Draft to appear in "Computational Auditory Scene Analysis", edited by DeLiang Wang and Guy J. Brown, John Wiley and sons, ISBN 0-471-45435-4, in press.

CHAPTER 1

MULTIPLE F0 ESTIMATION

1.1 INTRODUCTION

This chapter is about the estimation of multiple fundamental frequencies (F_0) from a waveform such as the compound sound of several people speaking at the same time, or several musical instruments playing together. That information may be needed to transcribe the music to a score, to extract intonation patterns for speech recognition, or as an ingredient for *computational auditory scene analysis*. The task of estimating the single F_0 of an isolated voice has motivated a surprising amount of effort over the years [45]. Work on the harder task of estimating multiple F_0s is now gaining momentum, fueled by progress in signal processing techniques on the one hand, and new applications such as interactive processing or indexing of music, multimedia and speech on the other.

A multiple F_0 estimation method is typically assembled from two elements: a single-voice F_0 estimator, and a voice-segregation scheme. Here "voice" is used in a wide sense to designate the periodic signal produced by a source (human voice, instrument sound, etc.). Some space is accordingly devoted the topic single voice F_0 estimation, but the reader should refer to the excellent treatise of Hess [45] for more details. Segregation techniques too are evoked, but the reader should follow pointers to other chapters of this book wherever possible.

A sound with a periodic waveform evokes a *pitch* that varies with F_0, the inverse of the period [87]. The pitch may be salient and musical as long as the F_0 is within about 30

ISBN 0-471-45435-4

Hz to 5 kHz [92, 105]. Sounds with the same period evoke the same pitch despite their diverse timbres: pitch can be understood as an equivalence class. The auditory system is able to extract the period despite very different waveforms or spectra of sounds at the ears. Explanations of how this is done have been elaborated since antiquity [27]. Modern theories can be classified into two families: pattern-matching and autocorrelation [27]. These theories are a source of inspiration for the development of F_0 estimation methods, that likewise can be organized according to a small number of basic principles, as we shall see in Sect. 1.3. Quite good solutions now exist for the task of single F_0 estimation [45, 31].

A musically inclined listener can often follow the melodic line of each instrument in a polyphonic ensemble. This implies that *several* pitches may be heard from a single compound waveform. Psychophysical data on this capability are fragmentary (e.g. [7, 8, 51]), so the limits of this capability, and the parameters that determine them, are not well known. This perceptual "proof of feasability" has nevertheless encouraged the search for algorithms for multiple F_0 estimation. Multiple F_0 estimation in essence involves *two* tasks: source separation and F_0 estimation. If the compound signal representing the mixture were separated into streams, then it would be a simple matter to derive an F_0 estimate from each stream using a single-voice estimation algorithm. On the other hand if F_0 estimates were known in advance, they could feed some of the separation algorithms described elsewhere in the book. This leads to a "chicken and egg" situation: estimation and segregation are each a prerequisite of the other, and a difficulty is to "bootstrap" this process.

There are other difficulties: the variety of signals and applications, the diversity of requirements and configurations that need evaluation, the existence of certain "degenerate" situations for which the problem is hard, etc. Many polyphonic estimation schemes have been proposed, and beginners in this field may be bewilderd by the wide range and sophistication of methods. Is this complexity really necessary? In this chapter I will try to show how most methods sprout from a few simple ideas. Once those are understood, the jungle of methods should seem less wild. The rest of the chapter reviews the main approaches to multiple F_0 estimation, trying wherever possible to extract the underlying insights and basic principles. A useful concept is that of "signal model".

1.2 SIGNAL MODELS

By definition a signal $x(t)$ is periodic *iff* there exists T such that:

$$x(t) = x(t+T), \forall t. \tag{1.1}$$

If there exists one such T there exist an infinity; the *period* is the smallest positive member of this set. Real signals differ from this description in various ways: they are of finite duration, their parameters may vary, there may be noise, etc. In this sense we speak of the periodic signal as a *model* that approximates signals found in the world. This model is parametrized by the period T (or its inverse F_0), and by the shape of the waveform over a period-sized interval: ($x(t)$ $t \in [0, T[$). It is useful in that it fits many sounds such as speech or musical sounds, because the parameter $F_0 = 1/T$ is a good predictor of musical pitch or speech intonation, and because that same parameter is a useful ingredient in acoustic scene analysis algorithms (e.g. Chapters 3 and 8).

An example of periodic signal is the *sinusoid* $x(t) = Acos(2\pi F_0 t + \phi)$. It is parametrized by the triplet (F_0, A, ϕ), where A is the amplitude and ϕ the starting phase. Sinusoids are useful in the context of linear systems: the output of a linear system for sinusoidal input is

another sinusoid of the same frequency but usually different amplitude and phase. Sinusoids (more precisely: complex exponentials) are *eigenvectors* of linear transforms. This property makes the sinusoid a very convenient model, as the effect of the linear system can be summarized by its effect on A and ϕ. The *sum of sinusoids* $x(t) = \sum_k A_k cos(2\pi f_k t + \phi_k)$ is useful for the same reason, as the effect the linear system is simply described by its effect on A_k and ϕ_k for all k.

A special case of the sum-of-sines model is the *harmonic complex* for which all component frequencies are multiples of a common fundamental frequency: $f_k = kF_0$. It is parametrized by specifying F_0, and (A_k,ϕ_k) for all k. The theorem of Fourier [35] states that this and the periodic signal model (Eq. 1.1) are equivalent, and fit exactly the same set of signals. Their parametrizations are related by the Fourier transform. F_0 estimation methods are divided into *time-domain* and *frequency-domain* according to whether they adopt one or the other of these signal models. Figures 1.1. (a-e) and 1.2. (a-e) show examples that illustrate both models. Estimation involves finding the parameter T of the periodic model, or the parameter F_0 of the harmonic model, that best fits the signal. Section 1.3 reviews a few simple ideas for doing so. Note that Fourier's theorem does *not* imply that there exists within the spectrum a component at F_0 with non-zero amplitude. Confusion on this point has led to much effort being diverted to solving the "missing fundamental" problem.

The periodic (or equivalently harmonic) signal is the most basic model involved in F_0 estimation, but other models may be of use. Examples are a periodic signal that varies slowly in amplitude or frequency, or a model of instrumental or voice production, or syntactic models of tone progression, etc. They are useful for two reasons: (a) the extra parameters allow a *better fit* to the signal and thus ease the estimation of F_0, and (b) other sources of knowledge may be brought in to *constrain* these parameters, again to get a more reliable estimate of F_0. That knowledge is either "hard-wired" into the algorithm, or else learned from the data at run time. There is a continuum between methods that process only information from the signal within the analysis frame, and those that bring in context, source models, grammars, expectations, etc.

1.3 SINGLE VOICE F_0 ESTIMATION

Before considering multiple voices, let us look at the simpler task of single voice F_0 estimation. Most polyphonic methods extend (or include) a single-voice method, and therefore schemes for that purpose are highly relevant. There are two basic approaches: spectral and temporal. In the former, a short-term Fourier transform is first applied to a frame of the waveform to obtain a spectrum, whereas in the latter the waveform is examined directly in the time domain. There are many variants of both approaches [45], but most flow from the same ideas. Note that most algorithms expect F_0 to vary over time and attempt to produce a *time series* of estimates, $F_0(t)$.

1.3.1 Spectral approach

Figure 1.1.(a) shows the short-term spectrum of a sinusoid. An obvious way to estimate its fundamental frequency is to measure the position of the *spectral peak*. However this procedure fails for the spectrum in Fig. 1.1.(b) that contains multiple peaks. A simple modification is to accept only the *largest peak*, but this algorithm fails for the spectrum in Fig. 1.1.(c) for which the largest peak falls on a multiple of F_0. A simple extension

is to select the *peak of lowest frequency* but this algorithm fails for the signal illustrated in Fig. 1.1.(d) for which the lowest peak falls on a higher harmonic (so-called "missing fundamental" waveform). Another cue, *spacing between partials* indicates the correct F_0 for this signal, but not for the signal illustrated in Fig. 1.1.(e).

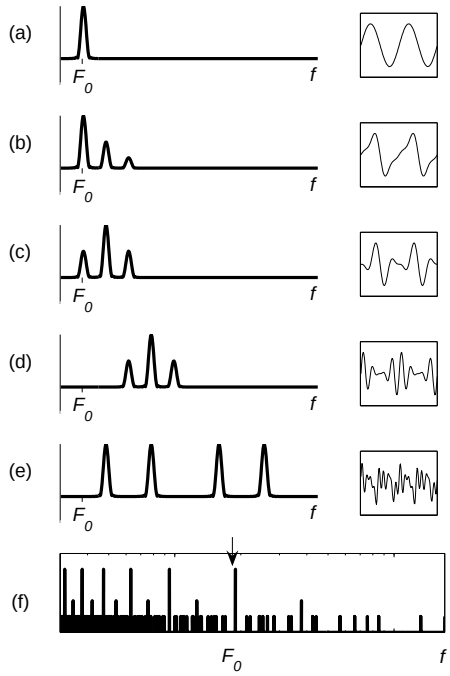

Figure 1.1. Spectra of simple signals that illustrate basic spectral F_0 estimation schemes. Corresponding waveforms are shown as insets to the right. The spectrum *peak* determines the F_0 of a pure tone (a) but a complex tone (b) has several such peaks. The *largest* peak determines the F_0 of the waveform in (b), but not (c). The *lowest frequency* peak determines the F_0 of the waveform of (c) but not (d). *Inter-partial spacing* determines the F_0 of (d) but not (e). The *Schroeder histogram* (f) determines the F_0 of the signal in (e) and of any periodic sound. The Schroeder histogram counts the subharmonics of every partial and accumulates them in a histogram. The cue to F_0 is the rightmost of the series of maximum values of this histogram (arrow). Note that the abscissa of (f) is logarithmic.

A final strategy that works for this signal and all others is *pattern matching*. For each peak in the spectrum, divide its frequency by successive positive integers and distribute the resulting values among the bins of a histogram (Fig. 1.1. (f)). The largest counts are found in bins at frequencies that divide all partial frequencies. There is an infinite series of such bins, that all have the same count but vanishingly small abscissas. All are situated at subharmonics of the *rightmost* bin of the series, and the position that bin thus indicates the fundamental. This idea was first applied to speech F_0 estimation by Schroeder [104], but it has earlier roots in "pattern matching" models of pitch perception ([20], see [21, 27] for reviews) that themselves evolved from the concept of "unconscious inference" of Helmholtz [123]. The idea has been proposed in many variants, such as the "spectral comb", "harmonic sieve" or "subharmonic summation" methods [78, 33, 44].

Most spectral F_0 estimation methods now use pattern-matching. Those that do not usually incorporate some form of preprocessing (non-linearity and/or filtering) to generate or enhance cues such as interpartial spacing, or a fundamental component. For example the method of [32] splits the signal over a bank of low-pass filters, selects the lowest frequency channel with significant power, and measures the frequency of its output. Filtering reduces the signal to a sinusoid, so that the strategy of Fig. 1.1. (a) can be applied to that output (see also [45]). Another recent example is the "TEMPO" algorithm of Kawahara et al. [61] which measures instantaneous frequency at the output of an array of bandpass filters. The channel that best responds to the fundamental is found on the basis of a "carrier-to-noise" measure. These algorithms are effective as long as the signal contains a sinusoidal component at F_0. Such is often, but not always, the case. If that partial is absent, as in Fig. 1.1. (d), it may be reintroduced by non-linear distortion (e.g. [101, 108]). Non-linear distortion is not without problems, as one can find cases where it instead *suppresses* the F_0 component (for example squaring a sinusoid would double its frequency and give an incorrect result).

Inter-partial spacing was used for example in the methods of Lahat et al. [70], Chilton and Evans [17], or Kunieda et al. [68] that calculate the autocorrelation of the positive-frequency part of the spectrum. Any two spectral components spaced by F_0 contribute to a peak at F_0 in the spectrum autocorrelation. As argued by Klapuri [64], the spacing between adjacent components determines the rate of beating between them, and thus it can also be measured in the time domain (see next section). Algorithms based on inter-partial spacing (or beats) fail if the spectrum is sparse, for example if it consists of a single component at F_0 (Fig. 1.1.(a)) or of components at non-contiguous frequencies (Fig. 1.1.(e)) but, again, one can use a non-linearity to reintroduce power at harmonic frequencies within the gaps.

The strength of spectral methods is that they benefit from the theoretical power of Fourier analysis, and from the efficiency of the Fast Fourier Transform (FFT) to implement them. A weakness is their dependency on the shape and size of the analysis window. These remain as "nuisance parameters" of the estimation. These pros and cons are discussed in more detail below in the context of multiple F_0 estimation. It may seem somewhat strange to go the trouble to split the signal into partials, and then apply pattern matching to find the period that is, after all, obvious in the time-domain waveform. This reasoning motivates time-domain methods.

1.3.2 Temporal approach

Figure 1.2. (a) shows the waveform of a sinusoid. An obvious way to measure its period is to measure the interval between "landmarks" such as waveform peaks. This simple algorithm fails for the waveform in (b) that has several peaks per period. A modification is to take the *largest* peak, but this would fail for this same waveform if it were negated, as it would then have two "largest" peaks per period. Positive-going zero-crossings would work for this waveform but fail for that of (c) that has many crossings (and peaks) per period, as a consequence of a relatively large proportion of high-frequency power. An option is to apply low-pass filtering (thin line), but this strategy fails for the waveform of (d), that lacks any low-frequency power. An option is to apply a non-linearity, for example full-wave rectification or squaring (thin line) and low-pass filter to extract the envelope. However this fails for the waveform of (e) for which the envelope period is half the waveform period.

A final strategy works for this and all other periodic waveforms: self-similarity across time. Each waveform sample may used, as it were, as a "landmark" to measure similarity for temporal spans of various sizes. For example, using the cross-product between waveforms

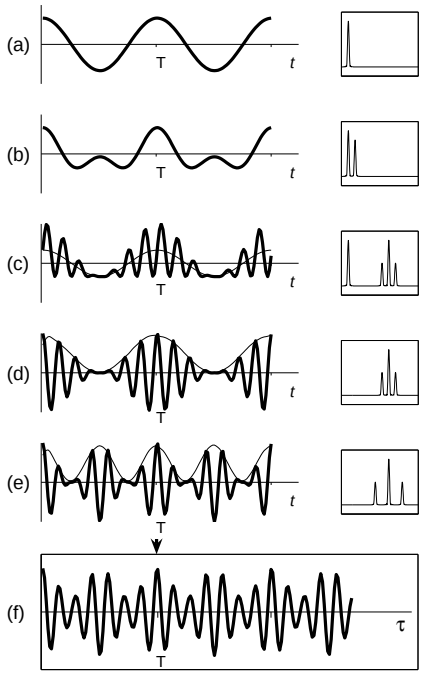

Figure 1.2. Waveforms of simple signals that illustrate time-domain F_0 estimation schemes. Corresponding spectra are shown as insets to the right. The interval between waveform *peaks* indicates the period for the pure tone (a), but the complex (b) has several peaks per period. The *largest peaks* work for complex (b), but would not work for the opposite waveform that has two "largest" peaks per period. Positive-going *zero-crossings* work for (b) but not (c). Low-pass filtering the signal in (c) would reduce it to its fundamental component (thin line) that has one peak or zero-crossing per period, but the waveform in (d) has no fundamental component. The *envelope* may be obtained by full-wave rectification (or some other non-linearity) followed by low-pass filtering (thin line). This works for (d), but the envelope of (e) oscillates at twice its F_0. The first major peak with non-zero lag τ (arrow) of the *autocorrelation function* (ACF) (f) can indicate the period of (e) or any other periodic waveform.

as a measure of similarity yields the familiar autocorrelation function (ACF) defined as:

$$r_t(\tau) = (1/W) \sum_{j=t+1}^{t+W} x(j)x(j+\tau) \tag{1.2}$$

where τ is the "lag" (or delay), t is time at which the calculation is made, and W is the size of the window over which the product is integrated. The purpose of integration is to ensure that the measure is stable over time. Figure 1.2. (f) shows the ACF of the waveform in Fig. 1.2. (e). The function has a series of global maxima at zero, at the period (arrow), and at all multiples of the period. The period is determined by scanning this pattern, starting at zero, and stopping at the first global maximum with non-zero abscissa. Autocorrelation was introduced for speech F_0 estimation by Rabiner [93], but Licklider [73] had earlier

suggested it to explain pitch perception, and the idea can be traced back even earlier [52], see review in [27].

Self-similarity methods such as the ACF can handle any periodic waveform. In contrast, strategies based on particular landmarks (peaks, etc.) must be associated with preprocessing to increase their salience or stability. For example, Dologlou and Carayannis [32] applied low-pass filtering to obtain a sinusoidal waveform with one peak per period, Howard [47] applied non-linear filtering to "simplify" the waveform, and Howard [48] applied a neural network to learn a mapping between the voiced speech waveform and the glottal pulses that produced it. Earlier examples are reviewed by Hess [45]. The difficulty is to ensure that (a) at least one landmark occurs per period, (b) no more than one occurs per period, and (c) the landmark's position does not jump around within the period. These goals are impossible to guarantee in the general case: for any type of marker one can find examples such that an infinitesimal change in waveform produces a jump in marker position.

A detail must be mentioned at this point. We defined the ACF as in Eq. 1.2, but it is quite common to find a slightly different definition:

$$r'_t(\tau) = (1/W) \sum_{j=t+1}^{t+W-\tau} x(j)x(j+\tau) \tag{1.3}$$

in which W is replaced by $W - \tau$ as the upper limit of summation. This is often referred to as the "short-term ACF", whereas the definition of Eq. 1.2 has been diversely called "running ACF", "autocovariance" or "cross-correlation" [50]. The advantage of Eq. 1.3 is that it allows efficient implementation by the FFT. Its drawback is that for large τ the statistic is integrated over a small window, and thus is less stable over time. Figure 1.3. illustrates both definitions. The "short-term" ACF is plotted in (b) and the corresponding "running" ACF in (c). Replacing $1/W$ by $1/(W - \tau)$ in Eq. 1.3 produces the so-called "unbiased" short-term ACF. In aspect it resembles the running ACF of Fig. 1.3.(c), but it is plagued by the same problem of insufficient temporal smoothing at large τ.

A useful variant of the ACF is the *squared-difference function* (SDF):

$$d_t(\tau) = (1/W) \sum_{j=t+1}^{t+W} (x(j) - x(j+\tau))^2 \tag{1.4}$$

which is simply the squared Euclidean distance between a chunk of signal of size W and a similar chunk time-shifted by τ. It is used for example by the cancellation model of [24], or the YIN method of [31]. Replacing Euclidean distance by city-block distance (sum of absolute values, instead of squares) would yield the well known AMDF, or average magnitude difference function [96]. ACF and SDF are related by the relation

$$d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau). \tag{1.5}$$

The two first terms are local estimates of signal power, and to the degree that they are constant as a function of τ (i.e. if W is large enough), autocorrelation and squared difference function carry the *same* information. The cue to the period for the SDF is a *dip* rather than a peak, as illustrated in Fig. 1.3. (d). The nice thing about the SDF, as we shall see in Sect. 1.4.2, is that it can be generalized to estimate multiple periods. Note that the relation between ACF and SDF in Eq. 1.5 holds only if the ACF is calculated as in Eq. 1.2.

The strength of temporal methods is their conceptual simplicity, close to the mathematical definition of periodicity. There is nevertheless a deep link between spectral and

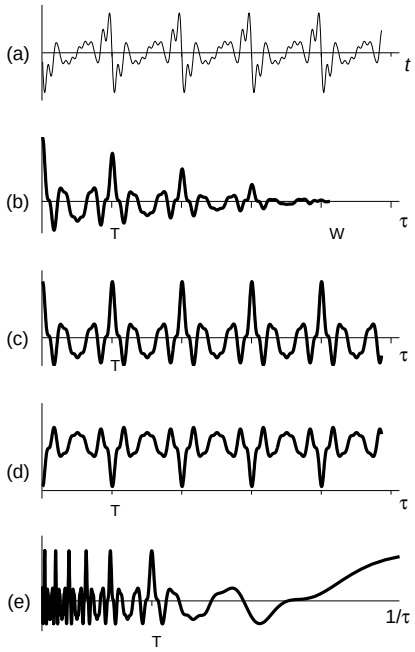

Figure 1.3. Illustration of the autocorrelation function. (a) Waveform of a periodic complex tone. (b) ACF calculated according to Eq. 1.3 ("short-term" ACF). Note that the function vanishes beyond $\tau = W$. (c) ACF calculated according to Eq. 1.2 ("running" ACF). (d) SDF. (e) Same ACF as in (c) but plotted as a function of an inverse log-lag scale ($log(1/\tau)$) [34]. Note the similarity of (e) with the Schroeder histogram plotted in Fig. 1.1. (f).

temporal methods, and in particular between pattern matching and the ACF. To understand this link, recall that according to the Wiener-Khinchine theorem, the ACF is the inverse Fourier transform (IFT) of the power spectrum. As the waveform is real and its power spectrum symmetrical, the IFT is equivalent to cross-corrrelation with a family of cosine functions. A cosine has regularly-spaced peaks at integer multiples of its period, and can be understood as a particular form of *harmonic template*. Thus the ACF can be seen as a form of pattern-matching. This parallel is obvious if the ACF is plotted as a function of a log-lag scale as proposed by Ellis [34] (compare Fig. 1.3.(e) with Fig. 1.1. (f)).

Based on this reasoning, useful variants of the ACF are obtained by replacing the IFT by convolution with periodic templates that have sharper peaks than cosines (to increase their spectral selectivity), or peaks that decrease in amplitude (to discount the contribution of partials of higher frequency or rank) [65, 66]. A problem with the ACF is that the power spectrum puts strong emphasis on high-amplitude portions of the spectrum, and thus is sensitive to the presence of strong harmonics. This is alleviated by taking the logarithm before the cosine transform to obtain the well-known *cepstrum* [85]. Raising to the power $1/2$ or $1/3$ has a similar balancing effect [56], as reviewed recently by Klapuri [65]. These details are of limited theoretical importance but they have an impact on performance, particularly when the method is used within the context of multiple F_0 estimation.

1.3.3 Spectrotemporal approach

A variant of the temporal approach, inspired by auditory processing, involves splitting the signal over a filterbank. Each channel is treated as a waveform function of time, rather than as a sample along a slowly-varying profile of spectral coefficients as in spectral methods. Each channel, dominated by a limited range of frequencies, is processed by time-domain methods as above, and the results are aggregated over channels. Typically, channel-wise ACFs may be added to obtain a summary autocorrelation function (SACF), as illustrated in Fig. 1.4.. The idea was originally proposed in the pitch perception model of Licklider [73] and further developed by Meddis and Hewitt [79] and others [110, 74, 12]. It was applied to F_0 estimation for example by [106, 22, 98].

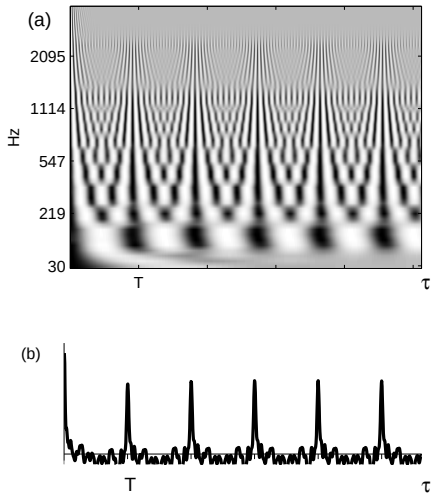

Figure 1.4. Spectrotemporal method of single-voice F_0 estimation. (a) Array of ACFs calculated within channels of a filterbank. The filters are 4th-order gammatone filters with bandwidths based on psychophysical estimates of auditory selectivity [82] and center frequencies spaced equally in terms of bandwidth. Each channel is amplitude-normalized before the ACF calculation. (b) Summary ACF (SACF). These plots were calculated from the same waveform as in Fig. 1.3. (a). The difference with Fig. 1.3. (c) is the result of amplitude-normalization that emphasizes low-amplitude portions of the spectrum.

There are several advantages of the spectro-temporal over the temporal approach. First, the weight of each channel may be adjusted to compensate for amplitude mismatches between spectral regions, that would otherwise be accentuated by the ACF [65]. Doing so is similar to the process of "spectral whitening" by inverse filtering that was applied in several early methods [45]. Second, channels dominated by noise, or by a competing source, can be discounted in the summary. We shall see how this can be put to use for multiple F_0 estimation. A third advantage was pointed out by Klapuri [64, 65]. If higher-frequency channels have larger bandwidths (as is the case for models of cochlear filtering), then adjacent partials of high order interact within those channels to create *beats*. Beat rate depends on inter-partial spacing, and for high-order partials it may provide a cue to F_0 that is more robust than the exact frequencies of the partials themselves, particularly if the spectrum is slightly inharmonic and/or F_0 varies with time. *Demodulation* of higher-

frequency channels (by a nonlinearity followed by low-pass filtering) allows cues from those beats to be incorporated into the SACF. Beats could actually be exploited without the filterbank, but what filtering buys in this context is to reduce the sensitivity of the beats to *phase* relations between partials that fall in different channels.

To summarize, many methods of single-voice F_0 estimation have been proposed. Estimation can be understood in terms of fitting a *model* to the waveform. The most basic model is that of a periodic signal (Sect. 1.2), but more complex models may be used, for example instrument models that specify in detail the spectrotemporal shape of a "note", or dynamic models that constrain the variation of F_0 over time, etc. An estimation error occurs when the signal fits the model for an inappropriate set of parameters. The art of F_0 estimation is to tweak the model (or the signal) to make such an erroneous fit less likely. This point of view is all the more useful in the case of multiple F_0 estimation.

1.4 MULTIPLE VOICE F_0 ESTIMATION

Several factors conspire to make multiple voice F_0 estimation more difficult than single voice F_0 estimation. Mutual overlap between voices weaken their pitch cues, and the cues must further must compete with cues to other voices. There exist degenerate situations where available information is ambiguous, as when the F_0s are in simple ratios. Also, the diversity of situations to be considered (number and type of sources, relative amplitudes and timing, etc.) makes progress harder to evaluate than in the single F_0 case.

The basic signal model is the *sum of periodic signals*. For example in the case of two voices, the observable signal $z(t)$ is the sum of signals $x(t)$ and $y(t)$ of periods T and U:

$$z(t) = x(t) + y(t), \quad x(t) = x(t+T), \quad y(t) = y(t+U), \quad \forall t \tag{1.6}$$

F_0 estimation consists of finding parameters T and U that best fit the signal z. More complex models are discussed later on.

Three basic strategies have been used. In the first, a single voice estimation algorithm is applied in the hope that it will find cues to several F_0s. In the second strategy (iterative estimation), a single-voice algorithm is applied to estimate the F_0 of *one* voice, and that information is then used to *suppress* that voice from the mixture so that the F_0s of the other voices can be estimated. Suppression of those voices in turn may allow the first estimate to be refined, and so on. In a third strategy (joint estimation) all the voices are estimated at the same time.

As an example of the first strategy, the speech separation system of Weintraub [125] searched the ACF for cues to multiple periods. In the system of Stubbs and Summerfield [115] the same was done for the cepstrum. It is rather challenging to make this strategy work. Looking at representations such as the Schroeder histogram of Fig. 1.1. (f) or the ACF of Fig. 1.2. (f), it is obvious that they already contain multiple cues even for a single voice. Distinguishing these from cues to *multiple* voices is bound to be hard. Schemes have been proposed to attenuate spurious cues [56, 117, 34], but the conditions under which they are successful appear to be limited. We will concentrate instead on the two other strategies: iterative and joint estimation. As before, approaches can be classified as *spectral*, *temporal*, and *spectrotemporal*.

1.4.1 Spectral approach

In a seminal paper, Parsons [89] calculated the short-term magnitude spectrum of mixed speech (sum of two talkers) over 51.2 ms windows, and applied Schroeder's subharmonic

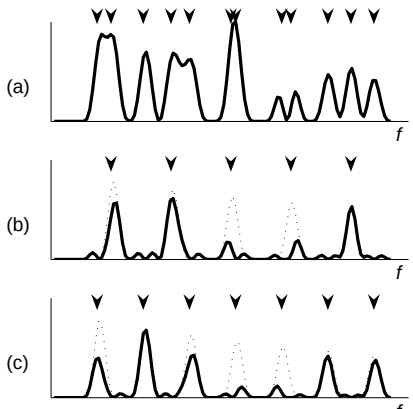

Figure 1.5. Spectral method of two-voice F_0 estimation, based on Parsons [89]. (a) Spectrum of the sum of two concurrent voices. A first F_0 estimate is derived from this spectrum and used to suppress one voice (voice A). (b) Thick line: result of suppressing voice A. The F_0 of voice B can be estimated from this remainder, and used to suppress that voice in turn. (c) Thick line: result of suppressing voice B. The arrows indicate the harmonic series of each voice, and the thin lines represent the spectra of the voices before mixing. Note that only part of the spectrum has been retrieved in each case.

histogram method, mentioned earlier, to determine the harmonic series that best matched the spectrum. A first F_0 was derived, spectral peaks that matched its harmonic series were removed from the spectrum, and a second F_0 was estimated from the remainder. The second voice could then be removed in turn to refine the estimate of the first. This process is illustrated in Fig. 1.5.. The aim of Parsons was voice separation, but F_0 extraction was a major subtask and his was one of the first multiple-F_0 estimation systems. Many, since Parsons, have proposed to apply harmonic templates iteratively to dissect the short-term spectrum [103, 114, 59, 129, 38, 5, 19, 55, 64, 100, 124, 121, 84]. These methods use the spectrum representation both as a source of cues to the F_0 of a voice, and as a substrate from which it is possible to *discount* those cues so that the other F_0s can be determined. In some methods the estimation and suppression steps are performed in sequence, in others they are performed jointly by fitting the compound spectrum to a model of overlapping spectra.

1.4.2 Temporal approach

Supposing the period T of one voice has been determined, that voice can be suppressed by applying to the compound waveform a time-domain comb-filter with impulse response $h_T(t) = \delta(t) - \delta(t - T)$. The impulse response and its power transfer function are illustrated in Fig. 1.6. (a) and (b). The transfer function has zeros at $1/T$ and all its multiples, and these can suppress all the partials of a voice with F_0=$1/T$. Tuning this filter to the period of voice A, that voice may be suppressed and the F_0 of voice B estimated. Tuning the filter to the period of voice B, the estimate of voice A may be refined. This process is illustrated in Fig. 1.6. (c-e).

The idea was first proposed by Frazier et al. [36] for voice separation, and later used for multiple F_0 estimation by de Cheveigné and others [23, 30, 56]. The period estimate may be obtained by any single-voice F_0 estimation method, for example by the ACF or SDF (Sect. 1.3.2). The latter option is of interest as the same operation (cancellation) serves in

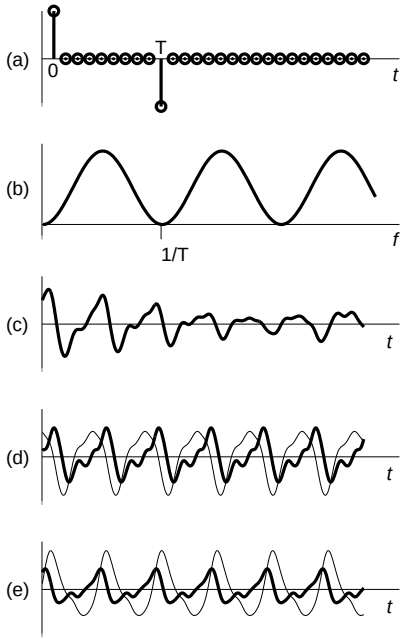

Figure 1.6. Temporal method of two-voice F_0 estimation (iterative). (a): Impulse response of time-domain comb-filter. (b) Power transfer function of the same filter. Zeros at multiples of $1/T$ cancel all harmonics of $F_0 = 1/T$. (c) Sum of two complex tones with F_0s one semitone apart (6%). A first F_0 estimate is derived from this waveform and used to suppress one voice (voice A). (d) Thick line: result of suppressing voice A. The F_0 of voice B can be estimated from this remainder and used to suppress voice B from the compound. (e) Thick line: result of suppressing voice B. The thin lines represent the complexes before mixing. Note that the filtered waveforms have the same period as voices A or B, respectively, but not the same shape.

turn to measure cues to the F_0 of a voice, and then to suppress them. Indeed, both steps may be performed jointly rather than in succession [23, 30, 29].

In the MMM method of [29], the period is found by forming the *double difference function* (DDF):

$$dd_t(\tau, \nu) = (1/W) \sum_{j=t+1}^{t+W} (x(j) - x(j+\tau) - x(j+\nu) + x(j+\tau+\nu))^2. \qquad (1.7)$$

It is easy to see that this function is zero for $(\tau, \nu) = (jT, kU)$ for all integers (j, k), and conversely if periods (T, U) are unknown they may be found by searching the (τ, ν) parameter space for the first minimum with non-zero coordinates. The function is illustrated in Fig. 1.7. for a mixture of two periodic sounds with periods that differ by two semitones (about 12%). Minima are visible at period multiples, as well as along the axes $\tau = 0$ and $\nu = 0$.

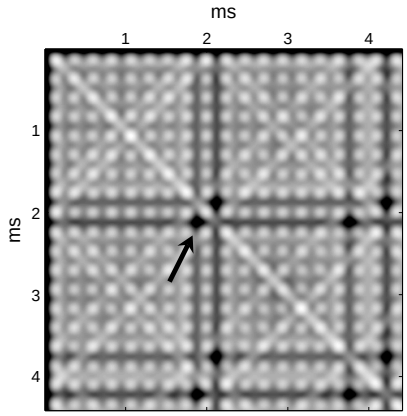

Figure 1.7. Temporal method of two-voice estimation (joint). Double difference function (DDF) in response to a mixture of two periodic sounds as a function of its two lag parameters, τ and ν. Darker means smaller. The coordinates of the minimum with smallest non-zero lag (arrow) indicate the periods T and U.

1.4.3 Spectrotemporal approach

A third approach, intermediate between spectral and temporal, is to split the waveform over a bank of band-pass filters (Fig. 1.8.). Meddis and Hewitt [80] extended their spectrotemporal model of single pitch perception [79] to explain voiced speech segregation, by using a cochlear filter bank to split acoustic information into channels belonging to either of two sources. ACFs calculated within each channel were initially summed across all channels to obtain a summary ACF (SACF) from which a dominant period was derived. Channels with peaks at that particular period were then assigned to the dominant voice, and the remaining channels used to estimate the identity of the second voice. Although not elaborated by the authors, a second *period* could also be estimated from those remaining channels. Channel selection had previously been proposed by Lyon [75] and Weintraub [125] for sound separation. The idea has since been used for multiple F_0 estimation by Wu et al. [128, 126] and others [49, 76, 72].

How do *spectral* and spectrotemporal methods compare? Both split the signal into spectral "elements" (spectrum bins in one case, filter channels in the other) on the basis of their spectral properties. However, whereas spectral methods assign bins according to their position along the frequency axis, spectrotemporal methods assign channels according to the periodicity that dominates them. They thus differ in resolution requirements: spectral methods must resolve individual partials, and this requires a long analysis window, whereas spectrotemporal methods need merely to isolate spectral regions dominated by one or the other voice (Fig. 1.4. (b, c)). Long analysis windows cannot be used if the signal is non-stationary: in that case spectrotemporal methods may have the advantage.

How do *temporal* and spectrotemporal methods compare? Both estimate F_0s based on temporal information. They differ in how the correlates of an unwanted voice are suppressed: channel-selection for the former, and comb-filtering for the latter. For signals that are perfectly periodic, comb-filtering provides perfect rejection, whereas the degree of rejection of most filterbanks is limited by the slope of filter characteristics. Nevertheless,

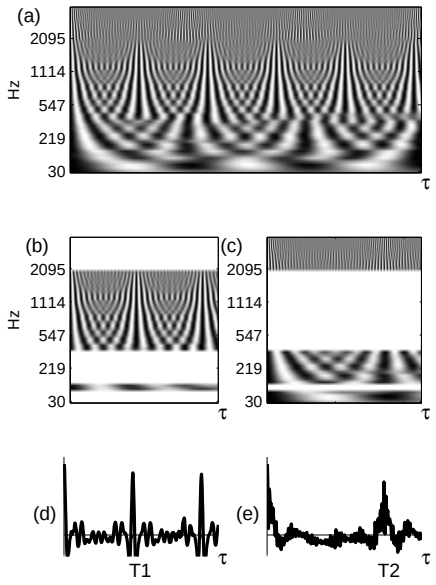

Figure 1.8. Spectrotemporal method of two-voice F_0 estimation. (a) Illustration of a spectrotemporal two-voice estimation algorithm. (a): Array of ACFs at the output of a filterbank in response to the sum of two periodic signals (synthetic vowels 'a' and 'i'). (b) ACFs of channels dominated by one voice. (c) ACFs of channels dominated by the other voice. (d) SACF calculated from channels dominated by the first voice. (e) SACF calculated from channels dominated by the second voice. The F_0s of both voices can be estimated from these SACFs.

channel-selection may be more effective in the presence of noise, or for slightly inharmonic sources for which harmonic cancellation is less effective. One might expect a *combination* of the two approaches (for example time-domain cancellation at the output of filterbank channels) to be most effective, but it seems that this idea remains to be fully explored.

For slightly inharmonic sources such as strings, or in the event of slight F_0 estimation errors or nonstationarity, it may be hard to segregate higher-order partials on the basis of their position relative to a F_0-based harmonic series. This is particularly the case for high-frequency components, and so spectral approaches may have difficulty making use of higher-order partials. Temporal approaches based on comb-filtering also run into problems in the same situation. However, the spectrotemporal approach allows the additional cue of *interharmonic spacing*. Spacing determines the beat rate between partials that interact within a channel, and that rate can be measured by applying a non-linearity to the filter output followed by low-pass filtering to isolate the low-frequency beat components [65, 66, 128]. For this to work, the channel must contain partials of only one voice, and for that the filters must be narrow compared to features of the spectral envelope (e.g. formants) of each sound. The ability to extract this extra cue gives the spectrotemporal approach an edge over spectral and temporal methods.

Various criteria may be used to recognize the channels that belong to a voice. For example Wu and colleagues [127, 128, 126] use heuristic quality criteria to eliminate channels dominated by noise. Hu and Wang [49] include cross-channel correlation to

group channels likely to belong to the same source. Klapuri [65, 66] discounts higher frequency channels in which partials may be unresolved, and thus dominated by beats at the "chord root" frequency. The chord root is a common subharmonic of the voices present. If it is high enough to fall within the search range, it may be mistaken for the F_0 of a primary voice.

1.5 ISSUES

This section deals with a number of "nitty-gritty" considerations that must be addressed for processing to be effective. Algorithms are sensitive to imperfections in the calculations, or to a mismatch between the signal model and the signal. It is important to distinguish between processing issues (for example spectral resolution) from application-dependent issues (for example imperfect periodicity, or noise). For multiple F_0 estimation, the "devil is in the details".

1.5.1 Spectral

The main issue with frequency-domain methods is *spectral resolution*. Supposing a temporal analysis window of duration D, short-term spectra are sampled in frequency with a resolution of $1/D$. This means that, according to Parseval's relation, signal power within the analysis window is partitioned among spectral coefficients. Spectral methods can use this partition to segregate voices and thus measure their F_0s. More precisely: if partials of a voice fall on multiples of $1/D$, that voice can be removed so as to estimate the other voice's F_0s. Such is the case only if that voice's F_0 is an integer multiple of $1/D$, unfortunately an unlikely event. In general there is a mismatch between partials and the frequency grid. This may interfere with estimation of F_0 of each voice and, more importantly, reduce the effectiveness of source suppression because each individual spectral coefficient contains power from several sources. Larger analysis windows allow finer spectral resolution, at the expense of temporal resolution and the ability to deal with time-varying signals. The need for power-of-two block sizes for FFT efficiency further restricts the choice of window size.

There are several ways to enhance spectral features. The short-term spectrum may be *interpolated* in the vicinity of spectral peaks (e.g. by fitting a smooth function such a parabola, or the Fourier transform of the analysis window, or a gaussian, etc.) [59, 118]. In place of the Fourier transform, the waveform may be fitted to a *sinusoidal model* (e.g. [107, 11]) or a sum of damped sinusoids (e.g. [116]). The complex spectra of successive bins may be paired to obtain an *instantaneous frequency* estimate for each frequency bin. This is then used - rather than bin position - as a measure of frequency of the power within the bin. Instantaneous frequency has been used for single voice F_0 estimation (e.g. [2, 3, 61, 4, 83]) and multiple voice F_0 estimation (e.g. [40, 9, 109]). Mapping power according to instantaneous frequency produces a spectrogram with sharper features than the Fourier spectrogram [16, 18, 61, 83, 43]. These techniques have been reviewed recently by Hainsworth [42] and Virtanen [118].

It is important to understand that these techniques improve the accuracy of cues to partials that are *resolved*, but do not address the problem of partials that are too close to have individual cues. Cues to partials that are close may undergo mutual distortion, or even merge into a single hybrid cue. To some extent, overlapping cues may be separated by modeling the superposition process. However the effectiveness of this operation is limited by uncertainty as to *phase relations* between partials (see further on). In addition

to these factors that relate purely to processing constraints, there are other factors related to stimulus imperfections, such as aperiodicity or noise, that contribute to make the compound spectrum difficult to partition among voices. Dual to spectral resolution is the problem of *temporal* resolution of spectral analysis, as determined by the size, shape and position of analysis windows. Kashino and colleagues [57] optimize the tradeoff between these conflicting constraints with the use of "snapshots", windows starting from a discontinuity such as note onset, and extending as far as the signal is stable.

To summarize, performance of spectral approaches is limited by spectral resolution, itself determined by the short analysis window size required to follow changes in the signal. Many techniques exist to overcome these limits, but (a) they add to conceptual complexity and difficulty of implementation, and (b) they are not always as effective as needed.

1.5.2 Temporal

Limited sampling resolution. The accuracy of of cues such as ACF peak position is limited by sampling resolution. Worse, suppression of a voice by comb-filtering may be imperfect, thus impairing the estimation of the other voices. Resolution of ACF peaks may be improved by three-point parabolic interpolation, as the vicinity of an ACF peak is well approximated as a sum of cosines, each of which can be expanded as a Taylor series with terms of even order. Interpolation refines the *value* at the peak, which determines whether it wins over competing peaks, and its *position* that determines the precise value of the period estimate. The same interpolation technique is applicable to the dip of the SDF (Eq. 1.4), and it may be extended to two-dimensional interpolation of the DDF pattern (Eq. 1.7) in the joint cancellation method: five samples (the minimum and its four immediate neighbors) constrain a paraboloid with no cross-terms from which the global minimum may be interpolated [29]. Interpolation is also needed for voice suppression. A voice with a non-integer period can be suppressed by applying a time-domain comb-filter with fractional delays, implemented either by an interpolating filter [69] or more simply, if less accurately, by linear interpolation.

Efficiency. Multiple F_0 estimation is computationally expensive, and it is important to understand the factors that determine the cost. Estimation involves *search* within the space of possible periods. Supposing N expected periods, the size of the space varies as $O(K^N)$, where K is the number of points at which each period dimension is sampled. Joint estimation methods (e.g. [29]) search this space exhaustively. Iterative methods (e.g. [30]) search a subset of size $0(KNk)$, where k is the number of iterations. Search is indifferent to permutation of lags, so cost may be reduced by a large factor by ordering lags as $\tau_1 < \tau_2 < \ldots < \tau_N$. The asymptotic trends however remain the same. Each lag dimension is typically sampled uniformly at the same resolution as the waveform, so $K = f_s \tau_{MAX}$, where f_s is the sampling rate and τ_{MAX} the largest expected period. Non-uniform sampling such as logarithmic (Fig. 1.3.) has also been proposed [34]. The appropriate degree of temporal integration also depends on τ_{MAX}. Specifically, the window of integration (W in Eq. 1.2, $W - \tau$ in Eq. 1.3) should be at least τ_{MAX} in order to guarantee the stability over time of F_0 estimates.

The short-term ACF, inverse Fourier transform of the short-term power spectrum, is best calculated by FFT. According to the previous reasoning the window size W in Eq. 1.3 should be at least equal to $2\tau_{MAX}$. The running ACF of Eq. 1.2 can likewise be calculated by FFT, as the inverse Fourier transform of the cross-spectrum between two

windowed chunks of signal of size W and $W + \tau_{MAX}$. The computational cost of an FFT of size N, $O(N \log N)$, is cheaper than the $O(N^2)$ cost of implementing Eqs. 1.2 or 1.3 directly. However, if it is necessary to repeat the calculation at a high frame rate, a recursive formula may be faster than the FFT. For example the formula $r_{t+1}(\tau) = r_t(\tau) - x(t)x(t+\tau) + x(t+W)x(t+W+\tau)$ updates the ACF at a frame rate equal to the waveform sampling rate. For exhaustive search Eq. 1.7 needs to be evaluated repeatedly. The cost of doing so may be reduced by applying a computational formula such as $d_t(\upsilon, \nu) = d_t(\upsilon) + d_{t-\nu}(\upsilon) + d_t(\nu) - d_t(\upsilon + \nu) - d_{t-\upsilon}(\nu - \upsilon) + d_{t-\upsilon}(\nu)$ in which the DDF is expressed as a linear combination of DFs. Similar formulae are available involving ACFs [29]. This leads to computational savings if the necessary DFs (or ACFs) are pre-calculated.

Efficiency considerations are important in that computational costs may prohibit otherwise effective schemes.

1.5.3 Spectrotemporal

Spectrotemporal methods use an initial filterbank to split the waveform into channels, each of which is then processed in the time domain. Selectivity requirements are less stringent than for spectral methods. Rather than *partials*, it is sufficient to resolve *spectral regions* dominated by one or another voice. Increasing filter selectivity allows off-frequency components belonging to noise or other voices to be better attenuated. However sharp skirts entail a long impulse response that may "smear" features over time, and thus limit the ability to track a time-varying voice. Also, if filters are narrow, more channels are required to cover the useful spectrum. The choice of filterbank is a tradeoff between these conflicting requirements.

A common choice is a filterbank with characteristics similar to the human ear (e.g. [127, 128, 65]). Auditory filters are typically modeled as gammatone filters for which efficient implementations exist (e.g. [111, 18, 46, 91]). Bandwidths are usually set according to estimates of human "critical bandwidth" [130] or "equivalent rectangular bandwidths" (ERB) [82] that are roughly constant below 1 kHz (about 50-100 Hz) and proportional to frequency beyond 1 kHz (about 10 %). There is no guarantee however that characteristics close to the human ear will ensure optimal multiple F_0 estimation. Indeed, Karjalainen and Tolonen [56, 117] used only two bands covering the regions below and above 1 kHz, and Goto [38, 40] likewise used filtering to separate a low-frequency region ($<$262 Hz) from which a bass line was derived, from a high-frequency region ($>$262 Hz) from which a melody line was derived. No studies seem to have searched for optimal filtering characteristics, whether theoretically or empirically. A system could conceivably incorporate *multiple* filter characteristics so as to satisfy a wider range of constraints [28].

A weakness of the spectrotemporal approach is the cost of processing multiple channels in parallel. Efficient schemes exist to implement processing that is functionally similar in the frequency domain via standard FFT-based methods [62].

1.6 OTHER SOURCES OF INFORMATION

Up to this point we reviewed methods that exploit only one source of information: the signal within the analysis frame. This information is fragile and fragmentary. Other sources of information may contribute both to improve the accuracy of a voice's F_0 estimate, and to better suppress that voice and estimate the others. This information is brought to bear via

models of what to expect of the signal. It is important to realize that, if a model does not fit the signal being treated, this process may instead increase the risk of error.

1.6.1 Temporal and spectral continuity

A common assumption is that voices change slowly. Continuity over time of F_0 estimates is exploited in post-processing algorithms [45] such as median-smoothing, dynamic programming, hidden-markov models (HMM, e.g. [128]), or multiple "agents" [40]. The value for the current frame given by the "bottom-up" algorithm is tested for consistency with past (or future) values. Proximity of value may be complemented by a measure of *quality* to give more weight to reliable estimates. Processing may occur post-hoc after the estimation stage, or else it may be integrated to the estimation algorithm itself (e.g. [119]). Estimation is improved directly, as a result of interpolating over errors and missing values, and also indirectly by (hopefully) increasing the likelihood that the voice is accurately suppressed so that other voice F_0s can be estimated.

In addition to continuity of F_0 tracks, the assumption that partial *amplitudes* vary smoothly can be used to track voices over instants when F_0s cross or fall into a ratio for which the separation task is ambiguous. A different but related assumption is that all partial amplitudes vary according to the same function of time (to a fixed factor) [129]. Granted this assumption, amplitude variations that do *not* follow this function may be assigned to beats between closely-spaced partials, and partial amplitudes can then be estimated from the minima and maxima of the beats [67, 121]. The assumption amounts to saying that the time-frequency envelope is the *outer product* of a spectral shape (common to all times) and a temporal shape (common to all frequencies). Spectrograms usually have more complex shapes, but techniques exist to decompose them into a sum of such simple envelopes [55, 13, 112]. The time-course of amplitude itself can be modeled as a sum of smooth basis functions such as gaussians or raised cosines [55, 19]. Cross-time dependencies can be modeled within the context of Bayesian models [124].

An assumption that has been used recently is *spectral smoothness*, that is, limited variation of partial amplitudes across the frequency axis [63, 129, 122, 5, 14, 71]. Many (but not all) musical instruments indeed have smooth spectral envelopes. Irregularity of the compound spectrum then signals the presence of multiple voices, and smoothness allows the contribution of a voice to *shared partials* to be discounted. For example if two voices are at an octave from each other, partials of even rank are the superposition of partials of both voices. Based on spectral smoothness, the contribution of the lower voice can be inferred from the amplitude of partials of odd rank, and subtracted to reveal the presence of the higher voice. The effectiveness of this strategy is nevertheless limited by uncertainty as to the relative *phase* of coinciding partials (see below). Spectral smoothness has also been used to reduce the likelihood of subharmonic errors [5, 63]. Beats between adjacent partials are strongest if the partials are of similar amplitude, and thus spectral smoothness enhances beat-related cues (e.g. [65]).

The effectiveness of the spectral smoothness assumption depends of course on its validity. If voices have irregular spectral envelopes, as in Fig. 1.1. (e), the assumption is likely instead to favor incorrect interpretations of the data. Some natural sources produce irregular spectra, such as the clarinet (for which even partials are weak) or the human voice (if harmonic spacing is wide relative to formant bandwidth), and of course there is no constraint at all on sounds produced electronically.

1.6.2 Instrument models

Similar to continuity constraints but more sophisticated are *instrument models*. These can be of a general nature, such as the source-filter model that grounds inverse-filtering methods [45], or models of naturally amortized sounds as sums of exponentially decreasing sinusoids [116]. They can also be instrument-specific such as the piano models of [86, 97]. They may be predetermined (for example from the physics of the instrument) or else learned from the data. Like spectral or temporal continuity, instrument models allow the contribution of one voice to be discounted from a compound spectrum. In some cases, a model of the *time-domain* waveform (or complex spectrum) of an instrumental note can be acquired [84, 9, 14] and used to perform exact subtraction. In general however, exact phase information is lacking. In that case, is usual to assume summation of *power*.

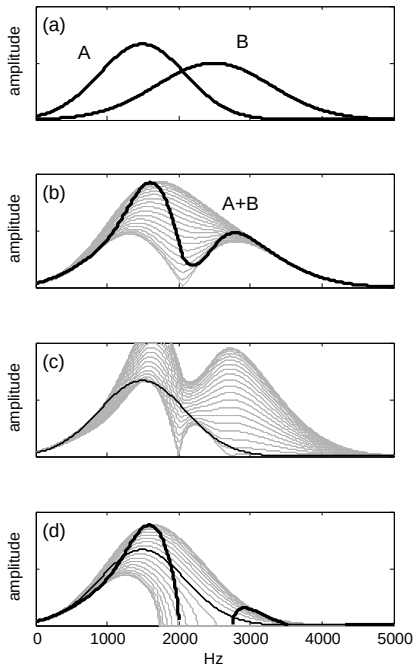

Figure 1.9. Phase dependency of spectral summation and subtraction. (a) Spectral envelopes of two sources A and B. (b) Gray: range of the possible phase-dependent amplitude of $A + B$. Thick line: a possible outcome. (c) Gray: range of values of A that might have produced the previous observation of $A + B$, given knowledge of B but no knowledge of the phase. Thin line: actual value of A. (d) Dotted line: value of A that would have given rise to the previous observation of $A + B$ on the assumption (incorrect) of summation of power, and given knowledge of B. Gray: range of possible estimates of A based on the assumption of power summation, for all possible values of $A + B$. Thin line: actual value of A. The density of gray lines gives an idea of the probability densities. These can be used as a basis for inference according to Bayesian methods.

At this point, it is worth examining more closely the issue of phase-dependent summation and subtraction. Fig. 1.9. (a) illustrates the magnitude spectral envelopes of two voices A and B with unknown phase, and Fig. 1.9. (b) plots in grey the range occupied by the phase-dependent magnitude of their sum $A+B$. It illustrates an annoying fact: vector summation with unknown phase can produce values scattered over a wide range (the spacing of the

curves gives an indication of the probability density). In black is the outcome for some arbitrary phase relation.

Suppose now that we wish to estimate A from this observation of $A + B$ and from prior knowledge of B (for example from a source model). The range of magnitudes of A that could have produced the observation is plotted in grey in Fig. 1.9. (c) (the correct value is in black). On the up side, this distribution is limited: we know for sure that A is *not* outside its range. On the down side, we do not know the exact value, and the range is rather wide. This illustrates a second annoying fact: knowledge of the magnitude spectra of the sum of two voices, and of one term, does not guarantee accurate estimation of the other term.

Lacking better knowledge, it is common to assume *quadrature phase* so that the power of the sum is the sum of the powers (this corresponds to its *expected* value, or average over trials with random phase). If we believe that such was the case here (i.e. we assume that the thick curve in Fig. 1.9. (b) is the sum of powers of A and B), then A can be inferred exactly as illustrated by the thick curve in Fig. 1.9. (d). Part of this curve is missing, indicating that no value of A could have produced the observed value with this interval. The range of estimates of A across all phase-dependent values of $A + B$ is plotted in grey. The thin black line indicates the correct value of A, and as the density of grey plots suggests, it is not a likely outcome. This illustrates a third annoying fact: the assumption of power summation rarely leads to an accurate estimate. Nevertheless, values are constrained by the probability distributions (illustrated by the density of gray lines), and can be further constrained by other sources of information using Bayesian methods (see below) to lead to accurate estimates.

Thus, instrument models are less effective than we would hope for factoring a compound spectrum. To obtain useful results there are at least three options: (a) use a model of the *waveform* or *complex spectrum* instead of the magnitude spectrum, (b) use a *sparse* representation (such as high-resolution spectrum) in which correlates of each voice occupy distinct dimensions so that vector summation does not occur, or (c) fall back on estimates of the *distributions* of values (as in Fig. 1.9.) to inform probabilistic models such as those discussed in the next section.

1.6.3 Learning-based techniques

Many recent methods may be loosely described as based on "machine learning". Chapter 4 treats these techniques in greater detail; they are evoked here briefly in respect to F_0 estimation. The idea is to learn the parameters of underlying models (for example instrument models) from a database, or to structure the observable waveform or spectrum according to a generative model. Methods differ according to their required assumptions, the nature of the information derived from the database to constrain the model parameters, and the methodological framework. In some cases the harmonic nature of voices is assumed before hand, in others it is "learned" from the data, and in yet others it is not needed for separation (i.e. the method would work with non-harmonic voices). Once the voices have been separated, their F_0s can be derived.

In *deconvolutive sparse coding* [120] the observed signal is assumed to be the sum of a relatively small number of recurring events such as musical notes. The waveform is modeled as a series of pulses (note onsets) convolved with time-frequency magnitude spectrograms (notes), and the method tries to derive both the pulse positions and the shape of the notes.

In *non-negative sparse coding* [119, 120, 1, 112] the power or magnitude spectrum is modeled as a sum of spectra of individual events that occur relatively rarely ("sparsely").

For [1, 119] each event has an invariant spectral shape with time-varying amplitude (outer product of time and spectrum shapes, see above), but for [120] events are allowed to have an arbitrarily complex spectrotemporal shape (nevertheless stereotyped from event to event). Notes from the same voice are treated as distinct events, and successive events then need to be clustered into "voices". In the absence of phase information, it is reasonable to assume summation of *power* spectra (see above), but magnitude spectra are reported to work as well or better [120]. Actually, the choice of power versus magnitude is important only if both sources have significant energy within the same time-frequency pixels. It is moot for sources in which energy is sparsely distributed within the time-frequency plane (as is often the case for speech or environmental sounds [99]). In that case, summation of spectra can just as well be modeled by taking their *maximum* within each pixel, and each pixel thus "belongs" to a different source. Non-negative decomposition has the advantage that components are guaranteed to be everywhere non-negative, as is appropriate for magnitude or power spectra. This is in contrast to methods such as deconvoluton, independent component analysis, or independent subspace analysis [13] that may well produce negative values.

Sagayama and colleagues [100] take a very different approach, by modeling the log-frequency *spectrum* as the result of convolution of a common harmonic spectrum by a number of "impulses" that each represents an individual F_0. Multiple F_0 estimates are obtained by deconvolution along the spectral axis. Saul and colleagues [102] perform a similar operation using nonnegative deconvolution in the spectral domain, and the log-lag ACF of [34] might allow a similar operation in the lag domain.

Bayesian approaches [58, 15, 59, 42, 113, 19, 94, 95, 124] treat the spectrum as the result of a number of "causes", and use a generative model to infer them from the observations [90]. Parameters of the generative model may be derived from previous knowledge, or they may be learned from isolated sounds [59, 109] or from repeated exposure to mixtures [90, 95]. An important distinction is between rule-based systems [77] and systems that undergo unsupervised learning [95, 99]. *Hidden markov models* are used to model transitions between numbers of voices [125, 128], values of F_0, or spectral shapes [5]. The models themselves may embody typical spectral shapes, simple transition probability rules (bigrams) or more complex grammars of music or language. Spectrum distributions are often modeled as gaussians, although there is reason to believe that typical distributions may have rather different shape (e.g. Fig. 1.9.).

1.7 ESTIMATING THE NUMBER OF SOURCES

Up to this point, no concern was given to finding the number of voices present within a mixture. This is a difficult aspect of F_0 estimation. Many studies ignore it and concentrate on the simpler task of producing some fixed number of estimates, regardless of the number of voices actually present. Human listeners too have difficulty counting voices. When asked to estimate the number of voices in mixtures of one to ten, subjects asymptoted at about three [60]. When asked to count voices in four-voice polyphony, musicians underestimated their number on half the trials [51].

Some signals are inherently ambiguous, and may be interpreted either as a single voice with low F_0, or as the sum of several voices with higher, harmonically-related F_0s. An algorithm tuned to find as many voices as possible (or to favor the shortest possible periods) may "dismember" a voice into subsets of partials. Tuned to find as few as possible (or the

longest possible periods), it may coalesce multiple voices. The voice count is accordingly over- or underestimated.

Iterative estimation methods typically apply a model at each iteration, and assign as much signal power to a voice as fits this model. Iteration continues on the remainder, and stops when the spectrum (or waveform) has been depleted of power. In the presence of *noise*, it may be hard to distinguish between residual noise and yet another source. It is necessary to set a threshold, and depending on its value the algorithm is prone to either miss voices, or to find spurious voices. The aperiodicity of "real-world" sources has an effect similar to noise.

In the method of [29], cancellation filters are applied successively to remove each periodic voice. The algorithm stops when application of a new filter reduces power by less than a criterion ratio. Klapuri [62] evaluates the "global weight" of the F_0 candidate derived from the residual after a voice has been suppressed, and stops the search if that weight falls below threshold. The system of Goto [40] uses "agents" that are created and terminated at each note onset and offset, while Martin [77] tracks hypotheses within a "blackboard" system [81, 90]. In nonnegative deconvolution [102, 109] the number of voices is given by the number of elements of the deconvolved matrix with amplitudes greater than some threshold. Wu and colleagues [127] use an HMM to model transitions between states of 0, 1 or 2 voices. Probabilities of state-to-state transitions, and value-to-value transitions within a state, were estimated from a speech database. The Bayesian system of Walmsley and colleagues [124] estimates the number of voices together with their F_0s. Information criteria (BIC, AIC) offer a principled basis to decide whether the complexity of a model (such as a sum of periodic voices) is justified by the data [53, 54]. Some music transcription systems detect note onsets and offsets; this information is also of use to determine the number of notes.

1.8 EVALUATION

Evaluation is perhaps the weakest aspect of research on multiple F_0 estimation. The many schemes suggested so far are difficult to rank on any common scale, for lack of widely accepted databases and metrics. Factors of difficulty are the diversity of target tasks (number of and nature of voices, application requirements, available knowledge, etc), the lack of databases, and the large size of the parameter space to be tested. The situation is complicated by paradoxical effects, for example the fact that random answers are more likely to be correct if the task is to estimate many voices (*a priori* a harder task). The authors of each algorithm tend to choose, as it were, a "niche" where the algorithm has the best chances of behaving favorably. Without universally established methodology, progress is difficult to measure. One can propose:

- Evaluation should be *differential*, proceeding by comparison of different algorithms (hopefully some of which are well-known and competitive) on the same task and database. Widely available databases (e.g. [41]) and software implementations of algorithms are among the most useful contributions that a researcher can make to this field.

- Ground truth should be available and agreed upon. For estimation of single voice F_0 this may involve speech labeled with laryngograph signals, or music produced from known scores. For polyphonic F_0 estimation the situation is easier, as it is usually easy to assemble polyphonic test data from monophonic material that has

been labeled by a well-known single-voice algorithm. The labels must nevertheless be made public to allow meaningful comparison.

- Evaluation should be designed to exercise each part (or knowledge source) individually. For example to assess the effectiveness of an instrument model, it is necessary to include also a condition where this model is disabled. In addition it is instructive to know how far an algorithm can go on purely bottom-up data. Estimation of the *number* of sources should be evaluated separately from estimation of their values. Evaluation should include conditions with synthetic data that fit perfectly the underlying model, as well as "real-world" data with imperfections.

- Evaluation should strive to distinguish between performance limits due to algorithm limitations (e.g. insufficient spectral resolution) from those due to imperfections of the material (e.g. inharmonicity or imperfect repeatability of natural sounds). Noise, if present, must be fully described. For example the mere specification of noise RMS is meaningless if the *bandwidth* of the noise is unknown.

Hopefully, as the field matures, evaluation methodology will converge onto a set of widely accepted procedures that will allow progress to be evaluated.

1.9 APPLICATION SCENARII

As mentioned in the Introduction, multiple F_0 estimation is useful for a wide range of applications. In *automatic transcription*, audio data are processed to obtain a score. In *score following* the score preexists and the task is simply to align it temporally with incoming music. These two tasks differ widely in difficulty, and the same is true of each of them applied to different musical genres. Some applications involve a simpler task, such as extraction of the "predominant F_0" [38, 41, 39], bass line [39], or melody [37, 88] from polyphonic audio. In interactive music applications, F_0 estimation may need to be performed in real time.

A relatively new application is *content-based indexing* of audio databases. In the often-cited task of "query by humming", a piece of music is retrieved on the basis of a short extract hummed or whistled (more or less accurately) by the user. A straightforward approach is to label polyphonic material with a multiple F_0 estimation algorithm, and compare the query to the multiple-track labels. However, a more effective alternative may be to match the query directly against the material, or against a "mid-level representation". Query-by-humming is just one example of a wide range of useful operations involving content-based indexing. An important issue is the scalability of such indexing representations [25, 10].

This book is interested in computational auditory scene analysis, for which F_0 is an important ingredient. It can be used in two ways: to help extract a periodic source from a background of interfering sources (themselves not necessarily periodic), or else to help extract a source (itself not necessarily periodic) from a background of periodic sources. The first strategy is termed "harmonic enhancement", the second "harmonic cancellation". Both are *a priori* possible, but it has been argued that the second (cancellation) is more readily used by the auditory system [23] and also more effective in many situations [26]. It is not always necessary to extract the F_0 of *all* sources: enhancement requires only the F_0(s) of the source(s) to be extracted, while cancellation requires only the F_0(s) of the interfering source(s). F_0 may be combined with other cues such as cross-channel correlation in a multi-microphone setup, and possibly jointly estimated with those cues [6].

The diversity of applications illustrates the fact that the term "multiple F_0 estimation" actually subsumes a wide variety of tasks. It underscores the difficulty of evaluation, and indeed, of defining the goal to be attained by future efforts. In principle, there is no hard limit on the number of F_0s that can be extracted from a complex mixture, supposing their values are not in simple ratios for which the task is undetermined. In practice, noise and imperfect periodicity will always conspire to make the task difficult.

1.10 CONCLUSION

To summarize, a wide range of methods may be found in the literature. While many may appear extremely sophisticated, most boil down to a handful of very simple ideas. A major divide is between spectral approaches based on the FFT, and temporal methods based directly on the waveform. Intermediate between them are spectrotemporal approaches inspired from models of auditory processing. Another divide is between methods that rely purely on information derived from the signal within a particular analysis frame, and those that incorporate other sources of knowledge. Among them one may distinguish methods based on learning from those based on "hardwired" rules. Within each method one can distinguish two logical steps: F_0 estimation of individual voices, and separation (or suppression) of voices so that the single-F_0 estimation may be effective. These two steps may be performed iteratively, each voice being estimated and suppressed in turn, or jointly, all steps being performed at once.

The difficulty of the task is its diversity, as there is little unity between the situations and applications that call for estimation. For example polyphonic *transcription* (production of a score from audio) and *score following* (alignment of a score to music) are of incommensurable difficulty, as are each of these tasks applied to different musical genres. A major obstacle is the lack of good evaluation methodology.

If I were to speculate on future progress in this field, I would suggest that it would come from (a) better evaluation, (b) better understanding of the basic issues involved in bottom-up estimation from the signal, and (c) principled incorporation of "high-level" sources of knowledge, most likely within a Bayesian framework.

1.11 ACKNOWLEDGEMENTS

Malcolm Slaney's AuditoryToolbox [111] was used for simulations. Thanks to Hideki Kawahara and the editors for useful comments on initial versions of this chapter.

REFERENCES

1. S.A. Abdallah and M.D. Plumbley. Polyphonic music transcription by non-negative sparse coding of power spectra. In *ISMIR*, pages 318–325, 2004.

2. T. Abe, T. Kobayashi, and S. Imai. Harmonics tracking and pitch extraction based on instantaneous frequency. In *IEEE-ICASSP*, pages 756–759, 1995.

3. T. Abe, T. Kobayashi, and S. Imai. The if spectrogram: a new spectral representation. In *International Conference on Simulation, Visualization and Auralization for Acoustic Research and Education*, pages 423–440, Tokyo, 1997.

4. Y. Atake, T. Irino, H. Kawahara, J. Lu, S. Nakamura, and K. Shikano. Robust fundamental frequency estimation using instantaneous frequencies of harmonic components. In *ICLSP*, volume II, pages 907–910, 2000.

5. F. Bach and M. Jordan. Discriminative training of hidden markov models for multiple pitch tracking. In *ICASSP*, volume V, pages 489–492, 2005.

6. A. Baskind and A. de Cheveigné. Pitch-tracking of reverberant sounds, application to spatial description of sound scenes. In *AES*, Banff Centre, Canada, 2003.

7. J.G. Beerends. *Pitches of simultaneous complex tones*. PhD thesis, Technical University of Eindhoven, 1989.

8. J.G. Beerends and A.J.M. Houtsma. Pitch identification of simultaneous diotic and dichotic two-tone complexes. *J. Acoust. Soc. Am.*, 85:813–819, 1989.

9. J.P. Bello Correa. *Towards the automated analysis of simple polyphonic music: a knowledge-based approach*. PhD thesis, University of London, 2003.

10. N. Bertin and A. de Cheveigné. Scalable metadata and quick retrieval of audio signals. In *ISMIR*, London, 2005.

11. P. Cano. Fundamental frequency estimation in the sms analysis. In *COST G6 Conference on digital audio effects*, pages 99–102, 1998.

12. P.A. Cariani and B. Delgutte. Neural correlates of the pitch of complex tones. i. pitch and pitch salience. *J. Neurophysiol.*, 76:1698–1716, 1996.

13. M.A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *ICMC*, pages 154–161, 2000.

14. Cauwenberghs. Monaural separation of independent acoustical components. In *IEEE Symp. Circuit and Systems (ISCAS)*, 1999.

15. A.T. Cemgil. *Bayesian music transcription*. PhD thesis, Radboud Universiteit Nijmegen, 2004.

16. F.J. Charpentier. Pitch detection using the short-term phase spectrum. In *ICASSP*, pages 113–116, 1986.

17. E Chilton and B.G. Evans. The spectral autocorrelation applied to the linear prediction residual of speech for robust pitch detection. In *IEEE-ICASSP*, pages 358–361, 1988.

18. M. P. Cooke. *Modeling auditory processing and organisation*. PhD thesis, Sheffield, Department of Computer Science, 1991.

19. M. Davy and S. Godsill. Bayesian harmonic models for musical signal analysis. In *Bayesian Statistics 7*, pages 105–124. Oxford University Press, Oxford, 2003.

20. E. de Boer. *On the "residue" in hearing*. PhD thesis, Amsterdam, 1956.

21. E. de Boer. On the "residue" and auditory pitch perception. In W.D. Keidel and W.D Neff, editors, *Handbook of sensory physiology, vol V-3*, pages 479–583. Springer-Verlag, Berlin, 1976.

22. A. de Cheveigné. Speech f0 extraction based on licklider's pitch perception model. In *ICPhS*, volume 4, pages 218–221, 1991.

23. A. de Cheveigné. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *J. Acoust. Soc. Am.*, 93:3271–3290, 1993.

24. A. de Cheveigné. Cancellation model of pitch perception. *J. Acoust. Soc. Am.*, 103:1261–1271, 1998.

25. A. de Cheveigné. Scalable metadata for search, sonification and display. In *International Conference on Auditory Display (ICAD 2002)*, pages 279–284, Kyoto (June 2002), 2002.

26. A. de Cheveigné. The cancellation principle in acoustic scene analysis. In P. Divenyi, editor, *Perspectives on Speech Separation*, pages 243–257. Kluwer, New York, 2004.
27. A. de Cheveigné. Pitch perception models. In C.J. Plack, A. Oxenham, R.R. Fay, and A.N. Popper, editors, *Pitch - Neural coding and perception*. Springer, New York, 2005.
28. A. de Cheveigné. Separable representations for cocktail party processing. In *Forum Acusticum (FA2005)*, Budapest, 2005.
29. A. de Cheveigné and A. Baskind. F0 estimation of one or several voices. In *Eurospeech*, pages 833–836, 2003.
30. A. de Cheveigné and H. Kawahara. Multiple period estimation and pitch perception model. *Speech Communication*, 27:175–185, 1999.
31. A. de Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111:1917–1930, 2002.
32. I Dologlou and G. Carayannis. Pitch detection based on zero-phase filtering. *Speech Comm.*, 8:309–318, 1990.
33. H. Duifhuis, L.F. Willems, and R.J. Sluyter. Measurement of pitch in speech: an implementation of goldstein's theory of pitch perception. *J. Acoust. Soc. Am.*, 71:1568–1580, 1982.
34. D. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, MIT, 1996.
35. J.B.J. Fourier. *Traité analytique de la chaleur*. Didot, Paris, 1820.
36. R.H. Frazier, S. Samsam, L.D. Braida, and A.V. Oppenheim. Enhancement of speech by adaptive filtering. In *Proc. IEEE ICASSP*, pages 251–253, 1976.
37. E. Gomez, S. Streich, B. Ong, R.P. Paiva, S. Tappert, and J.-M. Batke. A quantitative comparison of different approaches for melody extraction from polyphonic audio recordings. *IEEE Trans ASSP*, submitted.
38. M. Goto. A predominant-f0 estimation method for cd recordings: Map estimation using em algorithm for adaptive tone models. In *ICASSP*, volume V, pages 3365–3368, 2001.
39. M. Goto. A predominant-f0 estimation method for polyphonic musical audio signals. In *ICA*, volume II, pages 1085–1088, 2004.
40. M. Goto. A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43:311–329, 2004.
41. M. Goto, H. Hashiguchi, and S. Hayamizu. Rwc music database: popular, classic, and jazz music databases. In *ISMIR*, pages 287–288, 2002.
42. S. Hainsworth. *Techniques for the automated analysis of musical audio*. PhD thesis, University of Cambridge, 2003.
43. S.W. Hainsworth and P.J. Wolfe. Time-frequency reassignment for musical analysis. In *ICMC*, pages 14–17, 2001.
44. D.J. Hermes. Measurement of pitch by subharmonic summation. *J. Acoust. Soc. Am.*, 83:257–264, 1988.
45. W. Hess. *Pitch determination of speech signals*. Springer-Verlag, Berlin, 1983.
46. V. Hohmann. Frequency analysis and synthesis using a gammatone filterbank. *Acta Acustica united with Acustica*, 88:433 ˇ 442, 2002.
47. D.M. Howard. Peak-picking fundamental period estimation for hearing prostheses. *J. Acoust. Soc. Am.*, 86:902–910, 1989.
48. I Howard. *Speech fundamental period estimation using pattern classification*. PhD thesis, London, 1991.
49. G. Hu and D.L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans on Neural Networks*, 15:1135–1150, 2004.

50. X. Huang, A. Acero, and H.-W. Hon. *Spoken language processing*. Prentice Hall PTR, Upper Saddle River (NJ), 2001.

51. D. Huron. Voice denumerability in polyphonic music of homogenous timbres. *Music Perception*, 6:361–382, 1989.

52. C.H. Hurst. A new theory of hearing. *Proc. Trans. Liverpool Biol. Soc.*, 9:321–353 (and plate XX), 1895.

53. R.A. Irizarry. Local harmonic estimation in musical sound signals. *Journal of the American Statistical Association*, 96:357–367, 2001.

54. H. Kameoka, T. Nishimoto, and S. Sagayama. Separation of harmonic structures based on tied gaussian mixture model and information criterion for concurrent sounds. In *ICASSP*, volume IV, pages 297–300, 2004.

55. H. Kameoka, T. Nishimoto, and S. Sagayama. Audio stream segregation of multi-pitch music signal based on time-space clustering using gaussian kernel 2-dimensional model. In *ICASSP*, volume III, pages 5–8, 2005.

56. M. Karjalainen and T. Tolonen. Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis. In *ICASSP*, pages 929–932, 1999.

57. K. Kashino and S.J Godsill. Bayesian estimation of simultaneous musical notes based on frequency domain modelling. In *ICASSP*, volume IV, pages 305–308, 2004.

58. K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka. Organization of hierarchical perceptual sounds: Music scene analysis with autonomous processing modules and a quantitative information integration mechanism. In *IJCAI*, volume 1, pages 158–164, 1995.

59. K. Kashino, K. Nakadai, T. Kinoshita, and H Tanaka. Application of the bayesian probability network to music scene analysis. In D.F. Rosenthal and H.G. Okuno, editors, *Computational Auditory Scene Analysis*, pages 115–137. Lawrence Erlbaum Associates, 1998.

60. M. Kashino and T. Hirahara. How many concurrent talkers can we hear out? In *ASJ Autumn meeting (in Japanese)*, pages 467–468, 1995.

61. H. Kawahara, A. de Cheveigné, and R.D. Patterson. An instantaneous-frequency-based pitch extraction method for high quality speech transformation: revised tempo in the straight-suite. In *ICSLP*, page (Dec. 1998), 1998.

62. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Trans. ASSP*, 11:804–816, 2003.

63. A. Klapuri. Multipitch estimation and sound separation by the spectral smoothness principle. In *ICASSP*, 2001.

64. A. Klapuri. *Signal processing methods for the automatic transcription of music*. PhD thesis, Tampere, 2002.

65. A. Klapuri. Auditory-model based methods for multiple fundamental frequency estimation. In A. Klapuri and M. Davy, editors, *Signal processing methods for music transcription*, page in press. Springer, New York, 2005.

66. A. Klapuri. A perceptually-motivated multiple-f0 estimation method. In *IEEE WASPAA*, New Palz, New York, 2005.

67. A. Klapuri, T. Virtanen, and J.-M. Holm. Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals. In *COST-G6 Conference on digital audio effects*, Veron, Italy, 2000.

68. N. Kunieda, T. Shimamura, and J. Suzuki. Robust method of measurement of fundamental frequency by aclos - autocorrelation of log spectrum. In *ICASSP*, 1996.

69. T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine. Splitting the unit delay¯tools for fractional delay filter design. *IEEE Signal Proc. Mag.*, 13:30˘60, 1996.

70. M. Lahat, R.J. Niederjohn, and D.A. Krubsack. A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech. *IEEE Trans. ASSP*, 35:741–750, 1987.

71. R.J. Leistikow, H.D. Thornburg, J.O.III Smith, and J. Berger. Bayesian identification of closely-spaced chords from single-fram stft peaks. In *DAFX*, pages 5–8, 2004.

72. Y. Li and D. Wang. Detecting pitch of singing voice in polyphonic audio. In *ICASSP*, volume III, pages 17–20, 2005.

73. J.C.R. Licklider. A duplex theory of pitch perception. *Experientia*, 7:128–134, 1951.

74. R. Lyon. Computational models of neural auditory processing. In *IEEE ICASSP*, pages 36.1.(1–4), 1984.

75. R.F. Lyon. A computational model of binaural localization and separation. In W. Richards, editor, *Natural computation*, pages 319–327. MIT Press, Cambridge, Mass, 1983-1988. reprinted from Proc. ICASSP 83, 1148-1151.

76. K.D. Martin. Automatic transcription of simple polyphonic music: robust front end processing. Technical report, MIT Media Laboratory Perceptual Computing Section TR no 399., 1996.

77. K.D. Martin. A blackboard system for automatic transcription of simple polyphonic music. Technical Report 385, MIT Medial Laboratory Perceptual Computing, 1996.

78. P. Martin. Comparison of pitch detection by cepstrum and spectral comb analysis. In *IEEE ICASSP*, pages 180–183, 1982.

79. R. Meddis and M.J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification. *J. Acoust. Soc. Am.*, 89:2866–2882, 1991.

80. R. Meddis and M.J. Hewitt. Modeling the identification of concurrent vowels with different fundamental frequencies. *J. Acoust. Soc. Am.*, 91:233–245, 1992.

81. D. K. Mellinger. *Event formation and separation in musical sound*. PhD thesis, Stanford Center for computer research in music and acoustics, 1991.

82. B.C.J. Moore and B.R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.*, 74:750–753, 1983.

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

84. T. Nakatani, H.G. Okuno, and T. Kawabata. Residue-driven architecture for computational auditory scene analysis. In *IJCAI*, volume 1, pages 165–172, 1995.

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

86. L.I. Ortiz-Berenguer and Casajús-Quirós. Polyphonic transcription using piano modelin for spectral pattern recognition. In *DAFX*, pages 45–50, Hamburg, Germany, 2002.

87. A. Oxenham and C.J. Plack. *Pitch - Neural coding and perception*. Springer, New York, 2005.

88. R.P. Paiva, T. Mendes, and A. Cardoso. On the detection of melody notes in polyphonic audio. In *ISMIR*, 2005.

89. T.W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *J. Acoust. Soc. Am.*, 60:911–918, 1976.

90. M.D. Plumbley, S.A. Abdallah, J.P. Bello, M.E. Davies, G. Monti, and M.B. Sandler. Automatic music transcription and audio source separation. *Cybernetics and systems*, 33:603–627, 2002.

91. D. Pressnitzer and D. Gnansia. Real-time auditory models. In *ICMC*, Barcelona, Spain, 2005.

92. D. Pressnitzer, R. D. Patterson, and K. Krumbholz. The lower limit of melodic pitch. *Journal of the Acoustical Society of America*, 109:2074–2084, 2001.

93. L.R. Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE Trans. ASSP*, 25:24–33, 1977.

94. C. Raphael. Automatic transcription of piano music. In *ISMIR*, pages 15–19, 2002.

95. C. Raphael and J. Stoddard. Harmonic analysis with probabilistic graphical models. *Computer music journal*, 28:45–52, 2004.

96. M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, and H.J. Manley. Average magnitude difference function pitch extractor. *IEEE Trans. ASSP*, 22:353–362, 1974.

97. L. Rossi, G. Girolami, and M. Leca. Identification of polyphonic piano signals. *Acustica*, 83:1077–1084, 1997.

98. J. Rouat, C.Y. Liu, and D. Morissette. A pitch determination and voiced/unvoiced decision algorithym for noisy speech. *Speech Comm*, 21:191–207, 1997.

99. S. Roweis. One-microphone source separation. In *Advances in NIPS*, page 609˘616. MIT Press, Cambridge MA, 2000.

100. S. Sagayama, K. Takahashi, H. Kameoka, and T. Nishimoto. Specmurt anasylis: a piano-roll-visualization of polyphonic music signal by deconvolution of log-frequency spectrum. In *SAPA (ISCA tutorial and research workshop on statistical and perceptual audio processing)*, Jeju, Korea, 2004.

101. L.K. Saul, L.L. Lee, C.L. Isbel, and Y. LeCun. Real time voice processing with audiovisual feedback: toward autonomous agents with perfect pitch. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in informatioin processing systems 15*, pages 1205–1212. MIT Press, Cambridge, MA., 2003.

102. L.K. Saul, F. Sha, and D.D. Lee. Statistical signal processing with nonnegativity constraints. In *Eigth European Conference on Speech Communication and Technology*, volume 2, pages 1001–1004, 2003.

103. M.T.M. Scheffers. *Sifting vowels*. PhD thesis, Gröningen, 1983.

104. M.R. Schroeder. Period histogram and product spectrum: new methods for fundamental-frequency measurement. *J. Acoust. Soc. Am.*, 43:829–834, 1968.

105. C. Semal and L. Demany. The upper limit of musical pitch. *Music Perception*, 8:165–176, 1990.

106. S. Seneff. Pitch and spectral estimation of speech based on auditory synchrony model. In *IEEE ICASSP*, pages 36.2.1–4, 1984.

107. X. Serra. Musical sound modeling with sinusoids plus noise. In C. Roads, S. Pope, A. Picialli, and G. De Poli, editors, *Musical signal processing*. Swets & Zeitlinger, 1997.

108. F. Sha, J.A. Burgoyne, and L.K. Saul. Multiband statistical learning for f0 estimation in speech. In *ICASSP*, pages 661–664, 2004.

109. F. Sha and L.K. Saul. Real-time pitch determination of one or more signals by nonnegative matrix factorization. In L.K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Processing Systems 17*. MIT Press, Cambridge, MA, 2005.

110. M. Slaney. A perceptual pitch detector. In *ICASSP*, pages 357–360, 1990.

111. M. Slaney. An efficient implementation of the patterson-holdsworth auditory filter bank. technical report 35, Apple Computer, 1993.

112. P. Smaragdis. Discovering auditory objects through non-negativity constraints. In *Workshop on statistical and perceptual audio processing*, Jeju, Korea, 2004.

113. A.D. Sterian. *Model-based segmentation of time-frequency images for music transcription*. PhD thesis, The University of Michigan, 1999.

114. R.J. Stubbs and Q. Summerfield. Evaluation of two voice-separation algorithms using normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 84:1236–1249, 1988.

115. R.J. Stubbs and Q. Summerfield. Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 87:359–372, 1990.

116. H.D. Thornburg and R.J. Lestikow. An iterative filterbank approach for extracting sinusoidal parameters from quasi-harmonic sounds. In *IEEE WASPAA*, 2003.

117. T. Tolonen and M. Karjalainen. Computationally efficient multiplitch analysis model. *IEEE Trans Speech and Audio Processing*, 8:708–716, 2000.

118. T. Virtanen. *Audio signal modeling with sinusoids plus noise*. Master of science thesis, Tampere University of Technology, 2000.

119. T. Virtanen. Sound source separation using sparse coding with temporal continuity objective. In *ICMC*, pages 231–234, Singapore, 2003.

120. T. Virtanen. Separation of sound sources by convolutive sparse coding. In *ISCA tutorial and research workshop on statistical and perceptual audio process - SAPA 2004*, Jeju Korea, 2004.

121. T. Virtanen and A. Klapuri. Separation of harmonic sounds using multipitch analysis and iterative parameter estimation. In *IEEE WASPAA*, pages 83–86, 2001.

122. T. Virtanen and A. Klapuri. Separation of harmonic sounds using linear models for the overtone series. In *IEEE ICASSP*, volume 2, pages 1757–1760, 2002.

123. H. von Helmholtz. *On the sensations of tone (English translation A.J. Ellis, 1885, 1954)*. Dover, New York, 1877.

124. P.J. Walmsley, S. Godsill, and P.J.W. Rayner. Bayesian graphical models for polyphonic pitch tracking. In *Diderot Forum*, pages 1–26, Vienna, 1999.

125. M. Weintraub. *A theory and computational model of auditory monaural sound separation*. PhD thesis, Stanford, 1985.

126. M. Wu. *Pitch tracking and speech enhancement in noisy and reverberant environments*. PhD thesis, Ohio State University, 2003.

127. M. Wu, D. Wang, and G.J. Brown. A multi-pitch tracking algorithm for noisy speech. In *IEEE ICASSP*, volume I, pages 369–372, 2002.

128. M. Wu, D. Wang, and G.J. Brown. A multipitch tracking algorithm for noisy speech. *IEEE Trans. ASSP*, 11:229–241, 2003.

129. C. Yeh, A. Roebel, and X. Rodet. Multiple fundamental frequency estimation of polyphonic music signals. In *ICASSP*, volume III, pages 225–228, 2005.

130. E. Zwicker, G. Flottorp, and S.S. Stevens. Critical band width in loudness summation. *J. Acoust. Soc. Am.*, 29:548–557, 1957.