Fundamental frequency estimation of musical signals using a two-way mismatch procedure

Robert C. Maher
Department of Electrical Engineering and Center for Communication and Information Science, University of Nebraska, 209N WSEC, Lincoln, Nebraska 68588-0511

James W. Beauchamp
School of Music and Department of Electrical and Computer Engineering, University of Illinois, 2136 Music Bldg., 1114 W. Nevada, Urbana, Illinois 61801

(Received 28 June 1993; accepted for publication 16 November 1993)

Fundamental frequency (F_0) estimation for quasiharmonic signals is an important task in music signal processing. Many previously developed techniques have suffered from unsatisfactory performance due to ambiguous spectra, noise perturbations, wide frequency range, vibrato, and other common artifacts encountered in musical signals. In this paper a new *two-way mismatch* (TWM) procedure for estimating F_0 is described which may lead to improved results in this area. This computer-based method uses the quasiharmonic assumption to guide a search for F_0 based on the short-time spectra of an input signal. The estimated F_0 is chosen to minimize discrepancies between measured partial frequencies and harmonic frequencies generated by trial values of F_0. For each trial F_0, mismatches between the harmonics generated and the measured partial frequencies are averaged over a fixed subset of the available partials. A weighting scheme is used to reduce the susceptibility of the procedure to the presence of noise or absence of certain partials in the spectral data. Graphs of F_0 estimate versus time for several representative recorded solo musical instrument and voice passages are presented. Some special strategies for extending the TWM procedure for F_0 estimations of two simultaneous voices in duet recordings are also discussed.

PACS numbers: 43.75.Yy

INTRODUCTION

Many techniques for estimating the fundamental frequency, F_0, of monophonic quasiharmonic signals have been developed or proposed (Hess, 1983). Time-domain methods include multiplicative autocorrelation (Sondhi, 1968; Brown and Zhang, 1991), subtractive autocorrelation, i.e., the *average magnitude difference function* or *optimum comb* (Ross *et al.*, 1974; Moorer, 1974; Martin, 1982), and methods based on linear prediction (Markel and Gray, 1976; Rabiner and Schafer, 1978). Frequency-domain methods include the *cepstrum* (Noll, 1966), the *period histogram* (Schroeder, 1968; Piszczalski and Galler, 1979), *maximum likelihood* methods (Rife and Boorstyn, 1976; Wolcin, 1980), and, more recently, harmonic pattern matching procedures (Doval and Rodet, 1991; Brown, 1992; Brown and Puckette, 1993). The continuing research interest in this area is an indication that no completely successful system for fundamental frequency estimation for a wide range of audio signal types is yet available. Without a full, objective comparison of the various F_0 estimation procedures it is difficult to draw specific conclusions about the strengths and weaknesses of existing techniques. Nonetheless, algorithms which do not explicitly address the presence of noise, reverberation, and other common signal degradations are particularly difficult to evaluate for real-world analysis purposes.

Time-domain approaches make use of the assumed periodic nature of the input signal by identifying waveform features such as peaks, zero crossings, and other periodic structures. The time duration between the repetitive corresponding features is expected to be the waveform period, $1/F_0$. Other time-domain methods use an autocorrelation approach to identify waveform periods, based on the notion that we expect one cycle of a periodic signal to be highly correlated with the next. Similarly, frequency-domain techniques make use of the fact that the spectra of periodic time-domain signals exhibit quasiharmonic spectral structures manifested by regularly spaced peaks in the magnitude spectrum. Thus the frequency estimation problem becomes a task of determining the set of harmonic frequencies which, in some sense, best match the positions of the spectral peaks.

Other transform techniques, such as the cepstrum and linear predictive coding, take the process one step further by separating the power spectrum into an excitation component that varies relatively rapidly with frequency (harmonics of a relatively low fundamental), and a system function that varies relatively slowly with frequency (higher frequency formants). In the cepstrum the fundamental frequency estimation reduces to a problem of identifying the periodicity of the Fourier magnitude spectrum.

[a]Portions of this work were presented at the 124th Meeting of the Acoustical Society of America, New Orleans, LA [J. Acoust. Soc. Am. 92, 2429 (A) (1992)].

 0001-4966/94/95(4)/2254/10/$6.00

This paper concerns the development of more versatile and robust F_0 estimation procedures that are appropriate for use in monophonic and simple polyphonic situations, such as duets. In this paper we first present several of the research issues related to F_0 estimation. Next we describe the previously reported *two-way mismatch* F_0 estimation technique (Maher, 1990) to avoid the problems encountered in processing real signals containing noise and reverberation. Our implementation and intended applications are reviewed next, followed by a discussion of several examples. Finally, we conclude with a description of some additional research issues for the future.

I. RESEARCH ISSUES FOR F_0 ESTIMATION

Several problems plague researchers designing methods for automatic determination of fundamental frequency. First, most algorithms suffer from degraded performance when the amplitude of a signal is low, such as during the release of a musical note, primarily because of the diminished reliability with which the signal parameters can be measured under conditions of reduced signal-to-noise ratio. Second, there is an inherent ambiguity in estimating the F_0 for a series of partials, since two musical notes separated by one or more octaves share coincident partials. Estimators are particularly prone to octave errors when attempting to process musical instrument sounds with insufficiently strong F_0 components or in situations where the expected fundamental frequency spans a range larger than 1 oct. Finally, the general difficulty of accommodating the effects of nonideal signal characteristics (e.g., background noise, inharmonicity, and signal transients) means that there is no certainty that an algorithm that performs well on one input example will perform as well on *all* examples, even those that are ostensibly similar to the model. This makes validation of an estimation method quite difficult and complicates performance comparisons among various F_0 estimation techniques.

An additional difficulty occurs when attempting to process a recording made in a reverberant environment. Because of the reverberant character of the recording space, the acoustical signal captured by the microphone includes not only the direct sound of the instrument at a particular instant, but also echoes of previous notes that have not yet died away. Thus, the recorded signal actually contains a multiplicity of competing signals due to the reverberation. Although the reverberation problem could be virtually eliminated by using only close-miked or contact-miked recordings or by recording in an anechoic room, there are many situations where such pristine recordings are either unavailable or impractical. At least in the field of Western classical music, recording artists and music listeners prefer performances done in halls with significant amounts of reverberation; thus, many archival recordings that are of interest to musicological researchers are inherently contaminated with this artifact.

We have encountered yet another problem: For the case of F_0 detection of sounds having very few partials, we find that it is difficult for algorithms optimized for sounds having many harmonics to determine whether the measured components correspond to the fundamental frequency or to the second harmonic of an extremely weak fundamental. This did not affect our results for the musical instruments we processed, but it does cause unexpected problems with, for example, simple whistle input, which is primarily sinusoidal.

The research scenario becomes even more problematic in the case of polyphonic F_0 estimation, where the task is to identify the fundamental frequency of a musical voice in the presence of competing voices. While it may be tempting simply to apply a particular monophonic F_0 estimation technique to the polyphonic case, in practice this is usually doomed to failure. For example, monophonic methods based on the time-domain periodicity of the input signal must now contend with more than one periodicity, and worse, the frequency relationships of the simultaneous voices usually involve overlapping spectral components resulting in amplitude beating or cancellations among the partials. Comparable problems exist for most frequency-domain methods, since it is not clear which spectral components belong to which voice or may be caused by more than one voice.

II. RATIONALE AND IMPLEMENTATION

Our computer program for fundamental frequency analysis is designed to process time-varying spectral data produced by fixed-window (typically, 46 ms) short-time Fourier transform (STFT) analysis of an acoustic signal input (Allen and Rabiner, 1977). For each time frame (typically, 5.8 ms), this program saves magnitude spectrum peaks, henceforth referred to as *measured partials*, which have been refined in amplitude and frequency using a parabolic interpolation technique (Smith and Serra, 1987; Maher, 1990 and 1991). The theoretical accuracy of the partial frequency measurement is about 5% of the Fast Fourier transform (FFT) bin spacing, e.g., about 2 Hz with a 44.1-kHz sample rate and 1024-point FFT (Brown and Puckette, 1993). The window length must be chosen to trade-off the reduced time resolution available when using a long window with the reduced frequency resolution available with a short window. In order to resolve reliably the partials of a harmonic signal with fundamental frequency of F_0 Hz, it is necessary for the spectral bandwidth of the analysis window to be approximately $F_0/2$ Hz, corresponding to a Kaiser window about four waveform cycles long. Thus, the 46-ms typical window is satisfactory for F_0 above approximately 87 Hz. The primary underlying assumption for the measurement of F_0 is that the signal consists of a series of harmonic partials. However, we do not know which of the measured partials actually correspond to harmonics of the signal. Some "partials" may be caused by noise, reverberation, or other types of signal artifacts. Also, small (but important) uncertainties can occur in the estimates of the partial frequencies. Moreover, some low amplitude partials in the original signal may escape detection and thus may be missing in the spectral data. In short, we are confronted with the vagaries commonly associated with any *real* measurements on *real* signals.

A. Improving the reliability of the F_0 estimate: The mismatch error

In order to compensate for some of the inherent deficiencies in the spectral data, we have designed a method for F_0 detection called the *two-way mismatch* procedure (TWM). The procedure is reminiscent of maximum likelihood estimation, in that the measured spectrum is compared to a postulated harmonic spectral pattern. The TWM estimation procedure is based on the comparison of each measured sequence of partials from the STFT analysis (corresponding to a particular time frame) with predicted sequences of harmonic partials based on trial values of F_0. The discrepancy between the measured and predicted partials is referred to as the *mismatch error*. Of course, the mismatch error would be zero if a predicted F_0 were to match exactly the actual fundamental and the measured spectrum consisted solely of harmonic partials. However, the harmonics and partials would also "line up" for F_0's that are one or more octaves above and below the actual fundamental; thus even in the ideal case, some ambiguity occurs. Furthermore, in real situations, where noise and measurement uncertainty are present, the mismatch error will never be exactly zero.

Consider an example measured sequence of partials {200, 300, 500, 600, 700, 800} Hz. Choosing an F_0 of 100 Hz would give the predicted sequence of {100, 200, 300, 400, 500, 600, 700, 800}, where the predicted components at 100 and 400 Hz are not found in the measured data, but all the other measured components are accounted for. Selecting $F_0=50$ Hz also completely covers the measured partials, but many of the predicted partials (50, 100, 150, 250, 350, 400, 550, etc.) are not found in the measured sequence. Similarly, choosing $F_0=200$ Hz results in a predicted sequence of partials {200, 400, 600, 800}, which correctly predicts some of the measured partials but misses others. Therefore, some means of identifying the best match between predicted and measured partial frequencies is necessary.

Our solution has been to employ *two* mismatch error calculations. The first is based on the frequency difference between each partial in the measured sequence and its nearest harmonic neighbor in the predicted sequence. The second is based on the mismatch between each harmonic in the predicted sequence and its nearest partial neighbor in the measured sequence. The two measurements are not in general the same, as can be seen from the graphic depictions of Fig. 1. This two-way mismatch helps avoid octave errors by applying a penalty for partials that are present in the measured data but are *not* predicted, and also for partials whose presence in the measured data is predicted but do not actually appear in the measured sequence. The TWM approach also has the benefit that the effect of any spurious components or partials missing from the measurement can be counteracted by the presence of uncorrupted partials in the same frame.

Our algorithm for determining the TWM error, whose minimum determines the F_0 for each frame, is based on three considerations given as follows:

(a) The assumed harmonic relationship among the partials indicates that the frequency mismatch error, in Hz, between measured and predicted partial frequencies should be normalized by the frequency. Note that a mismatch of 10 Hz for components near 50 Hz is worse (20%) than a mismatch of 10 Hz for components near 5 kHz (0.2%).

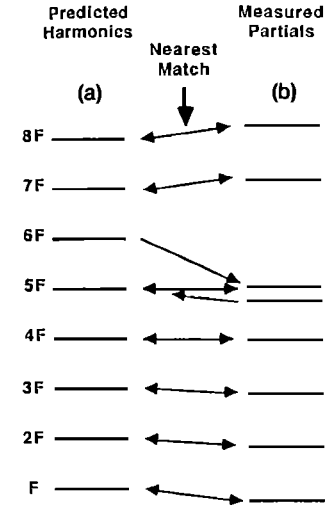


FIG. 1. The two-way mismatch error calculation is a two-step process where for each frame (a) each measured partial is compared to the nearest predicted harmonic, giving the measured-to-predicted error and (b) each predicted harmonic is compared to the nearest measured partial, giving the predicted-to-measured error. The total TWM error is a weighted combination of these two errors.

(b) Since the STFT spectrum measurement algorithm returns information with an approximately linear resolution, the fractional resolution, or Q, improves as frequency increases. Thus the higher partials have inherently better fractional resolution, which can help to improve the estimate of the corresponding fundamental frequency (Schroeder, 1968).

(c) Stronger partials generally have higher signal-to-noise ratio than weaker partials. Stronger partials are, therefore, assumed to have more reliable frequency estimates and their presence is less likely to be due to noise or other spurious events.

Based on the above guidelines, we have postulated an error weighting function, E_w. For measured partial n, E_w can be expressed as a function of the frequency *difference* between its frequency and the frequency of the nearest predicted harmonic (Δf_n), the measured *amplitude* and *frequency* of the partial (a_n and f_n), and the *maximum amplitude* of any partial in that frame ($A_{\max}$). Also, it is clear from the above considerations that the error function E_w should have the following properties:

Maximum error occurs when

(1) $\Delta f_n/f_n$ is large

or if

(2) $\Delta f_n/f_n$ is small and $a_n/A_{\max}$ is small.

Minimum error occurs when

(3) $\Delta f_n/f_n$ is small and $a_n/A_{\max}$ is large.

Any number of mathematical functions could be concocted to satisfy these conditions. We have designed a function, whose virtue is its simplicity, that has several coefficients which we have determined empirically for several cases

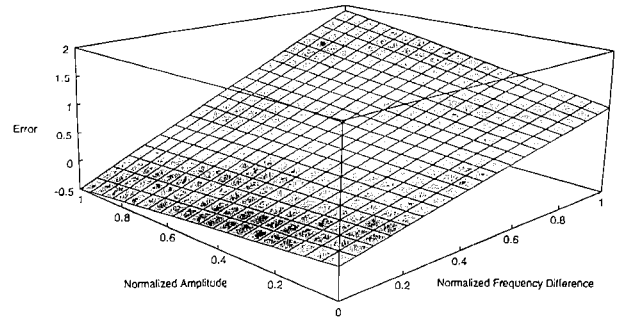


FIG. 2. An example error weighting function (E_w) used in the TWM procedure. The error penalty depends upon the amplitude of the measured partial normalized by the maximum amplitude partial in that frame, and the frequency difference between the measured partial frequency and the predicted harmonic frequency normalized by the measured partial frequency. For example, the best penalty (smallest error) occurs for a strong partial with small frequency difference.

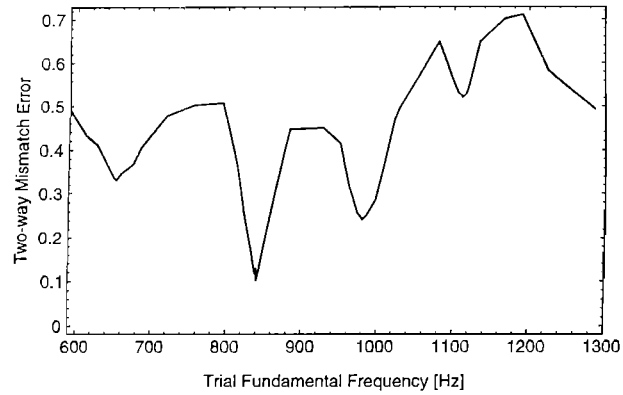


FIG. 3. Example TWM vs F_0 characteristic. The frequency corresponding to the minimum of this characteristic is the TWM estimate of the fundamental frequency (approximately 840 Hz in this case). Note that because many local minima are typically present a detailed search is required to identify the global minimum.

tried. A standard formulation that operates perfectly with every input example has not been found, although we present a set of coefficients that produces good results in practice. An example error weighting function is illustrated in Fig. 2.

The TWM calculation procedure for each frame can be summarized in the following form:

Step 1: Obtain K measured partials with amplitudes, A_k, and corresponding frequencies, f_k, from a particular STFT analysis frame; determine $A_{\max}=\max\{A_k\}$ and $f_{\max}=\max\{f_k\}$.

Step 2: Choose f_{fund}, the trial fundamental frequency (initially below the known frequency range of the input signal) and calculate the frequencies of N harmonics, $f_n=nf_{\text{fund}}$, where $N=\text{ceil}\{f_{\max}/f_{\text{fund}}\}$ is the smallest integer greater than $f_{\max}/f_{\text{fund}}$.

Step 3: For each f_n determine the corresponding partial frequency f_k that is closest to it; i.e., for each f_n choose f_k to minimize $\Delta f_n=|f_n-f_k|$. For k corresponding to the closest frequency, set $a_n=A_k$.

Step 4: Calculate the predicted-to-measured mismatch error according to the weighting formula:

$$\begin{aligned}\text{Err}_{p\to m}&=\sum_{n=1}^{N}E_w(\Delta f_n,f_n,a_n,A_{\max})\\&=\sum_{n=1}^{N}\Delta f_n\cdot(f_n)^{-p}+\left(\frac{a_n}{A_{\max}}\right)\\&\quad\times[q\Delta f_n\cdot(f_n)^{-p}-r].\end{aligned}\tag{1}$$

We have determined empirically that good values of p, q, and r are $p=0.5$, $q=1.4$, and $r=0.5$. This choice is consistent with the weighting properties, e.g., the minimum error occurs when $\Delta f_n/f_n$ is small **and** $a_n/A_{\max}$ is large.

Step 5: For each of the f_k determine the corresponding harmonic frequency f_n that is closest to it; i.e., for each f_k choose f_n to minimize $\Delta f_k=|f_n-f_k|$. For n corresponding to the closest frequency, set $a_k=A_n$.

Step 6: Calculate the measured-to-predicted mismatch error according to the weighting formula:

$$\begin{aligned}\text{Err}_{m\to p}&=\sum_{k=1}^{K}E_w(\Delta f_k,f_k,a_k,A_{\max})\\&=\sum_{k=1}^{K}\Delta f_k\cdot(f_k)^{-p}+\left(\frac{a_k}{A_{\max}}\right)\\&\quad\times[q\Delta f_k\cdot(f_k)^{-p}-r],\end{aligned}\tag{2}$$

where again the default values of p, q, and r are 0.5, 1.4, and 0.5, respectively. The total TWM error for the predicted f_{fund} is then given by

$$\text{Err}_{\text{total}}=\text{Err}_{p\to m}/(N)+\rho\text{Err}_{m\to p}/(K),\tag{3}$$

where a good value of ρ has been found empirically to be 0.33.

Finally, steps 2 through 6 are repeated for a series of trial fundamental frequencies spanning the known range of the input signal. The spacing between the trial fundamental frequencies can be chosen to achieve the required precision for the overall estimate. An example of $\text{Err}_{\text{total}}$ vs. f_{fund} is plotted in Fig. 3. Note that there are several local minima. The estimated F_0 for the frame is taken as the trial fundamental which yields the smallest value of $\text{Err}_{\text{total}}$.

Given the TWM error calculation procedure, the process of searching for the best value of F_0 for a given frame could follow one of several possible strategies. Our approach has been first to specify minimum and maximum frequencies for f_{fund} and then to perform a global search for the minimum mismatch error over the entire frequency range. This is done by iterating the trial frequency in equal-tempered semitone steps (i.e., by factors of 1.059 46) starting with the lowest frequency. However, this grid is often not fine enough to reveal the true TWM error minimum, so we then search in the vicinity of each local minimum with a progressively smaller step size until the change in $\text{Err}_{\text{total}}$ between steps drops below some arbitrary amount. Finally, the F_0 estimate corresponding to the least of the local minima is taken.

There are three user-specified parameters that can be changed to adjust the TWM procedure to suit the particular characteristics of a given input signal:

(1) *F_0 search range*: The minimum and maximum frequencies for the F_0 search are chosen to span the expected fundamental frequency range of the input signal. Choosing a larger range unnecessarily increases the computation time by requiring more steps in the search process and increases the possibility for errors.

(2) *Number of predicted partials N*: Selecting a large number of partials in the predicted sequence works best for recordings with minimal background noise and reverberation. Selecting a small number of predicted partials works best with noisy signals by limiting the effect of spurious partials in the measured data, but at the expense of lower resolution due to the lack of error averaging across many measured partial frequencies. The choice of N also depends upon the known or assumed spectral characteristics of the signal source. For most examples we have used $8 \leqslant N \leqslant 10$.

(3) *Exponent p in error function*: The p parameter adjusts the frequency-dependent weighting of the frequency difference calculation. A value of 0.5 (corresponding to a square-root weighting) has been found to work well on many examples, although $p=1.0$ has provided better results for highly reverberant recordings due to the reduced emphasis placed on the low level, high-frequency components.

As a simple example of the TWM calculation, Eqs. (1)–(3), consider the measured sequence of partials {200, 300, 500, 600, 700, 800} Hz mentioned previously. In this example we would like to determine, say, whether 50, 100, or 200 Hz is the best F_0 assuming all the measured partials are approximately equal in amplitude. Using 50 Hz in the TWM formulas results in $\mathrm{Err}_{p\to m}=122.58$, $\mathrm{Err}_{m\to p}=-3.0$, and $\mathrm{Err}_{\mathrm{total}}=7.49$. A 100-Hz F_0 yields $\mathrm{Err}_{p\to m}=32.0$, $\mathrm{Err}_{m\to p}=-3.0$, and $\mathrm{Err}_{\mathrm{total}}=3.83$. Finally, using 200 Hz gives $\mathrm{Err}_{p\to m}=10.0$, $\mathrm{Err}_{m\to p}=30.66$, and $\mathrm{Err}_{\mathrm{total}}=4.2$. In this case the minimum TWM error ($\mathrm{Err}_{\mathrm{total}}=3.83$) occurs for $F_0=100$ Hz, so 100 Hz is the fundamental frequency assigned to the measured sequence. Note that neither the predicted-to-measured nor measured-to-predicted errors acting alone can achieve an unambiguous F_0 choice. Although in this case the difference between the minima for 100 and 200 Hz is not huge, note that if we include a 100-Hz component in the measured spectrum the margin greatly improves. In this case the error for 200 Hz is about the same (4.0), whereas the error for 100 Hz drops to <1.0, i.e., the distinctiveness of the result depends on the degree to which F_0 can be interpreted unambiguously from the original spectrum.

B. The TWM procedure for duet signals

We have extended the TWM procedure to handle the case of two simultaneous voices. This has been done by modifying the procedure to include measurements of mismatches between the sequence of measured partial frequencies (for each frame) and a *pair* of trial harmonic patterns. The goal is to find the pair of fundamental frequencies which together best represent the partials found in each frame. A brief description is given next.

Two nonoverlapping frequency ranges are specified corresponding to the expected fundamental frequency ranges of the two musical voices. The nonoverlapping range requirement implies that either the lowest note of the upper voice is higher than the highest note of the lower voice for the whole duet, or else the duet has been divided into subsegments where this restriction is met. The TWM error calculation is performed as in the single voice case, except that the predicted sequence of partials is based upon two trial F_0 values: one chosen from the frequency range of the lower voice, the other from the range of the upper voice. The error minimization is accomplished by first stepping the trial F_0 for the lower voice across the search range while keeping the upper F_0 fixed at the middle of the upper search range. Once the lower sweep is completed, the lower F_0 is fixed at the frequency that resulted in the smallest error, then the upper F_0 is stepped across its search range in order to find the overall minimum. The resulting pair of F_0's are assumed to be the fundamental frequencies of the two duet voices.

Frequency tracking of duet signals has its own unique problems in addition to the problems already mentioned for the solo case. One problem stems from the limited frequency resolution of the STFT spectrum analysis: There is a high probability that certain original partials of the two voices will have frequencies close enough to one another that they will not be resolved by the STFT analysis, resulting in "collisions" of the partials. That is to say, whenever the frequencies of two partials differ by less than the STFT's resolution, such a collision may occur. A collision may appear as a single broad peak instead of the usual two distinct peaks in the spectrum, or it may be manifested as a narrow peak having a time-varying amplitude due to beating between the partials. Since many musical intervals in common-practice music, e.g., unisons, octaves, and fifths, result in deliberate coincidence of harmonics, partial collisions are a significant problem for duet frequency tracking. We have worked out procedures for "unwrapping" collisions in both of these cases, but the accuracies of these methods are limited when the amplitudes and frequencies of the partials are changing significantly from frame to frame.

Another problem is to somehow determine whether zero, one, or two voices are present at any given moment, since in general any of these conditions will frequently occur in duets. A complete solution to this problem would require an accurate segmentation of the duet into portions corresponding to {voice 1 only}, {voice 2 only}, {voice 1 and voice 2}, and {neither voice}. The segmentation task could be performed manually or through use of a yet-to-be developed automatic segmentation technique.

III. EXAMPLES AND DISCUSSION

We can best understand the strengths and limitations of the TWM frequency tracking procedure by observing its F_0 versus time performance for several different musical signals. Note that all of the examples given in this section show the unaltered output of the F_0 tracker. These examples are intended to demonstrate the underlying effective-

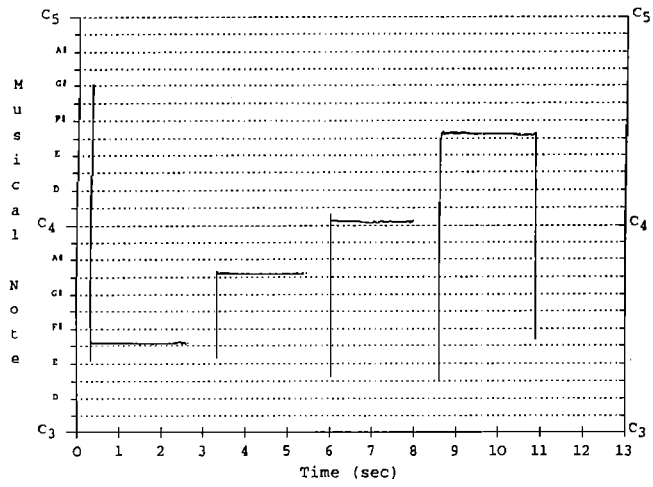


FIG. 4. TWM tracking results for a staccato arpeggio performed on a grand piano (F_3, A_3, C_4, and F_4). The ordinate axis is calibrated logarithmically as notes on the musical scale (A_4=440 Hz). Vertical "spikes" at the beginning or ending of each note are due to the nonharmonic character of the signal during rapid amplitude transients.

ness of the *raw* TWM procedure. Of course, the results could be improved by further procedures invoking rules of musical context or by making use of other specific knowledge about the signals being processed.

A. Piano signal without reverberant overlap

Figure 4 shows the TWM F_0 versus time output for a portion of an arpeggio sequence of notes played fortissimo on a grand piano (from track 39, index 3, SQAM compact disk, EBU, 1988). The recording for this example can be considered to be a "best case," since it contains very little noise, reverberation, or other degradations. The TWM results show the individual notes and rests of the arpeggio quite distinctly. The spikes are due to the initially uncertain fundamental frequencies of the piano notes, caused by brief, metallic thumps of hammers against strings. This effect is corroborated by an STFT spectrum analysis of the first two notes of the example (Fig. 5).

Another interesting feature of the frequency tracker output for piano tones is that a slight decrease in the F_0 estimate occurs during the decay of each note. This is probably caused by inharmonicities of the partials of piano tones; since high-frequency partials are normally "stretched" relative to lower ones (Fletcher, 1964) and because the TWM procedure seeks to find the best *harmonic* match to the measured partials, the initial positive inharmonicity tends to bias the F_0 estimate upward. This bias is reduced as the note decays, because the higher partials die out more quickly than the lower partials.

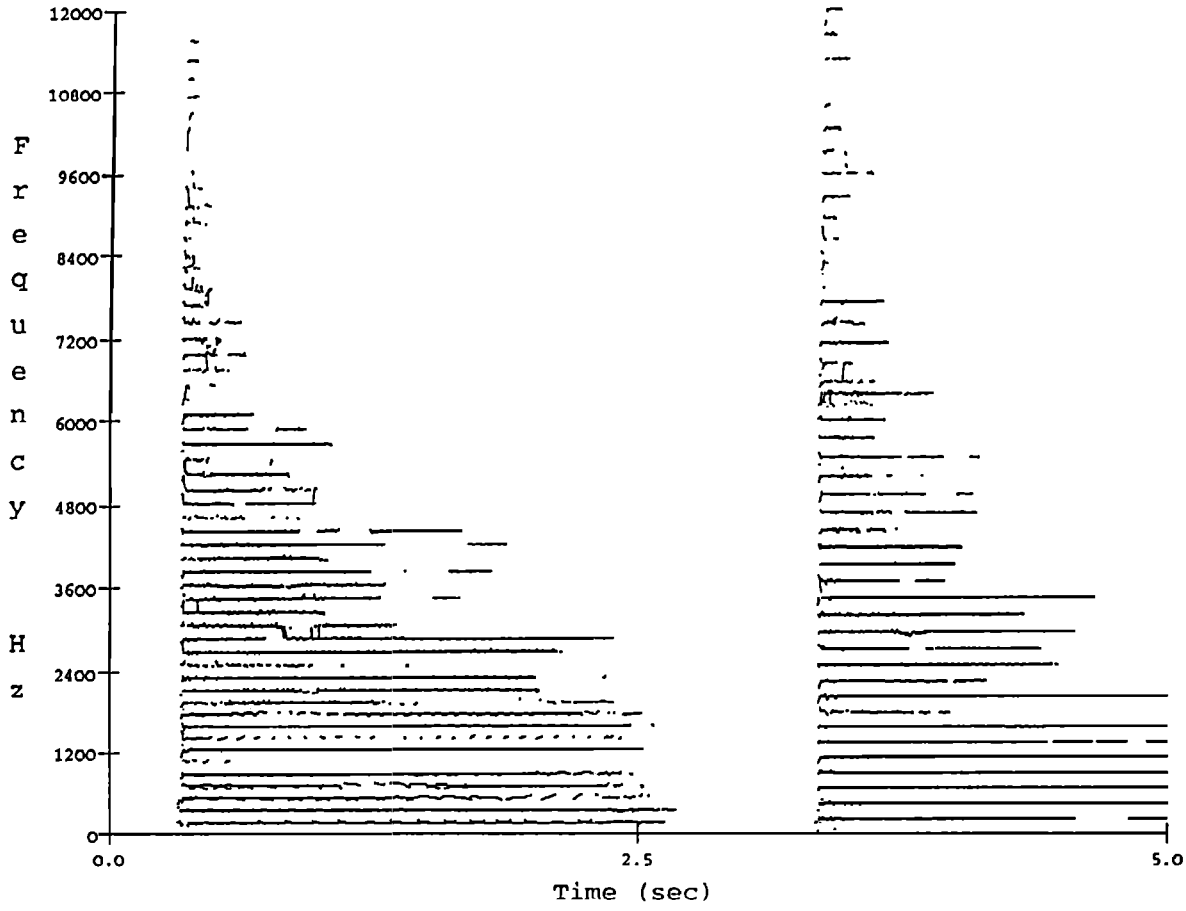


FIG. 5. A portion of the STFT frequency vs time analysis of the grand piano arpeggio signal used for the TWM calculation in Fig. 4. The rows of horizontal tracks correspond to the partials of the piano signal. The nonharmonic behavior at the onset of each note is due to the sound of the hammer striking the string or strings.

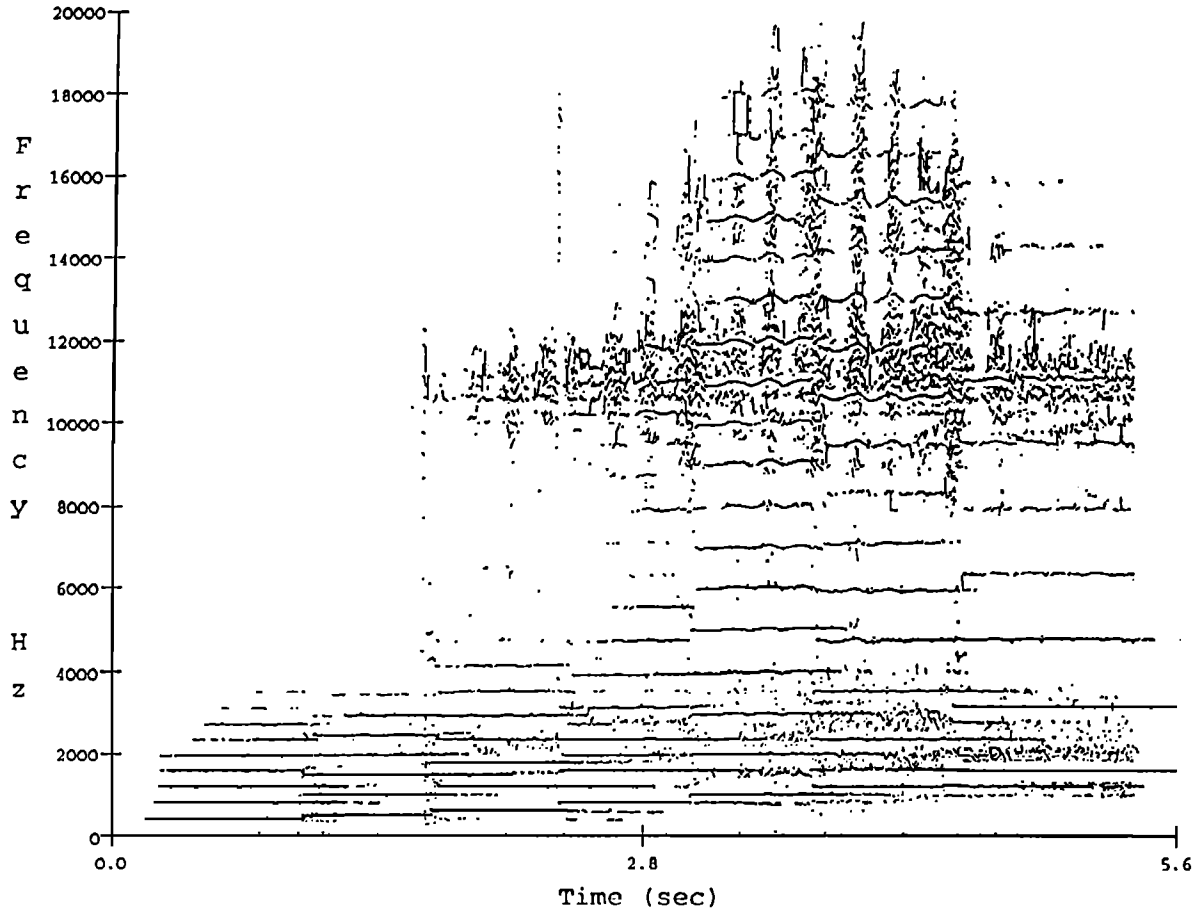


FIG. 6. STFT frequency vs time analysis of a legato arpeggio performed on a flute in reverberant surroundings. The harmonic behavior of the flute is observable as horizontal parallel lines. Noise due to the breathy quality of the flute and overlapping tracks at note boundaries are also apparent. Frequency vibrato can be identified, particularly in the upper partials.

B. Flute signal with reverberant overlap

Figure 6 shows the STFT spectrum analysis of a flute arpeggio (from track 13, index 1, SQAM compact disk, EBU, 1988). This recording was made with an omnidirectional microphone approximately 1 m from the instrument in a studio with 1.6-s reverberation time. The frequency versus time spectral data show discernible regions of parallel tracks corresponding to the harmonic partials of each note of the arpeggio. However, note that the tracks overlap from one note to the next due to the reverberant extension of each released note which continues after the next new note has begun. The flute's turbulent noise, or breathiness, is also visible, particularly in the 1- to 1.2-kHz region. Frequency vibrato is visible in the highest partials.

The corresponding frequency tracker output for the flute arpeggio (Fig. 7) shows the expected staircase shape. The frequency uncertainty of the transitions between many of the notes is due to the reverberation tail of the released note interfering with the F_0 estimation of the new note. Once the reverberation level drops below the level of the new note the TWM tracker locks onto the new F_0. Notice, however, that the TWM procedure is quite immune from effects of the turbulence noise in the flute signal.

The two spikes in frequency that occur during the first note of the flute arpeggio are due to an interesting feature of the flute spectrum during the note. Figure 8 shows the time-variant amplitudes of a few harmonics for the first note of the arpeggio. Note that there are pronounced amplitude fluctuations on each of the partials, due to the player's vibrato, and that the odd harmonics, including the

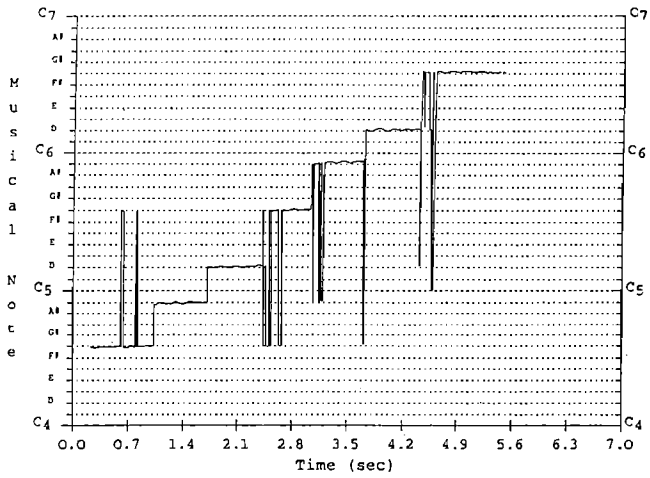


FIG. 7. TWM tracking results for the legato flute arpeggio with reverberation. The tracker is designed for a single-voice input signal, but the monophonic assumption is violated at note boundaries where the reverberation tail of the released note overlaps the onset of the next note, resulting in uncertain estimates of F_0.

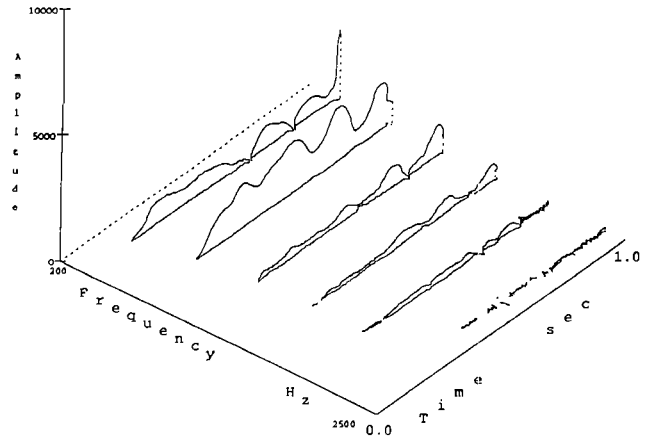


FIG. 8. Explanation of momentary octave jumps during first note (G_4) of flute arpeggio. Brief octave jumps occur when the amplitude of the odd partials becomes small due to the performer's vibrato: The missing odd partials result in an analyzed signal with F_0 equal to the second partial frequency.

fundamental, actually drop out at two points in the vibrato cycle. This causes the fundamental frequency effectively to jump an octave at these points. Thus, the spikes shown in the TWM tracker output are not spurious; the tracker accurately follows the brief octave jumps of the signal at those instants in time.

C. Violin signals: Synthetic, close miking, and reverberant

Figure 9 shows the musical score and TWM tracking results for a synthesized portion of Bach Partita III for

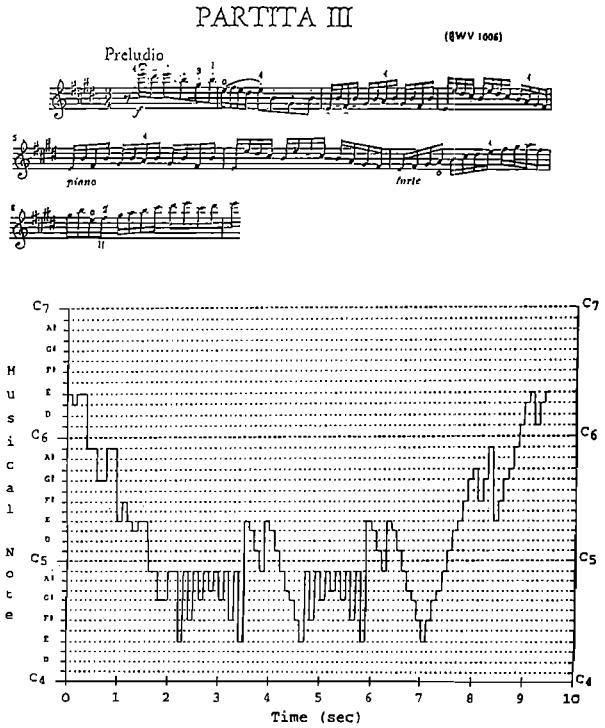


FIG. 9. Musical score and corresponding TWM tracking results for a synthesized performance of the Bach Partita III. The high quality of the F_0 estimate for the highly accurate, noise-free synthesized signal illustrates the best-case performance of the algorithm.

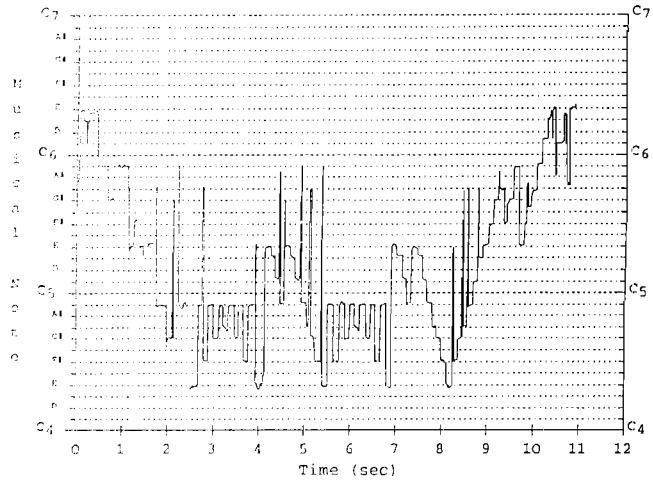


FIG. 10. TWM tracking results for a violin performance of the Bach Partita of Fig. 9, recorded in a nonreverberant studio with close miking. No explicit fundamental frequency is associated with the attack noise and bow scrape that occurs at the onset of several of the notes, resulting in spikes in the TWM results.

violin (BWV 1006) to demonstrate a best-case performance of the algorithm. This example was synthesized with exact note durations (no gaps), eight equal-amplitude harmonics, and exact equal-tempered note tuning. As expected, the TWM output shows excellent tracking capability with this pristine synthetic signal.

Figure 10 shows the TWM results for a studio recording of a real violin performance by Joel Smirnoff of the same partita obtained with close miking to minimize the effects of the recording room. The TWM tracker is able to follow the recorded performance quite well, but spikes and other fluctuations are present in the F_0 versus time graph. These effects are due to the characteristics of real violin signals. For example, noise, caused by the bow scraping the string as the performer attacks each note, results in frequency uncertainty at the beginning of each note. This should not be surprising, since in general noise violates the assumption of harmonicity on which the TWM procedure is based Also, the performer's vibrato and other stylistic performance expressions result in subtle frequency variations which occur during the individual notes, as shown in the graph.

Figure 11 shows the TWM results obtained from a reverberant recording (by violinist Itzhak Perlman) for the first two bars of the same Bach partita. The reverberation results in some uncertainty in the F_0 estimate, such as near 0.6, 1.2, and 1.5 s, where the TWM output briefly hops to the F_0 of the previous note. This occurs when the reverberated energy from the previous note or notes is comparable to the signal level of the current note, thereby introducing some ambiguity about the best monophonic F_0 estimate at that instant in time.

D. Soprano vocal signal with reverberant overlap

An example of the TWM procedure applied to vocal performance by a soprano voice is shown in Fig. 12 (from

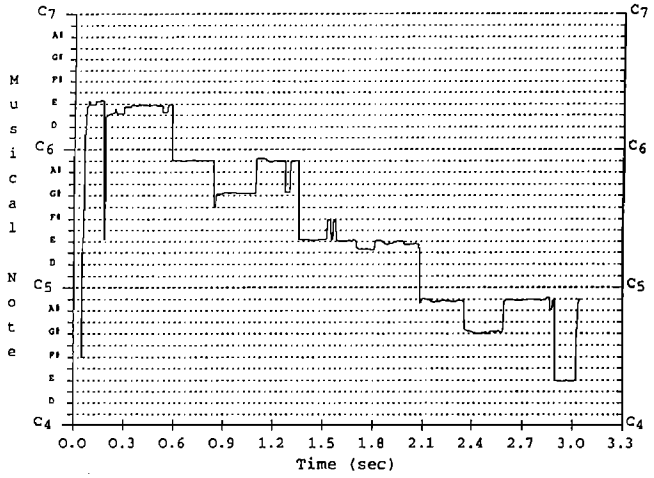


FIG. 11. TWM tracking results for a portion of a violin performance of the Bach Partita of Fig. 9, recorded in a reverberant auditorium. The tracking results are less accurate due to the nonmonophonic nature of the signal; reverberation from released notes is still present as each new note begins.

track 44, index 1, SQAM compact disk, EBU, 1988). The TWM output clearly indicates the shape and extent of the performer's vibrato, as well as the frequency behavior at the legato transitions within the performed excerpt.

E. Duet separation

An example of the duet separation procedure is depicted in Fig. 13. The duet consists of a soprano singing an arpeggio with vibrato while an alto sings a steady musical pitch. Figure 13(a) shows the analysis of the individual signals for reference, while Fig. 13(b) shows the TWM attempt for duet tracking. The major difficulty occurs when the fundamental frequencies of the two singers are close together, because of the numerous spectral collisions. For other examples of the duet separation procedure the reader is referred to previously published results (Maher, 1990).

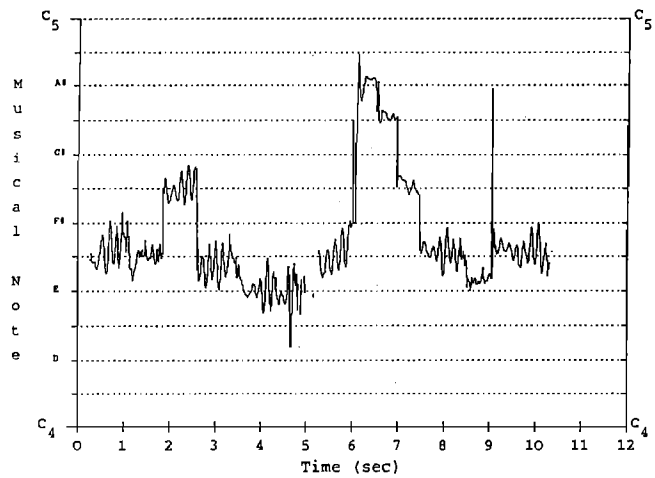
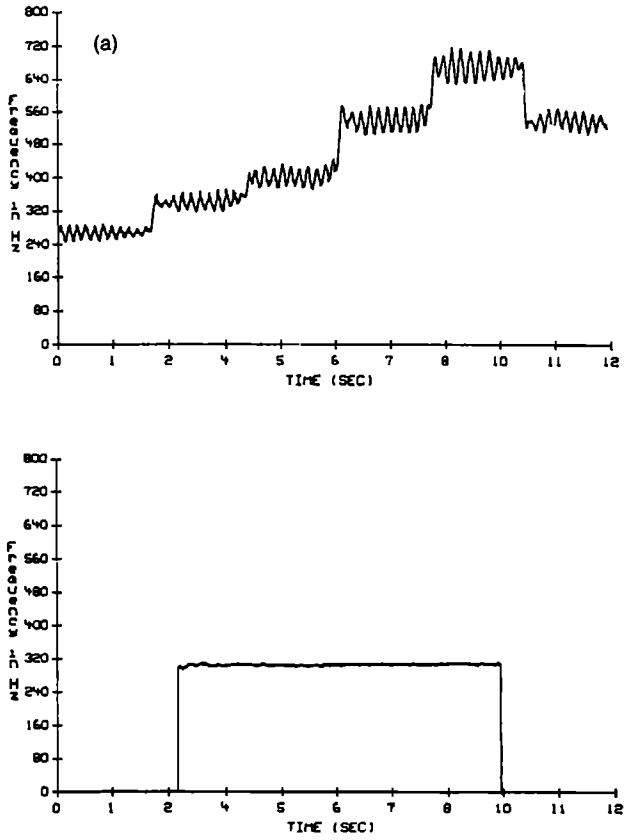


FIG. 12. TWM tracking results for a soprano vocal performance. The fluctuations in the F_0 estimate reflect the vibrato of the singer's voice.



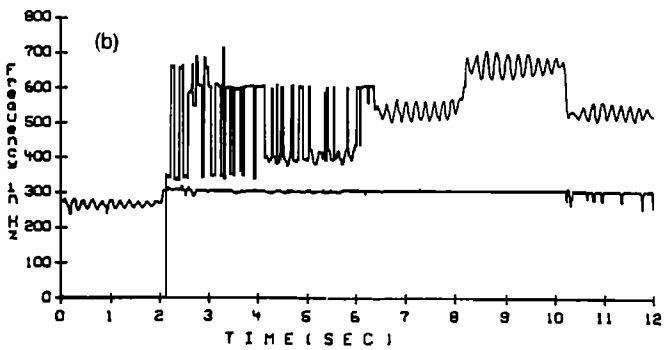


FIG. 13. Example TWM attempt for duet tracking. (a) For reference, the individual TWM results for a soprano (arpeggio with vibrato) and alto (single musical pitch) performance. (b) The TWM duet tracking results. Note the inadequate results when the two voices have similar fundamental frequencies.

IV. SUMMARY AND CONCLUSIONS

Thus far, results obtained using the two-way mismatch procedure for fundamental frequency estimation are encouraging and have been used successfully in several music signal processing projects involving analysis of musical sounds of variable frequency. The procedure has been applied to real signals corrupted by noise and reverberation with reasonable success, including examples where manual transcription would be quite difficult. The primary performance limitation of the TWM procedure is shared by most F_0 estimation techniques: nonharmonic signal components such as the bow scrape of a violin or the hammer strike of a piano. The interpretation and segmentation of the TWM

output has been left for a higher level process that makes use of musical context rules, *a priori* knowledge, and other heuristic decisions.

There has not been a great effort so far to minimize computation time because this investigation was carried out in a non-real-time research environment. The frequency search and mismatch calculation is time consuming, particularly where the signal fundamental varies rapidly and the F_0 estimation must be done many times per second, as in vibrato analysis. In these cases, the full TWM procedure (including the spectral analysis) implemented on an engineering work station requires more than 200 times real time, i.e., each second of sound requires more than 1 min of processing. An empirical relationship for the compute time to real-time ratio obtained for a NeXT 68040-based work station is

$$\mathrm{CT/RT} = 25 \cdot \mathrm{DWF} \cdot (\mathrm{FRO})^{1.7},$$

where DWF is the "dry–wet" factor (1 for a nonreverberant studio recording, and 1.3 for a reverberant recording), and FRO is the fundamental frequency search range in octaves. Thus, we are currently investigating methods to estimate the fundamental with coarse resolution first, then to employ the TWM calculation to identify the "best" fundamental with greater accuracy over a limited search range.

Finally, the problems associated with polyphonic F_0 estimation remain a significant area for research. The principal difficulty is due to overlapping spectral components among the multiple voices, requiring a source of additional *a priori* information (orchestration, number of voices, etc.) to enable the isolation of individual voices.

ACKNOWLEDGMENTS

This work was supported in part by a National Science Foundation Graduate Fellowship, and by grants from the Engineering Foundation (RI-A-91-11) and the University of Nebraska Research Council. Chris Kriese provided valuable assistance in preparing many of the figures used in this paper.

Allen, J. B., and Rabiner, L. R. (**1977**). "A unified approach to short-time Fourier analysis and synthesis," Proc. IEEE **65**, 1558–1564.

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.

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

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

Doval, B., and Rodet, X. (**1991**). "Estimation of fundamental frequency of musical sound signals," Proc. ICASSP **5**, 3657–3660.

EBU (European Broadcasting Union) (**1988**). "Sound quality assessment material recordings for subjective tests," Compact Disc, Technical Report No. 3253-E, Brussels, Belgium.

Fletcher, H. (**1964**). "Normal vibration frequencies of a stiff piano string," J. Acoust. Soc. Am. **36**, 203–209.

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

Maher, R. C. (**1990**). "Evaluation of a method for separating digitized duet signals," J. Audio Eng. Soc. **38**, 956–979.

Maher, R. C. (**1991**). "Sinusoidal additive synthesis revisited," in Proceedings of the 1991 Audio Engineering Society Convention, Preprint No. 3128.

Maher, R. C., and Beauchamp, J. W. (**1992**). "Frequency tracking of solo and duet passages using a harmonic two-way mismatch procedure," J. Acoust. Soc. Am. **92**, 2347 (abstract).

Markel, J. D., and Gray, A. H., Jr. (**1976**). *Linear Prediction of Speech* (Springer-Verlag, New York).

Martin, P. (**1982**). "Comparison of pitch detection by cepstrum and spectral comb analysis," Proc. IEEE ICASSP **1**, 180–183.

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

Noll, A. M. (**1966**). "Cepstrum pitch determination," J. Acoust. Soc. Am. **41**, 293–309.

Piszczalski, M., and Galler, B. A. (**1979**). "Predicting musical pitch from component frequency ratios," J. Acoust. Soc. Am. **66**, 710–720.

Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ).

Rife, D. C., and Boorstyn, R. R. (**1976**). "Multiple tone parameter estimation from discrete-time observations," Bell Syst. Tech. J. **155**, 1389–1410.

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. **ASSP-22**, 353–362.

Schroeder, M. R. (**1968**). "Period histogram and product spectrum: New methods for fundamental frequency measurement," J. Acoust. Soc. Am. **43**, 829–834.

Smith, J. O., and Serra, X. (**1987**). "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in Proceedings of the 1987 International Computer Music Conference, International Computer Music Association, San Francisco, CA., pp. 290–297.

Sondhi, M. M. (**1968**). "New methods of pitch extraction," IEEE Trans. Audio Electroacoust. **AU-16**, 262–266.

Wolcin, J. J. (**1980**). "Maximum a posteriori estimation of narrowband signal parameters," J. Acoust. Soc. Am. **68**, 174–178.