Musical fundamental frequency tracking using a pattern recognition method[a)]

Judith C. Brown
Physics Department, Wellesley College, Wellesley, Massachusetts 01281 and Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 12 March 1991; accepted for publication 15 May 1992)

In a recent article [J. C. Brown, "Calculation of a Constant Q Spectral Transform," J. Acoust. Soc. Am. **89**, 425–434 (1991)], the calculation of a constant Q spectral transform that gives a constant pattern in the log frequency domain for sounds with harmonic frequency components has been described. This property has been utilized in calculating the cross-correlation function of spectra of sounds produced by musical instruments with the ideal pattern, which consists of one's at the positions of harmonic frequency components. Therefore, the position of the best approximation to the "ideal" pattern for the spectra produced by these instruments has been determined, and in so doing the fundamental frequency for that sound has been obtained. Results are presented for scales produced by the piano, flute, and violin as well as for arpeggios played by a wide variety of instruments.

PACS numbers: 43.60.Gk, 43.75.Yy

INTRODUCTION

Previous work in the field of musical fundamental frequency tracking was reviewed recently in an article describing a frequency tracker operating in the time domain (Brown, 1991). Since this work is based on a calculation in the frequency domain, we will limit our background discussion to studies that have been done on musical systems in the frequency domain except where directly relevant to this work.

Most of the efforts at musical frequency tracking have taken place in the frequency domain (Terhardt, 1979; Terhardt *et al.*, 1982; Amuedo, 1985; Chafe and Jaffe, 1986) and have used a method of attack similar to that of the Schroeder (1968) histogram method. After the calculation of a fast Fourier transform, a hypothesis is asserted for each frequency component of all possible fundamental frequencies for which it could be a harmonic; i.e., each frequency component is divided by integers and the results are entered in a table. The entries are weighted, and a decision is made based on criteria involving the number of components and their weights. The frequency is chosen which most closely meets previously determined criteria.

A similar frequency tracker (Piszczalski and Galler, 1979) took ratios of pairs of components to form their hypotheses for the fundamental and then proceeded as above. Duifhuis *et al.* (1982) studied speech segments using a method which most closely approaches that of this article. Following an FFT, they kept a maximum of six peaks and then used a "harmonic sieve" to determine which of these peaks best fit the logarithmic spacing obtained with harmonic frequency components. This method was later refined by Scheffers (1983).

[a)] The nomenclature "fundamental frequency tracker" or "frequency tracker" is used rather than "pitch tracker" because the editor wishes to observe the psychoacoustical distinction between pitch as a perceived quantity and frequency as a physical quantity.

I. BACKGROUND

In a recent article (Brown, 1991), we described a calculation that serves as the basis for the frequency tracker which will be discussed. In this calculation, a spectral transform equivalent to a 1/24th octave filterbank is carried out every 15 ms on the digitized sound from a musical instrument. The frequency components thus have logarithmic spacing. As described (see Fig. 1 in Brown, 1991), for a sound with harmonic frequency components, these Fourier components have a spacing in the log frequency domain which is independent of the fundamental frequency. For example, the spacing between the fundamental and the second harmonic is log(2), that between the second and third components is log(3/2), and so on.

Since this pattern is constant for harmonic frequency components, the "pattern" with 1's at the appropriate spacings can be convolved (or cross correlated) with the spectral transform, and a maximum should occur at the position of the fundamental. Note that this method accomplishes the same purpose as the histogram method; the values of the cross-correlation function being similar to the the sums of table entries. Here, however, the values correspond to all possible harmonic components.

For example as the convolution is computed, each component of the harmonic pattern will fall on each component of the analyzed sound. This "asserts" the appropriate fundamental as a hypothesis (weighted with the value of the cross-correlation function at that point) but it is simultaneously asserting that same fundamental for each of the components of the sound which are in the appropriate position. So rather than dividing, choosing a weighting system, and keeping entries in a table, we obtain one number for each frequency corresponding to the sum of all the frequency components of the sound that are in the correct position to be multiples of that frequency (harmonics of that fundamental). Thus in a very elegant and complete way we are obtaining the results

 0001-4966/92/091394-09$00.80

that the previous researchers approached with the histogram method. There is a computational advantage as well in that we simply add the components with the appropriate spacing.

Finally this frequency tracker solves the problem of the "missing fundamental" in much the same manner as that hypothesized for humans. It is essentially comparing the harmonics present to a template and finding the best match. This is consistent with the pattern matching theory (Gerson and Goldstein, 1978) of human pitch perception.

II. RESULTS

Before presenting the results on musical instruments, it is instructive to consider this method with single frames. In general, the calculation was carried out on 15 ms of analyzed sound. In Fig. 1, the signal analyzed consisted of a sound generated in software with 20 harmonics of equal amplitude. The lower graph of Fig. 1 is the spectral transform of this signal in the log frequency domain. The pattern that is convolved with this transform consist of 1's with the same spacing as that of these peaks since this signal was synthesized to have the ideal shape. The cross-correlation function is given in the top of Fig. 1. Clearly the largest peak of the cross correlation is at the position of the fundamental for this idealized spectrum, and it can be chosen easily by a peak picker.

It should also be noted, however, that the cross-correlation function has peaks at the position of one-half of and two times the fundamental. These peaks give rise to octave errors (a problem for all frequency trackers). With our method, the source of the problem is that the even peaks of the pattern line up with the spectrum for the frequency an octave below that of the fundamental so, if enough peaks are included in the pattern, the cross-correlation function will have the same value at this position as at that of the true fundamental. For the frequency an octave over that of the fundamental (the second harmonic), the peaks of the pattern are aligned with the even peaks of the spectrum again giving rise to a large value of the cross-correlation function. The octave errors thus produced are the chief source of error for our frequency tracker. In a later section, we will describe a means of eliminating the error occurring on the second harmonic.

An example of the cross-correlation function for a digitized musical sound is given in Fig. 2 for the sound produced by a violin. This is a particularly favorable example as the spectrum includes a strong fundamental, and the cross correlation exhibits an even stronger and more unambiguous peak than for the synthesized data which were represented in Fig. 1. We will represent some of the calculations for the violin that are the most difficult to interpret after the general presentation of our results and the discussion of adjustable parameters for this frequency tracker.

With any pattern matching method, the cross correlation establishes most unambiguously the position of the pattern that is sought the closer in shape it is to the known "ideal" pattern. (Duda and Hart, 1973) Thus the number of components in the ideal pattern should match the average number of non zero Fourier components for a particular instrument. The effect of varying these components in the pattern is thus an adjustable parameter to be optimized for each instrument.

The graphs of Figs. 3–5 demonstrate the frequency tracking results on sounds produced by a flute, piano, and violin as examples of wind, keyboard, and string instruments, respectively. We have plotted midinote versus time where middle C (C4) is represented by midinote 60, and each semitone corresponds to a value of one midinote higher (or lower). Graphs of the spectra that were used for these calculations are found in Brown (1991). Each point in these graphs represents the peak of a cross-correlation calculation on an analysis frame similar to that of Fig. 2 corresponding to approximately 15 ms of sound. Each instrument is playing a scale so that perfect results would consist of a sequential set of horizontal lines rising by one or two midinotes corresponding to a half or a whole step in the scale. Thus errors made by the frequency tracker are easily distinguished as points off the appropriate horizontal line.

The spectrum of the sound produced by the flute consists of a fundamental and a few upper harmonics that diminish in amplitude in a regular fashion. In Fig. 3, we show the effect of varying the number of components in the harmonic pattern from three components to five where the optimum pattern consists of four harmonics. The flute is playing a C major scale from C4 to E6. Errors occur on note changes when there is more than one note present. Most of the errors for the flute occur on the lowest notes when there are too few components in the pattern (lowest curve). This is due to little energy in the lower harmonics for these notes.

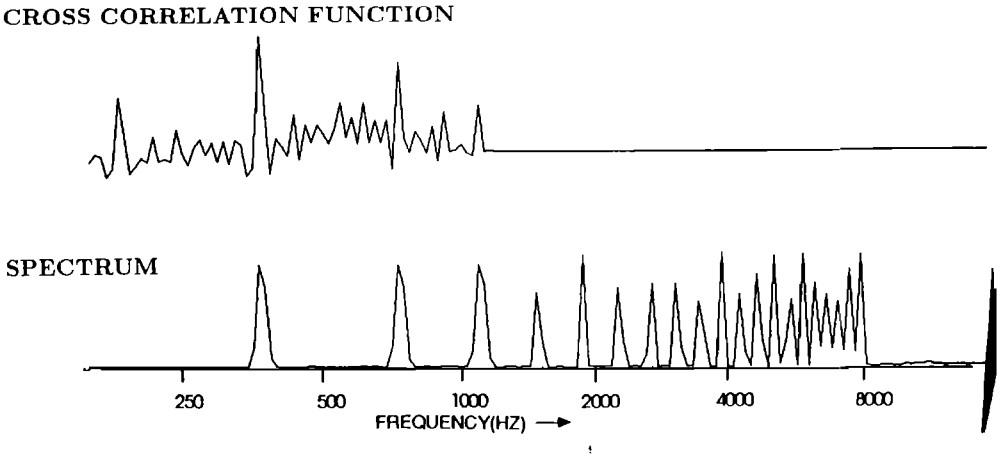

FIG. 1. Spectrum versus log frequency for a sound consisting of 20 harmonics of equal amplitude (below) and the cross-correlation function of this spectrum with a function consisting of 1's with the same spacing (above).

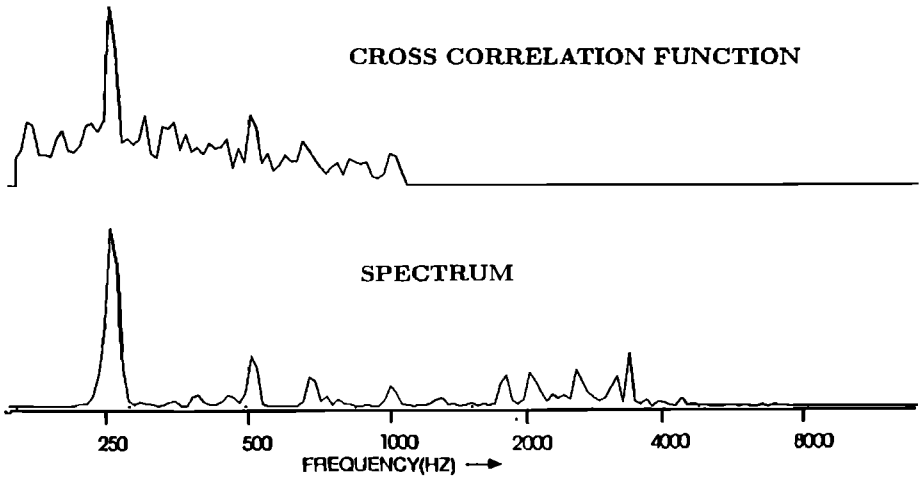

FIG. 2. Spectrum versus log frequency for the note C4 produced by a violin (below) and cross-correlation function (above) of this spectrum with the same function as that described for Fig. 1.

For the higher notes the errors on note changes are almost all low by an octave. These errors occur because, when the pattern has its lowest component at the position of an octave below the fundamental of the sound, the even components of the pattern pick up all the harmonics of the sound. The odd components of the pattern are in a position to pick up any sound from the previous note's decay, giving an advantage over the pattern at the position of the fundamental. Only by limiting the number of components in the pattern to match those of the particular instrument studied can this error be eliminated. In practice this is not a serious problem, as none of these errors occurred for more than one frame; they could be easily eliminated by having two frames agree.

The results for a piano sound are found in Fig. 4 where the qualitative behavior is similar to that of the flute, and the optimum number of components in the cross-correlation pattern is the same. The total number of errors for the piano, with the optimum number of components, is three as opposed to six for the flute.

Preliminary fundamental frequency tracking results on the violin were reported at the Syracuse meeting (Brown, 1989) of the Acoustical Society of America. The spectrum of

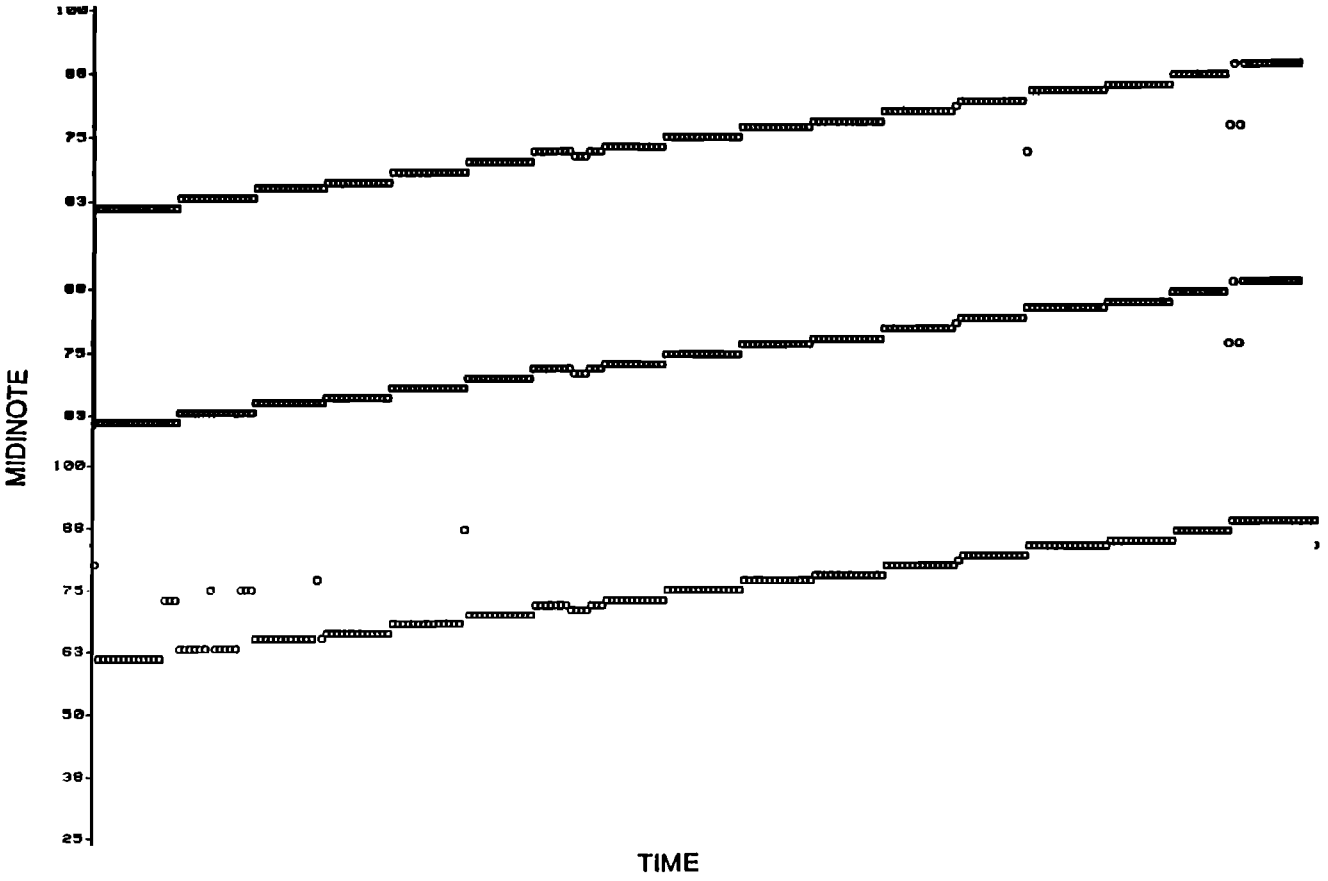

FIG. 3. Fundamental frequency tracking results for a flute scale from C4 to E6 using cross correlation with a pattern consisting of three components (below), four components (center), and five components (above).

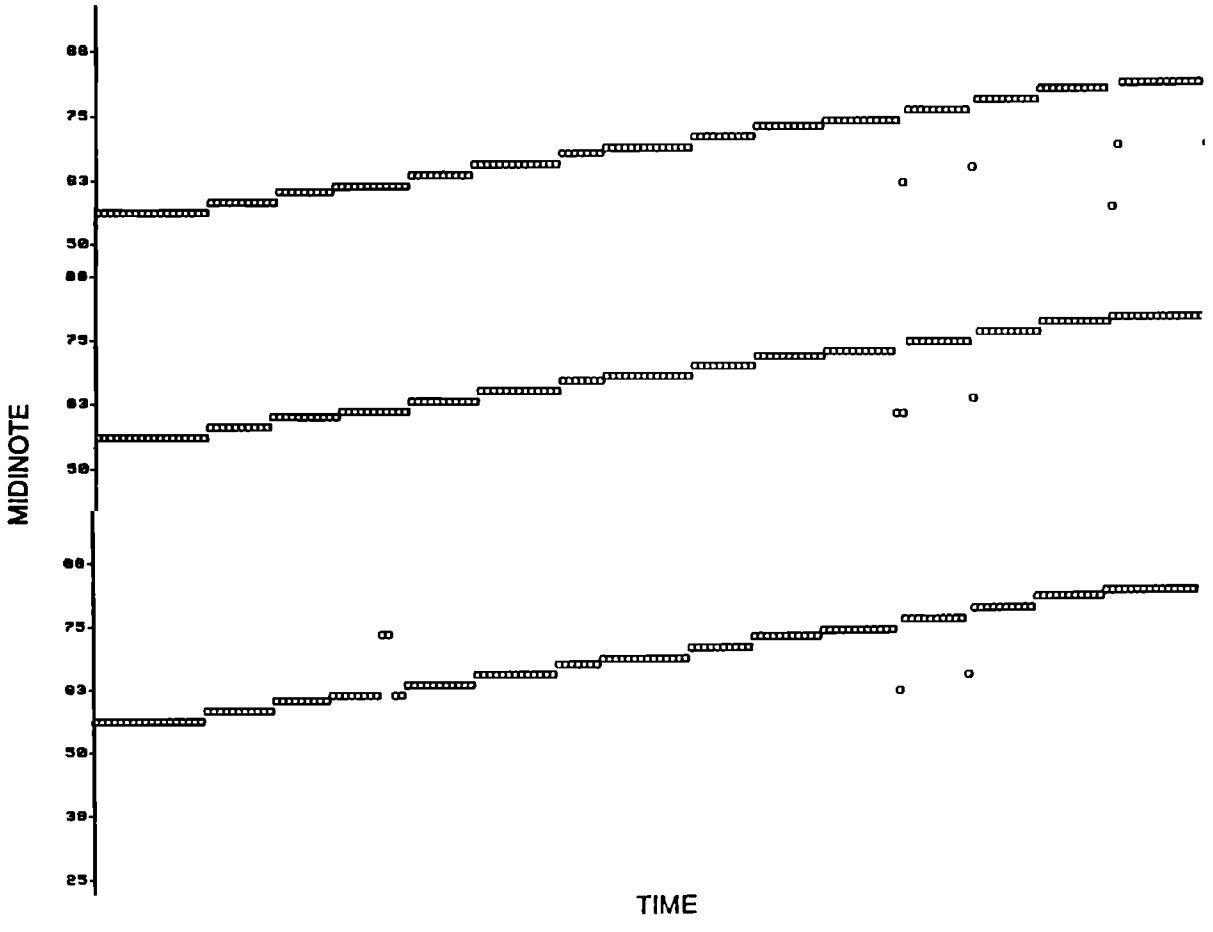

FIG. 4. Fundamental frequency tracking results for a piano scale from G3 to G5 using cross correlation with a pattern consisting of three components (below), four components (center), and five components (above).

of the violin is quite different from those of the flute and piano. There is a strong formant in the region of 3000 Hz, which gives rise to extremely strong upper harmonics. In Fig. 5, the frequency tracking results are found for this violin with 6 to 20 components in the cross-correlation pattern. The optimum number of components is 11, but the results are not terribly sensitive to this parameter as long as it is reasonably close to the optimum.

To clarify the source of errors with this method, we have chosen four frames for the violin from the region of the note transition from E5 to F♯5 (Fig. 6). All note transitions are difficult for a frequency tracker since there are two notes present. For each of these frames we have graphed the spectrum, the cross-correlation function, and the peak chosen to represent the fundamental frequency. Time increases from the bottom three graphs to the top three. The first two frames (lowest six graphs) are in error by an octave below the correct fundamental.

We have indicated the harmonics for F♯5 with small arrows. Harmonics three, four, and five fall into the formant region and are thus strongly amplified. Since positions on the pattern are closely spaced for the higher harmonics, when the pattern lines up on the (winning) frequency an octave below the frequency of E5, contributions from these higher harmonics from F♯5 are sufficient to make this the winning note. This happens again for the next frame. In the third and fourth frames (top six graphs) the higher harmonics of E5 have essentially died out, and F♯5 wins as it should.

Aside from the number of components in the cross-correlation pattern, the other adjustable parameter is the tuning of the center frequencies for the bins for the calculation of the spectrum. In Fig. 7, we have varied the tuning over a semitone in steps of 1% of the center frequency for the violin. For example, the curve marked 0.97 has center frequencies 3% lower than those in the curve marked 1.00. Note that the frequency tracker is very sensitive to this parameter with an obvious deterioriation in performance with even the small change to 0.99 from the optimum 1.00. It is suggested that this tuning analysis be carried out for each A to D converter used, since frequency shifts can occur during the process of sampling.

As a final test for this frequency tracker we have applied it to a "musical obstacle course" consisting of four ascending notes played very slowly (about 1 s per note) by each of a

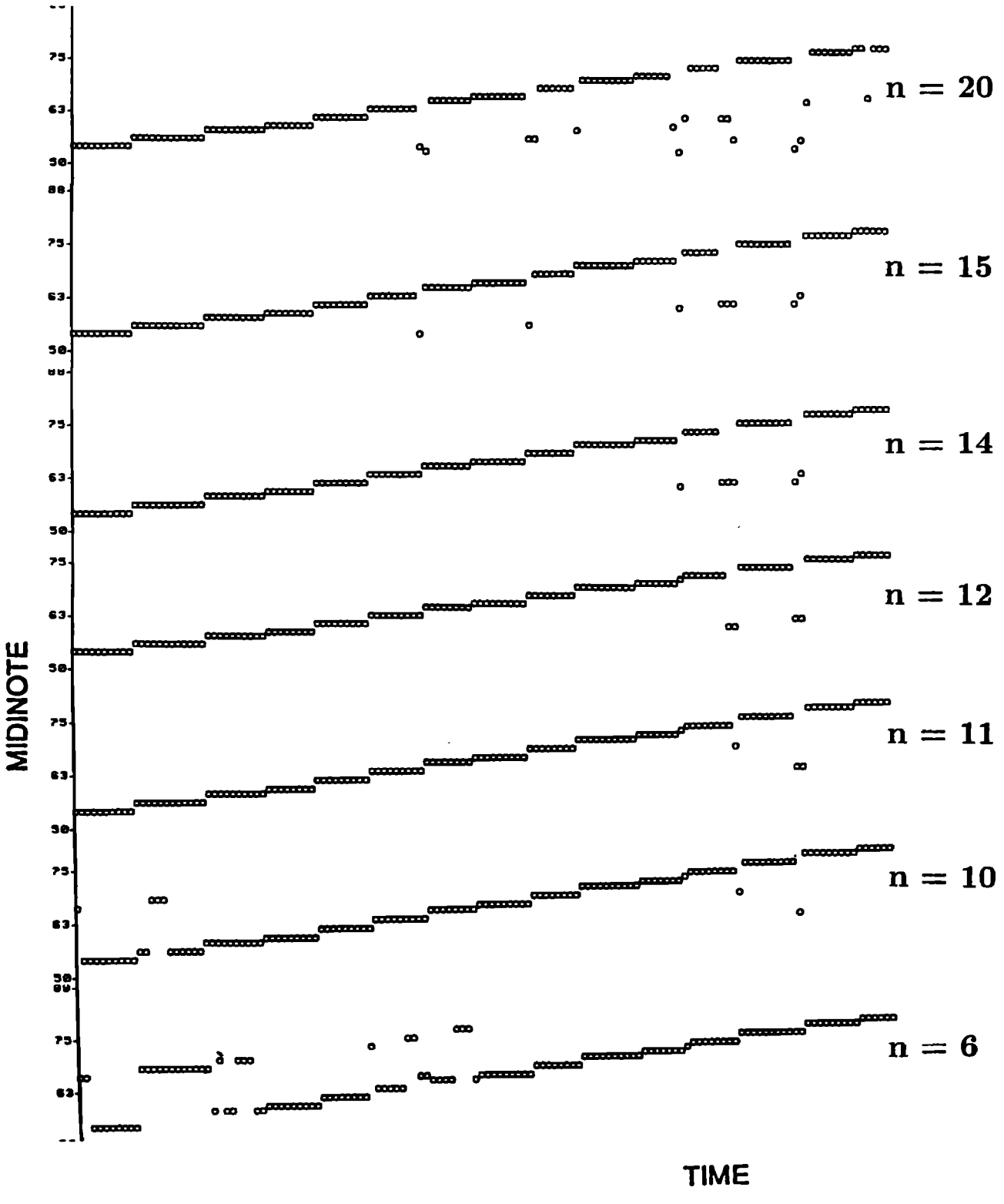

FIG. 5. Fundamental frequency tracking results for a violin scale from G3 to G5 using cross correlation with patterns consisting of components with the number varying from 6 to 20. The number of components is indicated on the curve.

violin, viola, cello, clarinet, alto sax, tenor sax, trumpet, trombone, and French horn. The cross-correlation pattern cannot be optimized for each instrument as was done previously so poorer results might be expected. A compromise of eight harmonics in the pattern was chosen, and a graph of the frequency tracking results is found in Fig. 8. The notes in the soundfile were spliced together so there is silence for a few frames between notes; thus these points off the horizon-

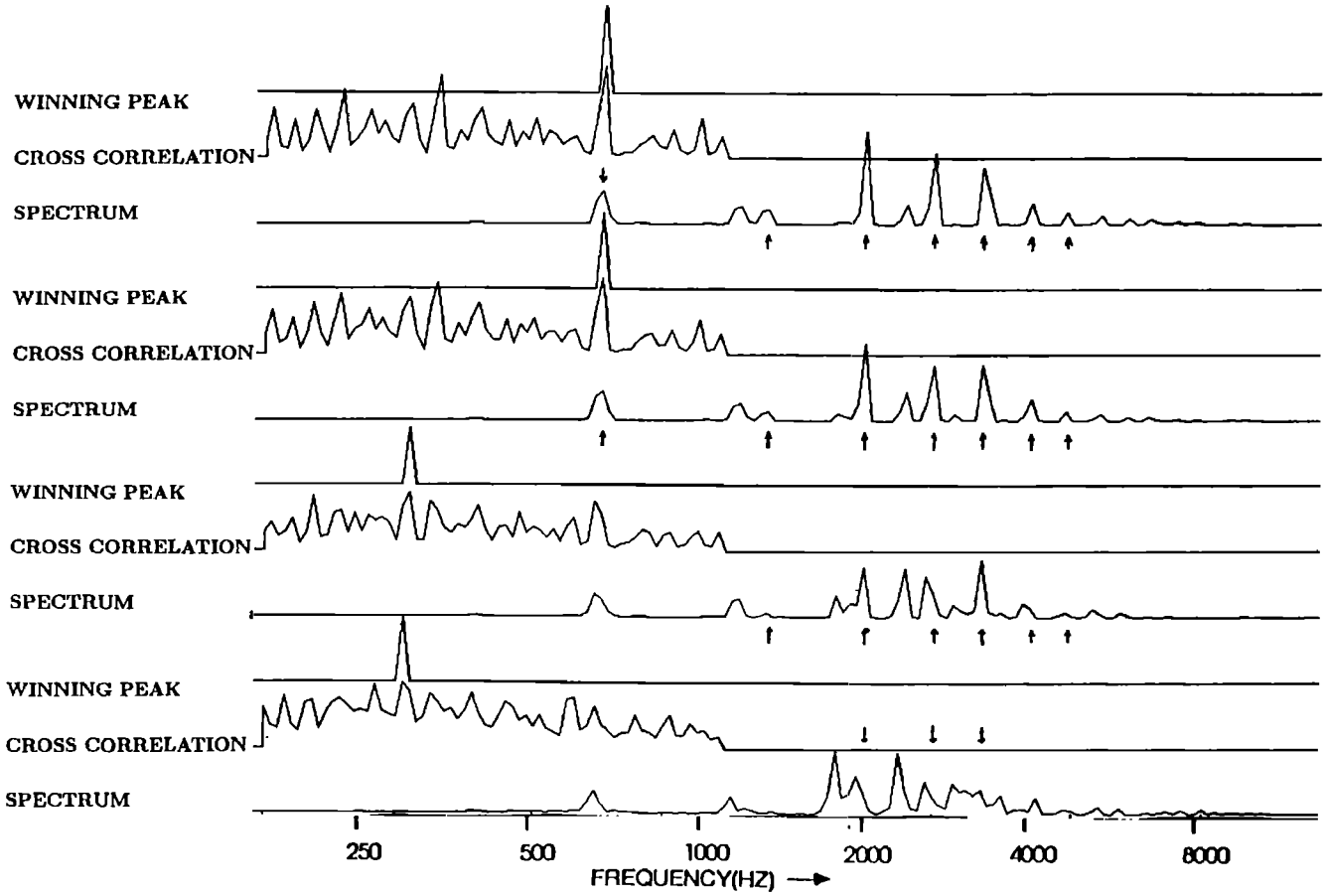

FIG. 6. Four frames for the violin transition from E5 to F#5. For each is graphed the spectrum versus log frequency, the cross-correlation function, and the peak chosen as the fundamental frequency.

tal lines between notes do not represent errors.

We have also analyzed these sounds with a frequency tracker using the method of narrowed inverted autocorrelation (Brown and Puckette, 1987; Brown *et al.*, 1989; Brown, 1991) in order to compare these two methods (Fig. 8). The autocorrelation results are found in Fig. 9. Figures 8 and 9 represent on the order of 2300 frames or discrete frequency tracking events and had to be sampled to be compressed on to a single graph. They are sampled at a rate of 4 to 1; so a given point represents the results of the analysis of four frames. Each note is held for approximately 1 s with a frame size of 16 ms which gives 64 events per note.

The autocorrelation frequency tracker misses fewer frames in the regions between notes because it has a mechanism for not reporting a note if the value of the the function differs sufficiently from its ideal value. The points off the horizontal lines for this method represent single errors. The points on the time axis are the ones which have been dropped. However, this frequency tracker had more difficulty with tuning problems (points just off the horizontal line or a split within a horizontal line into two lines a semitone apart). The pattern recognition frequency tracker does better during the playing of the note, i.e., the horizontal lines are unbroken with the exception of the first note. The points that are off during the silences could easily be dropped for both methods by adding an amplitude detector. This was not done, nor did we require two or more frames to agree, as is usual for frequency trackers, as we did not want to eliminate the bases for comparisons for each of the adjustable parameters.

III. DISCUSSION AND CONCLUSIONS

Essentially all errors with the pattern recognition method were octave errors. We are able to eliminate these errors on the high side by a rather ingenious method suggested by Steven Haflich. A cross-correlation pattern is used with component spacing corresponding to $2f$, $3f$, $4f$, $6f$,... that is, twice as many components as in the usual case but with alternating signs for these components. Thus the positive even components of this pattern will line up with harmonic spectral components as before; in this position the negative components of the pattern have no effect as they will fall between components of the sound. The value of the cross-correlation function will be the same at the position of the fundamental as with the previous pattern. Now however, when the pattern is aligned with its lowest component on the second harmonic (position of the octave up error) of the musical sound, each of the components of the pattern matches a component of the sound. Since every other component of the pattern has a different sign, the sum gives a low value of the cross-correlation function and eliminates this frequency as a candidate.

This method of suppressing the cross-correlation peak for the second harmonic worked extremely well as predict-

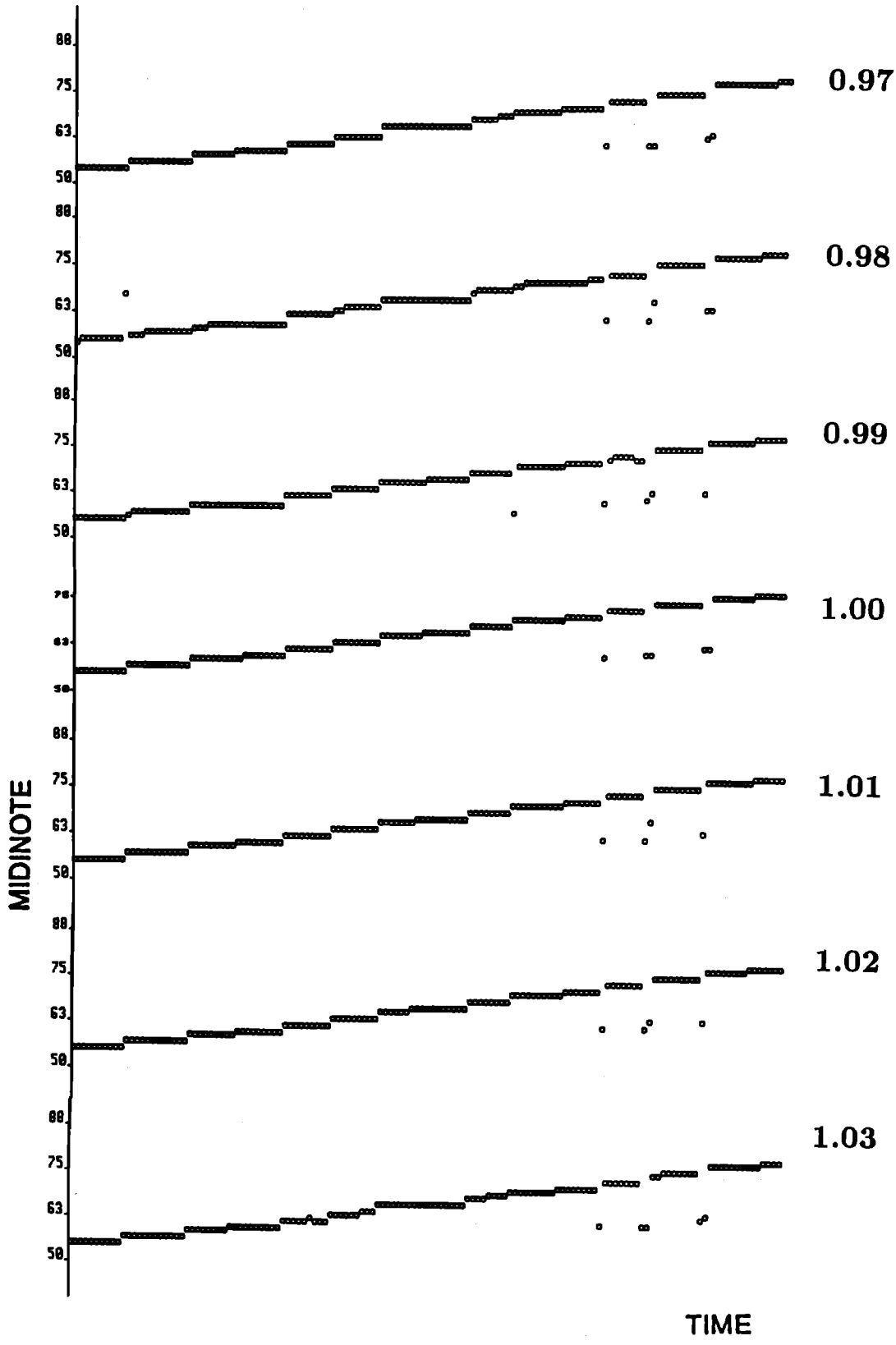

FIG. 7. Effect on the frequency tracking results of varying the center frequencies in the calculation of the spectrum versus log frequency. Tunings range from 97% to 103% of the standard tuning.

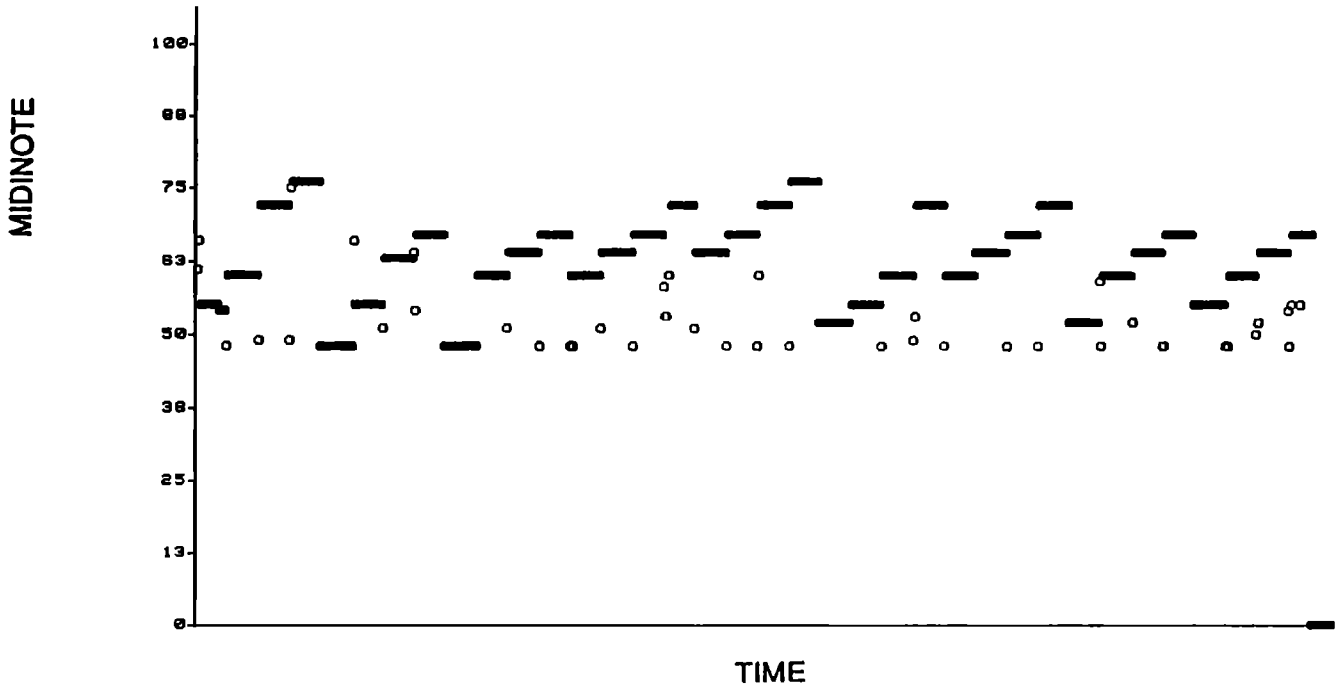

FIG. 8. Fundamental frequency tracking results using the method of pattern matching for nine instruments playing arpeggios.

ed. We did not include these results for two reasons. First the "optimized" frequency tracker worked extremely well on the sounds studied without it; and second, the disadvantage of this method is that it doubles the number of components in the pattern. This adds to the calculation time, and decreases the range of frequencies which can be examined.

Our pattern recognition method has produced excellent tracking results for the musical sounds in this study. While

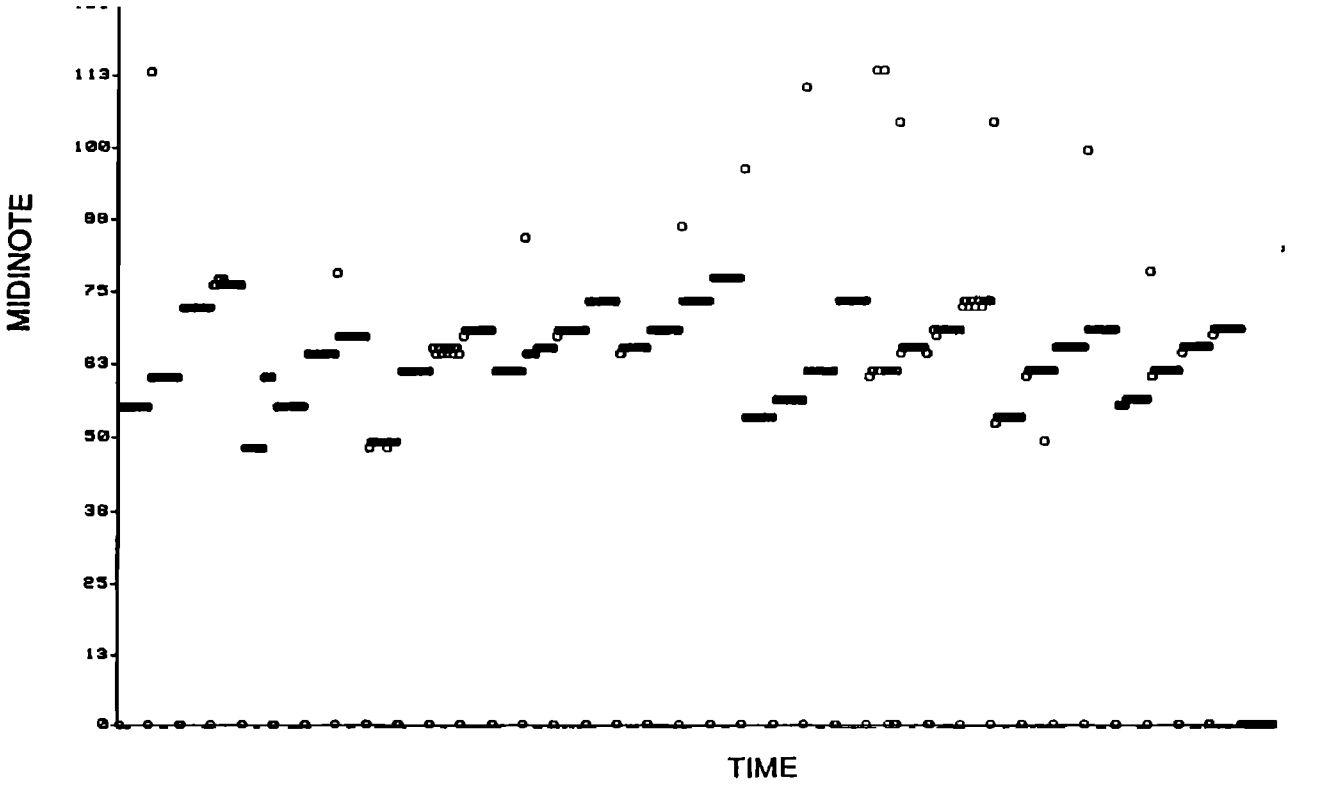

FIG. 9. Fundamental frequency tracking results using the method of narrowed autocorrelation for nine instruments playing arpeggios.

single examples of musical instruments can be atypical, the spectra of the sounds chosen varied from a simple spectrum consisting of a strong fundamental with a few higher harmonics to an extremely complex spectrum where the fundamental was often weak and most of the energy was concentrated in higher harmonics over 2000 Hz. The success of our frequency tracker in analyzing these sounds indicates the ability to deal with a wide variety of musical sounds. Finally, we emphasize that essentially perfect results could have been obtained for the sounds in this study simply by requiring that results from two successive frames agree.

ACKNOWLEDGMENTS

I am very grateful to Steven Haflich for suggesting this project as well as for many hours of extremely helpful conversations. I would like to thank the Marilyn Brachman Hoffman Committee of Wellesley College for a fellowship for released time during which much of this work was accomplished. I am indebted to IRCAM (Institut de Recherche et Coordination Acoustique/Musique) for their hospitality and the use of their facilities and to Wellesley College for a Sabbatical leave during which the writing took place. I would also like to thank Kenneth Malsky who created a soundfile for an experiment described in this paper.

Amuedo, J. (1985). "Periodicity Estimation by Hypothesis-Directed Search," ICASSP-IEEE International Conference on Acoustics, Speech, and Signal Processing, 395–398.

Brown, J. C., and Puckette, M. S. (1987). "Musical Information from a Narrowed Autocorrelation Function" Proceedings of the 1987 International Conference on Computer Music, Urbana, Illinois, 84–88.

Brown, J. C., and Puckette, M. S. (1989). "Calculation of a Narrowed Autocorrelation Function," J. Acoust. Soc. Am. **85**, 1595–1601.

Brown, J. C. (1989). "Musical Pitch Tracking Based on a Pattern Recognition Algorithm," J. Acoust. Soc. Am. Suppl. 1 **85**, S79.

Brown, J. C. (1991). "Calculation of a Constant Q Spectral Transform," J. Acoust. Soc. Am. **89**, 425–434.

Brown, J. C., and Zhang, B. (1991). "Musical Frequency Using the Methods of Conventional and 'Narrowed' Autocorrelation," J. Acoust. Soc. Am. **89**, 2346–2354.

Chafe, C., and Jaffe, D. (1986). "Source Separation and Note Identification in Polyphonic Music," Proc. ICASSP, Tokyo.

Duda, R. O., and Hart, P. E. (1973). *Pattern Classification and Scene Analysis* (Wiley, New York).

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.

Gerson, A., and Goldstein, J. L. (1978). "Evidence for a General Template in Central Optimal Processing for Pitch of Complex Tones," J. Acoust. Soc. Am. **63**, 498–510.

Piszczalski, M., and Galler, B. F. (1979). "Predicting Musical Pitch from Component Frequency Ratios," J. Acoust. Soc. Am. **66**, 710–720.

Scheffers, M. T. M. (1983). "Simulation of Auditory Analysis of Pitch: An Elaboration on the DWS Pitch Meter," J. Acoust. Soc. Am. **74**, 1716–1725.

Schoeder, M. R. (1968). "Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurements," J. Acoust. Soc. Am. **43**, 829–834.

Terhardt, E. (1979). "Calculating Virtual Pitch," Hear. Res. **1**, 155–182.

Terhardt, E., Stoll, G., and Swann, M. (1982). "Algorithms for Extraction of Pitch and Pitch Saliance from Complex Tonal Signals," J. Acoust. Soc. Am. **71**, 679–688.