Published as: Rossignol, S., Desain, P., and Honing, H. (2001) State-of-the-art in fundamental frequency tracking. *Proceedings of Workshop on Current Research Directions in Computer Music*, 244-254. Barcelona: UPF.

State-of-the-art in fundamental frequency tracking

Stéphane Rossignol, Peter Desain and Henkjan Honing

Music, Mind, Machine Group, NICI, University of Nijmegen, The Netherlands, http://www.nici.kun.nl/mmm

{S.Rossignol, desain, honing}@nici.kun.nl

Abstract

Pitch-tracking has been an important topic of research in speech and music research. Several methods have been proposed to obtain reliable f_0 trajectories from harmonic signals. The paper will review these. Some issues that are left are: how to evaluate and improve the quality and reliability of the pitch-tracking, and how to realize this in an automated method that can be use reliably and systematically on large data sets. In order to address these issues, we will focus on an approach that takes advantage of the availability of knowledge in trying to obtain more reliable and precise f_0 trajectories from monophonic and harmonic audio fragments. Two methods are compared that obtain reliable and precise f_0 trajectories from monophonic audio fragments. These trajectories can be used for the analysis and modeling of vibrato (frequency modulation) in music performance. The pitch extraction methods take advantage of the fact that the score, the timing (the performers synchronized with a piano accompaniment), the instrument and sometimes even the fingering is known.

1. Introduction

1.1 Fundamental frequency extraction

Robust systems that retrieve pitch information from musical performances are still hard to design. A very large number of methods have been developed (see for instance Hess 1983). We can classify pitch trackers into five general categories: autocorrelation, adaptive filter, time domain, frequency domain and models of the human ears (see Roads 1996).

Consider firstly the autocorrelation algorithms (Moorer 1975). These methods are most efficient at mid to low frequencies. In musical applications, the pitch range is broader.

Considering the adaptive filter methods (see Lane 1990), on pitch detector is based on the analysis of the difference between the filter output and the filter input. This difference must be close to zero. The band-pass filter center frequency is controlled by this difference. Another adaptive filter is based on the optimum comb method (Lane 1990). The goal is to minimize the output signal.

Considering the time domain methods, one type of pitch detector is based of the analysis of the zero-crossing points (Moorer 1975, Hermes 1992). Preprocessing by filters has to be performed, in order to solve the problem of the low-amplitude zero-crossings caused by high-frequency components.

A few pitch detectors exist in the frequency domain. Most of them are based on the analysis of the FFT spectrum, or of the cepstrum (Schafer and Rabiner 1970).

The methods based on auditory models combine frequency and temporal methods.

Most of the efforts have taken place in the frequency domain (see Brown 1992). In this paper we present methods working in the temporal and in the frequency domains (the FFT in the frequency domain; the Analytic Signal and the Teager-Kaiser methods in the temporal domain). It is shown that the frequency domain method is more efficient.

In order to obtain precise frequency trajectories, we must use local strategies, that is to say we have to use relatively short frames length. However, using the FFT spectrum, we must use frames which length have to be at least three times the period of the signal we want to detect. For a sine with a frequency of 440 Hz the frames length must be around 7 ms. That is to say, if the sampling rate is 11 kHz, 75 samples. Some alternatives to the FFT have been proposed in the literature. One of them is based on the Analytic Signal (Hess, 1983; Boashash, 1992; Wang, 1994). Another one is based on the Teager-Kaiser energy algorithm (Maragos, 1993; Vakman, 1996). For the first one only two samples are needed to estimate the instantaneous frequency and the instantaneous amplitude of a signal. For the second one, four samples are needed. But, for both of them, the signal is assumed a pure sine, which frequency and amplitude vary slowly in time. As the musical sounds in use are composed (i.e. composed of a sum a

harmonic sines) sounds, it is necessary to isolate each harmonic by band-pass filtering.
In our case, the score of the music is known and can be used as a guide. The pitch trackers described in this article are therefore referred to as knowledge-based pitch trackers.
A pitch tracker using knowledge is described in Scheirer (1995). One of the goals of the work presented in this article is to solve the problem of the transcription of polyphonic sounds. It is a score-aided transcription system. A comb-filter strategy, that is to say a not local strategy, is used. In this article, Scheirer says: "It seems on the surface that using the score to aid transcription is "cheating", or worse, useless - what good is it to build a system which extracts information you already know?". In our case, as the amplitude of the frequency modulation is assumed to be great, the score does not follow the frequency trajectory. The score-based pitch tracking is very useful to solve our specific problem.

1.2 Current Research

Pitch-tracking has been an important topic of research in speech and music research. Several methods have been proposed to obtain reliable f_0-trajectories from harmonic signals. The paper will review these. Some issues that are left are: how to evaluate and improve the quality and reliability of the pitch-tracking, and how to realize this in an automated method that can be use reliably and systematically on large data sets.
To address these issues, we will focus on an approach that takes advantage of the availability of knowledge in trying to obtain more reliable and precise f_0-trajectories from monophonic and harmonic audio fragments. It is a hard problem, especially, for instance, when sympathetic resonance of open strings in string instrument interfere with some harmonics of the main sound, or when transitions are so fast that tracks of different harmonics are connected. We will show that knowledge about the instrument and music played can be used to improve the results of the presented methods.
These methods are developed in the context of a larger project on the analysis and modeling of vibrato in music performance (Desain and Honing 1996; Timmers and Desain 2000). In order to model the vibrato during notes and in note transitions accurate f_0-trajectories are needed. For this a large systematic set of music performances was collected (see section 1.3). The setup of the data collection provides two kinds of knowledge. Firstly, "score" information is used such as pitch information and the predicted onset times, using the known tempo, is used (the latter makes it different from a score, hence the inverted comma's). f_0^s is used for instance to fit the length of the frames used for the band-pass filtering and for the f_n extraction (see Figures 2 and 4). During the data fusion stage, knowledge about the instrument can be used, like its spectral characteristics. Since sometimes a frequency trajectory is too noisy to be used, caused by, for example, a missing harmonic (e.g., in wind instruments) or sympathetic resonance (e.g., in string instruments).
We will examine here two alternative pitch extraction methods. Both are made-up of three stages. In the first stage, for both methods, the audio signal is fed through a band-pass filter bank. For each of the first N harmonics a time-varying band-pass filter is used which adjusts its length and central frequency according to the frequency information in the score, f_0^s. Information from the instrument is used to adjust the bandwidth to the pitch and to the speed of transitions. Thus, each harmonic is isolated, and N new sounds signals are obtained. The two following stages are not the same for the two methods. Considering the first method, in the second stage the frequency and energy trajectories are computed for each harmonic (peak tracking), using the signals obtained in the previous stage. In the final stage the f_i and amplitude trajectories obtained are merged to provide the optimal f_0 trajectory. Considering the other method, in the second stage, portions of the spectrum, centered on the frequency given by the score, are merged. In the third stage, the peak tracking is performed. During the data fusion stage, for both methods, instrument information is used to decide on the correct interpretation in situations where a higher harmonic is known to be a louder or more reliable source of f_0 information than the fundamental itself, or where the tracks of certain harmonics of certain fundamental frequencies are known to be distorted by sympathetic resonance. For the second method automatic techniques to detect the bad tracks have been implemented.
Next, we will describe the dataset that was used in the analyses, followed by the two pitch extraction methods (sections 2 and 3), completed by an evaluation and discussion of the results obtained (sections 4 and 5).

1.3 Data set of music performances

The dataset used in this paper consists of a large and systematically collected set of music performances of a single fragment of music performances by a variety of instruments. The fragment consists of the twenty first notes of "The Swan" of C. Saint-Saëns, performed along with a MIDI-controlled grand piano. This was used to control for the desired tempo, and as such allows for studying, for example, how vibrato is adapted to note duration. Seven instruments (cello,

Published as: Rossignol, S., Desain, P., and Honing, H. (2001) State-of-the-art in fundamental frequency tracking. *Proceedings of Workshop on Current Research Directions in Computer Music*, 244-254. Barcelona: UPF.

oboe, tenor, theremin, violin, soprano, and shakuhachi) played the melody in ten different tempos (54.5, 55.8, 57.1, 58.5, 60.0, 61.5, 63.2, 64.9, 66.7 and 68.8 beats per minute). And each performance was repeated six times to be able to check for consistency in performance. All this results in 420 recordings of which the f_0 trajectories had to be obtained. See Desain, Honing, Aarts and Timmers (2000) for more details.

An example is given Figure 1. The spectrogram, the score information in use (melody contours in straight lines) and the obtained frequency trajectories are shown.

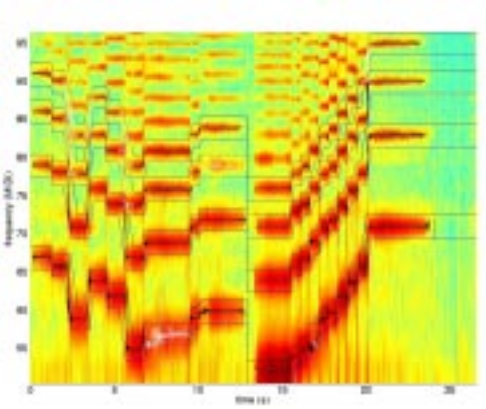


Figure 1: spectrogram, selection of bands using score information (melody contour in straight lines) and frequency trajectories obtained therein, for the cello (54.5 bpm)

2. Pitch-tracker A (fusion after peak detection)

2.1 Architecture

The analysis of f_0 from audio signals is composed of three stages. Firstly, the original audio signal is band-pass filtered. Thus, each harmonic is isolated (section 2.2), and N new sounds are obtained. Secondly, the frequency and the energy trajectories are computed for each harmonic, using the signals obtained at the previous stage of the analysis (section 2.3). Three methods to obtain these trajectories have been tested, with FFT as the preferred method. Thirdly, the f_n and A_n trajectories are mixed in order to provide the optimal f_0 trajectory (section 2.4).

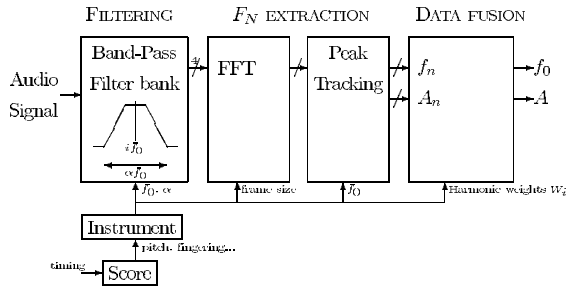


Figure 2. Architecture of pitch-tracker with fusion after peak detection

2.2 Filtering (phase 1)

In the first phase the appropriate harmonic needs to be selected. This is input for the f_N extraction phase. After this time-varying band-pass filtering, the amplitude of the harmonic we want to keep must be higher than the amplitude of all the other harmonics. Furthermore, the isolated harmonics are used for checking the quality and appropriate selection controlled by the score information (see section 4.3).

2.3 f_n extraction (phase 2)

Three harmonic trackers have been tested. The input signal considered for each of them is the sound obtained after the band-pass filtering. The results obtained with each of them for a simulated signal and for a true sound signal are shown in section 4. It is shown there why the last two methods have been rejected. The first method is based on the FFT spectrum (FFT method); the second one is based on the Analytic Signal (AS method): for more complete theoretical developments, examine the references Hess 1983, Boashash 1992, and Wang 1994; and the third one is based on the Teager-Kaiser energy algorithm (TK method): Maragos 1993 and Vakman 1996. These three algorithms are shown Figure 3.

The FFT method is a "frequency" domain strategy; and AS and TK methods are "temporal" domain strategies.

The results obtained with each of them for simulated signals are shown and compared in section 4.1.

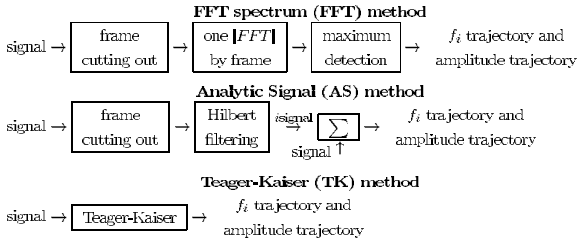


Figure 3: The three alternative harmonic trackers

Published as: Rossignol, S., Desain, P., and Honing, H. (2001) State-of-the-art in fundamental frequency tracking. *Proceedings of Workshop on Current Research Directions in Computer Music*, 244-254. Barcelona: UPF.

For the FFT method, the f_0^s knowledge is used to determine the length of the frames, which is equal to $M\, f_e/f_0^s$ samples (with $M \in [3\ 10]$). As the analysed sounds are cut into frames, this method is considered a "global strategy". But, as the length of the frames changes with f_0^s and as such provides us with an optimal size, knowledge allows us to improve the results.

For the AS method, the f_0^s knowledge is used to determine the length of the frames, which is equal to $M\, f_e/f_0^s$ samples (with $M \in [3\ 10]$). Due to the Hilbert filtering, we say that this method is "global". But to compute the "instantaneous frequency" only two complex samples are needed.

Considering the TK method, the instantaneous frequency is estimated as:

$$F = \arccos\left(1 - \frac{P[x(n) - x(n-1)]}{2P[x(n)]}\right) \quad (1)$$

where:

$$P[y(m)] = y^2(m) - y(m-1)y(m+1) \quad (2)$$

is the Teager-Kaiser operator; and where x are the sound samples.

A similar formula is available in order to estimate the instantaneous amplitude:

$$A = \sqrt{\frac{P[x(n)]}{1-\cos^2(F)}} \quad (3)$$

Knowledge is not used here.

As only four consecutive sound samples are needed to obtain an estimate of the frequency and of the amplitude, the TK method is considered a "local strategy". It is assumed that "the amplitude and the frequency do not vary too fast (time rate of change of value) or too greatly (range of value) in time compared to the carrier frequency" (Maragos and Kaiser 1993). These two conditions are related to the vibrato: the first one, to its frequency f_v (or vibrato rate), and the second one to its amplitude A_v (or vibrato extent). And the transitions have to be also relatively smooth.

2.4 Data fusion (phase 3)

The definition used is:

$$\hat{f}_0 = \frac{1}{N} \frac{1}{\sum_{i=1}^{N} A_i} \sum_{i=1}^{N} \frac{A_i\ f_i}{i} \quad (4)$$

where N is the number of harmonics taken into account, f_i is the frequency found for the i^{th} harmonic, and A_i is the amplitude of the i^{th} harmonic.

It can be noticed that for this first method, no information is coming from the box "Instrument" (see Figure 2).

A more refined method can be used:

$$\hat{f}_0 = \frac{1}{N} \frac{1}{\sum_{i=1}^{N} W_i\ A_i} \sum_{i=1}^{N} \frac{W_i\ A_i\ f_i}{s_i} \quad (5)$$

where W_i are the weights (information coming from the box "Instrument"). And where the parameters s_i describe the fact that for some instruments (e.g. string instruments) the harmonic are a little bit shifted in frequency. At the moment, the weights W_i are predefined. Some automatic methods have been studied. They are based on the results of a rating experiment in which listeners compared original and re-synthesized sound signals (see section 3.3 and 4.3).

3 Pitch-tracker B (fusion before peak detection)

3.1 Architecture

For the alternative pitch-tracker the analysis is also composed of three stages. The first stage is the same for both pitch-trackers. Secondly (section 3.2) portions of spectrum are extracted. Thirdly, these portions are merged, and the peak tracking performed. The weights *Wi* described in the section 2.4 can be taken into account. They weight the amplitude of the extracted portions of spectrums. In the other hand, a technique to automatically detect the bad notes has been implemented. Thus, the data fusion stage is completed by the automatic detection of bad tracks (section 3.3).

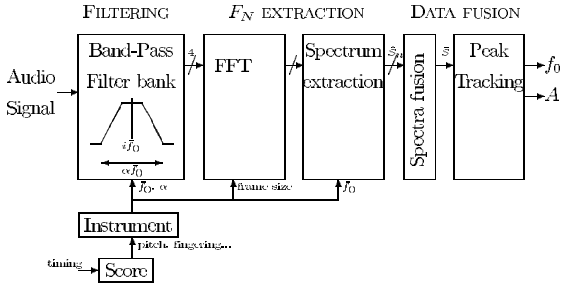


Figure 4: The whole system when the fusion is performed before the peak detection.

3.2 Spectra extraction (phase 2)

In the second stage transposed portions of the spectrum are combined. These portions correspond respectively to these frequency bands:

$$[f_0 - \Delta f \quad f_0 + \Delta f] \quad (6)$$

for the first harmonic, and

$$[2f_0 - 2\Delta f \quad 2f_0 + 2\Delta f] \quad (7)$$

for the second harmonic, etc. The bounds of each band correspond to the information given in the score. The bandwidth increases with the number of the

Published as: Rossignol, S., Desain, P., and Honing, H. (2001) State-of-the-art in fundamental frequency tracking. *Proceedings of Workshop on Current Research Directions in Computer Music*, 244-254. Barcelona: UPF.

considered harmonic. So, the width of the main lobe decreases with the number of the harmonic.

3.3 Data fusion with automatic detection of bad tracks (phase 3)

In the pitch-tracker discussed above weights were used (that had to be explicitly provided) to improve the quality of the data fusion. In this pitch tracker we incorporate an automatic method to rate the quality of the f_n's.

We analyze here the extracted portions of the spectrums, frame by frame. We inspect three measures.

The first one, M_1, is the ratio between the portion of energy around the maximum of the spectrum ($[f_{max} - \delta f \quad f_{max} + \delta f]$) and its whole energy. This portion is expected to be great when the analysed signal is a pure sine and when the score information is relevant.

The second one M_2 is the rate of change in the position of this maximum for two successive frames. When something disrupts the partial tracker, this rate is expected to be great.

The third one M_3 is the correlation between the spectrum around is maximum and its theoretical shape if the analyzed signal was a pure sine, witch constant amplitude and frequency.

The final measure of bad tracks detection is thus:

$$M_i = a_1 M_1 + a_2 (1- M_2) + a_3 M_3 \qquad (8)$$

The parameters a_1, a_2, a_3 and δf have to be optimized. This variable M is used instead of the W_i. Therefore, a value is obtained for each frame. It is not the case when the weights W_i are considered, which are defined note by note.

4. Results

Firstly, the performance of the three harmonics trackers is discussed. Secondly, the two whole pitch tracker systems are compared. Thirdly, the performance of the technique to automatically detect the bad tracks is analyzed. And fourthly, the performance of the whole system, using the second pitch-tracker, is shown.

4.1 Performance of the three harmonic trackers

4.1.1 Introduction

Four characteristics of the signal complicate the harmonic tracking. The first one is the vibrato (frequency and amplitude); the second one are transitions; the third one are neighboring harmonics; and the last one is the additive noise. The three methods do not behave in the same way at all. We showed these differences considering a simulated signal.

Three tests on a simulated signal have been performed. The parameters for this signal are equal to: fundamental frequency of the first note $f_0^{(a)} = 440\ Hz$, fundamental frequency of the second note $f_0^{(b)} = 493\ Hz$, transition moment $t_a = 0.71\ s$, transition speed $t_r = 0.003$, magnitude of the vibrato $A_v = 30\ Hz$, frequency of the vibrato $f_v = 5\ Hz$ and phase of the vibrato φ_v = 1.6 radian.

The transition is modeled as a hyperbolic tangent. Thus, the used model of the fundamental frequency trajectory (without vibrato) is:

$$f_0(t) = f_0^{(a)} + [\tanh(\frac{t - t_a}{t_r}) + 1]\frac{f_0^{(b)} - f_0^{(a)}}{2} \qquad (9)$$

So, finally, the signal model in use is: $s = \cos(\Phi_1 + \Phi^{(a)})$ with:

$$\Phi_1 = 2\Pi\{f_0^{(a)} t + ct + \frac{A_v}{f_v}\sin(2\Pi f_v t + \Phi_v) + t_r t[\log_e(\cosh(\frac{t - t_a}{t_r})) - \log_e(\cosh(-\frac{t_a}{t_r}))]\} \qquad (10)$$

where t is the time (in second), $c = \frac{f_0^{(b)} - f_0^{(a)}}{2}$, and $\Phi^{(a)}$ the phase at $t=0$.

It can be noticed that, for these tests, the disruptive parameters 1 and 2 concern the time rate of change of value and the range of value (see section 2.3). The length of the frames is constant. It has been chosen equal to *7 ms*, which is close to $3 f_e / f_0^s$ for the smallest fundamental frequency, $f_0^{(a)}$.

4.1.2 Behavior on sine signal with a transition and vibrato

In Figure 5 are shown the f_0 trajectories obtained for the whole sound. In Figure 6 are shown the f_0 trajectories during the transition. In both cases, four f_0 trajectories are plotted: the ideal f_0 trajectory, the f_0 trajectories obtained using the FFT method, the AS method and the TK method. It can be seen that the three harmonic trackers can follow the variation of the frequency well. However, the TK method shows some artefacts during the transition.

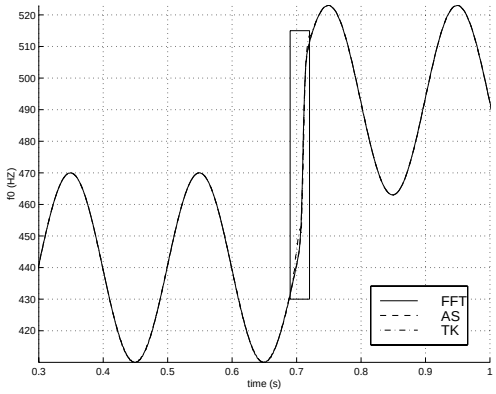


Figure 5: Results obtained with the three methods (frame length 7 ms)

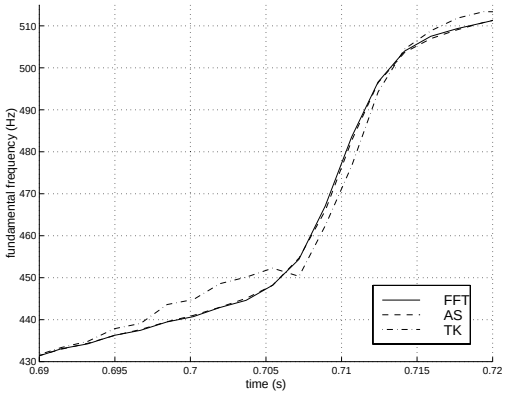


Figure 6: Close up of the transition shown in Figure 5

In Figure 7, are shown the results obtained when the length of the frames is fixed to 25 ms. This value is the value commonly used by the pitch trackers which do not use knowledge (see Brown and Puckette 1993). It can be demonstrated that, in the transition, the FFT method is less efficient when using a larger frame length.

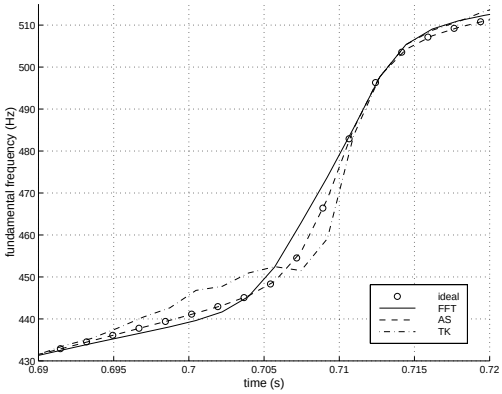


Figure 7: Results obtained with the three methods (frames length 25 ms)

4.1.3 Behavior with non-pure sine signals

In this case, the simulated signal is equal to:

$$s = \cos(\Phi_1 + \Phi^{(1)}) + \sum_{i=2}^{4} a_i \cos(i\Phi_1 + \Phi^{(i)}) \qquad (11)$$

It means that the higher harmonics are not completely removed. Their amplitudes are indicated by the parameter a_i. The results are shown in Figure 8.

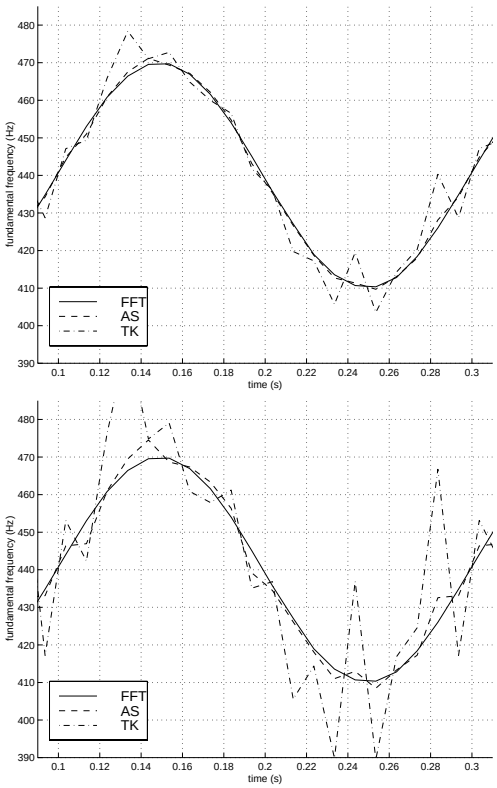


Figure 8: Behavior with non-pure sine signals. *$a_2=a_3=a_4=0.001$ (top), $a_2=a_3=a_4=0.003$ (bottom)*

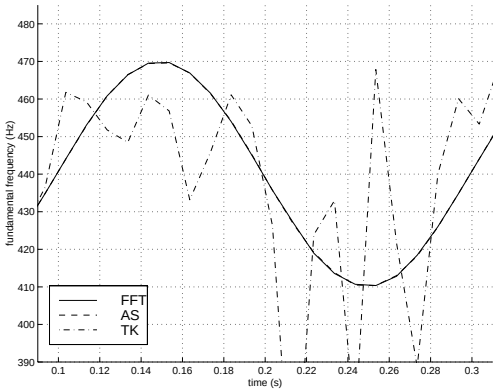
It can be seen that when the other harmonics are not removed well, the behavior of AS and TK methods is disturbed.

4.1.4 Behavior on noisy signals

In this case, the simulated signal is equal to:

$$s = \cos(\Phi_1 + \Phi^{(1)}) + b \qquad (12)$$

where b is a normal noise, with mean equal to 0 and standard deviation equal σ. The results are shown in Figure 9.



Published as: Rossignol, S., Desain, P., and Honing, H. (2001) State-of-the-art in fundamental frequency tracking. *Proceedings of Workshop on Current Research Directions in Computer Music*, 244-254. Barcelona: UPF.

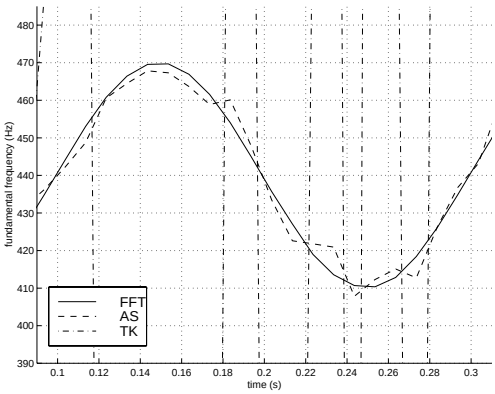


Figure 9: Behavior on noisy signals. $\sigma = 1e-5$, $\sigma = 3e-4$

Where there is noise, the AS and TK methods do not perform well.

4.1.5 Behavior of the FFT method

The goal is to show that the band-pass filtering is also necessary for the FFT based method.

We have to notice that the speed of a given transition increases with the number of the harmonic. For instance, let us consider two consecutive notes which fundamental frequencies are respectively *440 Hz* and *554.36 Hz*, and which are connected by a 50 milliseconds transition. For the first harmonic, during these 50 milliseconds, the jump in frequency is about *114 Hz*; and for the fourth harmonic, it is *457 Hz*. It is shown in Figures 6 and 7 that to adapt the length of the frames to the frequency allows the FFT based method to follow efficiently the frequency during the transitions.

These results are shown in Figures 10 and 11. The signal used is a simulated one. The model is described in the section 4.1.1. The amplitude of each harmonic is 1. The sound lasts 1 second. And we have: t_a = *0.5*, t_r = *0.012*, $\Phi_v = 0.9rad$, $f_0^{(a)}$ = *440 Hz* and $f_0^{(b)}$ = *554.36 Hz*.

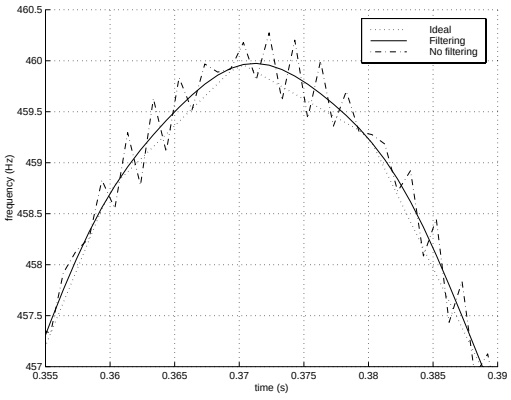


Figure 10: zoom of the f_0 trajectory (true trajectory, trajectory obtained with the pitch-tracker A, trajectory when the band-pass filtering is not performed)

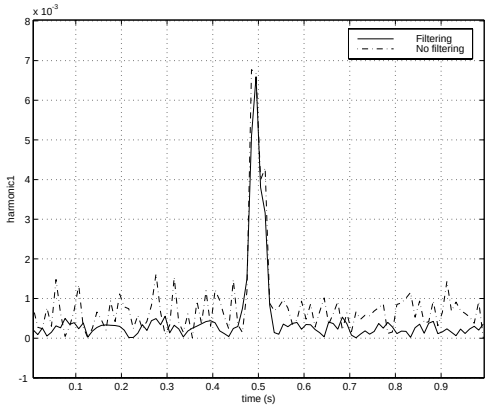
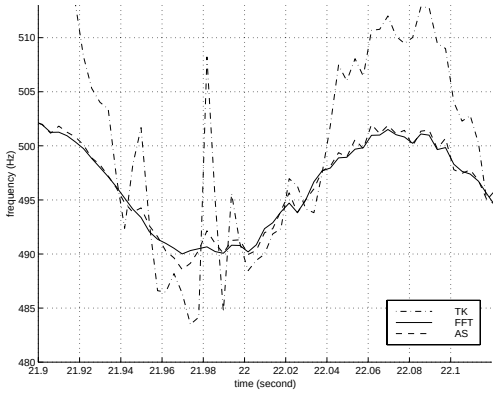


Figure 11: Relative difference between the true f_0 trajectory and the trajectory obtained with the pitch-tracker A and the trajectory when the band-pass filtering is not performed

4.1.6 Behavior of the three harmonic trackers on true sound signal

The top panel of Figure 12, shows the f_0 trajectories obtained for the first harmonic of the last note of the cello. As expected, the trajectory obtained with the AS is noisier than the result of FFT method, and the trajectory of the TK method even more. This is due to the fact that the analysed signal is not a pure sine (see the spectra shown in the bottom panel of Figure 12). After the band-pass filtering, the amplitude of the higher harmonics are respectively *[7.0 8.6e-3 7.2e-3 1.2e-2]*. The AS and TK methods need signals composed of a very dominant sine.



Published as: Rossignol, S., Desain, P., and Honing, H. (2001) State-of-the-art in fundamental frequency tracking. *Proceedings of Workshop on Current Research Directions in Computer Music*, 244-254. Barcelona: UPF.

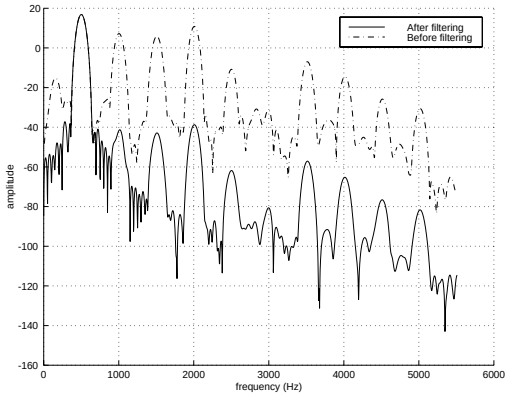


Figure12. Top panel: f_0 trajectories obtained with the three harmonic trackers. Cello (54.5 bpm, last note, first harmonic). Bottom panel: spectra of a frame of the original signal *[21.9s 21.92s]* and of the corresponding band-pass filtered signal

4.1.7 Discussion

A very efficient band-pass filtering stage is absolutely necessary for the AS and TK methods. For these two methods, the signal given to the harmonic trackers must be a pure sine with slowly varying amplitude and frequency. The FFT method seems to be the best, as the use of knowledge allows us to improve its performance. We decided therefore to use this method in our system.

4.2 Comparison of the two pitch trackers

4.2.1 Simulated signal

Here we give some results obtained for the evaluation of the two pitch tracker methods. The difference between these two methods concerns mainly the data fusion stage.

We use a simulated signal, composed of an harmonic component and of a disruptive component.

For the harmonic component, the fundamental frequency is constant (it means that there is only one note): $f_0 = 440\ Hz$; there is a vibrato: $f_v = 5\ Hz$, $Av = 20\ Hz$; and the amplitude of each harmonic is *1/15.*

The disruptive component is composed of an additional partial, which is close in frequency from the third harmonic.

For the disruptive partial, the frequency is $3f_0 + 150\ Hz$; the amplitude is *1.5/15* (notice that the amplitude of the disruptive partial is higher that the amplitude of the third harmonic) and there is a vibrato: $f_v = 4.9\ Hz$, $A_v = 29.4\ Hz$ (it is different of the vibrato presents on the harmonics).

Figure 13 shows the fundamental frequency trajectories for pitch-tracker A and B.

Here, clearly, the trajectory obtained using the alternative fusion method (i.e. pitch tracker B) is closer from the true trajectory than the one obtained using the first fusion method (i.e. pitch tracker A).

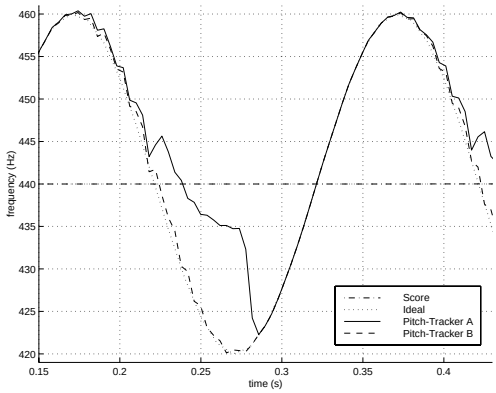


Figure 13: f_0 trajectories obtained with the two pitch trackers on a simulated signal

Next, we will look at the method of combining spectra method uses in pitch-tracker B on the simulated signal. This is illustrated in Figure 14. In this Figure, the spectra of each harmonic of the simulated signal are shown (indicated by S1, S2, S3 and S4).

It can be seen that the maximum for the third (labeled 3) spectrum does not occur in the same place (≈ 30) that for the 3 other spectrums (≈ 15).

We have also a trajectory labeled *mean*. It is the result of the fusion the four previous spectrums. It can be seen that the position of the maximum of this red spectrum is around 15. So, it is well positioned.

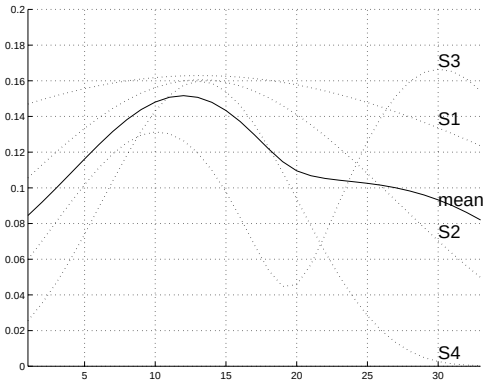


Figure 14: The spectra of the four harmonics in the simulated signal; and the combined spectra (X-axis: $\approx$ frequency (not in *Hz*); Y-axis: linear amplitude)

The second pitch-tracker is more robust to mistakes on W_i than the first one. Figure 14 illustrates that indeed, in the case of the presence of a more noisy harmonic, taking the average spectrum is a more realiable method.

Published as: Rossignol, S., Desain, P., and Honing, H. (2001) State-of-the-art in fundamental frequency tracking. *Proceedings of Workshop on Current Research Directions in Computer Music*, 244-254. Barcelona: UPF.

4.2.2 Instrumental sound

We will now demonstrate the workings of the two pitch trackers using a realistic example: a note of the cello (57 MIDI), for which we have a string resonance which disrupts the first harmonic (see Figure 1, between 6 and 10 seconds). For this sound, we obtain the results shown in Figure 15. The pitch tracker B is more robust.

In Figure 16, are given the f_0 trajectories obtained with the two pitch trackers when the weights W_i are taken into account (W_1=0). The results are very similar.

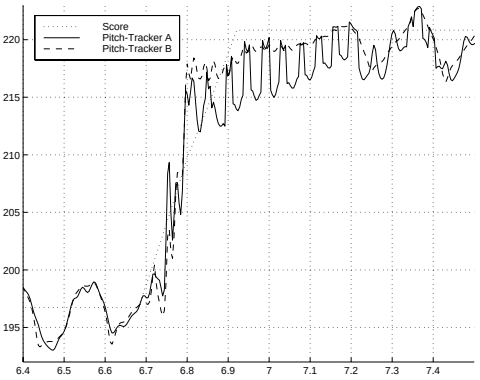


Figure 15: f_0 trajectories obtained for the beginning of a note of cello, for which there is a disruptive sympathetic resonance; the weights are not taken into account

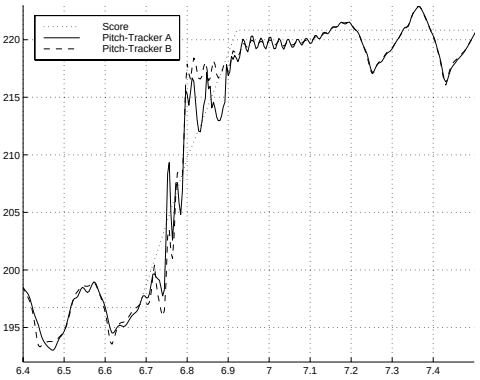
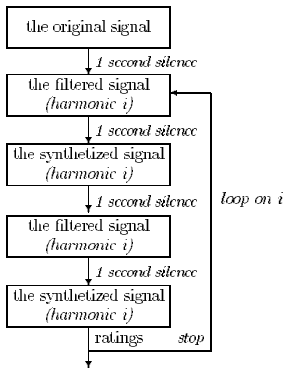


Figure 16: f_0 trajectories obtained for the beginning of a note of cello, for which there is a disruptive sympathetic resonance; the weights are taken into account

4.3 Evaluation of the automatic detection method

We conducted an experiment to get a better insight in quality and the relevancy of the processing by having participants listening to the filtered sounds and the resynthetised sounds obtained with pitch tracker A. The results of this experiment were used to improve and validate the automatic detection method used in pitch tracker B.

For this we use the dataset described in section 1.3, using a single performance of each instrument at tempo 60 BPM. Participants judged for each note in the selected fragment the filtered signal (i.e. the first four harmonics) and the signal resynthesised with the resulting f_n trajectories. First the original signal was presented, followed by four pairs of the filtered and synthesized harmonics, every time judging the *similarity* between the filtered and synthesized signal, and the *consistency* of the synthesized signal.



The goal of the similarity rating is to check the quality of the harmonic tracker. The filtered and the synthesized signals have to be similar (and they have to be harmonic). If they are different, it means that something is went wrong with extracting frequency trajectory (for example, caused by the noise in the signal is noisy, a resonance, etc.).

The goal of the consistency rating is to check if we can use the frequency trajectory during the data fusion stage. We cannot use the frequency trajectory if the note is not consistent, that is to say if more than one note is perceived. For instance caused by a jump in frequency in selecting two competing peaks.

Participants rated similarity on a three point scale, with 0 indicating that both the filtered and synthesized signal are different and 2 indicating that they are similar. They rated consistency on a two point scale, with 0 indicating that the signal is not consistent, and 1 that it was perceived as one note.

The rating were combined to a final measure as $r = r1/2*r2$.

Seven subjects participated to this experiment. The mean was computer over these seven subjects, and compared to the results obtained with the automatic method described in the section 3.2.

For the cello, the correlations between the mean within the subjects, and each subject are: [0.87 0.86 0.75 0.79 0.85 0.66 0.71]. The mean of these correlations is 0.78.

The correlation between the mean within the subjects, and the results obtained with the automatic method is 0.5.

4.4 Example of the full system

Finally, we will show an example of preferred method, pitch tracker B, the full system in operation. For this we return again to the example presented in the introduction (see Figure 1). In Figure 17, the spectrogram, the score information and the obtained frequency trajectories are plotted. In Figure 17, only the first harmonic is shown. In Figure 1, the first four harmonics are shown. The dotted score lines indicate that the corresponding harmonic is not taken into account. The corresponding portions of the frequency trajectories are plotted in white. The amplitude trajectory information is also taken into account. When the amplitude is too small, the corresponding portions of the frequency trajectories are not shown (see for instance the end of each harmonic, after 24 seconds).

As an example, for the long note at 57 MIDI pitch, we can see in Figure 1 that there is a resonance at the beginning of this note. So, the harmonic tracker fails for this part of the signal, as it can be seen. But, after the data fusion stage, the f_0 trajectory shown in Figure 17 is obtained. If we compare this trajectory to the frequency trajectory obtained for the first harmonic (Figure 1), the results have been clearly improved.

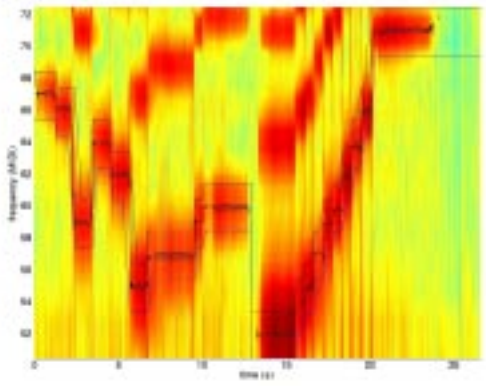


Figure 17: f_0 trajectory obtained for the cello (54.5 bpm); fusion before peak detection

5 Conclusion and prospects

In this paper, two efficient f_0 trackers, which use knowledge, have been presented and compared. Our future goal will be to provide models of the vibrato useful for music synthesis and composition. The first step of the analysis was to obtain "good" f_0 trajectories. The obtained f_0 trajectories can be used to analyse vibrato and portamento.

While our primary motivation of developing this knowledge-based method is to obtain precise f_0 information from the experimental data set, the idea to use knowledge in f_0 tracking can be useful for other computer music systems as well. For instance, when f_0 needs to be tracked in a live situation where score and timing information is available. The methods described in this paper can in principle be used for an efficient f_0 tracker that considers only those parts of the audio signal of the singer or instrumentalist to be followed that are relevant for f_0 tracking.

A measure of the voicing coefficient is also obtained. It allows us to detect, for instance, silences, noisy state part (noise component predominant over harmonic component), but also to check the quality of our processing.

An interesting extension of our pitch-trackers would be to use it to analyze polyphonic sounds. For example, when two harmonics, coming from two different instruments or voices, are to close, the corresponding trajectories (see Figure 2) or spectrum (see Figure 4) would not be used during the fusion stage.

6 Acknowledgements

This research is supported by the MOSART (Music Orchestration Systems in Algorithmic Research and Technology) European project, and by the Netherlands Organization for Scientific Research (NWO).

References

Boashash, Boualem, "Estimating and interpreting the instantaneous frequency of a signal", Proceedings of the IEEE, Volume 80, no. 4, pp. 539-568, 1992, April. IEEE.

Brown, Judith C. and Puckette, Miller S., "A high resolution fundamental frequency determination based on phase changes of the Fourier transform", Journal of the Acoustical Society of America, 1993, volume 94, no. 2, pp. 662-667

Brown, Judith C., "Musical fundamental frequency tracking using a pattern recognition method", Journal of the Acoustical Society of America, 1992, volume 92, no. 3, pages 1394 – 1402

Desain, P., Honing, H., Aarts, R. and Timmers, R., "Rhythmic Aspects of Vibrato (In P. Desain and W.

L. Windsor, Rhythm Perception and Production,) Swets & Zeitlinger, 2000, pp. 203-216

Desain, P. and Honing, H., "Modeling Continuous Aspects of Music Performance: Vibrato and Portamento", Proceedings of the International Music Perception and Cognition Conference, B. Pennycook & E. Costa-Giomi, CD-ROM, 1996

Hermes, D., "Pitch analysis", In M. Cooke and S. Beet, eds. *Visual Representation of Speech Signals*, New York, John Wiley and Sons, 1992

Hess, Wolfgang, "Pitch determination of speech signals", Springer-Verlag, 1983

Lane, J., "Pitch detection using a tunable IIR filter", Computer Music Journal, Volume 14, no. 3, pp. 46-59

Maragos, Petros and Kaiser, James K., "Energy separation in signal modulations with application to speech analysis", IEEE Transaction on Signal Processing, Volume 41, no. 10, pp. 3024-3050, 1993, October

Moorer, J. A., "On the segmentation and analysis of continuous musical sound", Ph-D thesis, Stanford University, Department of Music, 1975

Prame, Eric, "Measurements of the vibrato rate of ten singers", Journal of the Acoustical Society of America, 1994, October, 1979 - 1984

Prame, Eric, "Vibrato extent and intonation in professional Western lyric singing", Journal of the Acoustical Society of America, 1997, July, 616 - 621

Roads, Curtis, "The computer music tutorial", The MIT Press, Cambridge, Massachusetts, London, England, 1996

Rossignol, Stéphane, Desain, Peter and Honing, Henkjan, "Refined knowledge-based f_0 tracking: Comparing three frequency extraction methods", International Computer Music Conference, 2001, September

Schafer, R., and Rabiner, L., "System for automatic formant analysis of voiced speech", Journal of the Acoustical Society of America, Volume 47, no. 2, pp. 634-644

Scheirer, Eric, "Using musical knowledge to extract expressive performance information from audio recordings", IJCAI - Workshop on Computational Auditory Scene Analysis, 1995, August

Timmers, Renee and Desain, Peter, "Vibrato: the questions and answers from musicians and science", Proceedings of the International Conference on Music Perception and Cognition, Keele University, Department of Psychology, CD-ROM, 2000

Vakman, David, "On the AS, the TK energy algorithm and other methods for defining amplitude and frequency", IEEE Transaction on Signal Processing, Volume 44, no. 4, pp. 791-797, 1996, April

Wang, Avery Li-Chun, "Instantaneous and frequency-warped signal processing techniques for auditory source separation", Ph. D. thesis, Stanford University, 1994, August