YIN, a fundamental frequency estimator for speech and music[a)]

Alain de Cheveigné[b)]
Ircam-CNRS, 1 place Igor Stravinsky, 75004 Paris, France

Hideki Kawahara
Wakayama University

(Received 7 June 2001; revised 10 October 2001; accepted 9 January 2002)

An algorithm is presented for the estimation of the fundamental frequency (F_0) of speech or musical sounds. It is based on the well-known autocorrelation method with a number of modifications that combine to prevent errors. The algorithm has several desirable features. Error rates are about three times lower than the best competing methods, as evaluated over a database of speech recorded together with a laryngograph signal. There is no upper limit on the frequency search range, so the algorithm is suited for high-pitched voices and music. The algorithm is relatively simple and may be implemented efficiently and with low latency, and it involves few parameters that must be tuned. It is based on a signal model (periodic signal) that may be extended in several ways to handle various forms of aperiodicity that occur in particular applications. Finally, interesting parallels may be drawn with models of auditory processing. © *2002 Acoustical Society of America.* [DOI: 10.1121/1.1458024]

PACS numbers: 43.72.Ar, 43.75.Yy, 43.70.Jt, 43.66.Hg [DOS]

I. INTRODUCTION

The fundamental frequency (F_0) of a periodic signal is the inverse of its period, which may be defined as the smallest positive member of the infinite set of time shifts that leave the signal invariant. This definition applies strictly only to a *perfectly* periodic signal, an uninteresting object (supposing one exists) because it cannot be switched on or off or modulated in any way without losing its perfect periodicity. Interesting signals such as speech or music depart from periodicity in several ways, and the art of fundamental frequency estimation is to deal with them in a useful and consistent way.

The subjective pitch of a sound usually depends on its fundamental frequency, but there are exceptions. Sounds may be periodic yet ''outside the existence region'' of pitch (Ritsma, 1962; Pressnitzer *et al.*, 2001). Conversely, a sound may not be periodic, but yet evoke a pitch (Miller and Taylor, 1948; Yost, 1996). However, over a wide range pitch and period are in a one-to-one relation, to the degree that the word ''pitch'' is often used in the place of F_0, and F_0 estimation methods are often referred to as ''pitch detection algorithms,'' or PDA (Hess, 1983). Modern pitch perception models assume that pitch is derived either from the periodicity of neural patterns in the time domain (Licklider, 1951; Moore, 1997; Meddis and Hewitt, 1991; Cariani and Delgutte, 1996), or else from the harmonic pattern of partials resolved by the cochlea in the frequency domain (Goldstein, 1973; Wightman, 1973; Terhardt, 1974). Both processes yield the fundamental frequency or its inverse, the period.

Some applications give for F_0 a different definition, closer to their purposes. For voiced speech, F_0 is usually defined as the rate of vibration of the vocal folds. Periodic vibration at the glottis may produce speech that is less perfectly periodic because of movements of the vocal tract that filters the glottal source waveform. However, glottal vibration itself may also show aperiodicities, such as changes in amplitude, rate or glottal waveform shape (for example, the duty cycle of open and closed phases), or intervals where the vibration seems to reflect several superimposed periodicities (diplophony), or where glottal pulses occur without an obvious regularity in time or amplitude (glottalizations, vocal creak or fry) (Hedelin and Huber, 1990). These factors conspire to make the task of obtaining a useful estimate of speech F_0 rather difficult. F_0 estimation is a topic that continues to attract much effort and ingenuity, despite the many methods that have been proposed. The most comprehensive review is that of Hess (1983), updated by Hess (1992) or Hermes (1993). Examples of recent approaches are instantaneous frequency methods (Abe *et al.*, 1995; Kawahara *et al.*, 1999a), statistical learning and neural networks (Barnard *et al.*, 1991; Rodet and Doval, 1992; Doval, 1994), and auditory models (Duifhuis *et al.*, 1982; de Cheveigné, 1991), but there are many others.

Supposing that it can be reliably estimated, F_0 is useful for a wide range of applications. Speech F_0 variations contribute to prosody, and in tonal languages they help distinguish lexical categories. Attempts to use F_0 in speech recognition systems have met with mitigated success, in part because of the limited reliability of estimation algorithms. Several musical applications need F_0 estimation, such as automatic score transcription or real-time interactive systems, but here again the imperfect reliability of available methods is an obstacle. F_0 is a useful ingredient for a variety of signal processing methods, for example, F_0-dependent spectral envelope estimation (Kawahara *et al.*, 1999b). Finally, a fairly recent application of F_0 is as metadata for multimedia content indexing.

[a)]Portions of this work were presented at the 2001 ASA Spring Meeting and the 2001 Eurospeech conference.
[b)]Electronic mail: cheveign@ircam.fr

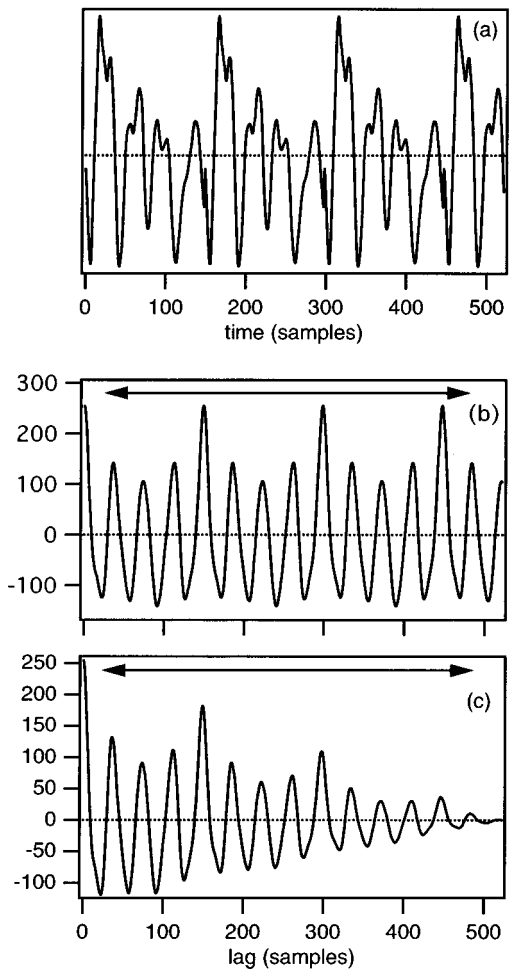

FIG. 1. (a) Example of a speech waveform. (b) Autocorrelation function (ACF) calculated from the waveform in (a) according to Eq. (1). (c) Same, calculated according to Eq. (2). The envelope of this function is tapered to zero because of the smaller number of terms in the summation at larger τ. The horizontal arrows symbolize the search range for the period.

The present article introduces a method for F_0 estimation that produces fewer errors than other well-known methods. The name YIN (from "yin" and "yang" of oriental philosophy) alludes to the interplay between autocorrelation and cancellation that it involves. This article is the first of a series of two, of which the second (Kawahara *et al.*, in preparation) is also devoted to fundamental frequency estimation.

II. THE METHOD

This section presents the method step by step to provide insight as to what makes it effective. The classic autocorrelation algorithm is presented first, its error mechanisms are analyzed, and then a series of improvements are introduced to reduce error rates. Error rates are measured at each step over a small database for illustration purposes. Fuller evaluation is proposed in Sec. III.

A. Step 1: The autocorrelation method

The autocorrelation function (ACF) of a discrete signal x_t may be defined as

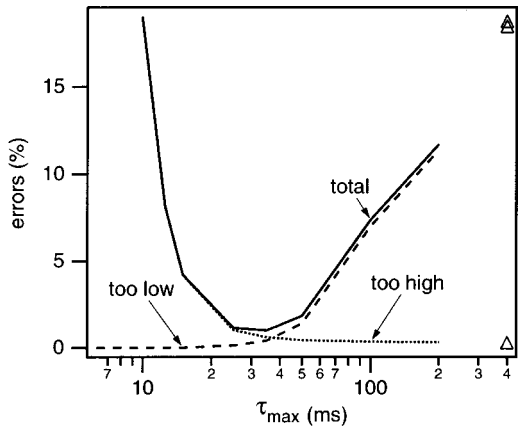

FIG. 2. F_0 estimation error rates as a function of the slope of the envelope of the ACF, quantified by its intercept with the abscissa. The dotted line represents errors for which the F_0 estimate was too high, the dashed line those for which it was too low, and the full line their sum. Triangles at the right represent error rates for ACF calculated as in Eq. (1) ($\tau_{\max}=\infty$). These rates were measured over a subset of the database used in Sec. III.

$$r_t(\tau)=\sum_{j=t+1}^{t+W} x_j x_{j+\tau}, \tag{1}$$

where $r_t(\tau)$ is the autocorrelation function of lag τ calculated at time index t, and W is the integration window size. This function is illustrated in Fig. 1(b) for the signal plotted in Fig. 1(a). It is common in signal processing to use a slightly different definition:

$$r'_t(\tau)=\sum_{j=t+1}^{t+W-\tau} x_j x_{j+\tau}. \tag{2}$$

Here the integration window size shrinks with increasing values of τ, with the result that the envelope of the function decreases as a function of lag as illustrated in Fig. 1(c). The two definitions give the same result if the signal is zero outside $[t+1, t+W]$, but differ otherwise. Except where noted, this article assumes the first definition (also known as "modified autocorrelation," "covariance," or "cross-correlation," Rabiner and Shafer, 1978; Huang *et al.*, 2001).

In response to a periodic signal, the ACF shows peaks at multiples of the period. The "autocorrelation method" chooses the highest non-zero-lag peak by exhaustive search within a range of lags (horizontal arrows in Fig. 1). Obviously if the lower limit is too close to zero, the algorithm may erroneously choose the zero-lag peak. Conversely, if the higher limit is large enough, it may erroneously choose a higher-order peak. The definition of Eq. (1) is prone to the second problem, and that of Eq. (2) to the first (all the more so as the window size W is small).

To evaluate the effect of a tapered ACF envelope on error rates, the function calculated as in Eq. (1) was multiplied by a negative ramp to simulate the result of Eq. (2) with a window size $W=\tau_{\max}$:

$$r''_t(\tau)=\begin{cases} r_t(\tau)(1-\tau/\tau_{\max}) & \text{if } \tau \leqslant \tau_{\max}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Error rates were measured on a small database of speech (see Sec. III for details) and plotted in Fig. 2 as a function of

TABLE I. Gross error rates for the simple unbiased autocorrelation method (step 1), and for the cumulated steps described in the text. These rates were measured over a subset of the database used in Sec. III. Integration window size was 25 ms, window shift was one sample, search range was 40 to 800 Hz, and threshold (step 4) was 0.1.

Version	Gross error (%)
Step 1	10.0
Step 2	1.95
Step 3	1.69
Step 4	0.78
Step 5	0.77
Step 6	0.50

$\tau_{\max}$. The parameter $\tau_{\max}$ allows the algorithm to be biased to favor one form of error at the expense of the other, with a minimum of total error for intermediate values. Using Eq. (2) rather than Eq. (1) introduces a natural bias that can be tuned by adjusting W. However, changing the window size has other effects, and one can argue that a bias of this sort, if useful, should be applied explicitly rather than implicitly. This is one reason to prefer the definition of Eq. (1).

The autocorrelation method compares the signal to its shifted self. In that sense it is related to the AMDF method (average magnitude difference function, Ross *et al.*, 1974; Ney, 1982) that performs its comparison using differences rather than products, and more generally to time-domain methods that measure intervals between events in time (Hess, 1983). The ACF is the Fourier transform of the power spectrum, and can be seen as measuring the regular spacing of harmonics within that spectrum. The cepstrum method (Noll, 1967) replaces the power spectrum by the log magnitude spectrum and thus puts less weight on high-amplitude parts of the spectrum (particularly near the first formant that often dominates the ACF). Similar ''spectral whitening'' effects can be obtained by linear predictive inverse filtering or center-clipping (Rabiner and Schafer, 1978), or by splitting the signal over a bank of filters, calculating ACFs within each channel, and adding the results after amplitude normalization (de Cheveigné, 1991). Auditory models based on autocorrelation are currently one of the more popular ways to explain pitch perception (Meddis and Hewitt, 1991; Cariani and Delgutte, 1996).

Despite its appeal and many efforts to improve its performance, the autocorrelation method (and other methods for that matter) makes too many errors for many applications. The following steps are designed to reduce error rates. The first row of Table I gives the gross error rate (defined in Sec. III and measured over a subset of the database used in that section) of the basic autocorrelation method based on Eq. (1) without bias. The next rows are rates for a succession of improvements described in the next paragraphs. These numbers are given for didactic purposes; a more formal evaluation is reported in Sec. III.

B. Step 2: Difference function

We start by modeling the signal x_t as a periodic function with period T, by definition invariant for a time shift of T:

$$x_t - x_{t+T} = 0, \quad \forall t. \tag{4}$$

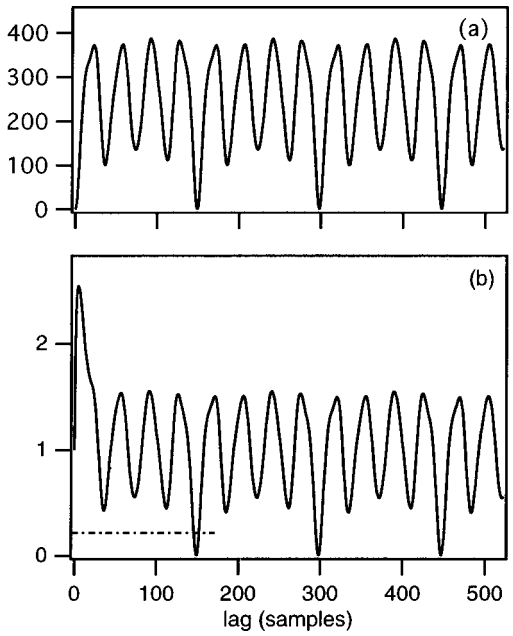

FIG. 3. (a) Difference function calculated for the speech signal of Fig. 1(a). (b) Cumulative mean normalized difference function. Note that the function starts at 1 rather than 0 and remains high until the dip at the period.

The same is true after taking the square and averaging over a window:

$$\sum_{j=t+1}^{t+W} (x_j - x_{j+T})^2 = 0. \tag{5}$$

Conversely, an unknown period may be found by forming the difference function:

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2, \tag{6}$$

and searching for the values of τ for which the function is zero. There is an infinite set of such values, all multiples of the period. The difference function calculated from the signal in Fig. 1(a) is illustrated in Fig. 3(a). The squared sum may be expanded and the function expressed in terms of the ACF:

$$d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau). \tag{7}$$

The first two terms are energy terms. Were they constant, the difference function $d_t(\tau)$ would vary as the opposite of $r_t(\tau)$, and searching for a minimum of one or the maximum of the other would give the same result. However, the second energy term also varies with τ, implying that maxima of $r_t(\tau)$ and minima of $d_t(\tau)$ may sometimes not coincide. Indeed, the error rate fell to 1.95% for the difference function from 10.0% for unbiased autocorrelation (Table I).

The magnitude of this decrease in error rate may come as a surprise. An explanation is that the ACF implemented according to Eq. (1) is quite sensitive to amplitude changes. As pointed out by Hess (1983, p. 355), an increase in signal amplitude with time causes ACF peak amplitudes to grow with lag rather than remain constant as in Fig. 1(b). This encourages the algorithm to choose a higher-order peak and make a ''too low'' error (an amplitude decrease has the opposite effect). The difference function is immune to this par-

ticular problem, as amplitude changes cause period-to-period dissimilarity to increase with lag in all cases. Hess points out that Eq. (2) produces a function that is less sensitive to amplitude change [Eq. (A1) also has this property]. However, using $d(\tau)$ has the additional appeal that this function is more closely grounded in the signal model of Eq. (4), and paves the way for the next two error-reduction steps, the first of which deals with ''too high'' errors and the second with ''too low'' errors.

C. Step 3: Cumulative mean normalized difference function

The difference function of Fig. 3(a) is zero at zero lag and often nonzero at the period because of imperfect periodicity. Unless a lower limit is set on the search range, the algorithm must choose the zero-lag dip instead of the period dip and the method must fail. Even if a limit is set, a strong resonance at the first formant (F1) might produce a series of secondary dips, one of which might be deeper than the period dip. A lower limit on the search range is not a satisfactory way of avoiding this problem because the ranges of F1 and F_0 are known to overlap.

The solution we propose is to replace the difference function by the ''cumulative mean normalized difference function:''

$$d'_t(\tau)=\begin{cases}1, & \text{if } \tau=0,\\ d_t(\tau)\Big/\left[(1/\tau)\sum_{j=1}^{\tau} d_t(j)\right] & \text{otherwise.}\end{cases} \tag{8}$$

This new function is obtained by dividing each value of the old by its average over shorter-lag values. It differs from $d(\tau)$ in that it starts at 1 rather than 0, tends to remain large at low lags, and drops below 1 only where $d(\tau)$ falls below average [Fig. 3(b)]. Replacing d by d' reduces ''too high'' errors, as reflected by an error rate of 1.69% (instead of 1.95%). A second benefit is to do away with the upper frequency limit of the search range, no longer needed to avoid the zero-lag dip. A third benefit is to normalize the function for the next error-reduction step.

D. Step 4: Absolute threshold

It easily happens that one of the higher-order dips of the difference function [Fig. 3(b)] is deeper than the period dip. If it falls within the search range, the result is a subharmonic error, sometimes called ''octave error'' (improperly because not necessarily in a power of 2 ratio with the correct value). The autocorrelation method is likewise prone to choosing a high-order peak.

The solution we propose is to set an absolute threshold and choose the smallest value of τ that gives a minimum of d' deeper than that threshold. If none is found, the global minimum is chosen instead. With a threshold of 0.1, the error rate drops to 0.78% (from 1.69%) as a consequence of a reduction of ''too low'' errors accompanied by a very slight increase of ''too high'' errors.

This step implements the word ''smallest'' in the phrase ''the period is the smallest positive member of a set'' (the previous step implemented the word ''positive''). The threshold determines the list of candidates admitted to the set, and can be interpreted as the proportion of aperiodic power tolerated within a ''periodic'' signal. To see this, consider the identity:

$$2(x_t^2+x_{t+T}^2)=(x_t+x_{t+T})^2+(x_t-x_{t+T})^2. \tag{9}$$

Taking the average over a window and dividing by 4,

$$\begin{aligned}1/(2W)&\sum_{j=t+1}^{t+W}(x_j^2+x_{j+T}^2)\\ &=1/(4W)\sum_{j=t+1}^{t+W}(x_j+x_{j+T})^2+1/(4W)\\ &\quad\times\sum_{j=t+1}^{t+W}(x_j-x_{j+T})^2.\end{aligned} \tag{10}$$

The left-hand side approximates the power of the signal. The two terms on the right-hand side, both positive, constitute a partition of this power. The second is zero if the signal is periodic with period T, and is unaffected by adding or subtracting periodic components at that period. It can be interpreted as the ''aperiodic power'' component of the signal power. With $\tau=T$ the numerator of Eq. (8) is proportional to aperiodic power whereas its denominator, average of $d(\tau)$ for τ between 0 and T, is approximately twice the signal power. Thus, $d'(T)$ is proportional to the aperiodic/total power ratio. A candidate T is accepted in the set if this ratio is below threshold. We'll see later on that the exact value of this threshold does not critically affect error rates.

E. Step 5: Parabolic interpolation

The previous steps work as advertised if the period is a multiple of the sampling period. If not, the estimate may be incorrect by up to half the sampling period. Worse, the larger value of $d'(\tau)$ sampled away from the dip may interfere with the process that chooses among dips, thus causing a gross error.

A solution to this problem is parabolic interpolation. Each local minimum of $d'(\tau)$ and its immediate neighbors is fit by a parabola, and the ordinate of the interpolated minimum is used in the dip-selection process. The abscissa of the selected minimum then serves as a period estimate. Actually, one finds that the estimate obtained in this way is slightly biased. To avoid this bias, the abscissa of the corresponding minimum of the *raw* difference function $d(\tau)$ is used instead.

Interpolation of $d'(\tau)$ or $d(\tau)$ is computationally cheaper than upsampling the signal, and accurate to the extent that $d'(\tau)$ can be modeled as a quadratic function near the dip. Simple reasoning argues that this should be the case if the signal is band-limited. First, recall that the ACF is the Fourier transform of the power spectrum: if the signal x_t is bandlimited, so is its ACF. Second, the ACF is a sum of cosines, which can be approximated near zero by a Taylor series with even powers. Terms of degree 4 or more come mainly from the highest frequency components, and if these are absent or weak the function is accurately represented by

lower order terms (quadratic and constant). Finally, note that the period peak has the same shape as the zero-lag peak, and the same shape (modulo a change in sign) as the period dip of $d(\tau)$, which in turn is similar to that of $d'(\tau)$. Thus, parabolic interpolation of a dip is accurate unless the signal contains strong high-frequency components (in practice, above about one-quarter of the sampling rate).

Interpolation had little effect on gross error rates over the database (0.77% vs 0.78%), probably because F_0's were small in comparison to the sampling rate. However, tests with synthetic stimuli found that parabolic interpolation reduced fine error at all F_0 and avoided gross errors at high F_0.

F. Step 6: Best local estimate

The role of integration in Eqs. (1) and (6) is to ensure that estimates are stable and do not fluctuate on the time scale of the fundamental period. Conversely, any such fluctuation, if observed, should not be considered genuine. It is sometimes found, for nonstationary speech intervals, that the estimate fails at a certain phase of the period that usually coincides with a relatively high value of $d'(T_t)$, where T_t is the period estimate at time t. At another phase (time t') the estimate may be correct and the value of $d'(T_{t'})$ smaller. Step 6 takes advantage of this fact, by ''shopping'' around the vicinity of each analysis point for a better estimate.

The algorithm is the following. For each time index t, search for a minimum of $d'_\theta(T_\theta)$ for θ within a small interval $[t-T_{\max}/2,\ t+T_{\max}/2]$, where T_θ is the estimate at time θ and $T_{\max}$ is the largest expected period. Based on this initial estimate, the estimation algorithm is applied again with a restricted search range to obtain the final estimate. Using $T_{\max}=25$ ms and a final search range of $\pm 20\%$ of the initial estimate, step 6 reduced the error rate to 0.5% (from 0.77%). Step 6 is reminiscent of median smoothing or dynamic programming techniques (Hess, 1983), but differs in that it takes into account a relatively short interval and bases its choice on quality rather than mere continuity.

The combination of steps 1–6 constitutes a new method (YIN) that is evaluated by comparison to other methods in the next section. It is worth noting how the steps build upon one another. Replacing the ACF (step 1) by the difference function (step 2) paves the way for the cumulative mean normalization operation (step 3), upon which are based the threshold scheme (step 4) and the measure $d'(T)$ that selects the best local estimate (step 6). Parabolic interpolation (step 5) is independent from other steps, although it relies on the spectral properties of the ACF (step 1).

III. EVALUATION

Error rates up to now were merely illustrative. This section reports a more formal evaluation of the new method in comparison to previous methods, over a compilation of databases of speech recorded together with the signal of a laryngograph (an apparatus that measures electrical resistance between electrodes placed across the larynx), from which a reliable ''ground-truth'' estimate can be derived. Details of the databases are given in the Appendix. The laryngograph F_0 estimate was derived automatically and checked visually, and estimates that seemed incorrect were removed from the statistics. This process removed unvoiced and also irregularly voiced portions (diplophony, creak). Some studies include the latter, but arguably there is little point in testing an algorithm on conditions for which correct behavior is not defined.

When evaluating the candidate methods, values that differed by more than 20% from laryngograph-derived estimates were counted as ''gross errors.'' This relatively permissive criterion is used in many studies, and measures the difficult part of the task on the assumption that if an initial estimate is within 20% of being correct, any of a number of techniques can be used to refine it. Gross errors are further broken down into ''too low'' (mainly subharmonic) and ''too high'' errors.

In itself the error rate is not informative, as it depends on the difficulty of the database. To draw useful conclusions, different methods must be measured on the same database. Fortunately, the availability of freely accessible databases and software makes this task easy. Details of availability and parameters of the methods compared in this study are given in the Appendix. In brief, postprocessing and voiced–unvoiced decision mechanisms were disabled (where possible), and methods were given a common search range of 40 to 800 Hz, with the exception of YIN that was given an upper limit of one-quarter of the sampling rate (4 or 5 kHz depending on the database).

Table II summarizes error rates for each method and database. These figures should not be taken as an accurate measure of the intrinsic quality of each algorithm or implementation, as our evaluation conditions differ from those for which they were optimized. In particular, the search range (40 to 800 Hz) is unusually wide and may have destabilized methods designed for a narrower range, as evidenced by the imbalance between ''too low'' and ''two high'' error rates for several methods. Rather, the figures are a sampling of the performance that can be expected of ''off-the shelf'' implementations of well-known algorithms in these difficult conditions. It is worth noting that the ranking of methods differs between databases. For example methods ''acf'' and ''nacf'' do well on DB1 (a large database with a total of 28 speakers), but less well on other databases. This shows the need for testing on extensive databases.

YIN performs best of all methods over all the databases. Averaged over databases, error rates are smaller by a factor of about 3 with respect to the best competing method. Error rates depend on the tolerance level used to decide whether an estimate is correct or not. For YIN about 99% of estimates are accurate within 20%, 94% within 5%, and about 60% within 1%.

IV. SENSITIVITY TO PARAMETERS

Upper and lower F_0 search bounds are important parameters for most methods. In contrast to other methods, YIN needs no upper limit (it tends, however, to fail for F_0's beyond one quarter of the sampling rate). This should make it useful for musical applications in which F_0 can become very high. A wide range increases the likelihood of ''finding'' an

TABLE II. Gross error rates for several F_0 estimation algorithms over four databases. The first six methods are implementations available on the Internet, the next four are methods developed locally, and YIN is the method described in this paper. See Appendix for details concerning the databases, estimation methods, and evaluation procedure.

	Gross error (%)					
Method	DB1	DB2	DB3	DB4	Average	(low/high)
pda	10.3	19.0	17.3	27.0	**16.8**	(14.2/2.6)
fxac	13.3	16.8	17.1	16.3	**15.2**	(14.2/1.0)
fxcep	4.6	15.8	5.4	6.8	**6.0**	(5.0/1.0)
ac	2.7	9.2	3.0	10.3	**5.1**	(4.1/1.0)
cc	3.4	6.8	2.9	7.5	**4.5**	(3.4/1.1)
shs	7.8	12.8	8.2	10.2	**8.7**	(8.6/0.18)
acf	0.45	1.9	7.1	11.7	**5.0**	(0.23/4.8)
nacf	0.43	1.7	6.7	11.4	**4.8**	(0.16/4.7)
additive	2.4	3.6	3.9	3.4	**3.1**	(2.5/0.55)
TEMPO	1.0	3.2	8.7	2.6	**3.4**	(0.53/2.9)
YIN	0.30	1.4	2.0	1.3	**1.03**	(0.37/0.66)

incorrect estimate, and so relatively low error rates despite a wide search range are an indication of robustness.

In some methods [spectral, autocorrelation based on Eq. (2)], the window size determines both the maximum period that can be estimated (lower limit of the F_0 search range), and the amount of data integrated to obtain any particular estimate. For YIN these two quantities are decoupled ($T_{\max}$ and W). There is, however, a relation between the appropriate value for one and the appropriate value for the other. For stability of estimates over time, the integration window must be no shorter than the largest expected period. Otherwise, one can construct stimuli for which the estimate would be incorrect over a certain phase of the period. The largest expected period obviously also determines the range of lags that need to be calculated, and together these considerations justify the well known rule of thumb: F_0 estimation requires enough signal to cover *twice* the largest expected period. The window may, however, be larger, and it is often observed that a larger window leads to fewer errors at the expense of reduced temporal resolution of the time series of estimates. Statistics reported for YIN were obtained with an integration window of 25 ms and a period search range of 25 ms, the shortest compatible with a 40 Hz lower bound on F_0. Figure 4(a) shows the number of errors for different window sizes.

A parameter specific to YIN is the threshold used in step 4. Figure 4(b) shows how it affects error rate. Obviously it does not require fine tuning, at least for this task. A value of 0.1 was used for the statistics reported here. A final parameter is the cutoff frequency of the initial low-pass filtering of the signal. It is generally observed, with this and other methods, that low-pass filtering leads to fewer errors, but obviously setting the cutoff below the F_0 would lead to failure. Statistics reported here were for convolution with a 1-ms square window (zero at 1 kHz). Error rates for other values are plotted in Fig. 4(c). In summary, this method involves comparatively few parameters, and these do not require fine tuning.

V. IMPLEMENTATION CONSIDERATIONS

The basic building block of YIN is the function defined in Eq. (1). Calculating this formula for every t and τ is computationally expensive, but there are at least two approaches to reduce cost. The first is to implement Eq. (1) using a recursion formula over time (each step adds a new term and subtracts an old). The window shape is then square, but a triangular or yet closer approximation to a Gaussian shape can be obtained by recursion (there is, however, little reason not to use a square window).

A second approach is to use Eq. (2) which can be calculated efficiently by FFT. This raises two problems. The first is that the energy terms of Eq. (7) must be calculated separately. They are not the same as $r'_t(0)$, but rather the

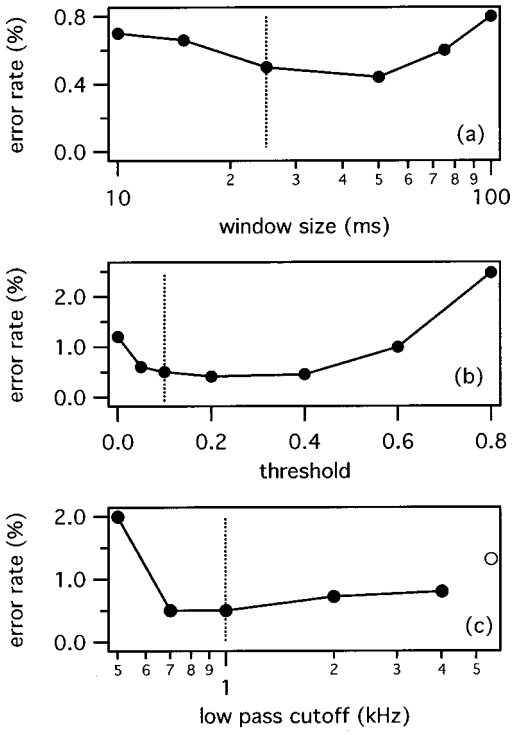

FIG. 4. Error rates of YIN: (a) as a function of window size, (b) as a function of threshold, and (c) as a function of low-pass prefilter cutoff frequency (open symbol is no filtering). The dotted lines indicate the values used for the statistics reported for YIN in Table II. Rates here were measured over a small database, a subset of that used in Sec. III. Performance does not depend critically on the values of these parameters, at least for this database.

sum of squares over the first and last $W-\tau$ samples of the window, respectively. Both must be calculated for each τ, but this may be done efficiently by recursion over τ. The second problem is that the sum involves more terms for small τ than for large. This introduces an unwanted bias that can be corrected by dividing each sample of $d(\tau)$ by $W-\tau$. However, it remains that large-τ samples of $d(\tau)$ are derived from a smaller window of data, and are thus less stable than small-τ samples. In this sense the FFT implementation is not as good as the previous one. It is, however, much faster when producing estimates at a reduced frame rate, while the previous approach may be faster if a high-resolution time series of estimates is required.

Real-time applications such as interactive music tracking require low latency. It was stated earlier that estimation requires a chunk of signal of at least $2T_{\max}$. However, step 4 allows calculations started at $\tau=0$ to terminate as soon as an acceptable candidate is found, rather than to proceed over the full search range, so latency can be reduced to $T_{\max}+T$. Further reduction is possible only if integration time is reduced below $T_{\max}$, which opens the risk of erroneously locking to the fine structure of a particularly long period.

The value $d'(T)$ may be used as a confidence indicator (large values indicate that the F_0 estimate is likely to be unreliable), in postprocessing algorithms to correct the F_0 trajectory on the basis of the most reliable estimates, and in template-matching applications to prevent the distance between a pattern and a template from being corrupted by unreliable estimates within either. Another application is in multimedia indexing, in which an F_0 time series may have to be down-sampled to save space. The confidence measure allows down-sampling to be based on correct rather than incorrect estimates. This scheme is implemented in the MPEG7 standard (ISO/IEC_JTC_1/SC_29, 2001).

VI. EXTENSIONS

The YIN method described in Sec. II is based on the model of Eq. (4) (periodic signal). The notion of model is insightful: an ''estimation error'' means simply that the model matched the signal for an unexpected set of parameters. Error reduction involves modifying the model to make such matches less likely. This section presents extended models that address situations where the signal deviates systematically from the periodic model. Tested quantitatively over our speech databases, none of these extensions improved error rates, probably because the periodic model used by YIN was sufficiently accurate for this task. For this reason we report no formal evaluation results. The aim of this section is rather to demonstrate the flexibility of the approach and to open perspectives for future development.

A. Variable amplitude

Amplitude variation, common in speech and music, compromises the fit to the periodic model and thus induces errors. To deal with it the signal may be modeled as a periodic function with time-varying amplitude:

$$x_{t+T}/a_{t+T}=x_t/a_t. \tag{11}$$

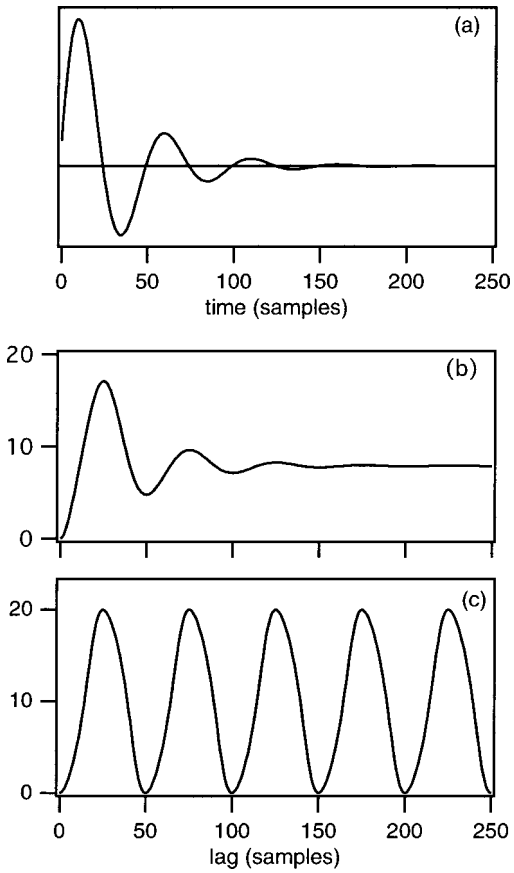

FIG. 5. (a) Sine wave with exponentially decreasing amplitude. (b) Difference function calculated according to Eq. (6) (periodic model). (c) Difference function calculated according to Eq. (12) (periodic model with time-varying amplitude). Period estimation is more reliable and accurate using the latter model.

If one supposes that the ratio $\alpha=a_{t+T}/a_t$ does not depend on t (as in an exponential increase or decrease), the value of α may be found by least squares fitting. Substituting that value in Eq. (6) then leads to the following function:

$$d_t(\tau)=r_t(0)[1-r_t(\tau)^2/r_t(0)r_{t+\tau}(0)]. \tag{12}$$

Figure 5 illustrates the result. The top panel displays the time-varying signal, the middle a function $d'(\tau)$ derived according to the standard procedure, and the bottom the same function derived using Eq. (12) instead of Eq. (6). Interestingly, the second term on the right of Eq. (12) is the square of the normalized ACF.

With two parameters the model of Eq. (12) is more ''permissive'' and more easily fits an amplitude-varying signal. However, this also implies more opportunities for ''unexpected'' fits, in other words, errors. Perhaps for that reason it actually produced a slight increase in error rates (0.57% vs. 0.50% over the restricted database). However, it was used with success to process the laryngograph signal (see the Appendix).

B. Variable F_0

Frequency variation, also common in speech and music, is a second source of aperiodicity that interferes with F_0 estimation. When F_0 is constant a lag τ may be found for

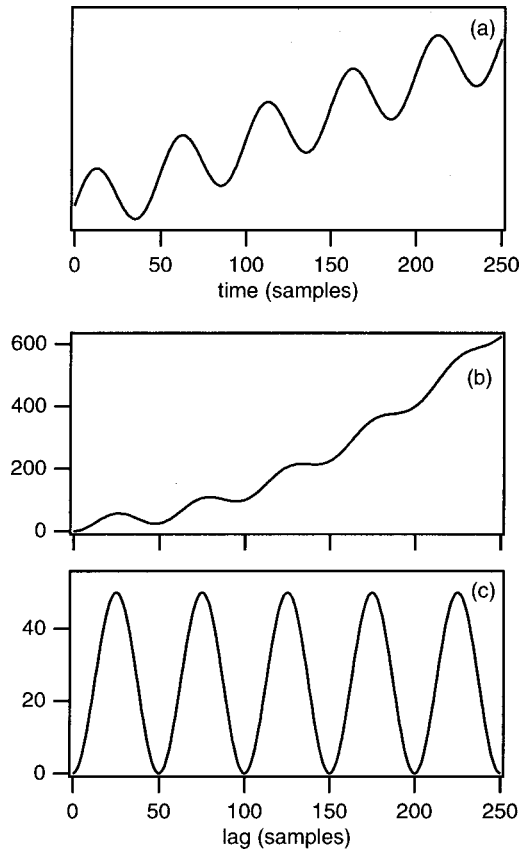

FIG. 6. (a) Sine wave with linearly increasing DC offset. (b) Difference function calculated according to Eq. (6). (c) Difference function calculated according to Eq. (13) (periodic model with DC offset). Period estimation is more reliable and accurate using the latter model.

which $(x_j - x_{j+\tau})^2$ is zero over the whole integration window of $d(\tau)$, but with a time-varying F_0 it is identically zero only at one point. On either side, its value $(x_j - x_{j+\tau})^2$ varies quadratically with distance from this point, and thus $d(\tau)$ varies with the cube of window size, W.

A shorter window improves the match, but we know that the integration window must not be shortened beyond a certain limit (Sec. IV). A solution is to split the window into two or more segments, and to allow τ to differ between segments within limits that depend on the maximum expected rate of change. Xu and Sun (2000) give a maximum rate of F_0 change of about ± 6 oct/s, but in our databases it did not often exceed ± 1 oct/s (Fig. 10). With a split window the search space is larger but the match is improved (by a factor of up to 8 in the case of two segments). Again, this model is more easily satisfied than that of Eq. (4), and therefore may introduce new errors.

C. Additive noise: Slowly varying DC

A common source of aperiodicity is additive noise which can take many forms. A first form is simply a time-varying ''DC'' offset, produced for example by a singer's breath when the microphone is too close. The deleterious effect of a DC ramp, illustrated in Fig. 6(b), can be eliminated by using the following formula, obtained by setting the derivative of $d_t(\tau)$ with respect to the DC offset to zero:

$$d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau) + \left[\sum_{j=t+1}^{t+W} (x_j - x_{j+\tau})\right]^2 \quad (13)$$

as illustrated in Fig. 6(c).

Again, this model is more permissive than the strict periodic model and thus may introduce new errors. For that reason, and because our speech data contained no obvious DC offsets, it gave no improvement and instead slightly increased error rates (0.51% vs 0.50%). However, it was used with success to process the laryngograph signal, which had large slowly varying offsets.

D. Additive noise: Periodic

A second form of additive noise is a concurrent periodic sound, for example, a voice or an instrument, hum, etc. Except in the unlucky event that the periods are in certain simple ratios, the effects of the interfering sound can be eliminated by applying a comb filter with impulse response $h(t) = \delta(t) - \delta(t+U)$ where U is the period of the interference. If U is known, this processing is trivial. If U is unknown, both it and the desired period T may be found by the joint estimation algorithm of de Cheveigné and Kawahara (1999). This algorithm searches the (τ, ν) parameter space for a minimum of the following difference function:

$$dd_t(\tau, \nu) = \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau} - x_{j+\nu} + x_{j+\tau+\nu})^2. \quad (14)$$

The algorithm is computationally expensive because the sum must be recalculated for all pairs of parameter values. However, this cost can be reduced by a large factor by expanding the squared sum of Eq. (14):

$$\begin{aligned} dd_t(\tau, \nu) = {} & r_t(0) + r_{t+\tau}(0) + r_{t+\nu}(0) + r_{t+\tau+\nu}(0) \\ & - 2r_t(\tau) - 2r_t(\nu) + 2r_t(\tau+\nu) \\ & + 2r_{t+\tau}(\nu-\tau) - 2r_{t+\tau}(\nu) - 2r_{t+\nu}(\tau). \end{aligned} \quad (15)$$

The right-hand terms are the same ACF coefficients that served for single period estimation. If they have been precalculated, Eq. (15) is relatively cheap to form. The two-period model is again more permissive than the one-period model and thus may introduce new errors. As an example, recall that the sum of two closely spaced sines is equally well interpreted as such (by this model), or as an amplitude-modulated sine (by the periodic or variable-amplitude periodic models). Neither interpretation is more ''correct'' than the other.

E. Additive noise: Different spectrum from target

Suppose now that the additive noise is neither DC nor periodic, but that its spectral envelope differs from that of the periodic target. If both long-term spectra are known and stable, filtering may be used to reinforce the target and weaken the interference. Low-pass filtering is a simple example and its effects are illustrated in Fig. 4(c).

If spectra of target and noise differ only on a short-term basis, one of two techniques may be applied. The first is to split the signal over a filter bank (for example, an auditory model filter bank) and calculate a difference function from

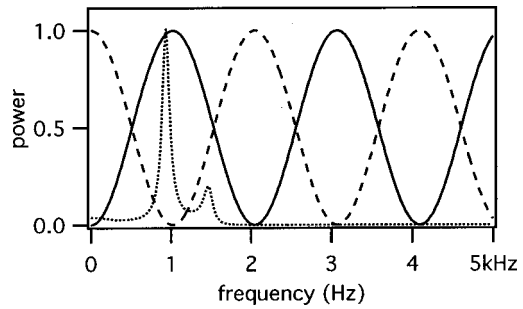

FIG. 7. Power transfer functions for filters with impulse response $\delta_t - \delta_{t+\tau}$ (full line) and $\delta_t + \delta_{t+\tau}$ (dashed line) for $\tau = 1$ ms. To reduce the effect of additive noise on F_0 estimation, the algorithm searches for the value of τ and the sign that maximize the power of the periodic target relative to aperiodic interference. The dotted line is the spectrum of a typical vowel.

each output. These functions are then added to obtain a summary difference function from which a periodicity measure is derived. Individual channels are then removed one by one until periodicity improves. This is reminiscent of Licklider's (1951) model of pitch perception.

The second technique applies an adaptive filter at the input, and searches jointly for the parameters of the filter and the period. This is practical for a simple filter with impulse response $h(t) = \delta(t) \pm \delta(t+V)$, where V and the sign determine the shape of the power transfer function illustrated in Fig. 7. The algorithm is based on the assumption that some value of V and sign will advantage the target over the interference and improve periodicity. The parameter V and the sign are determined, together with the period T, by searching for a minimum of the function:

$$\begin{aligned} dd'_t(\tau,\nu) = {} & r_t(0) + r_{t+\tau}(0) + r_{t+\nu}(0) + r_{t+\tau+\nu}(0) \\ & \pm 2r_t(\tau) - 2r_t(\nu) \mp 2r_t(\tau+\nu) \\ & \mp 2r_{t+\tau}(\nu-\tau) - 2r_{t+\tau}(\nu) \pm 2r_{t+\nu}(\tau), \end{aligned} \tag{16}$$

which (for the negative sign) is similar to Eq. (15). The search spaces for T and V should be disjoint to prevent the comb-filter tuned to V from interfering with the estimation of T. Again, this model is more permissive than the standard periodic model, and the same warnings apply as for other extensions to that model.

F. Additive noise: Same spectrum as target

If the additive noise shares the same spectral envelope as the target on an instantaneous basis, none of the previous methods is effective. Reliability and accuracy can nevertheless be improved if the target is stationary and of sufficiently long duration. The idea is to make as many period-to-period comparisons as possible given available data. Denoting as D the duration, and setting the window size W to be at least as large as the maximum expected period, the following functions are calculated:

$$d_k(\tau) = \sum_{j=1}^{D-kW} (x_j - x_{j-\tau})^2, \quad k = 1,...,D/W. \tag{17}$$

The lag (τ) axis of each function is then "compressed" by a factor of $D/W-k$, and the functions are summed:

$$d(\tau) = \sum_{k=1}^{D/W} d_k(\tau/(D/W-k)). \tag{18}$$

This function is the sum of $(D/W)(D/W-1)/2$ differences. For $\tau \neq T$ each difference includes both a deterministic part (target) and a noise part, whereas for $\tau = T$ they only include the noise part. Deterministic parts add in phase while noise parts tend to cancel each other out, so the salience of the dip at $\tau = T$ is reinforced. Equation (18) resembles (with different coefficients) the "narrowed autocorrelation function" of Brown and Puckette (1989) that was used by Brown and Zhang (1991) for musical F_0 estimation, and by de Cheveigné (1989) and Slaney (1990) in pitch perception models.

To summarize, the basic method can be extended in several ways to deal with particular forms of aperiodicity. These extensions may in some cases be combined (for example, modeling the signal as a sum of periodic signals with varying amplitudes), although all combinations have not yet been explored. We take this flexibility to be a useful feature of the approach.

VII. RELATIONS WITH AUDITORY PERCEPTION MODELS

As pointed out in the Introduction, the autocorrelation model is a popular account of pitch perception, but attempts to turn that model into an accurate speech F_0 estimation method have met with mitigated success. This study showed how it can be done. Licklider's (1951) model involved a network of delay lines (the τ parameter) and coincidence-counting neurons (a probabilistic equivalent of multiplication) with temporal smoothing properties (the equivalent of integration). A previous study (de Cheveigné, 1998) showed that excitatory coincidence could be replaced by inhibitory "anti-coincidence," resulting in a "cancellation model of pitch perception" in many regards equivalent to autocorrelation. The present study found that cancellation is actually more effective, but also that it may be accurately implemented as a sum of autocorrelation terms.

Cancellation models (de Cheveigné, 1993, 1997, 1998) require both excitatory and inhibitory synapses with fast temporal characteristics. The present study suggests that the same functionality might be obtained with fast excitatory synapses only, as illustrated in Fig. 8. There is evidence for fast excitatory interaction in the auditory system, for example in the medial superior olive (MSO), as well as for fast inhibitory interaction, for example within the lateral superior olive (LSO) that is fed by excitatory input from the cochlear nucleus, and inhibitory input from the medial trapezoidal body. However, the limit on temporal accuracy may be lower for inhibitory than for excitatory interaction (Joris and Yin, 1998). A model that replaces one by the other without loss of functionality is thus a welcome addition to our panoply of models.

Sections VI D and VI E showed how a cascade of subtractive operations could be reformulated as a sum of autocorrelation terms. Transposing to the neural domain, this suggests that the cascaded cancellation stages suggested by de Cheveigné and Kawahara (1999) to account for multiple pitch perception, or by de Cheveigné (1997) to account for

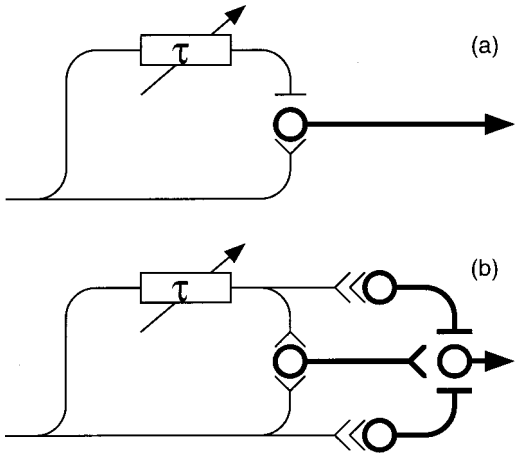

FIG. 8. (a) Neural cancellation filter (de Cheveigné, 1993, 1997). The gating neuron receives excitatory (direct) and inhibitory (delayed) inputs, and transmits any spike that arrives via the former unless another spike arrives simultaneously via the latter. Inhibitory and excitatory synapses must both be fast (symbolized by thin lines). Spike activity is averaged at the output to produced slowly varying quantities (symbolized by thick lines). (b) Neural circuit with the same properties as in (a), but that only requires fast excitatory synapses. Inhibitory interaction involves slowly varying quantities (thick lines). Double ''chevrons'' symbolize that output discharge probability is proportional to the square of input discharge probability. These circuits should be understood as involving many parallel fibers to approximate continuous operations on probabilities.

concurrent vowel identification, might instead be implemented in a single stage as a neural equivalent of Eq. (15) or (16). Doing away with cascaded time-domain processing avoids the assumption of a succession of phase-locked neurons, and thus makes such models more plausible. Similar remarks apply to cancellation models of binaural processing (Culling and Summerfield, 1995; Akeroyd, 2000; Breebart *et al.*, 2001).

To summarize, useful parallels may be drawn between signal processing and auditory perception. The YIN algorithm is actually a spin-off of work on auditory models. Conversely, addressing this practical task may be of benefit to auditory modeling, as it reveals difficulties that are not obvious in modeling studies, but that are nevertheless faced by auditory processes.

VIII. DISCUSSION

Hundreds of F_0 estimation methods have been proposed in the past, many of them ingenious and sophisticated. Their mathematical foundation usually assumes periodicity, and when that is degraded (which is when smart behavior is most needed) they may break down in ways not easy to predict. As pointed out in Sec. II A, seemingly different estimation methods are related, and our analysis of error mechanisms can probably be transposed, mutatis mutandis, to a wider class of methods. In particular, every method is faced with the problem of trading off too-high versus too-low errors. This is usually addressed by applying some form of bias as illustrated in Sec. II A. Bias may be explicit as in that section, but often it is the result of particular side effects of the algorithm, such as the tapering that resulted with Eq. (2) from limited window size. If the algorithm has several parameters, credit assignment is difficult. The key to the success of YIN is

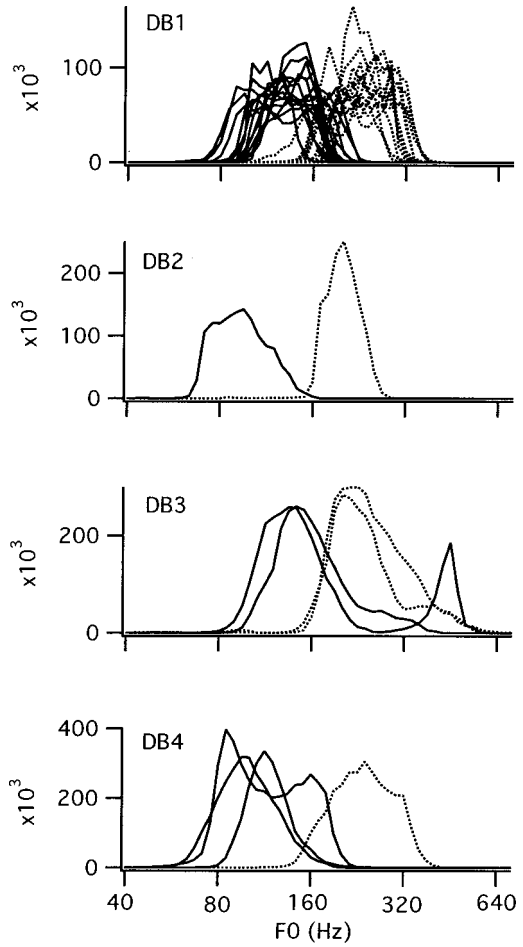

FIG. 9. Histograms of F_0 values over the four databases. Each line corresponds to a different speaker, either male (full lines) or female (dotted lines). The bin width is one semitone ($\frac{1}{12}$ of an octave). The skewed or bimodal distributions of database 3 are due to the presence of material pronounced in a falsetto voice.

probably step 3 that allows it to escape from the bias paradigm, so that the two types of error can be addressed independently. Other steps can be seen as either preparing for this step (steps 1 and 2) or building upon it (steps 4 and 6).

Parabolic interpolation (step 5) gives subsample resolution. Very accurate estimates can be obtained using an interval of signal that is not large. Precisely, to accurately estimate the period T of a perfectly periodic signal, *and* to be sure that the true period is not instead greater than T, at least

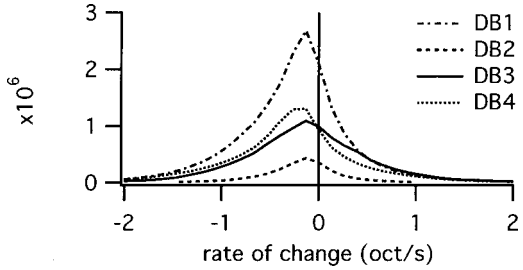

FIG. 10. Histograms of rate of F_0 change for each of the four databases. Each line is an aggregate histogram over all speakers of the database. The rate of change is measured over a 25-ms time increment (one period of the lowest expected F_0). The bin width is 0.13 oct/s. The asymmetry of the distributions reflects the well-known declining trend of F_0 in speech.

$2T+1$ samples of data are needed. If this is granted, there is no theoretical limit to accuracy. In particular, it is not limited by the familiar uncertainty principle $\Delta T \Delta F = \text{const}$.

We avoided familiar postprocessing schemes such as median smoothing (Rabiner and Schafer, 1978) or dynamic programming (Ney, 1982; Hess, 1983), as including them complicates evaluation and credit assignment. Nothing prevents applying them to further improve the robustness of the method. The aperiodicity measure $d'(T)$ may be used to ensure that estimates are corrected on the basis of their reliability rather than continuity *per se*.

The issue of voicing detection was also avoided, again because it greatly complicates evaluation and credit assignment. The aperiodicity measure $d'\tau$ seems a good basis for voicing detection, perhaps in combination with energy. However, equating voicing with periodicity is not satisfactory, as some forms of voicing are inherently irregular. They probably still carry intonation cues, but how they should be quantified is not clear. In a companion paper (Kawahara *et al.*, in preparation), we present a rather different approach to F_0 estimation and glottal event detection, based on instantaneous frequency and the search for fixed points in mappings along the frequency and time axes. Together, these two papers offer a new perspective on the old task of F_0 estimation.

YIN has been only informally evaluated on music, but there are reasons to expect that it is appropriate for that task. Difficulties specific to music are the wide range and fast changes in F_0. YIN's open-ended search range and the fact that it performs well without continuity constraints put it at an advantage over other algorithms. Other potential advantages, yet to be tested, are low latency for interactive systems (Sec. V), or extensions to deal with polyphony (Sec. VI D). Evaluation on music is complicated by the wide range of instruments and styles to be tested and the lack of a well-labeled and representative database.

What is new? Autocorrelation was proposed for periodicity analysis by Licklider (1951), and early attempts to apply it to speech are reviewed in detail by Hess (1983), who also traces the origins of difference-function methods such as the AMDF. The relation between the two, exploited in Eq. (7), was analyzed by Ney (1982). Steps 3 and 4 were applied to AMDF by de Cheveigné (1990) and de Cheveigné (1996), respectively. Step 5 (parabolic interpolation) is a standard technique, applied for example to spectrum peaks in the F_0 estimation method of Duifhuis *et al.* (1982). New are step 6, the idea of combining steps as described, the analysis of why it all works, and most importantly the formal evaluation.

IX. CONCLUSION

An algorithm was presented for the estimation of the fundamental frequency of speech or musical sounds. Starting from the well-known autocorrelation method, a number of modifications were introduced that combine to avoid estimation errors. When tested over an extensive database of speech recorded together with a laryngograph signal, error rates were a factor of 3 smaller than the best competing methods, without postprocessing. The algorithm has few parameters, and these do not require fine tuning. In contrast to most other methods, no upper limit need be put on the F_0 search range. The method is relatively simple and may be implemented efficiently and with low latency, and may be extended in several ways to handle several forms of aperiodicity that occur in particular applications. Finally, an interesting parallel may be drawn with models of auditory processing.

ACKNOWLEDGMENTS

This work was funded in part by the Cognitique program of the French Ministry of Research and Technology, and evolved from work done in collaboration with the Advanced Telecommunications Research Laboratories (ATR) in Japan, under a collaboration agreement between ATR and CNRS. It would not have been possible without the laryngograph-labeled speech databases. Thanks are due to Y. Atake and co-workers for creating DB1 and making it available to the authors. Paul Bagshaw was the first to distribute such a database (DB2) freely on the Internet. Nathalie Henrich, Christophe D'Alessandro, Michèle Castellengo, and Vu Ngoc Tuan kindly provided DB3. Nick Campbell offered DB4, and Georg Meyer DB5. Thanks are also due to the many people who generously created, edited, and distributed the software packages that were used for comparative evaluation. We offer apologies in the event that our choice of parameters did not do them justice. Thanks to John Culling, two anonymous reviewers, and the editor for in-depth criticism, and to Xuejing Sun, Paul Boersma, and Axel Roebel for useful comments on the manuscript.

APPENDIX: DETAILS OF THE EVALUATION PROCEDURE

1. Databases

The five databases comprised a total of 1.9 h of speech, of which 48% were labeled as regularly voiced. They were produced by 48 speakers (24 male, 24 female) of Japanese (30), English (14), and French (4). Each included a laryngograph waveform recorded together with the speech.

(1) DB1: Fourteen male and 14 female speakers each spoke 30 Japanese sentences for a total of 0.66 h of speech, for the purpose of evaluation of F_0-estimation algorithms (Atake *et al.*, 2000). The data include a "voiced–unvoiced" mask that was not used here.

(2) DB2: One male and one female speaker each spoke 50 English sentences for a total of 0.12 h of speech, for the purpose of evaluation of F_0-estimation algorithms (Bagshaw *et al.*, 1993). The database can be downloaded from the URL ⟨http://www.cstr.ed.ac.uk/~pcb/fda_eval.tar.gz⟩.

(3) DB3: Two male and two female speakers each pronounced between 45 and 55 French sentences for a total of 0.46 h of speech. The database was created for the study of speech production, and includes sentences pronounced according to several modes: normal (141), head (30), and fry (32) (Vu Ngoc Tuan and d'Alessandro, 2000). Sentences in fry mode were not used for evaluation because it is not obvious how to define F_0 when phonation is not periodic.

TABLE III. Gross error rates measured using alternative ground truth. DB1: manually checked estimates derived from the laryngograph signal using the TEMPO method of Kawahara *et al.* (1999b). DB2 and DB5: estimates derived independently by the authors of those databases.

Method	Gross error (%)		
	DB1	DB2	DB5
pda	9.8	14.5	15.1
fxac	13.2	14.9	16.1
fxcep	4.5	12.5	8.9
ac	2.7	7.3	5.1
cc	3.3	6.3	8.0
shs	7.5	11.1	9.4
acf	0.45	2.5	3.1
nacf	0.43	2.3	2.8
additive	2.16	3.4	3.7
TEMPO	0.77	2.8	4.6
YIN	0.29	2.2	2.4

(4) DB4: Two male speakers of English and one male and one female speaker of Japanese produced a total of 0.51 h speech, for the purpose of deriving prosody rules for speech synthesis (Campbell, 1997).
(5) DB5: Five male and five female speakers of English each pronunced a phonetically balanced text for a total of 0.15 h of speech. The database can be downloaded from ⟨ftp://ftp.cs.keele.ac.uk/pub/pitch/Speech⟩.

Ground-truth F_0 estimates for the first four databases were extracted from the laryngograph signal using YIN. The threshold parameter was set to 0.6, and the schemes of Secs. VI A and VI C were implemented to cope with the large variable DC offset and amplitude variations of the laryngograph signal. Estimates were examined together with the laryngograph signal, and a reliability mask was created manually based on the following two criteria: (1) any estimate for which the F_0 estimate was obviously incorrect was excluded and (2) any remaining estimate for which there was evidence of vocal fold vibration was included. The first criterion ensured that all estimates were correct. The second aimed to include as many ''difficult'' data as possible. Estimate values themselves were not modified. Estimates had the same sampling rate as the speech and laryngograph signals (16 kHz for DB1, DB3, and DB4, 20 kHz for database DB2). Figures 9 and 10 show the range of F_0 and F_0 change rate over these databases.

It could be argued that applying the same method to speech and laryngograph data gives YIN an advantage relative to other methods. Estimates were all checked visually, and there was no evidence of particular values that could only be matched by the same algorithm applied to the speech signal. Nevertheless, to make sure, tests were also performed on three databases using ground truth not based on YIN. The laryngograph signal of DB1 was processed by the TEMPO method of Kawahara *et al.* (1999a), based on instantaneous frequency and very different from YIN, and estimates were checked visually as above to derive a reliability mask. Scores are similar (Table III, column 2) to those obtained previously (Table II, column 2). Scores were also measured for DB2 and DB5, using reference F_0 estimates produced by the authors of those databases using their own criteria. The ranking of methods is similar to that found in Table III, suggesting that the results in that table are not a product of our particular procedures.

2. Reference methods

Reference methods include several methods available on the Internet. Their appeal is that they have been independently implemented and tuned, are representative of tools in common use, and are easily accessible for comparison purposes. Their drawback is that they are harder to control, and that the parameters used may not do them full justice. Other reference methods are only locally available. Details of parameters, availability and/or implementation are given below.

ac: This method implements the autocorrelation method of Boersma (1993) and is available with the Praat system at ⟨http://www.fon.hum.uva.nl/praat/⟩. It was called with the command ''To Pitch (ac)...0.01 40 15 no 0.0 0.0 0.01 0.0 0.0 800.''

cc: This method, also available with the Praat system, is described as performing a cross-correlation analysis. It was called with the command: ''To Pitch (cc)... 0.01 40 15 no 0.0 0.0 0.01 0.0 0.0 800.''

shs: This method, also available with the Praat system, is described as performing spectral subharmonic summation according to the algorithm of Hermes (1988). It was called with the command: ''To Pitch (shs)...0.01 40 4 1700 15 0.84 800 48.''

pda: This method implements the eSRPD algorithm of Bagshaw (1993), derived from that of Medan *et al.* (1991), and is available with the Edinburgh Speech Tools Library at ⟨http://www.cstr.ed.ac.uk/⟩. It was called with the command: ''pda input_file -o out-put_file -L -d 1 -shift 0.001-length 0.1-fmax 800-fmin 40-lpfilter 1000 -n 0.'' Examination of the code suggests that the program uses continuity constraints to improve tracking.

fxac: This program is based on the ACF of the cubed waveform and is available with the Speech Filing System at ⟨http://www.phon.ucl.ac.uk/resource/sfs/⟩. Examination of the code suggests that the search range is restricted to 80–400 Hz. It provides estimates only for speech that is judged ''voiced,'' which puts it at a disadvantage with respect to programs that always offer an estimate.

fxcep: This program is based on the cepstrum method, and is also available with the Speech Filing System. Examination of the code suggests that the search range is restricted to 67–500 Hz. It provides estimates only for speech that is judged ''voiced,'' which puts it at a disadvantage with respect to programs that always offer an estimate.

additive: This program implements the probabilistic spectrum-based method of Doval (1994) and is only locally available. It was called with the command: ''additive -0 -S input_file -f 40 -F 800 -G 1000 -X -f0ascii -I 0.001.''

acf: This program calculates the ACF according to Eq. (1) using an integration window size of 25 ms, multiplied by a linear ramp with intercept $T_{\max}=35$ ms (tuned for best performance over DB1), and chooses the global maximum between 1.25 to 25 ms (40 to 800 Hz).

nacf: As ''acf'' but using the normalized ACF according to Eq. (12).

TEMPO: This program implements the instantaneous frequency method developed by the second author (Kawahara *et al.*, 1999a).

YIN: The YIN method was implemented as described in this article with the following additional details. Equation (1) was replaced by the following variant:

$$r_t(\tau) = \sum_{j=t-\tau/2-W/2}^{t-\tau/2+W/2} x_j x_{j+\tau}, \tag{A1}$$

which forms the scalar product between two windows that shift symmetrically in time with respect to the analysis point. The window size was 25 ms, the threshold parameter was 0.1, and the F_0 search range was 40 Hz to one quarter the sampling rate (4 or 5 kHz depending on the database). The window shift was 1 sample (estimates were produced at the same sampling rate as the speech waveform).

3. Evaluation procedure

Algorithms were evaluated by counting the number of estimates that differed from the reference by more than 20% (gross error rate). Reference estimates were time shifted and downsampled as necessary to match the alignment and sampling rate of each method. Alignment was determined by taking the minimum error rate over a range of time shifts between speech-based and laryngograph-based estimates. This compensated for time shifts due to acoustic propagation from glottis to microphone, or implementation differences. Some estimation algorithms work (in effect) by comparing two windows of data that are shifted symmetrically in time with respect to the analysis point, whereas others work (in effect) by comparing a shifted window to a fixed window. An F_0-dependent corrective shift should be used in the latter case.

A larger search range gives more opportunities for error, so search ranges must be matched across methods. Methods that implement a voicing decision are at a disadvantage with respect to methods that do not (incorrect ''unvoiced'' decisions count as gross errors), so the voicing decision mechanism should be disabled. Conversely, postprocessing may give an algorithm an advantage. Postprocessing typically involves parameters that are hard to optimize and behavior that is hard to interpret, and is best evaluated separately from the basic algorithm. These recommendations cannot always be followed, either because different methods use radically different parameters, or because their implementation does not allow them to be controlled. Method ''pda'' uses continuity constraints and postprocessing. The search range of ''fxac'' was 80–400 Hz, while that of ''fxcep'' was 67–500 Hz, and these two methods produce estimates only for speech that is judged voiced. We did not attempt to modify the programs, as that would have introduced a mismatch with the publicly available version. These differences must be kept in mind when comparing results across methods.

Abe, T., Kobayashi, T., and Imai, S. (**1995**). ''Harmonics tracking and pitch extraction based on instantaneous frequency,'' Proc. IEEE-ICASSP, pp. 756–759.

Akeroyd, M. A., and Summerfield, A. Q. (**2000**). ''A fully-temporal account of the perception of dichotic pitches,'' Br. J. Audiol. **33**(2), 106–107.

Atake, Y., Irino, T., Kawahara, H., Lu, J., Nakamura, S., and Shikano, K. (**2000**). ''Robust fundamental frequency estimation using instantaneous frequencies of harmonic components,'' Proc. ICLSP, pp. 907–910.

Bagshaw, P. C., Hiller, S. M., and Jack, M. A. (**1993**). ''Enhanced pitch tracking and the processing of *F*0 contours for computer and intonation teaching,'' Proc. European Conf. on Speech Comm. (Eurospeech), pp. 1003–1006.

Barnard, E., Cole, R. A., Vea, M. P., and Alleva, F. A. (**1991**). ''Pitch detection with a neural-net classifier,'' IEEE Trans. Signal Process. **39**, 298–307.

Boersma, P. (**1993**). ''Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,'' Proc. Institute of Phonetic Sciences **17**, 97–110.

Breebart, J., van de Par, S., and Kohlrausch, A. (**2001**). ''Binaural processing model based on contralateral inhibition. I. Model structure,'' J. Acoust. Soc. Am. **110**, 1074–1088.

Brown, J. C., and Puckette, M. S. (**1989**). ''Calculation of a 'narrowed' autocorrelation function,'' J. Acoust. Soc. Am. **85**, 1595–1601.

Brown, J. C., and Zhang, B. (**1991**). ''Musical frequency tracking using the methods of conventional and 'narrowed' autocorrelation,'' J. Acoust. Soc. Am. **89**, 2346–2354.

Campbell, N. (**1997**). ''Processing a Speech Corpus for CHATR Synthesis,'' in Proc. ICSP (International Conference on Speech Processing).

Cariani, P. A., and Delgutte, B. (**1996**). ''Neural correlates of the pitch of complex tones. I. Pitch and pitch salience,'' J. Neurophysiol. **76**, 1698–1716.

Culling, J. F., and Summerfield, Q. (**1995**). ''Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay,'' J. Acoust. Soc. Am. **98**, 785–797.

de Cheveigné, A. (**1989**). ''Pitch and the narrowed autocoincidence histogram,'' Proc. ICMPC, Kyoto, pp. 67–70.

de Cheveigné, A. (**1990**), ''Experiments in pitch extraction,'' ATR Interpreting Telephony Research Laboratories technical report, TR-I-0138.

de Cheveigné, A. (**1991**). ''Speech *f*0 extraction based on Licklider's pitch perception model,'' Proc. ICPhS, pp. 218–221.

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

de Cheveigné, A. (**1996**). ''Speech fundamental frequency estimation,'' ATR Human Information Processing Research Laboratories technical report, TR-H-195.

de Cheveigné, A. (**1997**). ''Concurrent vowel identification. III. A neural model of harmonic interference cancellation,'' J. Acoust. Soc. Am. **101**, 2857–2865.

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

de Cheveigné, A., and Kawahara, H. (**1999**). ''Multiple period estimation and pitch perception model,'' Speech Commun. **27**, 175–185.

Doval, B. (**1994**). ''Estimation de la fréquence fondamentale des signaux sonores,'' Université Pierre et Marie Curie, unpublished doctoral dissertation (in French).

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

Goldstein, J. L. (**1973**). ''An optimum processor theory for the central formation of the pitch of complex tones,'' J. Acoust. Soc. Am. **54**, 1496–1516.

Hedelin, P., and Huber, D. (**1990**). ''Pitch period determination of aperiodic speech signals,'' Proc. ICASSP, pp. 361–364.

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

Hermes, D. J. (**1993**). ''Pitch analysis,'' in *Visual Representations of Speech Signals*, edited by M. Cooke, S. Beet, and M. Crawford (Wiley, New York), pp. 3–25.

Hess, W. (**1983**). *Pitch Determination of Speech Signals* (Springer-Verlag, Berlin).

Hess, W. J. (**1992**). ''Pitch and voicing determination,'' in *Advances in Speech Signal Processing*, edited by S. Furui and M. M. Sohndi (Marcel Dekker, New York), pp. 3–48.

Huang, X., Acero, A., and Hon, H.-W. (**2001**). *Spoken Language Processing* (Prentice–Hall, Upper Saddle River, NJ).

ISO/IEC_JTC_1/SC_29 (**2001**). ''Information Technology—Multimedia

Content Description Interface—Part 4: Audio,'' ISO/IEC FDIS 15938-4.
Joris, P. X., and Yin, T. C. T. (**1998**). ''Envelope coding in the lateral superior olive. III. Comparison with afferent pathways,'' J. Neurophysiol. **79**, 253–269.
Kawahara, H., Katayose, H., de Cheveigné, A., and Patterson, R. D. (**1999a**). ''Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of *F*0 and Periodicity,'' Proc. EUROSPEECH **6**, 2781–2784.
Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (**1999b**). ''Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,'' Speech Commun. **27**, 187–207.
Kawahara, H., Zolfaghari, P., and de Cheveigné, A. (in preparation). ''Fixed-point-based source information extraction from speech sounds designed for a very high-quality speech modifications.''
Licklider, J. C. R. (**1951**). ''A duplex theory of pitch perception,'' Experientia **7**, 128–134.
Medan, Y., Yair, E., and Chazan, D. (**1991**). ''Super resolution pitch determination of speech signals,'' IEEE Trans. Acoust., Speech, Signal Process. **39**, 40–48.
Meddis, R., and Hewitt, M. J. (**1991**). ''Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification,'' J. Acoust. Soc. Am. **89**, 2866–2882.
Miller, G. A., and Taylor, W. G. (**1948**). ''The perception of repeated bursts of noise,'' J. Acoust. Soc. Am. **20**, 171–182.
Moore, B. C. J. (**1997**). *An Introduction to the Psychology of Hearing* (Academic, London).
Ney, H. (**1982**). ''A time warping approach to fundamental period estimation,'' IEEE Trans. Syst. Man Cybern. **12**, 383–388.
Noll, A. M. (**1967**). ''Cepstrum pitch determination,'' J. Acoust. Soc. Am. **41**, 293–309.
Pressnitzer, D., Patterson, R. D., and Krumbholz, K. (**2001**). ''The lower limit of melodic pitch,'' J. Acoust. Soc. Am. **109**, 2074–2084.
Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ).
Ritsma, R. J. (**1962**). ''Existence region of the tonal residue. I,'' J. Acoust. Soc. Am. **34**, 1224–1229.
Rodet, X., and Doval, B. (**1992**). ''Maximum-likelihood harmonic matching for fundamental frequency estimation,'' J. Acoust. Soc. Am. **92**, 2428–2429 (abstract).
Ross, M. J., Shaffer, H. L., Cohen, A., Freudberg, R., and Manley, H. J. (**1974**). ''Average magnitude difference function pitch extractor,'' IEEE Trans. Acoust., Speech, Signal Process. **22**, 353–362.
Slaney, M. (**1990**). ''A perceptual pitch detector,'' Proc. ICASSP, pp. 357–360.
Terhardt, E. (**1974**). ''Pitch, consonance and harmony,'' J. Acoust. Soc. Am. **55**, 1061–1069.
Vu Ngoc Tuan, and d'Alessandro, C. (**2000**). ''Glottal closure detection using EGG and the wavelet transform,'' in Proceedings 4th International Workshop on Advances in Quantitative Laryngoscopy, Voice and Speech Research, Jena, pp. 147–154.
Wightman, F. L. (**1973**). ''The pattern-transformation model of pitch,'' J. Acoust. Soc. Am. **54**, 407–416.
Xu, Y., and Sun, X. (**2000**). ''How fast can we really change pitch? Maximum speed of pitch change revisited,'' Proc. ICSLP, pp. 666–669.
Yost, W. A. (**1996**). ''Pitch strength of iterated rippled noise,'' J. Acoust. Soc. Am. **100**, 3329–3335.