8

Auditory Model-Based Methods for Multiple Fundamental Frequency Estimation

Anssi Klapuri

Institute of Signal Processing, Tampere University of Technology,
Korkeakoulunkatu 1, 33720 Tampere, Finland
Anssi.Klapuri@tut.fi

8.1 Introduction

This chapter describes fundamental frequency (F0) estimation methods that make use of computational models of human auditory perception and especially pitch perception. At the present time, the most reliable music transcription system available is the ears and the brain of a trained musician. Compared with any artificial audio processing tool, the analytical ability of human hearing is very good for complex mixture signals: in natural acoustic environments, we are able to perceive the characteristics of several simultaneously occurring sounds, including their pitches [49]. It is therefore quite natural to pursue automatic music transcription and multiple F0 estimation by investigating what happens in the human listener. Here the term multiple F0 estimation means estimating the F0s of several concurrent sounds.

Fundamental frequency is the measurable physical counterpart of *pitch*. In Chapter 1, pitch was defined as the perceptual attribute of sounds which allows them to be ordered on a frequency-related scale extending from low to high. More exactly, the pitch of a sound was said to be the frequency of a sine wave that is matched to a target sound by human listeners. The importance of pitch for hearing in general is indicated by the fact that the auditory system tries to assign a pitch frequency to almost all kinds of acoustic signals. Not only sinusoids and periodic signals have a pitch, but even noise signals of various kinds can be consistently matched with a sinusoid of a certain frequency. For a steeply lowpass- or highpass- filtered noise signal, for example, a weak pitch is heard around the spectral edge. Amplitude modulating a random noise signal causes a pitch perception corresponding to the modulation frequency. Also, the sounds of bells and vibrating membranes have a pitch, although their waveform is not clearly periodic and their spectrum does not have a regular structure. A complete review of this 'zoo of pitch effects' can be found in [275], [474], [297]. The auditory system seems to be strongly inclined towards using a single frequency value to summarize certain aspects of sound

events. Computational models of pitch perception attempt to replicate this phenomenon.

Practical multiple F0 estimation methods have a slightly different purpose than pitch perception models. The set of acoustic signals of interest is narrower since the physical concept of F0 is defined only for periodic and nearly periodic sounds. Also, the evaluation criteria are different: multiple F0 estimation methods are judged based on their reliability in the given task, F0 estimation in a mixture signal, whereas an auditory model should faithfully reproduce the mechanisms and the behaviour of the auditory system.

For musical sounds, the F0 and the perceived pitch are practically equivalent. However, there are ambiguous situations such as the octave ambiguity, where it is not clear if the F0 of a sound is x Hz or half or twice that value. From the music transcription point of view, it would be desirable to solve these ambiguities so that the estimated F0 would correspond to the perceived pitch. This is one of the reasons why auditory model-based methods have been employed. Other reasons include the aim of achieving robustness for diverse kinds of musical sounds (these are discussed in Section 8.2) and obtaining a good time/F0 resolution by using a time-frequency decomposition similar to that in human hearing. The advantages and disadvantages of auditory model-based methods are summarized later in this chapter. In general, perceptually motivated methods have been quite successful in audio content analysis.

The primary focus of this chapter is on practical multiple F0 estimation and not so much on auditory modelling. More comprehensive introductions to pitch perception models can be found in [297], [522], [132]. Also, the emphasis is laid on *multiple* F0 estimation methods: some perceptually motivated methods are omitted that are purported to be useful for single F0 estimation in noisy speech signals. The aim of this chapter is twofold: to give a compact description of pitch perception models so that the reader will be able to develop auditorily motivated analysis methods of his own and, secondly, to describe already-existing multiple F0 estimators that are based on and motivated by these models.

This chapter is organized as follows. Section 8.2 discusses the basic acoustic characteristics of pitched musical sounds and how these can be used to compute the F0 of the sounds. Section 8.3 describes computational models of pitch perception. Section 8.4 introduces music transcription systems which use an auditorily model as a 'front end'. That is, the systems apply a perceptually-motivated data representation but the emphasis is laid on the inference that follows the auditory modelling stage, instead of proposing changes to the auditory model itself. Section 8.5 describes multiple F0 estimation methods which extend or modify pitch perception models in order to make them better applicable to F0 estimation in polyphonic music signals. In the end, two algorithms are described which can be directly used for this purpose. Finally, Section 8.6 summarizes the main conclusions.

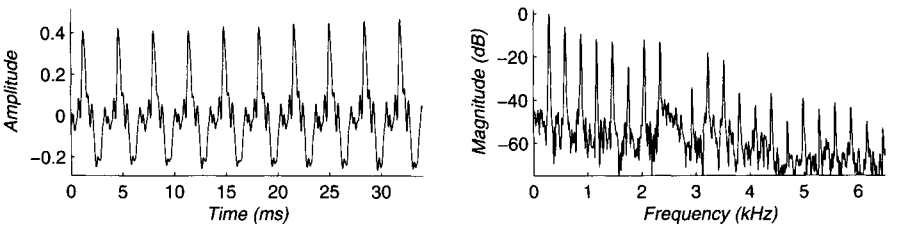

Fig. 8.1. A harmonic sound in the time and frequency domains. The example represents a violin sound with fundamental frequency 290 Hz and fundamental period 3.4 ms.

8.2 Musical Sounds and F0 Estimation

This section discusses the acoustic characteristics of pitched musical sounds and F0 estimation when the sounds are presented in isolation. This provides the background for describing pitch perception models and multiple F0 estimation methods in the subsequent sections.

8.2.1 Pitched Musical Sounds

Musical sounds usually consist of several frequency components. The relative amplitudes of the overtone partials and their time evolution determines the timbre of the sound. Here we are primarily interested in the *frequencies* of the partials since F0 estimation methods try to normalize away the timbre information. From this point of view, pitched musical sounds can be divided into two main classes: sounds that are harmonic and sounds that are not. The methods to be described in this chapter are concerned with both of these.

Most Western musical instruments produce harmonic sounds.[1] These sounds have a spectral structure where the dominant frequency components, called *harmonics*, are approximately regularly spaced. Figure 8.1 illustrates a harmonic sound in the time and frequency domains. The F0 of the sound is the inverse of its time-domain period and the frequency spacing between the overtone partials corresponds approximately to the F0. Usually the overtone components are not perceived separately but only the pitch and the timbre of the entire sound are heard.

For an ideal harmonic sound, the frequencies of the overtone partials are integer multiples of the F0. However, it should be noted that the spectra of harmonic sounds are not always perfectly harmonic; the higher-order overtones of plucked and struck string instruments deviate slightly from their ideal

[1] More exactly, all instruments in the chordophone and aerophone families (see Table 6.1 on p. 167).

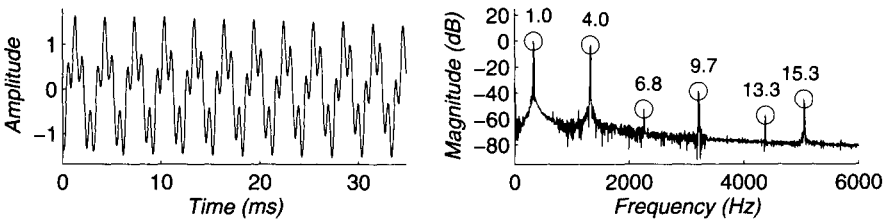

Fig. 8.2. A vibraphone sound (F0 330 Hz) illustrated in the time and frequency domains. In the right panel, the frequencies of the most dominant spectral components are shown in relation to the F0.

harmonic positions. For these classes of instruments, the partial frequencies obey the formula

$$f_j = jF\sqrt{1 + B(j^2 - 1)}, \tag{8.1}$$

where F is the fundamental frequency, $j = 1, 2, \ldots$ is the partial index, and B is an inharmonicity factor [193, p. 363]. Typical values of B are of the order 10^{-4} or 10^{-3} for the middle pitch range of the piano, for example. This makes the higher-order partials gradually shift upwards in frequency, but the structure of the spectrum is in general very similar to that in Fig. 8.1, and the sounds can be classified as harmonic. The inharmonicity is due to the stiffness of real strings, which contributes a restoring force along with the string tension [193], [315].

Figure 8.2 shows an example of a sound which does not belong to the class of harmonic sounds although it is nearly periodic in the time domain and has a clear pitch. In Western music, mallet percussion instruments are a case in point: these instruments produce pitched sounds which are not harmonic. The most common instruments in this family are the marimba, the vibraphone, the xylophone, and the glockenspiel. The sound production mechanism in all of these is a vibrating bar. A bar of uniform thickness with free ends has vibration modes whose frequencies are not in integral ratios. However, by making the bar thinner at the middle of its length, the overtones can be tuned. The first overtone of the marimba and the vibraphone is typically tuned to be four times the F0 and that of the xylophone to be three times the F0.

8.2.2 Basic Principles of F0 Estimation

There are a large number of different methods for monophonic F0 estimation [289]. Comparative evaluations of these can be found e.g. in [535], [290], [134]. The aim of this section is not to make an exhaustive coverage of these, but merely to point out the main acoustic features that different algorithms are built upon: time-domain periodicity and frequency-domain periodicity, and to provide a few representative examples of each approach.

The majority of F0 estimation methods are based on measuring the periodicity of an acoustic signal in the time domain (see e.g. [618], [135]). This

makes sense, since all the pitched musical sounds described above are periodic or almost periodic in the time domain. As reported in [134], quite accurate single F0 estimation can be achieved simply by an appropriate normalization of the short-time autocorrelation function (ACF), defined as

$$r(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} x(n)x(n+\tau). \tag{8.2}$$

The F0 of the signal $x(n)$ can be computed as the inverse of the lag τ that corresponds to the maximum of $r(\tau)$ within a predefined range. To avoid detecting an integer multiple of the period, short lags have to be favoured over longer ones.

An implicit way of measuring time-domain periodicity is to match a *harmonic pattern* to the signal in the frequency domain. According to the Fourier theorem, a periodic signal with period τ can be represented with a series of sinusoidal components at the frequencies j/τ, where j is a positive integer. This can be observed for the musical sounds in Figs. 8.1 and 8.2. Algorithms that are based on frequency-domain harmonic pattern matching have been proposed in [153], [54], [428], for example.

Another class of F0 estimators measure the periodicity of the Fourier spectrum of a sound [384], [380]. These methods are based on the observation that a harmonic sound has an approximately periodic magnitude spectrum, the period of which is the F0. In its simplest form, the autocorrelation function $\rho(m)$ over an N-length magnitude spectrum is calculated as

$$\rho(m) = \frac{2}{N} \sum_{k=0}^{N/2-m-1} |X(k)||X(k+m)|. \tag{8.3}$$

In the above formula, any two frequency components with a certain spectral interval m support the corresponding F0. The spectrum can be arbitrarily shifted without affecting the output value. An advantage of this is that the calculations are somewhat more robust against the imperfect harmonicity of plucked and struck string instruments since the intervals between the overtone partials do not vary as much as their absolute frequencies deviate from the harmonic positions. However, in its pure form this approach has more drawbacks than advantages. In particular, estimating low F0s is not reliable since the F0 resolution of the method is linear whereas the time-domain ACF leads to $1/F$ resolution.

An interesting difference between the F0 estimators in (8.2) and (8.3) is that measuring the periodicity of the time-domain signal is prone to errors in F0 halving because the signal is periodic at twice the fundamental period too, whereas measuring the periodicity of the magnitude spectrum is prone to errors in F0 doubling because the spectrum is periodic at twice the F0 rate, too. The two approaches can be combined using an auditory model, as will be described in Section 8.3.2.

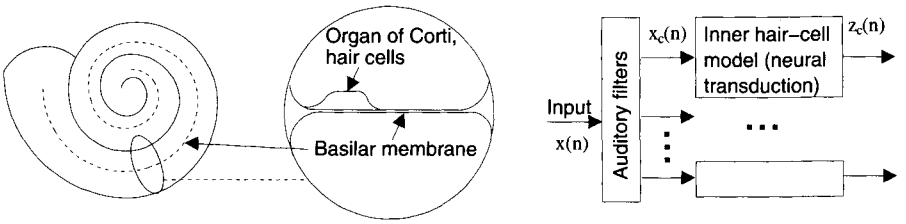

Fig. 8.3. An illustration of the cochlea (*left*) and its cross-section (*middle*). The right panel shows a rough computational model of the cochlea.

8.3 Pitch Perception Models

This section describes computational models of pitch perception and discusses the advantages that an auditory model-based method may have in multiple F0 estimation.

The human auditory system can be divided into two main parts: peripheral hearing and the auditory cortex in the brain. Both of these play an important part in pitch perception. The peripheral part consists of the outer ear, the middle ear, and the inner ear. The first two of these essentially contribute to directional hearing and impedance matching of sound. From the pitch analysis point of view, the interesting part starts from the inner ear, where there is an organ called the *cochlea.*

The cochlea is a sophisticated organ where pressure variations are transformed into properly coded neural impulses in the auditory nerve. Physiologically, the cochlea is a long, coiled, tubular structure which is filled with liquid and tapers towards its end (see Fig. 8.3). The cochlea is divided into two main sections by the *basilar membrane* that runs its entire length. When the mechanical vibrations of the eardrum are transmitted via the middle ear to the inner ear, hydraulic pressure waves are caused in the cochlea and the basilar membrane starts to vibrate. The waves propagate along the basilar membrane so that high frequencies peak in amplitude (resonate) near the beginning and low frequencies get their largest amplitude at the far end.

On the basilar membrane, there is the *organ of Corti* which contains two types of *hair cells.* Outer hair cells are active elements which contribute to the resolution of the cochlear frequency analysis, making different places along the basilar membrane more sharply tuned to their characteristic frequencies than they would be by the acoustic properties of the membrane alone. Inner hair cells register the movement of the basilar membrane. They respond to mechanical displacement by generating nerve impulses into the auditory nerve fibres that are attached to them and lead to the brain [680].

Computational models of the cochlea comprise two main parts which can be summarized as follows (see Fig. 8.3):

1. An acoustic input signal is passed though a bank of bandpass filters, called *auditory filters*, which model the frequency selectivity of the inner ear.

Typically about 100 filters are used with centre frequencies uniformly distributed on a nearly logarithmic frequency scale (details in Section 8.3.1). The outputs of individual filters simulate the mechanical movement of the basilar membrane at different points along its length.

2. The signal at each band, or *auditory channel*, is processed to model the transform characteristics of the inner hair cells which produce neural impulses in the auditory nerve. In signal processing terms, this involves three main characteristics: compression and level adaptation, half-wave rectification, and lowpass filtering (details in Section 8.3.2).

In the following, the acoustic input signal is denoted by $x(n)$ and the impulse response of an auditory filter by $g_c(n)$, where c is the channel index. The output of the auditory filter at channel c is denoted by $x_c(n)$ and functions as an input to the second step. The output of the inner hair cell model is denoted by $z_c(n)$ and represents the probability of observing a neural impulse at channel c.

The processing mechanisms in the brain can be studied only indirectly and are therefore not as accurately known. Typically the relative merits of different models are judged according to their ability to predict the perception of human listeners for various acoustic stimuli in psychoacoustic tests. Different theories and models of the central auditory processing will be summarized in Section 8.3.3, but in all of them, the following two processing steps can be distinguished:

3. Periodicity analysis of some form takes place for the signals $z_c(n)$ within the auditory channels. Phase differences between channels become meaningless.
4. Information is integrated across channels.

In the above processing chain, the auditory nerve signal $z_c(n)$ represents a nice 'interface' between the Steps 2 and 3 and thus between the peripheral and central processes. The signal in the auditory nerve has been directly measured in cats and in some other mammals and this is why the stages 1 and 2 are quite well known. Computational models of the peripheral hearing can approximate the auditory-nerve signal quite accurately, which is a great advantage since an important part of the processing already takes place at these stages. However, central processes and especially Step 3 are (arguably) even more crucial in pitch perception. The above four steps are now described in more detail.

8.3.1 Cochlear Filterbank

Frequency analysis is an essential part of the cochlear processing. Frequency components of a complex sound can be perceived separately and are coded independently in the auditory nerve (in distinct nerve fibres) provided that their frequency separation is sufficiently large [473]. This frequency analysis

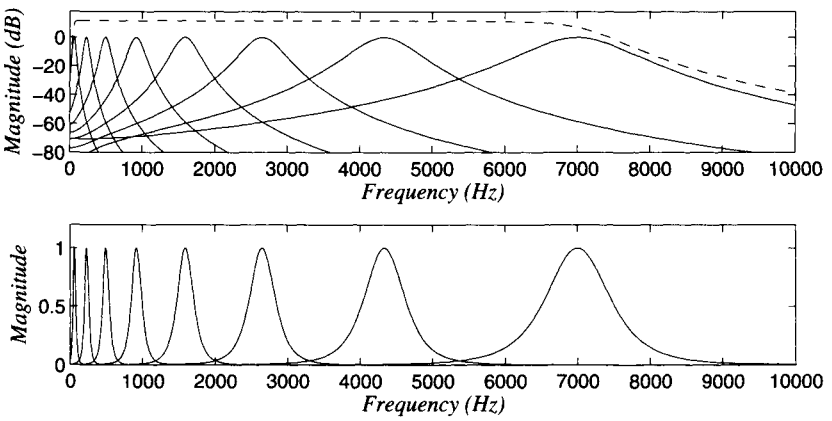

Fig. 8.4. Frequency responses of a few auditory filters shown on the logarithmic (*top*) and on the linear magnitude scale (*bottom*). The dashed line in the upper panel shows the summary response of the filterbank when 70 auditory filters are distributed between 60 Hz and 7 kHz.

can be modelled with a bank of linear bandpass filters: Figure 8.4 shows an example of such a filterbank.

The bandwidths and the shape of the power response of the auditory filters have been studied using the *masking* phenomenon [192], [499]. Masking refers to a situation where an audible sound becomes inaudible in the presence of another, louder sound. In particular, if the distance between two spectral components is less than a so-called *critical bandwidth*, one easily masks the other. The situation can be thought of as if the components would go to the same auditory filter, or to the same channel in the auditory nerve. If the frequency separation is larger, the components are coded independently and are both audible.

The bandwidths of the auditory filters can be conveniently expressed using the equivalent rectangular bandwidth (ERB) concept. The ERB of a filter is defined as the bandwidth of a perfectly rectangular filter which has a unity magnitude response in its passband and an integral over the squared magnitude response which is the same as for the specified filter. The ERB bandwidths b_c of the auditory filters have been found to obey

$$b_c = 0.108 f_c + 24.7 \text{ Hz}, \tag{8.4}$$

where f_c is the centre frequency of the filter at channel c [473].

The centre frequencies of the auditory filters are typically assumed to be uniformly distributed on a critical-band scale. This frequency-related scale is derived by integrating the inverse of (8.4), which yields

$$\xi(f) = 21.4 \log_{10}(0.00437 f + 1). \tag{8.5}$$

In the above expression, f denotes frequency in Hertz and $\xi(f)$ gives the critical-band scale. When f varies between 0 Hz and 20 kHz, $\xi(f)$ varies between 0 and 42. Intuitively, this means that approximately 42 critical bands (or auditory filters) would fit within the range of hearing if the passbands of the filters were non-overlapping and rectangular in shape. Conversion from the critical-band scale back to Hertz units is given by

$$f(\xi) = 229 \times (10^{\xi/21.4} - 1). \tag{8.6}$$

For example, let us distribute 70 filters uniformly on the critical-band scale between 100 Hz and 10 kHz. Using (8.5), we find that the corresponding frequency boundaries on the critical-band scale are 3.36 and 35.3, respectively, and that the distance between each two centre frequencies has to be $(35.3 - 3.36)/69 = 0.463$ on this scale. The centre frequencies on the critical-band scale can then be converted to Hertz units using (8.6).

When a lot of auditory filters are uniformly distributed on the scale $\xi(f)$, power responses of the filters sum approximately to a flat response, as indicated by the dashed line in Fig. 8.4. Typically about 100 filters are used to obtain a good sampling of centre frequencies along the cochlea and a sufficiently flat summary response. Note that in this case, the passbands of the filters overlap considerably. In F0 estimation, only the filters up to about 5 to 8 kHz need to be used, as the most significant harmonic components are below this.

The time-domain impulse responses of the auditory filters have been studied using a so-called reverse correlation method. In the study by de Boer and de Jongh [128], the ear of a cat was stimulated with white noise and the resulting action potentials of individual auditory nerve fibres were recorded simultaneously. Using the input signal and the recorded train of neural impulses, the impulse response of the corresponding auditory filter was derived. The impulse response relates the input signal to the *firing probability* of the nerve fibre under study, that is, to the probability of an inner hair cell generating an impulse to the fibre.

A so-called gammatone filter provides an excellent fit to the experimentally found impulse responses. The filter is defined by its impulse response as [502]

$$g_c(t) = at^{n-1}e^{-2\pi bt} \times \cos(2\pi f_c t + \theta), \tag{8.7}$$

where the normalization factor $a = (2\pi b)^n/\Gamma(n)$ ensures a unity response at the centre frequency, $\Gamma(n)$ is the gamma function, and the parameter value $n = 4$ leads to a shape of the power response that matches best with real auditory filters. The parameter $b = 1.019b_c$ is used to control the bandwidth of the filter.

Figure 8.5 illustrates the impulse responses of two gammatone filters with centre frequencies 100 Hz and 1.0 kHz, and with bandwidths obtained from (8.4). The impulse response consists of a sinusoidal tone at the centre

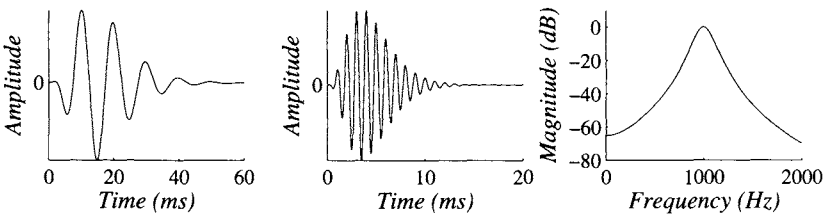

Fig. 8.5. Impulse responses of two gammatone filters with centre frequencies 100 Hz (*left*) and 1.0 kHz (*middle*). The frequency response of the latter filter is shown on the right.

frequency of the filter, f_c, windowed with a function that is precisely the gamma distribution from statistics. Frequency responses of several gammatone filters are shown in Fig. 8.4.

The gammatone filters can be implemented efficiently using a cascade of four second-order IIR filters. A detailed description of the design of the filterbank and the corresponding source code can be found in the technical report by Slaney [591].

8.3.2 Mechanical-to-Neural Transduction

Inner hair cells (IHC) are the elements which convert the mechanical motion of the basilar membrane into firing activity in the auditory nerve. Each IHC rests at a certain point along the basilar membrane and thus follows its movement at this position. Correspondingly, in the computational models the output of each auditory filter is processed by an IHC model.

The IHCs produce neural impulses, or 'spikes', which are binary events. However, since there is a large population of the cells, it is conventional to model the firing *probability* as a function of the basilar membrane movement. Thus the input to an IHC model comes from the output of an auditory filter, $x_c(n)$, and the output of the IHC model represents the time-varying firing probability denoted by $z_c(n)$.

Several computational models of the IHCs have been proposed. An extensive comparison of eight different models was presented by Hewitt and Meddis in [291]. In the evaluation, the model of Meddis [456] outperformed the others by showing only minor discrepancies with the empirical data and by being also one of the most efficient computationally. An implementation of this model is available in the AIM [501] and HUTear [273] auditory toolboxes, for example.

A problem with the realistic IHC models is that they depend critically on the absolute level of their input signal. The dynamic range of the model of Meddis [456], for example, is only 25 dB and the firing rate saturates at the 60 dB level. This limitation of individual IHCs is real, and it seems that the auditory system uses a population of IHCs with different dynamic ranges to

achieve the good intensity discrimination performance over a dynamic range of about 120 dB [523, pp. 137–142]. This has not been included in the computational models of the individual IHCs [291].

For the above-described reason and for the sake of simplicity, many practical systems have replaced a realistic IHC model by a cascade of (i) compression, (ii) half-wave rectification, and (iii) lowpass filtering [171], [327], [677], [354]. As mentioned in the beginning of this section, these are the main characteristics of the IHCs. An advantage of doing this is that the behaviour of the overall system becomes easier to analyse and the signal-level dependency is removed. As a disadvantage, the longer-term level adaptation properties of more realistic IHC models are lost. This is also the approach followed here: instead of going into the details of realistic IHC models, we analyse the basic characteristics of the IHC in order to understand their function in pitch perception and practical F0 estimation.

(i) The compression step has taken slightly different forms in different implementations, but a common theme in all of these has been to scale the sub-band signals $x_c(n)$ inversely proportional to their variance. Ellis scaled the variances of the sub-band signals to unity [171]. Klapuri generalized this approach by scaling the sub-band signals by a factor $\sigma_c^{\nu-1}$, where σ_c is the standard deviation of $x_c(n)$ and $0 \leq \nu \leq 1$ is a compression coefficient [354]. Tolonen and Karjalainen omitted compression at sub-bands but pre-whitened the spectrum of an input signal using inverse warped-linear-prediction filtering, which leads to a very similar result [627].

(ii) Half-wave rectification (HWR) is the clearly non-linear processing step in the mechanical-to-neural transduction. It is defined as

$$\mathrm{HWR}(x) = \begin{cases} x, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{8.8}$$

As simple as it seems, rectification within the sub-bands plays an important part in pitch perception and in practical F0 estimation. In particular, it allows a synthesis of the time and the frequency-domain periodicity analysis methods introduced in (8.2) and (8.3), respectively.

Figure 8.6 illustrates the HWR operation for a narrow-band signal which consists of five overtones of a harmonic sound. Most importantly, the rectification generates spectral components which correspond to the frequency *intervals* between the input partials. The spectral components generated below 1 kHz represent the amplitude envelope of the input signal, as shown in the lowest panels. A signal that consists of more than one frequency component exhibits periodic fluctuations, *beating*, in its time-domain amplitude envelope. That is, the partials alternately amplify and cancel each other out, depending on their phase. The rate of beating caused by each pair of frequency components depends on their frequency difference and, for a harmonic sound, the frequency interval corresponding to the F0 dominates.

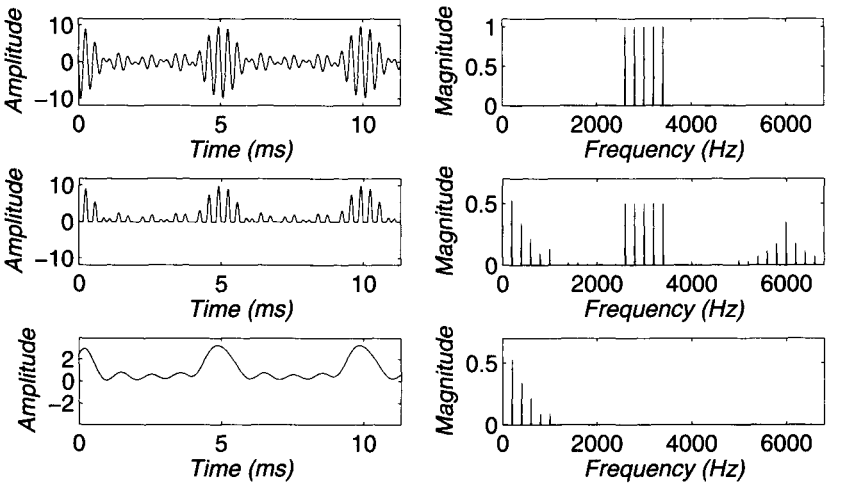

Fig. 8.6. Upper panels show a signal consisting of the overtone partials 13–17 of a sound with F0 200 Hz (fundamental period 5 ms) in the time and frequency domains. Middle panels illustrate the signal after half-wave rectification. Lower panels show the result of lowpass filtering the rectified signal with a 1 kHz cut-off.

The complex Fourier spectrum $Y(k)$ of a rectified signal $y(n) = \mathrm{HWR}(x(n))$ can be approximated by

$$\hat{Y}(k) = \frac{\sigma_x}{\sqrt{8\pi}}\delta(k) + \frac{1}{2}X(k) + \frac{1}{\sigma_x\sqrt{8\pi}} \sum_{j=-N/2+k}^{N/2-k} X(j)X(k-j), \qquad (8.9)$$

where $\delta(k)$ is the unit impulse function, and $X(k)$ and σ_x are the complex Fourier spectrum and the standard deviation of $x(n)$, respectively [353, p. 38], [117]. The approximation assumes that $x(n)$ is a zero-mean Gaussian random process but it is sufficiently accurate for signals such as that in Fig. 8.6, too. On the right-hand side of (8.9), the first term is a dc-component, the second term represents the spectrum of the input signal, and the last term, the convolution of the spectrum $X(k)$ with itself, represents the beating components of the amplitude-envelope spectrum. In addition, the last term generates a harmonic distortion spectrum centred on twice the centre frequency of the input narrow-band signal $x(n)$ in Fig. 8.6. Periodicity analysis of the resulting signal in the time domain (see the next subsection) leads to a combined use of the time and frequency domain periodicity because the rectified signal consists of both the input partials and partials that correspond to their difference frequencies.

Another important property of the HWR is that a series of partials with approximately *uniform amplitudes* cause strong beating. This is because the magnitude of beating caused by each two frequency components is determined by the smaller of the two amplitudes. In the spectrum of a harmonic sound, each pair of neighbouring harmonics contributes to the beating at the

fundamental-frequency rate, but the 'minimum amplitude' property filters out individual higher-amplitude partials. This phenomenon is well known in hearing: if the amplitude of one of the overtones of a harmonic sound rises clearly above the others, it is perceptually segregated and stands out as an independent sound [49]. In computational multiple F0 estimation, this is a desirable characteristic since it makes the F0 computations more immune to the partials of other, co-occurring sounds. Especially when processing the higher overtones of a sound, this partly prevents stealing the energy of the partials of other sounds.

(iii) Lowpass filtering the rectified signal can be used to balance the weight between the amplitude envelope versus the input narrow-band signal. Most systems have used a fixed low-order lowpass filter with a cut-off frequency around 1 kHz at all channels. The sub-band signal after compression, rectification, and lowpass filtering is denoted by $z_c(n)$.

8.3.3 Periodicity Analysis at Sub-Bands and Cross-Band Integration

The auditory nerve signal, modelled by $z_c(n)$, $c = 1, \ldots, C$, is further processed in the brain. Although the central processing mechanisms are not accurately known, it has been convincingly shown that periodicity analysis of some kind takes place within each auditory channel and the results are then combined across channels to yield a pitch perception [457], [67]. This amount of knowledge is already very useful and almost carries us to a situation where only parameter optimization is left in order to process pitch in a way similar to that of the human brain.

The first pitch model of the above-described type was proposed by Licklider [409]. He proposed to computed short-time autocorrelation functions $r_c(\tau)$ within the auditory channels c and to derive pitch from the resulting two-dimensional $(c \times \tau)$ representation. This became known as the 'duplex theory' of pitch perception because it involved both frequency analysis (by the cochlear filterbank) and autocorrelation analysis. Further development with this class of models was made by Lyon [422], Weintraub [663], and Slaney and Lyon [594].

Meddis and Hewitt implemented Licklider's model using a gammatone filterbank and a realistic IHC model and carried out extensive simulations to investigate if the pitch estimate of the model agreed with human listeners for various audio signals [457]. The authors computed ACFs within the auditory channels as

$$r_c(n, \tau) = \sum_{i=0}^{n} z_c(n-i) z_c(n-i-\tau) w(i), \tag{8.10}$$

where $z_c(n)$ is the output of the IHC model in channel c and at time n, $r_c(n, \tau)$ is the ACF, and an exponentially decaying window function $w(i) = (1/\Omega) e^{-i/\Omega}$ was applied to give more emphasis to the most recent

samples [457], [459].[2] It should be noted that the data structure at this stage was three dimensional ($c \times \tau \times n$). Across-channel information integration was then done simply by summing across channels, resulting in a summary ACF

$$s(n, \tau) = \sum_c r_c(n, \tau). \tag{8.11}$$

Pitch at time n was estimated by searching the highest peak in $s(n, \tau)$ within a predefined lag range [457, p. 2884].

Meddis and Hewitt demonstrated that the model was able to predict the perceived pitch for a large set of test stimuli used previously in psychoacoustic tests [457]. Moreover, Meddis and O'Mard later noted that the implementation is a special case of a more general model consisting of four stages: (i) cochlear bandpass filtering, (ii) half-wave rectification and lowpass filtering, (iii) within-channel periodicity extraction, and (iv) across-channel aggregation of periodicity estimates [459]. This became known as the *unitary model* of pitch perception because the single model was capable of simulating a wide range of pitch perception phenomena. Different variants of the unitary model have been used since then in a number of signal analysis systems [171], [133], [627], [677].

Cariani and Delgutte carried out a direct experiment to find out the characteristics in the auditory nerve signals that correlate with the perceived pitch [67]. Instead of using a simulated cochlea, the authors studied the signal in the auditory nerve of a cat in response to complex acoustic waveforms. They found that the *time intervals* between neural spikes are particularly important in encoding pitch. The authors computed histograms of time intervals between both successive and non-successive impulses in individual auditory nerve fibres, and summed the histograms of 507 fibres to form a pooled histogram. What the authors noticed was that, for a diverse set of audio signals, the perceived pitch correlated strongly with the most frequent interspike interval in the pooled histogram at any given time [67]. This suggests that the pitch of these signals could result from central auditory processing mechanisms that analyse interspike interval patterns. Computational models of the cochlea do not produce discrete neural spikes but rather real-valued signals $z_c(n)$, which represent the probability of a neural firing (in different nerve fibres). However, Cariani and Delgutte noted that the interspike interval codes are closely related to autocorrelation operations [67, p. 1712]. For a real-valued signal, ACF can replace the interval histogram.

Despite the above strong evidence, it seems that the ACF is not *precisely* the mechanism used for periodicity estimation in the central auditory system, but some experimental and neurophysiological findings contradict the ACF (see e.g. [322] and the brief summary in [131, p. 1262]). Meddis and Hewitt, for example, used the ACF but wanted to 'remain neutral about the exact

[2] In practice, the windowing and summing can be implemented very efficiently using a leaky integrator.

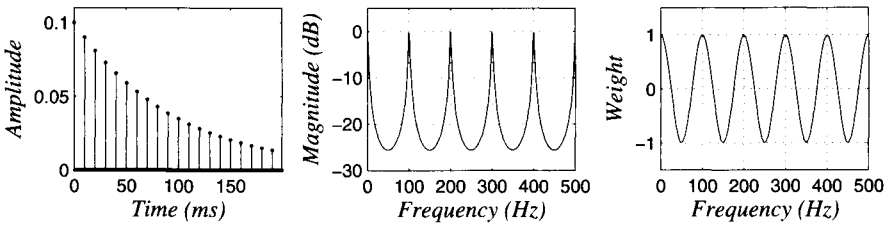

Fig. 8.7. The impulse response (*left*) and frequency response (*middle*) of a comb filter with the feedback delay of 10 ms and feedback gain 0.9. For comparison, the right panels shows the power response of the ACF for 10 ms lag.

mechanism by which temporal information is extracted from the activity of the auditory nerve fibres' [457, p. 2879].

A number of alternative mechanisms to the ACF have been proposed [500], [68], [66], [132]. Although none of these really surpass the modelling power of the ACF for a large class of signals, *comb filter*-like solutions have been proposed by several authors and are therefore discussed in the following. The output of a comb filter for an input signal $z_c(n)$ is given by

$$y_c(n,\tau) = (1-\alpha)z_c(n) + \alpha y_c(n-\tau,\tau), \tag{8.12}$$

where τ is the feedback delay and $0 < \alpha \leq 1$ is the feedback gain.

Figure 8.7 shows the impulse response and the frequency response of a comb filter with a feedback delay $\tau = 10$ ms. For comparison, the power reponse of the ACF for the corresponding lag τ is shown in the rightmost panel.[3] As can be seen, the comb filter is more sharply tuned to the harmonic frequencies of the period candidate and no negative weights are applied between these.

Periodicity analysis with comb filters can be accomplished by invoking a bank of such filters with different feedback delays τ and by computing locally time-averaged powers at the outputs of the filters. Figure 8.8 illustrates the output powers of a bank of comb filter for a couple of test signals. In the case of a periodic signal, all comb filters that are in rational-number relations to the period of the sound show response to it, as seen in panel (b).

A bank of comb filters has been proposed for auditory processing e.g. by Cariani [66, Eq. (1)], who used the filterbank to separate concurrent vowels with different F0s. Cariani also proposed a non-linear mechanism which consisted of an array of delay lines, each associated with its characteristic delay and a non-linear feedback mechanism instead of the linear one in (8.12). Periodic sounds were reported to be captured by the corresponding delay loop and thus became segregated from the mixture signal. The *strobed temporal*

[3] As a non-linear operation, the ACF does not have a frequency response. However, since the ACF of a time-domain signal is the inverse Fourier transform of its power spectrum, the power response of the ACF can be depicted for a single period value.

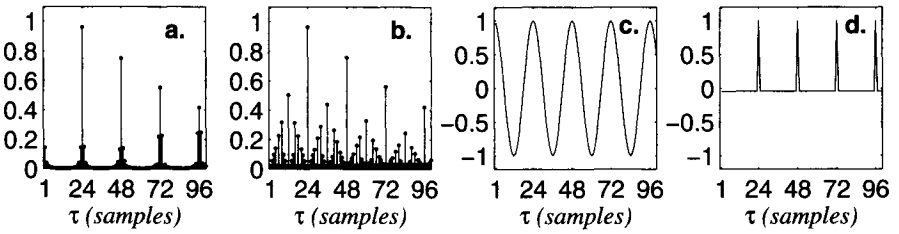

Fig. 8.8. Normalized output powers of a bank of comb filters for (**a**) a sinusoidal with 24-sample period and (**b**) an impulse train with the same period. The feedback delays of the filters are shown on the x-axis and all the feedback gains were 0.9. The panels (**c**) and (**d**) show the ACFs of the same signals, respectively.

integration (STI) mechanism of Patterson [502], [500, p.186] is closely related to comb filters too, although the relation is less direct and full details of the method are beyond the scope of this chapter.

In all the above-described models, the across-band integration has received a rather small role. For example, Meddis and Hewitt [457] and Cariani and Delgutte [67] suggest simply summing the autocorrelation functions or periodicity histograms across channels (see (8.11)). More complex ways of integrating the information across channels have been proposed, though. These will be discussed in more detail in Section 8.5.1, in connection with the estimation of multiple pitches. In particular, a technique called *channel selection* will be discussed which attempts to identify the spectro-temporal regions that represent the target sound and to reject the channels which contain noise or interference. Here it suffices to note that the across-channel information integration takes place in the central auditory system and may thus employ almost any complex technique. One curious consequence of this is that the pitch of a sound can be perceived even when two overtone partials of the sound are fed to the different ears of a listener [298].

8.4 Using an Auditory Model as a Front End

This section discusses music transcription systems which use an auditory model as a front end. That is, the systems apply a perceptually motivated data representation but the emphasis is laid on higher-level processing instead of proposing changes to the auditory model itself. Section 8.5 will discuss systems which do the latter and, as will be seen, often some practical modifications are needed in order to make the models more robust in polyphonic music signals. However, putting transcription systems under these two sections primarily serves the purpose of presentation instead of representing two clear categories.

The intermediate data representations employed between an input signal and the transcription result are of great importance. An appropriate representation facilitates the design of algorithms that use it and often improves

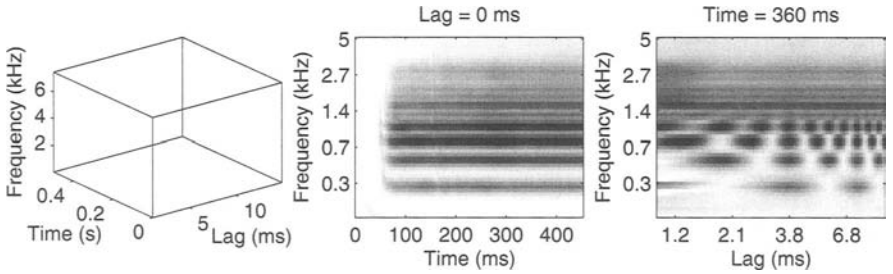

Fig. 8.9. Illustration of the log-lag correlogram of Ellis [171]. Input signal in this case was a trumpet sound with F0 260 Hz (fundamental period 3.8 ms). The left panel illustrates the three-dimensional correlogram volume. The middle panel shows the zero-lag face of the correlogram which is closely related to the power spectrogram. The right panel shows one time slice of the volume, from which the summary ACF can be obtained by summing over frequency.

the analysis result in practice. The idea of using the same data representation as the human auditory system is therefore very appealing. The aim of this section is to investigate the advantages and disadvantages of doing this and to introduce a few auditory-model implementations that have been employed. For this purpose, three different music transcription systems are briefly introduced. A discussion of other mid-level data representations in acoustic signal analysis can be found in [173] and in Chapter 3.

8.4.1 Martin's Transcription System

Martin proposed a system for transcribing piano performances of four-voice Bach chorales [440], [439]. As a front end of his system, Martin used the log-lag correlogram model of Ellis [171] which is closely related to the unitary model of Meddis and Hewitt described above. A bank of 40 gammatone filters was applied, the output of each filter was half-wave rectified and lowpass filtered, and then subjected to autocorrelation analysis. Specific to Ellis's model is that the within-channel ACFs are computed only for a set of logarithmically distributed lag values, 48 lags per an octave. This makes it computationally feasible to estimate the ACFs continuously over time and not just in discrete frames. For each lag τ and channel c, the signal $\hat{r}_c(n,\tau) = z_c(n)z_c(n-\tau)$ is computed and then lowpass filtered in the time dimension, analogous to (8.10). Summary ACFs are obtained by normalizing each ACF by the value at lag zero and by summing across channels. Figure 8.9 illustrates Ellis's model.

Martin utilized the good time resolution of Ellis's model by tracking summary ACF peaks through time and by combining temporally continuous peaks into musical notes. Simple pruning mechanisms were introduced to eliminate spurious subharmonic peaks in the summary ACF.

The overall system of Martin's was a complex inference architecture (a blackboard) where knowledge about the spectral structure of harmonic sounds was combined with rules governing tonal music and with heuristic techniques. Support for different F0s was sought for in the summary ACF and then combined with the power envelope information to create note hypotheses. Much of the innovative work was put into developing an extendable software architecture which allowed the integration of various types of processing modules to the system.

The first version of Martin's system simply used a time-frequency spectrogram as its input [440], but later the author switched to using the auditory model [439]. Interestingly, Martin mentions a specific reason for switching to an auditorily motivated data representation: he suspected that the log-lag correlogram would facilitate the detection of notes in an octave relationship without introducing explicit instrument models. Although some evidence for this was presented, no extensive simulations were carried out to support this conclusion. Also, Martin reported that the correlogram representation indicated chord roots very clearly and that the analysis did not require resolving individual higher-order harmonic partials in the spectrum [439, p. 10]. Although Martin's transcription system was never formally evaluated, it was among the first systems to be able to process signals with more than two simultaneous sounds and thus had a strong influence on subsequent research.

8.4.2 Auditory Scene Analysis Approach of Godsmark and Brown

Godsmark and Brown proposed a system for modelling the auditory scene analysis (ASA) function in humans, that is, our ability to perceive and recognize individual sound sources in mixture signals [215]. The authors used music signals as their test material. ASA is usually viewed as a two-stage process where a mixture signal is first *decomposed* into time-frequency components of some kind, and these are then *grouped* to their respective sound sources. In humans, the grouping stage has been found to depend on various acoustic properties of the components, such as their harmonic frequency relationships, common onset times, or synchronous frequency modulation [49].

Godsmark and Brown used the auditory model of Cooke [100] for the decomposition stage. This auditory model also uses a bank of gammatone filters at its first stage. Notable in Cooke's model is that rectification and lowpass filtering are not applied at the filterbank outputs but only the compression and level adaptation properties of the IHCs are modelled, amounting to an auditorily motivated bandwise gain control. Thus the overall model can actually be viewed as a sophisticated way of extracting sinusoidal components from an input signal, instead of being a complete and realistic model of the auditory periphery. The frequency of the most prominent sinusoidal component at the output of each auditory filter is tracked through time using median-smoothed instantaneous-frequency estimation [100, p. 36] and, in addition, the instantaneous amplitudes of the components are calculated. Since the passbands of

the gammatone filters overlap, usually several adjacent filters show response to the same frequency component. This redundancy is removed by combining the outputs of adjacent channels so as to form 'synchrony strands' which represent the time-frequency behaviour of dominant spectral components in the input signal.

The main focus in the work of Godsmark and Brown was on developing a computational architecture which would facilitate the integration of different spectral organization (grouping) principles [215]. The synchrony strands were used as the elementary units that were grouped to sound sources. The authors reported that these were particularly suitable for modelling the ASA because the temporal continuity of the strands is made explicit and they are sufficiently few in number to perform the grouping for every strand.[4] Godsmark and Brown computed various acoustic features for each strand and then performed grouping according to onset and offset synchrony, time-frequency proximity, harmonicity, and common frequency movement.

Godsmark and Brown evaluated their model by investigating its ability to segregate polyphonic music into its constituent melodic lines. This included both multiple F0 estimation and organization of the resulting notes into melodic lines according to the applied musical instruments. The latter task was carried out by computing pitch and timbre proximities between successive sounds. Although transcription accuracy as such was not the main goal, promising results were obtained for musical excerpts with polyphonies ranging from one to about four simultaneous sounds.

8.4.3 Marolt's Transcriber for Piano Music

Marolt proposed a system for the automatic transcription of piano music [434]. His system was composed of two main parts: a partial tracking module and a note recognition module. Input to the partial tracking part was provided by a model of the peripheral hearing where an input signal was passed through a bank of 200 gammatone filters and the output of each filter was processed by Meddis's IHC model [456]. Adaptive oscillators were then used to track partials at the outputs of the IHC models, one oscillator per channel. The oscillators employed were similar to those proposed by Large and Kolen in [391], locking their period and phase to the incoming signal. In order to track harmonically related partials, the oscillators were interconnected to *oscillator nets*, one per each candidate musical note.

Time-delay neural networks (NNs) were trained to recognize musical notes at the output of the partial tracking module. Each NN was specifically trained to recognize a certain piano note in its input. The input to the NNs consisted of the outputs of all the oscillator networks in a few recent time frames and of the amplitude envelopes at the outputs of the auditory filterbank. Supervised learning with a large amount of piano music was used to train the NNs.

[4] Cooke designed his model exactly for this purpose: to support the grouping activities in ASA [100, p. 14].

Good transcription results were reported for a test set of three real and three synthesized piano performances. Concerning the use of the auditory model, Marolt reported that the compression and level adaptation properties of Meddis's IHC model were important to the system as they reduced the dynamic range of the signal and thus enabled the system to track small-amplitude partials.

8.4.4 Summary of Using an Auditory Front End

Specific advantages of using a perceptually motivated data representation were reported in the above systems. Martin observed that the log-lag correlogram is a good indicator of chord roots and that the analysis with the model does not require resolving individual higher-order harmonics, allowing a better time resolution. Some evidence for detecting two notes in an octave relationship was presented. Godsmark and Brown reported that the model of Cooke was particularly suitable for computational ASA since it produced temporally continuous sinusoidal components which were relatively few in number. Marolt reported that the dynamic compression and level adaptation properties of Meddis's IHC model facilitated the use of small-amplitude partials in the analysis. Finally, an important feature of auditory models that is not explicitly mentioned by any of the above authors is that the compression properties of the IHC models remove timbral information efficiently and thus make the models more robust for different musical instruments.

The disadvantages of employing an auditory model were not specifically reported. However, compared to the use of the Fourier spectrum, for example, it is fair to say that the computational load of an auditory model is significantly higher and that the output of the model is not as straightforward to interpret and understand.

8.5 Computational Multiple F0 Estimation Methods

The pitch perception models described in Section 8.3 are not sufficient as such for accurate multiple F0 estimation in real-world music signals. The purpose of this section is to describe different approaches to extending the models so that they become applicable in the present task.

The most obvious shortcoming of the pitch perception models is that they typically account for a single pitch only. Several pitches in a mixture signal cannot be detected simply by picking several local maxima in the summary ACF, for example. The models have been tested using very diverse kinds of acoustic signals but usually not with sound mixtures. Another shortcoming, related to the first one, is that the models are not robust in polyphonic signals. Even the global maximum of the summary ACF does not necessarily correspond to any of the actual pitches in a mixture signal; certain pitch relationships can confuse the model. In a typical situation, the constituent notes

of a musical chord match the overtones of a non-existing chord root and the highest peak in the summary ACF indicates the chord root instead of one of the component sounds.[5] Further, the pitch models do not address robustness against additive noise: drum sounds often accompany the pitched sounds in music. Finally, the computational complexity of the models is rather high since they involve periodicity analysis at a large number of sub-bands.

On the other hand, there are several issues that are quite efficiently dealt with using a pitch model. These were summarized in Section 8.4.4 above.

In the following, a number of different methods are described that aim at overcoming the above-mentioned shortcomings. Some of these were designed for two-speaker speech signals but are included here in order to cover the substantial amount of work done in the analysis of multiple-speaker speech signals. This is followed by a more detailed description of two multiple F0 estimation methods for music signals. It should be noted that the main interest in this section is not to model hearing but to address the practical task of multiple F0 estimation.

8.5.1 Multiple F0 Estimation in Speech Signals

Multiple F0 estimation is closely related to sound separation. An algorithm that is able to estimate the F0 of a sound in the presence of other sounds is, in effect, also assigning the respective spectral components to their sound sources [49, p. 240]. Separation of speech from interfering speech for the purpose of its automatic recognition is an important area of sound separation. Here we look at methods that have utilized pitch information to carry out this task. A couple of state-of-the-art methods are described, with the aim of discussing the basic mechanisms that have been used to extend an auditory model to process multiple pitches.

Multiple F0 estimation in speech signals is in many ways a more constrained task than in music: the F0 range is limited to about three octaves and the described methods attempt to estimate only two simultaneous F0 tracks. However, the described basic mechanisms are not restricted to speech signals, and many of them can be generalized to the case of more than two simultaneous sounds.

Channel Selection

Meddis and Hewitt extended their pitch model (see p. 241) to simulate the human ability to identify two concurrent vowels with different F0s [458]. The proposed method included a template-matching process to recognize the vowels too, but here only the F0 estimation part is summarized. It consists of the following steps:

[5] Examples of such chords are the major triad and the interval of a perfect fifth.

1. The pitch model of Meddis and Hewitt is applied [457]. This involves a bank of gammatone filters, Meddis's IHC simulation, within-channel ACF computation, and across-channel summing. The highest peak in the summary ACF within a predefined lag range is used to estimate the F0 of the more dominant sound.
2. Individual channel ACFs that show a peak at the period of the first detected F0 are removed. If more than 80% of the channels get removed, only one F0 is judged to be present and the algorithm terminates.
3. The ACFs of the remaining channels are combined into a new summary ACF from which the F0 of the other vowel is derived.

The authors did not give statistics on the F0 estimation accuracy, but reported clear improvements in vowel recognition as the F0 difference of the two sounds was increased from zero to one semitone or beyond.

Time-Domain Cancellation

The above channel selection scheme can be seen as an instance of a more general iterative approach where F0 estimation is followed by the cancellation of the detected sound from the mixture, and the estimation is then repeated for the residual signal. This generalization was pointed out by de Cheveigné, who further proposed that the cancellation can take place in the time domain [129], [130]. When the period τ_0 of one sound in the mixture has been found, the sound can be removed by applying a cancellation filter with the impulse response

$$h_{\tau_0}(n) = \delta(n) - \delta(n - \tau_0), \tag{8.13}$$

where $\delta(n)$ is the unit impulse function. Convolving an input signal $x(\tau)$ with $h_{\tau_0}(n)$ yields $h_{\tau_0}(n) \otimes x(n) = x(n) - x(n - \tau_0)$ and, if the detected sound is perfectly periodic, the above filter completely removes it from the mixture. As a side-effect, however, the filter also removes the partials of other sounds that coincide with those of the sound being cancelled. Also, a more sophisticated filter is needed to cancel a sound whose period is not precisely a multiple of the sampling interval [382].

An advantage of the time-domain cancellation is that it is not bound to the resolution of the cochlear filterbank and, in principle, it works even when all the channels are dominated by a single period. The filtering can be done directly for the input signal or within the channels of an auditory model. These two are equivalent unless the within-channel filtering is done after the non-linear IHC simulation stage.

De Cheveigné used the cancellation principle for the actual F0 estimation, too. He proposed to calculate a squared difference function (SDF) which is defined for an input signal $x(n)$ as

$$\mathrm{SDF}(n, \tau) = \sum_{i=0}^{N-1} (x(n - i) - x(n - i - \tau))^2, \tag{8.14}$$

where N is the analysis frame size [131].[6] By expanding the square, it can be seen that $\mathrm{SDF}(n,\tau) = E(n) + E(n-\tau) - 2r(n,\tau)$, where $E(n)$ denotes the signal power at time n and $r(n,\tau)$ is the ACF. Thus the SDF and the ACF are functionally equivalent, and period estimation can be carried out by searching for minima in the SDF instead of maxima in the ACF. De Cheveigné also proposed a joint cancellation model, where two cancellation filters with periods τ_A and τ_B were applied in a cascade so as to cancel two periodic sounds. By computing the power of the resulting signal as a function of the two periods, the F0s were found by locating the minumum of the two-dimensional function [129], [133].

De Cheveigné evaluated both the iterative and the joint F0 estimation method for mixtures of two-voiced speech segments [129]. The iterative algorithm was reported to produce estimates which were correct within 3% accuracy in 86% of the frames and the exhaustive joint estimator produced correct estimates in 90% of the frames. Computational complexity is a drawback of the joint estimator.

Channel and Peak Selection

Wu, Wang, and Brown proposed an algorithm for tracking the F0s of two simultaneous speakers, taking particular interest in noise robustness [677]. Their method employed a computational model of the peripheral auditory system, after which the channels significantly corrupted by noise were excluded. From the remaining channels, ACF peaks were selected so that peaks judged to give misleading information were rejected. This led to an intermediate data representation which consisted of only the lag values and channel labels of the selected ACF peaks (discarding peak amplitudes). The information was then processed using statistical models.

In more detail, the channel and peak selection process was the following. First, a gammatone filterbank was applied and the resulting channels were classified as 'low-frequency' or 'high-frequency' channels depending on whether their centre frequency was below or above 800 Hz. Normalized ACFs were then computed for the low-channel signals directly and for the amplitude envelopes of the high-channel signals. Low channels were selected (i.e., included in further computations) if the highest peak of the normalized ACF exceeded a given threshold value. High-frequency channels were selected if the shapes of the normalized ACFs computed in 16 ms and in 32 ms frames were sufficiently similar. Peak selection, in turn, consisted of two main rules. First, an acceptable peak (peak not due to noise) was required to show a submultiple peak at twice its lag value. At high-frequency channels, envelope beating at the F0 rate was assumed and, therefore, subharmonics of any peak higher than a threshold value were removed. Full details can be found in [677].

[6] The SDF is closely related to the average magnitude difference function (AMDF) that has been used to estimate the F0 of speech [549]. The AMDF is obtained by summing absolute values instead of their squares in (8.14).

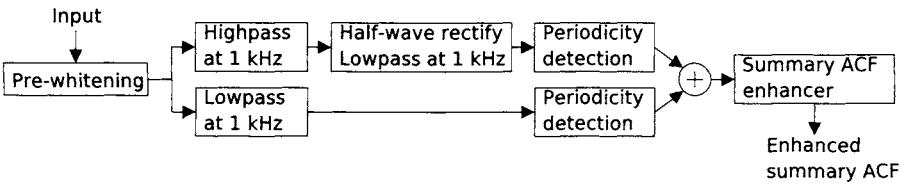

Fig. 8.10. Block diagram of the pitch analysis method proposed by Karjalainen and Tolonen [627]. ©2005 IEEE, reproduced here by permission.

The remaining channels and peaks were subjected to statistical modelling. Using clean speech as training material, the difference $\delta_c = \tau_c - \tau_0$ between a true (annotated) fundamental period τ_0 and the period of the closest selected peak τ_c at channel c was studied. The statistical distribution of δ_c was used to determine the likelihood of the observed peaks at channel c given a fundamental period candidate τ. Different observation likelihood functions were defined for the cases of zero, one, and two F0s (two F0s were jointly estimated). Finally, a hidden Markov model was employed to model the dynamic aspects of the F0 contours. This included both the continuity of the F0 tracks and jump probabilities between the state spaces of zero, one, or two F0s.

In evaluations, Wu et al. used ten voiced utterances to generate mixtures of two voices. These were mixed with realistic noise signals, including harmonic interference and interfering speech signals. Five utterances were used for training and five for testing. Good results were reported for this database and an implementation of the method is publicly available [677].

8.5.2 Multiple F0 Estimator of Karjalainen and Tolonen

Karjalainen and Tolonen proposed a computationally efficient version of the unitary pitch model (see p. 241) and extended it to the multiple F0 estimation of musical sounds. [327], [627] Figure 8.10 shows the block diagram of their method. The most obvious difference from the original auditory model is that the method divides an input signal into two channels only, below and above 1 kHz, and then analyses the periodicity of the low-channel signal and of the envelope of the high-channel signal. Despite the drastic reduction in computation load compared to the unitary pitch model, many important characteristics of the model were preserved.

The method included several features to address practical robustness issues. Robustness against timbral variation (different musical instruments for example) was achieved by *pre-whitening* the input signal using inverse warped-linear-prediction filtering [272]. In essence, this flattens the spectral energy distribution but does not affect the spectral fine structure.

Periodicity analysis in the method of Karjalainen and Tolonen was carried out using a *generalized ACF*, originally proposed by Indefrey et al. in [306]. According to the Wiener–Khintchine theorem, the ACF of a time-domain

signal $\mathbf{x}$ is the inverse Fourier transform of its power spectrum [276, p. 334]. The generalized ACF, then, is defined as

$$\hat{r}(\tau) = \mathrm{IDFT}(|\mathrm{DFT}(\mathbf{x})|^{\alpha}), \tag{8.15}$$

where DFT and IDFT denote the discrete Fourier transform and its inverse, and α is a free parameter which determines the frequency domain compression.[7] The standard ACF is obtained by substituting $\alpha = 2$. Definition of the *cepstrum* of $\mathbf{x}$ is analogous to ACF and is obtained by replacing the second power with the logarithm function. The difference between the ACF and cepstrum-based F0 estimators is quantitative: raising the magnitude spectrum to the second power emphasizes spectral peaks in relation to noise but, on the other hand, further aggravates spectral peculiarities of the target sound. Applying the logarithm function causes the opposite for both. And indeed, ACF-based F0 estimators have been reported to be relatively noise immune but sensitive to formant structures in speech, and vice versa for cepstrum-based methods [535]. As a trade-off, Karjalainen and Tolonen suggested using the value $\alpha = 0.67$.

Extension to multiple F0 estimation was achieved by cancelling subharmonics in the summary ACF (SACF) by clipping the SACF to positive values, time-scaling it to twice its length, and by subtracting the result from the original clipped SACF. This cancellation operation was repeated for time-scaling factors up to about five. From the resulting *enhanced SACF*, all F0s were picked without iterative estimation and cancellation. In more detail, the enhancing procedure was as follows:

Algorithm 8.1: Enhancing Procedure of Karjalainen and Tolonen

1. The enhanced SACF $\tilde{s}(\tau)$ is initialized to be equal to the SACF $s(\tau)$. The scaling factor m is initialized to value 2.
2. The original SACF is time-scaled to m times its length and the result is denoted by $s_m(\tau)$. Using linear interpolation,

$$s_m(\tau) = s(d) + \frac{\tau - md}{m}\left(s(d+1) - s(d)\right), \tag{8.16}$$

   where $d = \lfloor \tau/m \rfloor$ and $\lfloor \cdot \rfloor$ denotes rounding towards negative infinity.
3. The enhanced SACF is updated as

$$\tilde{s}(\tau) \leftarrow \max(0, \tilde{s}(\tau) - \max(0, s_m(\tau))). \tag{8.17}$$

4. Increment m by 1. If m is smaller than 6, return to Step 2.

The above enhancing procedure is surprisingly efficient in removing spurious peaks from the SACF and in revealing more than one F0 in it. Also,

[7] In practice, the analysis frame $\mathbf{x}$ has to be zero-padded to twice its length before the first transform.

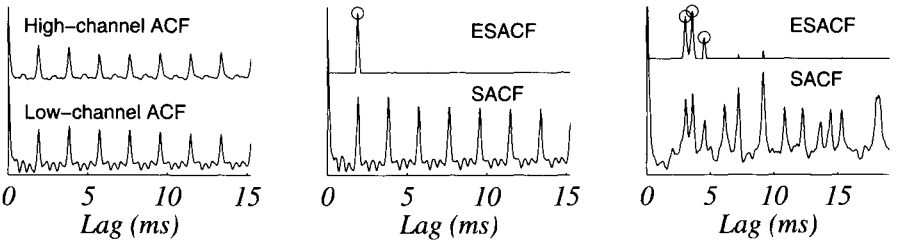

Fig. 8.11. Left: The ACFs at the low and the high channel for a violin sound (F0 523 Hz). Middle: SACF and enhanced SACF for the same sound. Right: SACF and enhanced SACF for a major triad chord played by the trumpet (F0s 220 Hz, 277 Hz, and 330 Hz). The circles indicate the correct fundamental periods.

it partly solves the 'chord root' problem mentioned in the beginning of Section 8.5 since the enhancing procedure scales the true F0 peaks to the position of the chord root and, if a note does not truly appear at the root, the spurious peak becomes cancelled. The only place where care has to be taken is in setting values of the original SACF to zero in the lag range $[0, f_s/1000 \text{ Hz}]$ before the enhancing (here f_s denotes the sampling rate). This ensures that the values on the $\tau = 0$ hill do not spread and wipe away important information. Zeroing the mentioned lags causes no harm for the analysis since the algorithm cannot detect F0s above 1 kHz.

Figure 8.11 illustrates the enhancing procedure for an isolated sound and for a musical chord. As mentioned by Martin [439], the SACF indicates the non-existing F0 of the chord root in the latter case. After enhancing, however, the true F0s are revealed.

Overall, the method of Karjalainen and Tolonen is quite accurate and it has been described in sufficient detail to be exactly implementable based on [627] and on the Matlab toolbox for frequency-warped signal processing by Härmä et al. [272]. A drawback of the method as stated by the authors is that it is 'not capable of simulating the spectral pitch' [627, p. 713], i.e., the pitch of a sound whose first few harmonics are above 1 kHz. In practice, the method is most accurate for F0s below about 600 Hz. Later, Karjalainen and Tolonen also proposed an iterative approach to multiple F0 estimation using the described simplified auditory model [328].

8.5.3 Multiple F0 Estimator of Klapuri

Klapuri's multiple F0 estimator for music signals was originally described in [353, Ch. 4] and later improved and simpified in [354]. The method consists of a model of the peripheral auditory system followed by a periodicity analysis mechanism where F0s are iteratively estimated and cancelled (Fig. 8.12).

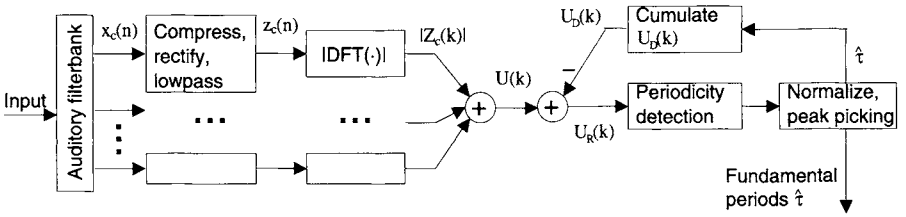

Fig. 8.12. Block diagram of the multiple F0 estimator of Klapuri [354]. ©2005 IEEE, reproduced here by permission.

Model of the Peripheral Auditory System

In the peripheral hearing model, an input signal was first passed through a bank of gammatone filters with centre frequencies uniformly distributed on the critical-band scale (see (8.5)) between 60 Hz and 5.2 kHz. A total of 72 filters were employed using the implementation of Slaney [591].

Hair cell transduction was modelled by compressing, half-wave rectifying, and lowpass filtering the sub-band signals. The compression was implemented by simulating the full-wave νth law compression (FWC), which is defined as

$$\mathrm{FWC}(x) = \begin{cases} x^{\nu}, & x \geq 0, \\ -(-x)^{\nu}, & x < 0. \end{cases} \tag{8.18}$$

For a narrow-band signal, such as the output of an auditory filter, the effect of the FWC within the passband of the filter can be accurately modelled by simply scaling the signal with a factor

$$\gamma_c = a(\sigma_c)^{\nu-1}, \tag{8.19}$$

where σ_c is the standard deviation of the signal at channel c and the scalar a depends on ν but is common to all channels and can thus be omitted [353, p. 37]. In addition to the scaling mentioned, FWC generates small-amplitude distortion components at odd multiples of the channel centre frequency. These were avoided by using the model (8.19) instead of (8.18) directly.

The FWC provides a single parameter ν which determines the degree of spectral whitening applied on an input signal. The scaling factors γ_c normalize the variances of the sub-band signals towards unity when $0 \leq \nu \leq 1$. Here, the value $\nu = 0.33$ was applied.

The compressed sub-band signals were half-wave rectified by constraining negative values to zero. As shown in Fig. 8.6, this generates spectral components near zero frequency and on twice the channel centre frequency. The rectified signal at each channel was steeply lowpass filtered with a cut-off frequency 1.5 times the channel centre frequency in order to attenuate the distortion spectrum at twice the centre frequency but to pass the sub-band signal along with its amplitude envelope spectrum. The rectified and lowpass filtered signals $z_c(n)$ were then subjected to periodicity analysis.

Periodicity Analysis

The periodicity analysis mechanism proposed by Klapuri is best understood by comparing it with the ACF-based method employed by Meddis and Hewitt (see p. 241). Short-time ACF estimates within the channels can be efficiently computed as $r_{c,n}(\tau) = \text{IDFT}(|Z_{c,n}(k)|^2)$, where IDFT denotes the inverse Fourier transform and $Z_{c,n}(k)$ is the Fourier transform of $z_c(n)$ computed in a time frame that is centred at time n and zero-padded to twice its length before the transform. The within-band ACFs are then summed to obtain the summary ACF, $s_n(\tau) = \sum_c r_{c,n}(\tau)$.

Because the IDFT and the summing are linear operations, their order can be reversed and we can write $s_n(\tau) = \text{IDFT}(S_n(k))$, where $S_n(k) = \sum_c |Z_{c,n}(k)|^2$. The spectra of real-valued (audio) signals are conjugate symmetric and the IDFT can therefore be written out as

$$s_n(\tau) = \text{IDFT}(S_n(k)) = \frac{2}{K} \sum_{k=0}^{K/2-1} \cos\left(\frac{2\pi\tau k}{K}\right) S_n(k), \tag{8.20}$$

where K is the length of the transform frame after the zero-padding.

Klapuri made three modifications to (8.20). First, as seen in Fig. 8.12, magnitude spectra were summed across channels instead of power spectra. Analogous to the generalized ACF in (8.15), it was observed that raising the magnitude spectra to the second power accentuates timbral peculiarities that cannot be completely removed by bandwise compression in polyphonic signals. Therefore, within-band magnitude spectra were summed to obtain a summary magnitude spectrum (SMS),

$$U(k) = \sum_c |Z_c(k)|, \tag{8.21}$$

where the time index n has been omitted to simplify the notation in the following. The SMS functioned as an intermediate data representation and all the subsequent processing took place using it only.

Figure 8.13 illustrates the bandwise magnitude spectra $|Z_c(k)|$ for a saxophone sound. As can be seen, the within-channel rectification maps the contribution of higher-order partials to the position of the F0 and its few multiples in the spectrum. Most importantly, the degree to which an individual overtone partial j is mapped to the position of the fundamental increases as a function of j. This is because the auditory filters become wider at higher frequencies and the partials thus have larger-magnitude neighbours with which to generate the difference frequencies (beating) in the envelope spectrum. Klapuri's method was largely based on this observation, as will be explained below.

The second modification concerned the function $\cos(\cdot)$ in (8.20), which can be seen as a harmonic template that picks overtone partials of the frequency K/τ in the spectrum (see the rightmost panel of Fig. 8.7 on p. 243). The function was replaced by a response that is more sharply tuned to the frequencies

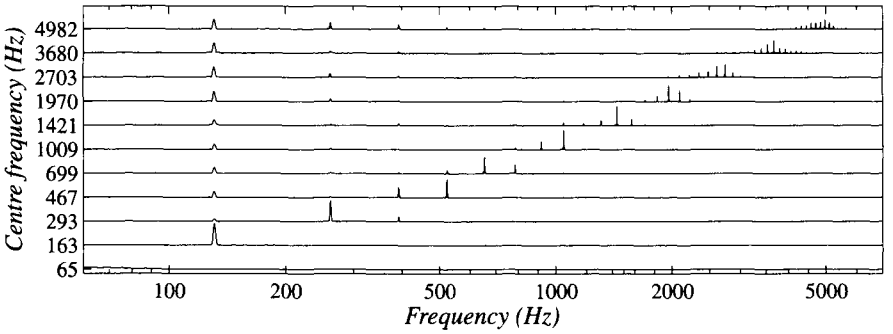

Fig. 8.13. The spectra $|Z_c(k)|$ at a few channels for a tenor saxophone sound (F0 131 Hz).

of the harmonic overtones of a F0 candidate and employs no negative weights between the partials. In practice, the frequency response resembled that of a comb filter shown in Fig. 8.7. This modification alleviates the interference of other, co-occurring sounds. Moreover, instead of pointwise multiplying the complete spectrum $U(k)$ with a comb filter response and then summing, it was found sufficient to sum up spectral components near the positions of the peaks of the comb-filter response (see (8.22) below). This led to a very efficient implementation computationally and is closely related to the *harmonic selection* methods reviewed by de Cheveigné in [129], and to the harmonic transform of Walmsley et al. [657].

The relative strength, or *salience*, $\lambda(\tau)$ of a fundamental period candidate τ was calculated in Klapuri's system as

$$\lambda(\tau) = \frac{f_s}{\tau} \sum_{j=1}^{\tau/2} \left(\max_{k \in \kappa_{j,\tau}} [H_{\mathrm{LP}}(k) U(k)] \right), \tag{8.22}$$

where f_s denotes the sampling rate and the factors f_s/τ and $H_{\mathrm{LP}}(k)$ are related to the third modification to be explained later. The set $\kappa_{j,\tau}$ defines a narrow range of frequency bins in the vicinity of the jth overtone partial of the F0 candidate f_s/τ. More exactly, $\kappa_{j,\tau} = [k_{j,\tau}^{(0)}, k_{j,\tau}^{(1)}]$, where

$$k_{j,\tau}^{(0)} = \lfloor jK/(\tau + \Delta\tau/2) \rfloor + 1, \tag{8.23}$$

$$k_{j,\tau}^{(1)} = \max(\lfloor jK/(\tau - \Delta\tau/2) \rfloor, k_{j,\tau}^{(0)}). \tag{8.24}$$

In the above formulas, K is the transform length and the scalar $\Delta\tau = 1$ denotes spacing between successive period candidates τ. A uniform sampling of lag values was used, analogous to the ACF. Equations (8.23)–(8.24) define the sets $\kappa_{j,\tau}$ so that, for a fixed partial index j, all the spectral components belong to the range of at least one period candidate τ, and the ranges of adjacent period candidates cannot overlap by more than one frequency bin.

The third modification in (8.22) compared to (8.20) is that individual partials in the sum in (8.22) are weighted by $f_s/\tau \times H_{\mathrm{LP}}(k)$, where the lowpass response is

$$H_{\mathrm{LP}}(k) = \frac{1}{0.108 f_s k/K + 24.7}. \tag{8.25}$$

By comparison with (8.4), it is easy to notice that this is the reciprocal of the bandwidth of an auditory filter centred at frequency bin k. The factor $f_s/\tau \times H_{\mathrm{LP}}(k)$ can therefore be written as $F(\tau)/b_c(jF(\tau))$, where $F(\tau) = f_s/\tau$ is the F0 of the period candidate τ (i.e., the frequency interval between its overtones) and $b_c(jF(\tau))$ is the width of an auditory filter centred at its jth overtone. The ratio of these two was interpreted as the *resolvability* of the partial j [353, p. 45]. The lower-order overtones of a harmonic sound are resolved into separate auditory channels, whereas the higher-order overtones go to the same auditory channel with their neighbours and their frequencies cannot be perceived separately (resolved). Actually the lowpass filter $H_{\mathrm{LP}}(k)$ would belong to the within-band IHC modelling stage but, since the filter is the same for all channels, it is equivalent to apply it after the channels have been combined. The higher the centre frequency of an auditory channel, the more the filter attenuates the spectrum at the passband of the auditory filter and thus gives it a smaller weight in relation to the envelope spectrum, which is around zero frequency and not much affected. This corresponds to the fact that, at higher auditory channels, the neural firing activity more and more follows the amplitude envelope of the sub-band signal and not its fine structure—this is directly related to the concept of resolvability. Discrete categorization into 'low' and 'high' channels is not needed.

The degree of resolvability as modelled above (and thus the weight of a partial in the sum in (8.22)) is approximately inversely proportional to the harmonic index j when τ is fixed. As a consequence, the sum in (8.22) can be limited to $j \approx 20$ since weights beyond this are relatively small.

Taken together, the computation of the salience function $\lambda(\tau)$ can be seen as a process where partials are picked from harmonic positions of the spectrum $U(k)$, their magnitudes are weighted by the estimated resolvability $f_s/\tau \times H_{\mathrm{LP}}(k)$, and then summed. What makes all the difference is that the within-channel rectification maps the contribution of higher-order partials to the position of the fundamental and its few multiples in the spectra $Z_c(k)$, and the degree to which an individual overtone partial j is mapped to the position of the fundamental increases as a function of j, as explained above. As a consequence, the whole harmonic series of a sound contributes to its salience, despite the weighting with resolvability.

The above-described benefit of bandwise rectification cannot be overemphasized. Assigning the higher-order partials to their respective sound sources in polyphonic music signals is a nightmare. The rectification operation accomplishes this 'automatically' by mapping the support from higher-order harmonics to the position of F0 and its few multiples in $U(k)$. Figure 8.14

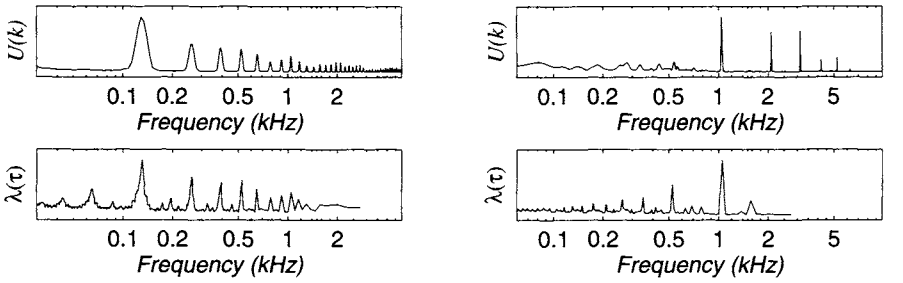

Fig. 8.14. The upper panels show the summary magnitude spectrum $U(k)$ for a saxophone sound with F0 131 Hz (*left*) and a violin sound with F0 1050 Hz (*right*). The lower panels show the corresponding salience functions $\lambda(\tau)$.

illustrates the calculation of $\lambda(\tau)$ for the saxophone sound shown in Fig. 8.13, and for a violin sound with the F0 1050 Hz.

Iterative Estimation and Cancellation

The global maximum of the function $\lambda(\tau)$ was found to be a robust indicator of one of the correct F0s in polyphonic signals. As with most F0 estimators, however, the next-highest salience was often assigned to half or twice that of the first detected F0. Similarly to de Cheveigné (see p. 250), Klapuri employed an iterative technique where F0 estimation was followed by the cancellation of the detected sound from the mixture and the estimation was then repeated for the residual signal. Algorithm 8.2 summarizes the applied technique [354].

Algorithm 8.2: Multiple F0 Estimator of Klapuri

1. A *residual SMS* $U_{\mathrm{R}}(k)$ is initialized to be equal to $U(k)$. A summary spectrum of all detected sounds, $U_{\mathrm{D}}(k)$, is initialized to zero.
2. A fundamental period $\hat{\tau}$ is estimated using $U_{\mathrm{R}}(k)$ and (8.22).
3. Harmonic selection is carried out for the found period $\hat{\tau}$ according to (8.22)–(8.24). However, instead of summing up the magnitude values, the precise frequency and amplitude of each partial is estimated and used to calculate its magnitude spectrum at the few surrounding frequency bins.
4. The magnitude spectrum of the jth partial is weighted by $f_{\mathrm{s}}/\tau \times H_{\mathrm{LP}}(k_j)$ and added to the corresponding position of $U_{\mathrm{D}}(k)$ which represents the cumulative spectrum of all the detected sounds.
5. The residual SMS is recalculated as

$$U_{\mathrm{R}}(k) \leftarrow \max(0, U(k) - dU_{\mathrm{D}}(k)), \tag{8.26}$$

   where $d = 0.5$ controls the amount of the subtraction and is a free parameter of the algorithm.
6. Return to Step 2.

An important characteristic of the Step 4 is that, before adding the partials of a detected sound to $U_{\mathrm{D}}(k)$, they are weighted by their resolvability in the same manner as at the F0 detection stage. As a consequence, the higher-order partials are not entirely removed from the mixture spectrum when the residual $U_{\mathrm{R}}(k)$ is formed. This principle is important in order not to corrupt the sounds that remain in the residual and have to be detected at the coming iterations. The described weighting limits the effect of the cancellation to the lowest harmonics but, as explained above, the higher-order harmonics have been mapped to the position of the fundamental by the rectification and are thus effectively cancelled, too.

8.5.4 Results

Simulation experiments were carried out to evaluate the performance of the method of Tolonen and Karjalainen [627] and that of Klapuri [354]. Implementations of the method of Wu et al. [677] and Marolt [434] are publicly available too, but these would have required a specific experimental setup since the former was designed to process continuous two-speaker speech signals and the latter to transcribe piano music only.

The acoustic material consisted of samples from the McGill University Master Samples collection [487], the University of Iowa website,[8] IRCAM Studio Online,[9] and of independent recordings for the acoustic guitar. There were altogether 32 different musical instruments, comprising brass and reed instruments, strings, flutes, the piano, the guitar, and mallet percussion instruments. The total number of samples (individual notes) was 2842.

Semi-random sound mixtures were generated by first allotting an instrument and then a random note from its playing range. This was repeated to get the desired number of simultaneous sounds, which were then mixed with equal mean-square levels. One thousand test cases were generated for mixtures of one, two, four, and six sounds.

One analysis frame immediately after the onset[10] of the sounds was fed to the multiple F0 method. The number of F0s to extract, i.e., the polyphony, was given along with the mixture signal. A correct F0 estimate was defined to deviate less than 3% from the nominal F0 of the sound, making it round to a correct note on the Western musical scale. Two different error rates were computed. *Multiple F0* estimation error rate was defined as the percentage of all F0s that were not correctly detected in the input signals. In *predominant F0* estimation, only one F0 in the mixture was being estimated and it was defined to be correct if it matched the correct F0 of any of the component sounds.

[8] University of Iowa samples: theremin.music.uiowa.edu/MIS.html

[9] IRCAM Studio Online: soleil.ircam.fr

[10] The onset of the sounds was defined to be at the point where the waveform reached one third of its maximum value during the first 200 ms of its playing.

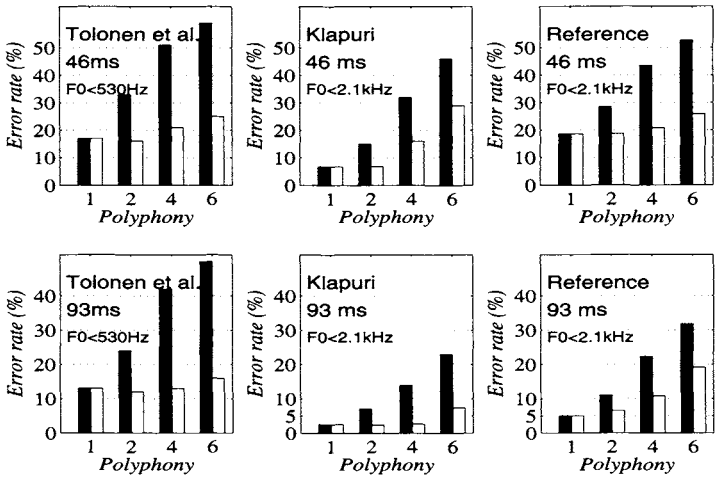

Fig. 8.15. F0 estimation error rates as a function of the number of concurrent sounds (polyphony) for the method of Tolonen and Karjalainen [627], the method of Klapuri [354], and the reference method [351]. The black bars and the white bars show the multiple F0 and the predominant F0 estimation error rates, respectively. The upper panels show the results for a 46 ms analysis frame and the lower panels for a 93 ms frame.

The left-hand panels of Fig. 8.15 show the error rates for the method of Tolonen and Karjalainen in 46 ms and 93 ms analysis frames. The F0 range in these experiments was limited to the three octaves between 65 Hz and 520 Hz, because the accuracy of the method was found to degrade rapidly above 600 Hz (see Section 8.5.2). The black bars show the multiple F0 estimation error rates and the white bars show the predominant F0 estimation error rates. The global maximum of the enhanced SACF was used for the latter purpose. The method performed robustly in polyphonic mixtures, and especially the predominant F0 estimation error rates remained reasonably low even in short time frames and in rich polyphonies. Taking into account the computational efficiency (faster than real-time) and conceptual simplicity of the method, the results are very good.

The middle panels of Fig. 8.15 show the error rates for the method of Klapuri [354]. The first detected F0 was used for the predominant F0 estimation. In these experiments, the pitch range was limited to five octaves between 65 Hz and 2.1 kHz. The method performs robustly in all cases and is very accurate, especially in the 93 ms analysis frame. Computational complexity is a drawback of this method. The calculations are clearly slower than real-time on a 2-GHz desktop computer, the most intensive part being the cochlear filterbank and the within-band DFT calculations.

The right-hand panels of Fig. 8.15 show the error rates for a state-of-the-art reference method proposed by Klapuri in [351]. This method is based on

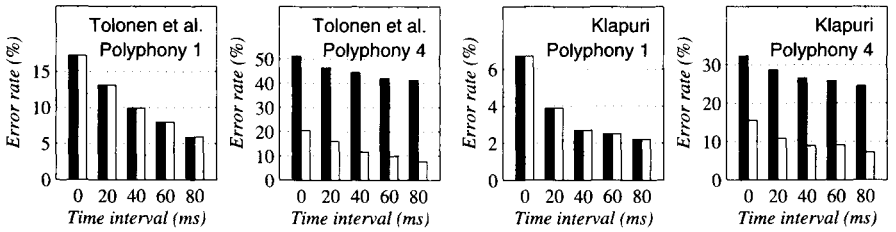

Fig. 8.16. Error rates as a function of the interval between the sound onset and the beginning of a 46 ms analysis frame. The two panels on the left show results for the method of Tolonen and Karjalainen [627] and the two panels on the right for the method of Klapuri [354]. The black bars show multiple F0 estimation error rates and the white bars show predominant F0 estimation error rates.

spectral techniques instead of an auditory model and is therefore a good point of comparison. The test cases given to Klapuri's method [354] and the reference method were identical. It was observed that the reference method requires quite a long analysis frame to resolve and process the overtones of low-pitched sounds, and mallet percussion instruments could not be reliably analysed. In addition to the differences in handling the higher-order overtones, a factor involved is that the frequency resolution of the Fourier spectrum is linear, whereas time-domain periodicity analysis within the auditory channels leads to $1/f$ frequency resolution, which enables more accurate analysis at the lower end of the logarithmic scales applied in music. The reference method is conceptually (technically) the most complex among the three.

An important factor in the above results is that the analysis frames were positioned immediately after the onsets of the sounds. Figure 8.16 shows the error rates of the two methods as a function of the time interval between the sound onset and the beginning of the analysis frame. As can be seen, the error rates improve clearly as the interval increases, and especially the predominant F0 estimation error rates shrink to about a third of the initial values after 80 ms of the onset. This is because the noisy beginning transients of many sounds die off rapidly and F0 estimation becomes easier thereafter. In music signals, however, notes are often short and such an offset cannot be applied. In Fig. 8.15, maximally realistic simulations were of interest and thus a zero offset was applied. Figure 8.16 shows results only for the 46 ms analysis frame, but the general trend is similar (although less pronounced) for the longer frame.

8.5.5 Summary of the Multiple F0 Estimation Methods

The beginning of this section listed several issues where the pitch perception models fall short of being practically applicable multiple F0 estimators. This section summarizes and discusses the various technical solutions that were proposed as improvements.

Two main approaches can be distinguished among the techniques used to extend a single-pitch model to the estimation of multiple pitches: the iterative estimation-and-cancellation approach and the joint estimation approach. Most methods fall into the former category: F0 estimation is done using the summary ACF, for example, and the F0 found is then cancelled before deciding the next one. Meddis and Hewitt performed the cancellation by removing the auditory channels associated with the first detected pitch [458]. De Cheveigné employed within-channel cancellation filtering in the time domain [130]. Klapuri subtracted the partials of a detected sound in the frequency domain and removed only the lower-order partials entirely [354].

Joint estimation methods were proposed by de Cheveigné [129], Karjalainen and Tolonen [327], and Wu et al. [677]. Among these, the method of de Cheveigné was not actually based on an auditory model, but the method applied two cancellation filters in a cascade and searched for such cancellation-filter periods that the output power was minimized. Karjalainen and Tolonen enhanced the summary ACF so that all F0s could be directly extracted from the result. In the method of Wu et al., the distribution of the peaks in the sub-band ACFs was statistically modelled in the cases of zero, one, or two pitches.

The limited robustness of the pitch perception models in polyphonic signals is another important problem addressed by the multiple F0 estimation methods. The chord-root detection problem was mentioned as an example of this. The SACF enhancing technique of Karjalainen and Tolonen [627] is rather efficient in this respect, as illustrated in Fig. 8.11. Klapuri addressed the problem by applying the lowpass response in (8.25), which suppresses the support of higher-order partials to the chord root unless the series of partials has sufficiently uniform amplitudes so as to generate strong beating at the fundamental rate. This is usually not the case if the partials are due to several different sounds (component F0s of a chord). Also, the use of harmonic selection in the frequency domain alleviated the interference of other sounds since the spectrum between the partials was not used in salience calculations. Iterative estimation and cancellation methods that estimate the first F0 directly from the summary ACF suffer from its robustness limitations [458], [130].

Robustness for different sound sources (different musical instruments) is a very important aspect in F0 estimation. Here the pitch perception models are readily very efficient. Meddis's hair-cell model compresses the sub-band signals and results in spectral whitening, that is, removal of timbral information to some extent [457], [456]. Ellis [171] and Klapuri [354] carried out this function by scaling the sub-band signals inversely proportional to their variance. Karjalainen and Tolonen pre-processed the input signals by inverse warped-linear-prediction filtering. This had the advantage that a multi-channel filterbank was not needed [327], [272]. An advantage of all these is that they flatten the spectral energy distribution without raising the noise floor in relation to spectral peaks. The latter happens for example in cepstrum pitch

detection, where the logarithm function is applied bin-by-bin to the magnitude spectrum [535]. The system of Wu et al. [677] is interesting, since in this method, whitening would not have any effect at all because only the lag-values of within-channel ACF peaks are retained and not their amplitudes.

Noise robustness was not discussed in depth in this chapter. In music, percussive instruments and recording imperfections cause noise-like interference for the F0 estimation. Particular emphasis on this issue was laid by Wu et al., who performed channel and peak selection so as to avoid the spectro-temporal regions that were severely corrupted by noise. The authors remarked that they essentially treated multiple F0 tracking and noise robustness as a single problem [677, p. 240]. Karjalainen and Tolonen selected the generalized ACF power so as to make a compromise between noise robustness and spectral flattening [327].

Computational complexity of the pitch models was significantly reduced only in the method of Karjalainen and Tolonen [327]. In the other methods, the most time-consuming operation is typically the peripheral filterbank and the periodicity analysis within channels, usually leading to computation times which are 10 to 100 times slower than that of the method of Karjalainen and Tolonen. Ellis computed the within-channel ACFs only for a set of logarithmically distributed lag values, which allowed the use of a very good time resolution without causing a prohibitative computational load [171]. In the iterative methods, the peripheral analysis usually has to be computed only once [458], [354].

8.6 Conclusions

Pitch perception models and practical F0 estimators address slightly different tasks and are judged according to different criteria. The former should faithfully represent the mechanisms of the human auditory system, whereas the latter are expected to perform accurate multiple F0 estimation by any means available. The main focus of this chapter was on the practical side. However, the two aspects have significantly influenced and benefited each other and this is one reason to study auditory modelling.

Many characteristics of human pitch perception can be traced to the peripheral stages of hearing, as discussed in Section 8.3. In this sense, auditory models have a lot to say about the intermediate data representations used in acoustic signal analysis. A particularly important principle in an auditorily motivated analysis is that the higher-order overtones of a sound are processed collectively within each auditory channel; estimation and separation of individual higher-order partials is not attempted. The cochlear filterbank is 'fair' for different F0 values in this respect since the first few harmonic partials of all F0s are resolved into separate auditory channels, whereas the harmonics above about 10 go to the same channel along with their neighbours and generate amplitude envelope beating at the fundamental rate. This is an efficient

mechanism for dealing with the higher-order overtones in complex polyphonic data. The other advantages and disadvantages of a perceptually motivated data representation were discussed in Section 8.4.4.

Compared with the peripheral stages of hearing, at least an equally important part of pitch perception takes place in the brain (Steps 3 and 4 in the overview on p. 235). These stages are not yet well understood and thus there is a larger variance in the proposed practical techiques as well. The biggest defect of the existing pitch perception models from the music transcription viewpoint is that they have been designed to process isolated sounds instead of polyphonic signals. Different techniques for transforming a pitch model into a multiple F0 estimator were described in Section 8.5. Both iterative methods and joint estimation methods were discussed, and different ways of cancelling a detected F0 from the mixture signal were described. Periodicity analysis techniques were presented that applied the ACF [458], [130], the generalized ACF [627], statistical modelling of the ACF peaks [677], adaptive oscillators [434], or simulation of comb filters in the frequency domain [354]. For now, none of the described methods can claim to be the 'right' or the optimal one, but they provide a wealth of technical solutions and approaches to build upon.