** Signal frame (45 ms); **(b)** impulse response $h(n)$ of the vocal tract, estimated by linear prediction; **(c)** correlation function $c(n)$; **(d)** envelope $e(n)$ (dotted line) and high-pass filtered envelope; **(e)** product of $c(n)$ and the filtered envelope

frequency and can be calculated [10.94] via the DFT:

$$\tau_{\mathrm{G}}(m) = \mathrm{Re}\left(\frac{\tilde{S}(m)}{S(m)}\right) ; \qquad (10.24)$$

$$S(m) = \mathrm{DFT}\{s(n)\} ; \quad \tilde{S}(m) = \mathrm{DFT}\{ns(n)\}$$

If a frame contains a single impulse at position $k = n_0$, this will produce a constant group delay of n_0 for all frequency indexes m. If there is additional noise in the frame, one can expect that at least the average group delay (averaged over all frequencies) equals n_0. If we now compute the average group delay for each sample of the time signal $s(n)$, it will show a negative-going zero crossing (with a slope of -1) when the impulse is at the starting position of the frame. GCIs mark nonstationarities in the signal and show up as impulse-like structures, for instance, in an LP residual. This makes the method useful for GCI detection, and there are algorithmic shortcuts that allow the average group delay to be computed without needing two DFTs per sample. *Brookes* et al. [10.95] investigate several weighted measures for the average group delay and find that these measures are quite robust against additive noise; however, they critically depend on the length of the time-domain window used for group delay measurement. If there is no impulse in the frame, the average group delay will vary around zero; if there is more than one impulse, the results are unpredictable. In a cascaded implementation using an auxiliary short-term PDA, when an estimate for T_0 is available, it is straightforward to adjust the window length according to this estimate.

Another method, rather simple to implement, is based on singular value decomposition. *Ma* et al. [10.96] show that the Frobenius norm $||\mathbf{S}||_{\mathrm{F}}$ of an $N \times (K+1)$ matrix $\mathbf{S}$, being

$$||\mathbf{S}||_{\mathrm{F}} = \sqrt{\sum_{n=1}^{N}\sum_{k=1}^{K+1} s_{nk}^2} \quad \text{with} \qquad (10.25)$$

$$\mathbf{S} = (s_{nk};\ 1 \le n \le N;\ 1 \le k \le K+1)$$

equals the product of the squared singular values of the singular value decomposition (SVD) of $\mathbf{S}$,

$$||\mathbf{S}||_{\mathrm{F}} = \sqrt{\sum_{k=1}^{K+1} \sigma_k^2} \, . \qquad (10.26)$$

SVD is known to model linear dependencies of the rows and columns of the associated matrices. GCIs provide new information and cannot be covered by linear modeling. Hence, singular values tend to be large when the pertinent speech data matrix representing a signal frame,

$$\mathbf{S} = \begin{pmatrix} s(K) & s(K-1) & \cdots & s(0) \\ s(K+1) & s(K) & \cdots & s(1) \\ \vdots & \vdots & \ddots & \vdots \\ s(K+N-1) & s(K+N-2) & \cdots & s(N-1) \end{pmatrix}$$

contains a glottal impulse. (We assume $N > K$ and full column rank of the matrix $\mathbf{S}$ [10.96].) The singular values become largest when the glottal impulse is found in the first row of $\mathbf{S}$. So the Frobenius norm gives an algorithmic shortcut to the costly SVD of $\mathbf{S}$ which would have to be performed otherwise. This algorithm had a couple of forerunners that are well described in [10.96].

Other methods to determine the GCI involve neural networks with appropriate training [10.63], wavelet functions [10.97], simplified inverse filtering [10.77], nonlinear filtering [10.98] or statistical evaluation of the nonstationarity of the signal at the GCI [10.99]. *Mousset* et al. [10.77] evaluated several of these methods using the Keele database for PDA evaluation [10.81]. They used the analysis scheme proposed by *Rabiner* et al. [10.47] and extended it by two error classes suitable for any time-domain PDA: (1) insertion of a GCI marker where no GCI is present, and (2) miss (deletion) of a GCI that should have been detected. Their results show that the methods evaluated are about equivalent but show some sensitivity to the length of the respective time windows and to the speakers.

10.7.2 Multiple Pitch Determination

Simultaneous tones with different frequencies require some frequency-domain processing since, in this domain, peaks resulting from different tones show up at their respective frequencies. The same holds for the lag domain of a double-spectral-transform PDA when the nonlinear distortion in the frequency is of local nature, such as in computing a cepstrum or an autocorrelation function. Hence PDAs that explicitly or implicitly involve a Fourier transform will have this property, as was shown in [10.33] using simultaneous speech from two talkers as well as musical signals. Although not explicitly stated in the literature, active modeling would also allow multipitch tracking.

Such PDAs have been used in several configurations for speaker separation. The usual procedure is that the PDA determines pitch for one of the speakers; then a speech enhancement procedure (e.g., spectral subtraction) is applied to remove this speaker from the signal. The PDA then determines the pitch of the sec-

ond speaker. The problem with this configuration is the correct assignment of the signal to the two speakers, particular when the signal from a speaker is unvoiced. *De Cheveigné* [10.100] describes a PDA designed for this purpose that estimates two pitches simultaneously using a cascaded comb filter and a two-dimensional AMDF estimation procedure.

Multiple pitch determination, particularly for music, is a wide field that would justify a chapter of its own. In this section we can give only a small selection of examples. For more-comprehensive surveys of the activities in this field, the reader is referred to *De Cheveigné* [10.100] or *Klapuri* [10.101].

Goto's algorithm [10.102] for music signals focuses on extracting *leading voices* from a polyphonic signal, in this case a melody line and a bass line which occupy different and non-overlapping F_0 ranges (32–260 Hz versus 260–4100 Hz). The PDA works in the frequency domain. It basically estimates the fundamental frequency of the most predominant harmonic structure corresponding to the melody or bass line. It simultaneously takes into account all possibilities for F_0 and treats the input spectrum as if it contains all possible harmonic structures with different weights (amplitudes). It regards a probability density function (PDF) of the input frequency components as a weighted mixture of harmonic-structure tone models (represented by PDFs) of all possible pitches and simultaneously estimates both their weights corresponding to the relative dominance of every possible harmonic structure and the shape of the tone models by maximum a-posteriori probability (MAP) estimation regarding their prior distribution. It then considers the maximum-weight model as the most predominant harmonic structure and obtains its fundamental frequency. A multiple-agent architecture evaluates the temporal continuity of the estimates.

The PDA uses a logarithmic frequency scale corresponding to the musical notation unit *cent*,

$$\frac{f}{\text{cent}} = 1200 \log_2 \frac{f/\text{Hz}}{440 \cdot 2^{\frac{3}{12}-5}} , \qquad (10.27)$$

where 1 cent equals 1/100 of a tempered semitone so that an octave consists of 1200 cents. Each tone model corresponds to a trial fundamental frequency p and provides a harmonic structure; the individual harmonics are modeled as weighted one-dimensional Gaussian distribution. The weights of all models are estimated simultaneously, where the same frequency can be shared by different harmonics of different tone models. The tone model with maximum weight yields the estimate for the predominant F_0. A multiple-agent architecture performs a temporal track across frames and gives the most stable trajectory as the final result. It consists of a salience detector that picks promising F_0 candidates, and a number of agents that interact to allocate the salient peaks among themselves according to peak closeness. Each agent has its own penalty record, and it gets a penalty when no suitable peak can momentarily be allocated to it. If a penalty threshold is exceeded, the agent is terminated. If a peak cannot be assigned to a running agent, a new agent is created. The final output is determined on the basis of which agent has the highest reliability and greatest total power along the trajectory of the peak it is tracking. Detection rates for melody and bass are reported as 80–90% and depend on the respective tune being processed.

The PDA by *Tolonen* and *Karjalainen* [10.103] uses a perception-oriented double-transform PDA based on the unitary model of pitch perception [10.6]. The original model in [10.6] employs a large number of critical-band filters (spaced at much less than a critical band). In each channel the signal is half-wave rectified and low-pass filtered, and its short-term ACF is determined. The ACFs of all channels are then added to give an estimate of pitch. In Tolonen and Karjalainen's PDA the signal is first filtered by an optional pre-whitening filter and then split into only two subbands with a cutoff frequency of 1000 Hz. The high-frequency channel is then rectified and smoothed. The lag-domain representations obtained by the double-spectral-transform principle for each subband are added up to the so-called summary ACF. This function is then enhanced. It is first clipped from under so that only positive values are retained. Then it is time stretched by a factor of 2, 3, etc. and subtracted from the original summary ACF, and again only positive values are retained. From this enhanced summary ACF significant maxima are sought. For musical sound analysis, relatively long windows (up to 180 ms) are employed. The algorithm works well when the simultaneous tones in the input signal do not differ too much in amplitude. The method is limited to fundamental frequencies below 1000 Hz [10.101].

Klapuri [10.101], besides giving a comprehensive survey of PDAs for multipitch determination, developed two PDAs for this task. One of these is based on the unitary model of pitch perception [10.6], with some modification that makes the PDA more reliable and computationally less costly. The other is based on an iterative frequency-domain technique. The predominant pitch is detected, then the corresponding harmonic structure is removed from the spectrum, and the procedure is repeated for the next predominant pitch. The

iteration is stopped when the energy of the harmonic structure drops below a given threshold.

Determination of the predominant harmonic structure follows a perception-oriented approach that resembles Terhardt's virtual pitch model [10.4] in that it is tolerant toward inharmonicity. The spectrum is subdivided into 18 subbands with overlapping triangular transfer functions which add to unity for adjacent subbands. In each subband salience estimates are made for each trial fundamental frequency p in the measuring range,

$$L(p,k) = \max_m \left(c(m,i) \sum_i^{(\text{in subband } k)} S(m+ip) \right) , \tag{10.28}$$

where m stands for a possible (small) offset due to inharmonicity, and c specifies a weighting function that depends on m and the harmonic number i. The salience estimates are then added over all subbands with the additional possibility of taking into account shift of higher harmonics toward higher frequencies, as it sometimes occurs with musical instruments. One of the trial frequencies p emerges as predominant, and the pertinent harmonic amplitude spectrum is smoothed. To remove the tone, the smoothed harmonic structure is subtracted from the overall spectrum.

Special attention is given to the problem of several tones with harmonic frequency ratios. The smoothing procedure helps to solve this problem. Think of two tones with a frequency ratio of 1:3. All the harmonics of the higher tone coincide with harmonics of the lower one, but we can expect that every third harmonic has an outstanding amplitude. It is likely that the lower tone will be detected first. The smoothing procedure equalizes these amplitude fluctuations so that the higher tone will be preserved when the lower one is removed.

The PDA by *Kameoka* et al. [10.104] performs simultaneous multipitch extraction in the frequency domain based on a statistical model given by

$$\{\Theta\} = \{\mu(k), w(k), \sigma | k = 1, \ldots, K\} . \tag{10.29}$$

The model consists of a set of K harmonic structures. Each of these is described by a tied Gaussian mixture (which corresponds to *Goto*'s PDA [10.102], see above). Its vector $\boldsymbol{\mu}$ of mean values stands for the frequencies of the partials and has only one degree of freedom, i. e., its fundamental frequency. The weight $w(k)$ stands for the predominance of the k-th harmonic structure; the variance is kept constant for the sake of simplicity. The model is initialized, and the parameters are iteratively improved using a log likelihood criterion and the expectation maximization algorithm [10.105]. As the true number of simultaneous tones in the signal is unknown, the model order K is initialized too high and becomes part of the optimization. Akaike's information criterion (AIC), which specifies a tradeoff between the order of a model and its accuracy,

$$\begin{aligned} \text{AIC} = {} & 2 \cdot (\text{number of free parameters}) \\ & - 2 \cdot (\text{maximum of log likelihood}) \end{aligned} \tag{10.30}$$

is used to determine the optimal value of K. A harmonic structure is abandoned when its weight becomes low or when it is placed between two other structures that move toward each other and have higher weights. If a harmonic structure is abandoned, the number of free parameters decreases by two (frequency and weight), and the likelihood gets worse. Whenever a harmonic structure is to be abandoned, AIC is computed. The iteration is stopped when AIC has reached a minimum.

The role of gross pitch determination errors in this context differs from that in speech. In speech we have the special problem of octave errors (cf. Sect. 10.5.2) which are to be strictly avoided. In polyphonic music, on the other hand, two instruments frequently play an octave apart from each other, and should then both be detected. *Klapuri*'s smoothing procedure [10.101] explicitly takes into account two tones at any harmonic interval. Nevertheless it is reported [10.101, 104] that the PDAs have problems when notes are played at certain musical intervals, among them the octave.

10.7.3 Instantaneousness Versus Reliability

It is always desirable to get the estimate from a PDA instantaneously. Processing time depends on two factors: (1) computational complexity of the PDA and speed of the device running it, and (2) latency, i. e., the amount of signal required to get a reliable estimate plus computing time. A PDA runs in real time if the processing time required is less than the elapsed time, say, between two successive frames. As today's computers are able to run even complex PDAs in real time, we can put aside the first issue. With the amount of signal required, however, there may be a hidden problem. An ordinary short-term PDA needs (at least) two complete pitch periods to detect periodicity in an orderly manner. If the spectral transform is done in one step for the whole range of F_0, latency will be twice the longest possible period in the window. For speech, with F_0 ranging from 50 Hz for a deep male voice up to, say, 1000 Hz for a child in spontaneous speech, this means a lag of at least 40 ms from

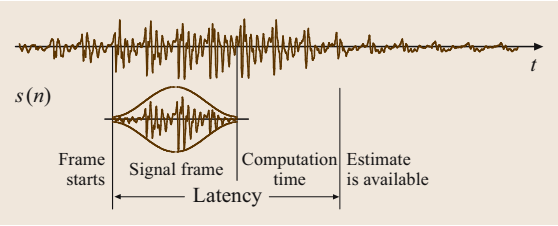

Fig. 10.12 The latency problem. Latency is defined as the elapsed time from the beginning of the frame until the estimate is available

the beginning of the window until we obtain an estimate, not including processing time. In vocoder telephony this may just be tolerable, but this will be a problem for online capturing of music in a MIDI application. Of course the latency will go down when the lower end of the frequency range gets higher. One period of the lowest tone of an ordinary guitar lasts about 12 ms, while that of a violin lasts about 5 ms. However, if the PDA applies postprocessing, for instance, pitch contour tracking by dynamic programming methods [10.17], latencies will go up drastically. It goes without saying that such methods cannot be applied when tough time constraints are given.

How can this problem be solved? If reliability requires a short-term analysis PDA, we must speed it up. If we can assume that the signal is nearly sinusoidal, we can apply time-domain methods, which may be more instantaneous.

Several attempts have been made to speed up short-term analysis PDAs. It has been known for a long time that distance functions, such as the AMDF, can work with short frames. Since no Fourier transform is involved, the function can be computed synchronously as time runs, and so we may be able to obtain estimates of T_0 in little more than the duration of the period to be measured. Estimates of short periods will thus be available sooner than those of long ones. Such nonstationary approaches do not only exist for distance functions. For instance, *Talkin*'s PDA ([10.17]; Sect. 10.2.1) uses a nonstationary autocorrelation function. *Yoo* and *Fujinaga* [10.106] performed some practical experiments with several hardware and software PDAs for MIDI capture on a violin and found latencies of 15–90 ms. For a MIDI application this may already be unacceptable.

If the signal is near-sinusoidal, time-domain pitch determination can be very fast. The lag between two consecutive zero crossings or between a maximum and the consecutive minimum of a sinusoid equals half a period, and only a quarter of a period elapses between an extreme (maximum or minimum) and the adjacent zero crossing. Even faster methods needing fewer signal samples can be conceived, leading to a kind of *anytime* algorithm that can yield a rather imprecise estimate almost instantaneously but is able to refine this estimate when given more time. (Remember, however, that in MIDI applications a tone command can be given only once; if it is in error, the wrong tone will come out.) In this case the problem is: how can we obtain a signal close to a sinusoid? In speech this is unrealistic due to the transfer function of the vocal tract, where lip radiation introduces a zero at $f = 0$ that strongly attenuates the first harmonic. In music, however, such an approach can be of interest when an instrument yields a signal that is almost sinusoidal, or when the capturing microphone or sensor is placed on the instrument in such a way that it performs as a low-pass filter.

10.8 Concluding Remarks

The problem of pitch determination and fundamental frequency tracking is long-standing, known, and yet unsolved in a general sense. However, for most applications, algorithms that yield good and acceptable solutions have been developed. Applications in speech and music have moved away from the vocoder toward prosody recognition, automatic melody detection, acoustic data retrieval, computational auditory scene analysis, and high-precision analysis of speech synthesis corpora. New challenges for the development of PDAs, among others, include high-precision pitch period determination, processing of signals with multiple pitches, and PDAs with very short latencies.

References

10.1 W.J. Hess: *Pitch Determination of Speech Signals – Algorithms and Devices* (Springer, Berlin, Heidelberg 1983)

10.2 R.J. McAulay, T.F. Quatieri: Speech analysis/synthesis based on a sinusoidal representation, IEEE Trans. Acoust. Speech Signal Process. **34**, 744–754 (1986)

10.3 E. Zwicker, W.J. Hess, E. Terhardt: Erkennung gesprochener Zahlworte mit Funktionsmodell und Rechenanlage, Kybernetik **3**, 267–272 (1967), (in German)

10.4 E. Terhardt: Calculating virtual pitch, Hearing Res. **1**, 155–182 (1979)

10.5 R. Plomp: *Aspects of Tone Sensation* (Academic, London 1976)

10.6 R. Meddis, L. O'Mard: A unitary model for pitch perception, J. Acoust. Soc. Am. **102**, 1811–1820 (1997)

10.7 K.J. Kohler: 25 Years of Phonetica: Preface to the special issue on pitch analysis, Phonetica **39**, 185–187 (1992)

10.8 W.J. Hess, H. Indefrey: Accurate time-domain pitch determination of speech signals by means of a laryngograph, Speech Commun. **6**, 55–68 (1987)

10.9 W.J. Hess: Pitch and voicing determination. In: *Advances in Speech Signal Processing*, ed. by M.M. Sondhi, S. Furui (Dekker, New York 1992), p.3–48

10.10 A.M. Noll: Cepstrum pitch determination, J. Acoust. Soc. Am. **41**, 293–309 (1967)

10.11 L.R. Rabiner: On the use of autocorrelation analysis for pitch detection, IEEE Trans. Acoust. Speech Signal Process. **25**, 24–33 (1977)

10.12 E. Terhardt, G. Stoll, M. Seewann: Algorithm for extraction of pitch and pitch salience from complex tonal signals, J. Acoust. Soc. Am. **71**, 679–688 (1982)

10.13 M.S. Harris, N. Umeda: Difference limens for fundamental frequency contours in sentences, J. Acoust. Soc. Am. **81**, 1139–1145 (1987)

10.14 J. 't Hart: Differential sensitivity to pitch distance, particularly in speech, J. Acoust. Soc. Am. **69**, 811–822 (1981)

10.15 H. Duifhuis, L.F. Willems, R.J. Sluyter: Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception, J. Acoust. Soc. Am. **71**, 1568–1580 (1982)

10.16 D.J. Hermes: Measurement of pitch by subharmonic summation, J. Acoust. Soc. Am. **83**, 257–264 (1988)

10.17 D. Talkin: A robust algorithm for pitch tracking (RAPT). In: *Speech Coding and Synthesis*, ed. by B. Kleijn, K. Paliwal (Elsevier, Amsterdam 1995), p.–495–518

10.18 P. Hedelin, D. Huber: Pitch period determination of aperiodic speech signals, Proc. IEEE ICASSP (1990) pp. 361–364

10.19 H. Hollien: On vocal registers, J. Phonet. **2**, 225–243 (1974)

10.20 N.P. McKinney: *Laryngeal Frequency Analysis for Linguistic Research* (Univ. Michigan, Ann Arbor 1965), Res. Rept. No. 14

10.21 H. Fujisaki, K. Hirose, K. Shimizu: A new system for reliable pitch extraction of speech, Proc. IEEE ICASSP (1986), paper 34.16

10.22 M.M. Sondhi: New methods of pitch extraction, IEEE Trans. Acoust. Speech Signal Process. **26**, 262–266 (1968)

10.23 J.D. Markel: The SIFT algorithm for fundamental frequency estimation, IEEE Trans. Acoust. Speech Signal Process. **20**, 149–153 (1972)

10.24 V.N. Sobolev, S.P. Baronin: Investigation of the shift method for pitch determination, Elektrosvyaz **12**, 30–36 (1968), in Russian

10.25 J.A. Moorer: The optimum comb method of pitch period analysis of continuous digitized speech, IEEE Trans. Acoust. Speech Signal Process. **22**, 330–338 (1974)

10.26 T. Shimamura, H. Kobayashi: Weighted autocorrelation for pitch extraction of noisy speech, IEEE Trans. Speech Audio Process. **9**, 727–730 (2001)

10.27 A. de Cheveigné, H. Kawahara: YIN, a fundamental frequency estimator for speech and music, J. Acoust. Soc. Am. **111**, 1917–1930 (2002)

10.28 K. Hirose, H. Fujisaki, S. Seto: A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag, Proc. IEEE ICASSP (1992) pp. 149–152

10.29 D.E. Terez: Robust pitch determination using nonlinear state-space embedding, Proc. IEEE ICASSP (2002)

10.30 C.M. Rader: Vector pitch detection, J. Acoust Soc. Am. **36**(C), 1463 (1964)

10.31 L.A. Yaggi: *Full Duplex Digital Vocoder* (Texas Instruments, Dallas 1962), Scientific Report No.1, SP14-A62; DDC-AD-282986

10.32 Y. Medan, E. Yair, D. Chazan: Super resolution pitch determination of speech signals, IEEE Trans. Signal Process. **39**, 40–48 (1991)

10.33 M.R. Weiss, R.P. Vogel, C.M. Harris: Implementation of a pitch-extractor of the double spectrum analysis type, J. Acoust. Soc. Am. **40**, 657–662 (1966)

10.34 H. Indefrey, W.J. Hess, G. Seeser: Design and evaluation of double-transform pitch determination algorithms with nonlinear distortion in the frequency domain, Proc. IEEE ICASSP, Vol. 2 (1985), paper 11.12

10.35 P. Martin: Comparison of pitch detection by cepstrum and spectral comb analysis, Proc. IEEE ICASSP (1982) pp. 180–183

10.36 V.T. Sreenivas: *Pitch estimation of aperiodic and noisy speech signals* (Indian Institute of Technology, Bombay 1982), Diss., Department of Electrical Engineering, Indian Institute of Technology

10.37 M.R. Schroeder: Period histogram and product spectrum: new methods for fundamental-frequency measurement, J. Acoust. Soc. Am. **43**, 819–834 (1968)

10.38 P. Martin: A logarithmic spectral comb method for fundamental frequency analysis, Proc. 11th Int. Congr. on Phonetic Sciences Tallinn (1987), paper 59.2

10.39 P. Martin: WinPitchPro – a tool for text to speech alignment and prosodic analysis, Proc. Speech Prosody 2004 (2004) pp. 545–548, http://www.isca-speech.org/archive/sp2004 and http://www.winpitch.com

10.40 J.C. Brown, M. Puckette: A high-resolution fundamental frequency determination based on phase changes of the Fourier transform, J. Acoust. Soc. Am. **94**, 662–667 (1993)

10.41 J.C. Brown: Musical fundamental frequency tracking using a pattern recognition method, J. Acoust. Soc. Am. **92**, 1394–1402 (1992)

10.42 F. Charpentier: Pitch detection using the short-term phase spectrum, Proc. IEEE ICASSP (1986) pp. 113–116

10.43 M. Lahat, R.J. Niederjohn, D.A. Krubsack: A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech, IEEE Trans. Acoust. Speech Signal Process. **35**, 741–750 (1987)

10.44 B. Doval, X. Rodet: Estimation of fundamental frequency of musical sound signals, Proc. IEEE ICASSP (1991) pp. 3657–3660

10.45 T. Abe, K. Kobayashi, S. Imai: Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency, Proc. ICSLP'96 (1996) pp. 1277–1280, http://www.isca-speech.org/archive/icslp_1996

10.46 T. Nakatani, T. Irino: Robust and accurate fundamental frequency estimation based on dominant harmonic components, J. Acoust. Soc. Am. **116**, 3690–3700 (2004)

10.47 L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, C.A. McGonegal: A comparative study of several pitch detection algorithms, IEEE Trans. Acoust. Speech **24**, 399–423 (1976)

10.48 L. Arévalo: *Beiträge zur Schätzung der Frequenz gestörter Schwingungen kurzer Dauer und eine Anwendung auf die Analyse von Sprachsignalen* (Ruhr-Universität, Bochum 1991), Diss. in German

10.49 A.M. Noll, A. Michael: Pitch determination of human speech by the harmonic product spectrum the harmonic sum spectrum and a maximum likelihood estimate, Symp. Comput. Process. Commun. **19**, 779–797 (1970), ed. by the Microwave Inst., New York: Univ. of Brooklyn Press

10.50 D.H. Friedman: Pseudo-maximum-likelihood speech pitch extraction, IEEE Trans. Acoust. Speech Signal Process. **25**, 213–221 (1977)

10.51 R.J. McAulay, T.F. Quatieri: Pitch estimation and voicing detection based on a sinusoidal speech model, Proc. IEEE ICASSP (1990) pp. 249–252

10.52 A. Moreno, J.A.R. Fonollosa: Pitch determination of noisy speech using higher order statistics, Proc. IEEE ICASSP (1992) pp. 133–136

10.53 B.B. Wells: Voiced/Unvoiced decision based on the bispectrum, Proc. IEEE ICASSP (1985) pp. 1589–1592

10.54 J. Tabrikian, S. Dubnov, Y. Dickalov: Speech enhancement by harmonic modeling via MAP pitch tracking, Proc. IEEE ICASSP (2002) pp. 3316–3319

10.55 S. Godsill, M. Davy: Bayesian harmonic models for musical pitch estimation and analysis, Proc. IEEE ICASSP (2002) pp. 1769–1772

10.56 C.A. McGonegal, L.R. Rabiner, A.E. Rosenberg: A subjective evaluation of pitch detection methods using LPC synthesized speech, IEEE Trans. Acoust. Speech Signal Process. **25**, 221–229 (1977)

10.57 C. Hamon, E. Moulines, F. Charpentier: A diphone synthesis system based on time-domain prosodic modifications of speech, Proc. IEEE ICASSP (1989) pp. 238–241

10.58 D.M. Howard: Peak-picking fundamental period estimation for hearing prostheses, J. Acoust. Soc. Am. **86**, 902–910 (1989)

10.59 I. Dologlou, G. Carayannis: Pitch detection based on zero-phase filtering, Speech Commun. **8**, 309–318 (1989)

10.60 W.J. Hess: An algorithm for digital time-domain pitch period determination of speech signals and its application to detect F0 dynamics in VCV utterances, Proc. IEEE ICASSP (1976) pp. 322–325

10.61 T.V. Ananthapadmanabha, B. Yegnanarayana: Epoch extraction of voiced speech, IEEE Trans. Acoust. Speech Signal Process. **23**, 562–569 (1975)

10.62 L.O. Dolanský: An instantaneous pitch-period indicator, J. Acoust. Soc. Am. **27**, 67–72 (1955)

10.63 I.S. Howard, J.R. Walliker: The implementation of a portable real-time multilayer-perceptron speech fundamental period estimator, Proc. EUROSPEECH-89 (1989) pp. 206–209, http://www.isca-speech.org/archive/eurospeech_1989

10.64 W.J. Hess: A pitch-synchronous digital feature extraction system for phonemic recognition of speech, IEEE Trans. Acoust. Speech Signal Process. **24**, 14–25 (1976)

10.65 A. Davis, S. Nordholm, R. Togneri: Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold, IEEE Trans. Audio Speech Lang. Process. **14**, 412–424 (2006)

10.66 L.J. Siegel, A.C. Bessey: Voiced/unvoiced/mixed excitation classification of speech, IEEE Trans. Acoust. Speech Signal Process. **30**, 451–461 (1982)

10.67 S. Ahmadi, A.S. Spanias: Cepstrum-based pitch detection using a new statistical V/UV classification algorithm, IEEE Trans. Speech Audio Process. **7**, 333–338 (1999)

10.68 B.M. Lobanov, M. Boris: Automatic discrimination of noisy and quasi periodic speech sounds by the phase plane method, Soviet Physics – Acoustics 16, 353–356 (1970) Original (in Russian) in Akusticheskiy Zhurnal **16**, 425–428 (1970)

10.69 E. Fisher, J. Tabrikian, S. Dubnov: Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model, IEEE Trans. Audio Speech Lang. Process. **14**, 502–510 (2006)

10.70 B.S. Atal, L.R. Rabiner: A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition, IEEE Trans. Acoust. Speech Signal Process. **24**, 201–212 (1976)

10.71 O. Fujimura: An approximation to voice aperiodicity, IEEE Trans. Acoust. Speech Signal Process. **16**, 68–72 (1968)

10.72 A.K. Krishnamurthy, D.G. Childers: Two-channel speech analysis, IEEE Trans. Acoust Speech Signal Process. **34**, 730–743 (1986)

10.73 K.N. Stevens, D.N. Kalikow, T.R. Willemain: A miniature accelerometer for detecting glottal waveforms and nasalization, J. Speech Hear. Res. **18**, 594–599 (1975)

10.74 V.R. Viswanathan, W.H. Russell: *Subjective and objective evaluation of pitch extractors for LPC and harmonic-deviations vocoders* (Bolt Beranek and Newman, Cambridge 1984), MA: Report No. 5726

10.75 A.J. Fourcin, E. Abberton: First applications of a new laryngograph, Med Biol Illust **21**, 172–182 (1971)

10.76 D.G. Childers, M. Hahn, J.N. Larar: Silent and voiced/Unvoiced/Mixed excitation (four-way) classification of speech, IEEE Trans. Acoust. Speech Signal Process. **37**, 1771–1774 (1989)

10.77 E. Mousset, W.A. Ainsworth, J.A.R. Fonollosa: A comparison of several recent methods of fundamental frequency and voicing decision estimation, Proc. ICSLP'96 (1996) pp. 1273–1276, http://www.isca-speech.org/archive/icslp_1996

10.78 D.A. Krubsack, R.J. Niederjohn: An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech, IEEE Trans. Signal Process. **39**, 319–329 (1991)

10.79 Y. Xu, X. Sun: Maximum speed of pitch change and how it may relate to speech, J. Acoust. Soc. Am. **111**, 1399–1413 (2002)

10.80 B.G. Secrest, G.R. Doddington: Postprocessing techniques for voice pitch trackers, Proc. IEEE ICASSP (1982) pp. 172–175

10.81 F. Plante, G.F. Meyer, W.A. Ainsworth: A pitch extraction reference database, Proc. EUROSPEECH'95 (1995) pp. 837–840, http://www.isca-speech.org/archive/eurospeech_1995

10.82 H. Kawahara, H. Katayose, A. de Cheveigné, R.D. Patterson: Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity, Proc. EUROSPEECH'99 (1999) pp. 2781–2784, http://www.isca-speech.org/archive/eurospeech_1999

10.83 L.R. Rabiner, M.R. Sambur, C.E. Schmidt: Applications of nonlinear smoothing algorithm to speech processing, IEEE Trans. Acoust. Speech Signal Process. **23**, 552–557 (1975)

10.84 P. Specker: A powerful postprocessing algorithm for time-domain pitch trackers, Proc. IEEE ICASSP (1984), paper 28B.2

10.85 F. Itakura: Minimum prediction residual applied to speech recognition, IEEE Trans. Acoust. Speech Signal Process. **23**, 67–72 (1975)

10.86 Y.R. Wang, I.J. Wong, T.C. Tsao: A statistical pitch detection algorithm, Proc. IEEE ICASSP (2002) pp. 357–360

10.87 Y. Sagisaka, N. Campbell, N. Higuchi (eds.): *Computing prosody. Computational models for processing spontaneous speech* (Springer, New York 1996)

10.88 P. Bagshaw: *Automatic prosodic analysis for computer aided pronunciation teaching* (Univ. of Edinburgh, Edinburgh 1993), PhD Thesis http://www.cstr.ed.ac.uk/projects/fda/Bagshaw_PhDThesis.pdf

10.89 R.J. Baken: *Clinical Measurement of Speech and Voice* (Taylor Francis, London 1987)

10.90 A. Askenfelt: Automatic notation of played music: The Visa project, Fontes Artis Musicae **26**, 109–120 (1979)

10.91 Y.M. Cheng, D. O'Shaughnessy: Automatic and reliable estimation of glottal closure instant and period, IEEE Trans. Acoust. Speech Signal Process. **37**, 1805–1815 (1989)

10.92 W.J. Hess: Determination of glottal excitation cycles in running speech, Phonetica **52**, 196–204 (1995)

10.93 W.J. Hess: Pitch determination of acoustic signals – an old problem and new challenges, Proc. 18th Intern. Congress on Acoustics, Kyoto (2004), paper Tu2.H.1

10.94 B. Yegnanarayana, R. Smits: A robust method for determining instants of major excitations in voiced speech, Proc. IEEE ICASSP (1995) pp. 776–779

10.95 M. Brookes, P.A. Naylor, J. Gudnason: A quantitative assessment of group delay methods for identifying glottal closures in voiced speech, IEEE Trans. Audio Speech Language Process. **14**, 456–466 (2006)

10.96 C.X. Ma, Y. Kamp, L.F. Willems: A Frobenius norm approach to glottal closure detection from the speech signal, IEEE Trans. Speech Audio Process. **2**, 258–265 (1994)

10.97 L. Du, Z. Hou: Determination of the instants of glottal closure from speech wave using wavelet transform, http://www.icspat.com/papers/329mfi.pdf
10.98 K.E. Barner: Colored L-ℓ filters and their application in speech pitch detection, IEEE Trans. Signal Process. **48**, 2601–2606 (2000)
10.99 J.L. Navarro-Mesa, I. Esquerra-Llucià: A time-frequency approach to epoch detection, Proc. EUROSPEECH'95 (1995) pp. 405–408, http://www.isca-speech.org/archive/eurospeech_1995
10.100 A. de Cheveigné: Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing, J. Acoust. Soc. Am. **93**, 3279–3290 (1993)
10.101 A.P. Klapuri: *Signal processing methods for the automatic transcription of music* (Tampere Univ. Technol., Tampere 2004), Ph.D. diss. http://www.cs.tut.fi/sgn/arg/klap/klap_phd.pdf
10.102 M. Goto: A predominant F0-estimation method for polyphonic musical audio signals, Proc. 18th Intern. Congress on Acoustics Kyoto (2004), paper Tu2.H.4
10.103 T. Tolonen, M. Karjalainen: A computationally efficient multipitch analysis model, IEEE Trans. Speech Audio Process. **8**, 708–716 (2000)
10.104 H. Kameoka, T. Nishimoto, S. Sagayama: Separation of harmonic structures based on tied Gaussian mixture model and information criterion for concurrent sounds, Proc. IEEE ICASSP (2004), paper AE-P5.9
10.105 A.P. Dempster, N.M. Laird, D.B. Rubin: Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. B **39**, 1–38 (1977)
10.106 L. Yoo, I. Fujinaga: A comparative latency study of hardware and software pitch-trackers, Proc. 1999 Int. Computer Music Conf. (1999) pp. 36–40