Refined Knowledge-Based f_0 Tracking: Comparing Three Frequency Extraction Methods

Stéphane Rossignol, Peter Desain and Henkjan Honing
Music Mind Machine Group, NICI, University of Nijmegen, The Netherlands
{S.Rossignol, desain, honing}@nici.kun.nl
www.nici.kun.nl/mmm

Abstract

This paper presents a new method to obtain reliable and precise f_0-trajectories from monophonic audio fragments that can be used for the analysis and modeling of vibrato *in music performance. The proposed f_0 tracking method takes advantage of the fact that the score, the performance timing, the instrument and sometimes even the fingering are known.*

1 Introduction

Obtaining accurate f_0 information from audio data is a hard problem, especially when, for example, sympathetic resonance of open strings in string instruments interfere with some harmonics of the main sound, or when transitions are so fast that tracks of different harmonics become indistinguishable. This paper presents an elegant method to obtain reliable and precise f_0-trajectories from monophonic audio fragments of harmonic sounds.

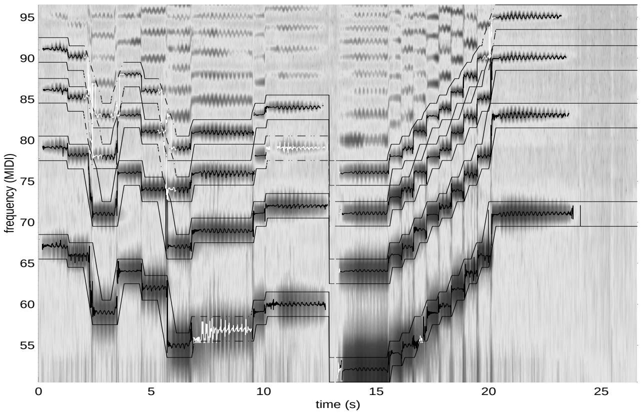

Figure 1: The spectrogram, the selection of bands using score information (melody contours in straight lines) and the frequency trajectories obtained therein for the cello (54.5 bpm)

The method is 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 were collected. The data set consists of seven instruments that performed the first phrase of "The Swan" of C. Saint-Saëns (see figure 2) along with a MIDI-controlled grand piano. These performances were each repeated six times (to check for consistency in performance) in ten different tempi (to get, e.g., an insight in how vibrato is adapted to note duration). See Desain, Honing, Aarts, and Timmers (2000) for more details.

An example is presented in figure 1. The spectrogram, the score information in use and the obtained frequency trajectories are shown. In figure 2, the corresponding score is given.

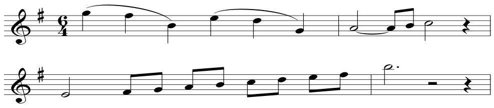

Figure 2: First phrase of "The Swan" of C. Saint-Saëns

Two kinds of knowledge are used. Firstly, "score" information is used, such as pitch information and predicted onset times. The performers synchronized with a piano accompaniment, such that onset times can be estimated. The fundamental frequency, f_0^s, is used to adjust the width and central frequency of the band-pass filter and for f_N extraction. When the frequency trajectory for each harmonic has been tracked, the information obtained is fusioned in order to obtain a final f_0-trajectory. During this data fusion step, a second kind of knowledge is used, which concerns characteristics of the considered instrument. Sometimes a frequency trajectory is too noisy to be usable during the fusion step. This can be due to the fact that the amplitude of the harmonic is low, or caused by a missing harmonic, etc.

Next, we will describe the full system, followed by a discussion of the results.

2 The f_0 tracker

2.1 Description of the complete system

The analysis of f_0 from the audio signals is executed in three steps. In the first step the audio signal is fed through a band-pass filter bank (see section 2.2). For each of the first N harmonics a time-varying band-pass filter is used which adjusts its length to f_0^s. Information from the instrument is used to adjust the bandwidth to the speed of transitions. Thus, each harmonic is isolated, and N new sound signals are obtained. In the second step, the frequency and the amplitude trajectories are computed for each harmonic, using the signals obtained at the previous step of the analysis (see section 2.3). Three alternatives have been tested, with FFT as the preferred method. In the final step the f_i and amplitude trajectories obtained are combined to provide the optimal f_0-trajectory (see section 2.4). Here the instrument information is used to decide on the correct interpretation in situations where a higher harmonic is known to be a louder or a 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.

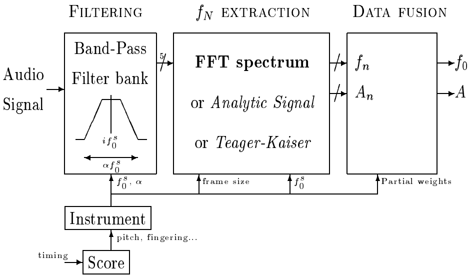

Figure 3: f_0 tracker

2.2 Filtering (phase 1)

The time-varying band-pass filtering selects the appropriate harmonic. This is input for the f_N extraction phase (see section 2.3). The isolated harmonics were also used in a listening session to check for the quality of the band-pass filtering stage. These informal tests indicated that it is not necessary to compute a new filter for every sound sample; to update the filter every $10ms$ is sufficient.

2.3 f_N extraction (phase 2)

In order to obtain precise frequency trajectories, local strategies must be used, that is to say we have to use relatively short frame lengths. 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 $440Hz$, the frames length must be around $7ms$. That is to say, if the sampling rate is $11kHz$, 75 samples. Some alternatives to the FFT have been proposed in the literature. One of them is based on the Analytic Signal (AS method): (Hess 1983; Boashash 1992; Wang 1994). Another one is based on the Teager-Kaiser energy algorithm (TK method): (Maragos and Kaiser 1993; Vakman 1996). The three algorithms are shown in figure 4.

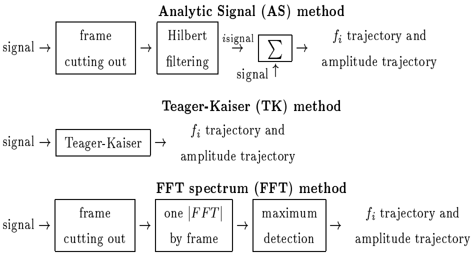

Figure 4: Three alternative harmonic trackers

For the AS method only two samples are needed to estimate the instantaneous frequency and amplitude of a signal. For the TK method, four samples are needed. But, for both of them, to perform efficiently the f_N extraction, the signal is assumed a pure sine which frequency and amplitude vary slowly in time. So, in our case, as the musical sounds in use are composed sounds (i.e. composed of a sum of harmonic sines), it is necessary to isolate each harmonic by band-pass filtering.

These 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 3 and 4, respectively. It is shown there why the first two methods have been rejected.

For the Analytic Signal (AS) method, the f_0^s knowledge is used to determine the length of the frames, which is equal to Mf_e/f_0^s samples. Due to the Hilbert filtering, we say that this method is "global". But to compute the "instantaneous frequency" only two complex samples are needed.

In the Teager-Kaiser energy algorithm (TK), the instantaneous frequency is estimated as:

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

where $P[y(m)] = y(m)^2 - y(m-1)y(m+1)$ is the TK energy operator; and where x are the sound samples. A similar formula is available in order to estimate the instantaneous amplitude. For more details about the TK method, see Maragos and Kaiser (1993). 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).

For the Fast Fourier spectrum (FFT) method, the f_0^s knowledge is used to determine the length of the frames, which is equal to Mf_e/f_0^s samples. 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.

2.4 Data fusion (phase 3)

The definition used is:

$$\hat{f_0} = \frac{1}{N} \frac{1}{\sum_{i=1}^{N} w_i A_i} \sum_{i=1}^{N} \frac{w_i A_i f_i}{i}$$

where N is the number of harmonics taken into account, f_i is the frequency found for the ith harmonic, A_i is the amplitude of the ith harmonic, and the weights w_i are information coming from the box "Instrument" (figure 3). At the moment, w_i are predefined. Automatic methods to determine these weights will be implemented in the future. They will be based on the results of a rating experiment in which listeners compared original and resynthetised sounds.

3 Performance of the three frequency trackers – On simulated sounds

Four characteristics of the signal complicate harmonic tracking. The first one is the vibrato (frequency and amplitude); the second one are transitions; the third one are neighbouring 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 have been performed. The parameters for this signal are equal to: fundamental frequency of the first note $f_0^a = 440Hz$, fundamental frequency of the second note $f_0^b = 493Hz$, transition moment $T_m = 0.71s$, transition speed $T_s = 0.003$, magnitude of the vibrato $A_v = 30Hz$, frequency of the vibrato $f_v = 5Hz$ and phase of the vibrato $\phi_v = 1.6rad$. The signal model in use is: $s = \cos(\phi_1)$ with:

$$\phi_1 = 2\pi f_0^a t + 2\pi c t + \frac{A_v}{f_v}\sin\left(2\pi f_v t + \phi_v\right) +$$
$$2\pi T_s t\left[\log_e\left(\cosh\left(\frac{t - T_m}{T_s}\right)\right) - \log_e\left(\cosh\left(-\frac{T_m}{T_s}\right)\right)\right]$$

where t is time in second, and $c = \left(f_0^b - f_0^a\right)/2$. The first two disruptive parameters 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 $7ms$, which is close to $3f_e/f_0^s$ for the smallest fundamental frequency, f_0^a.

3.1 Behaviour on sine signal with a transition and a vibrato

The left panel of figure 5 shows the f_0-trajectories obtained for the whole sound; in the right panel the f_0-trajectories during the transition are shown. In both cases, four f_0-trajectories are plotted: the ideal f_0-trajectory, and the f_0-trajectories obtained using the AS (dash-dot line), the TK (dotted line) and the FFT (dashed line) methods. 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. A frame length of $25ms$ is commonly used by f_0 trackers which do not use knowledge (cf. Brown and Puckette 1993). It can be demonstrated that the FFT method is less efficient when using a larger frame size.

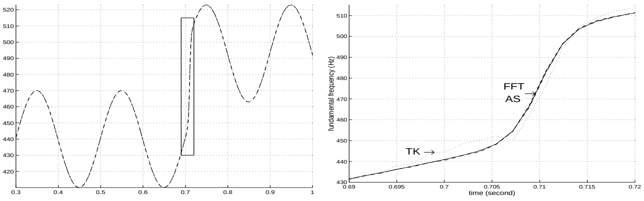

Figure 5: Results of the three methods (frame length $7ms$); TK (dotted line) ; AS (dash-dot line); FFT (dashed line). Right panel: close up

3.2 Behaviour with non-pure sine signals

In this case, the simulated signal is equal to:

$$s = \cos\left(\phi_1 + \varphi^1\right) + \sum_{i=2}^{4} a_i \cos\left(i\phi_1 + \varphi^i\right)$$

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

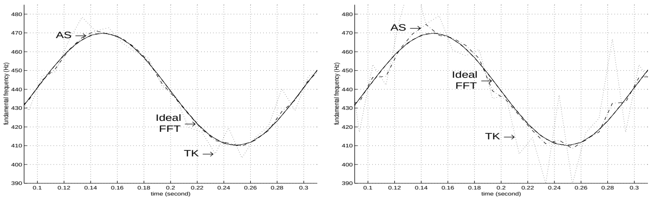

Figure 6: Behaviour with non-pure sine signals: $a_2 = a_3 = a_4 = 0.001$ (left), $a_2 = a_3 = a_4 = 0.003$ (right); the ideal f_0 trajectory is plotted in solid line

3.3 Behaviour on noisy signals

In this case, the simulated signal is equal to $s = \cos\left(\phi_1 + \varphi^1\right) + b$, where b is a normal noise, with mean equal to 0 and standard deviation equal to σ. The results are shown in figure 7. When there is noise, the AS and TK methods do not perform well.

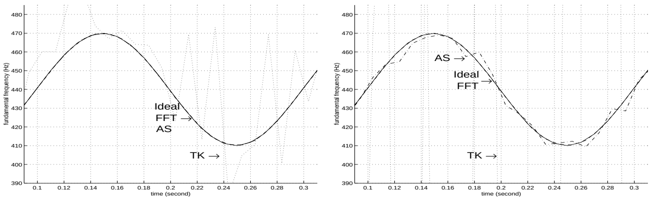

Figure 7: Behaviour on noisy signals: $\sigma = 1e^{-5}$ (left), $\sigma = 3e^{-4}$ (right)

3.4 Discussion

The last two tests (sections 3.2 and 3.3) show that for the AS and TK methods the signal must be a pure sine with slowly varying amplitude and frequency. A very efficient band-pass filtering step is absolutely necessary for these two alternative methods. The FFT method seems to be the best, as the use of knowledge allows to improve its performance. We decided therefore to use this method in our final system.

4 Performance of the three frequency trackers – On true sound signals

The left panel of figure 8, 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 more noisy 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 right side of the figure 8). 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.

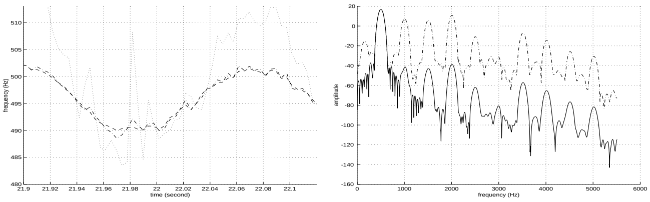

Figure 8: Left panel: f_0-trajectories obtained with the three methods. Cello (54.5 bpm, last note, first harmonic). Right panel: spectra of a frame of the original signal [$21.9s$ $21.92s$] (solid line) and of the corresponding band-pass filtered signal (dash-dot line)

An extra advantage of using FFT is that the peak tracking can be done after combining the transposed spectra of the individual harmonics. The results of performing peak tracking before and after fusion were compared. The latter improves the results considerably.

5 Example: performance of the complete system for the cello

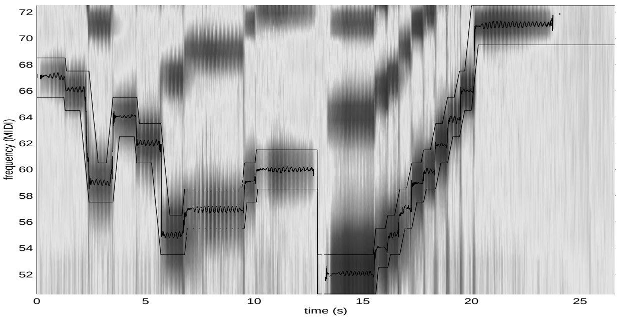

Figure 9: f_0 trajectory obtained with the complete system for the cello (54.5 bpm)

Results obtained for the cello are shown in figures 1 and 9. In figure 1, the first four harmonics are shown. Figure 9 shows the spectrogram, the score information and the f_0-trajectory. As an example, for the long note at 57 MIDI pitch, we can see 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 in figure 1. But, after the data fusion step, the f_0-trajectory shown in figure 9 is obtained. If we compare this trajectory to the frequency trajectory obtained for the first harmonic (figure 1), the results have clearly been improved.

6 Summary and conclusion

In this paper, an efficient and elegant f_0 tracker is presented, that can be used to analyse vibrato and portamento. It uses knowledge of the music and the instrument.

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 method 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.

7 Acknowledgment

This research is supported by the European MOSART IHP network (see `http://www.diku.dk/research-groups/musinf/mosart/`) and by the Netherlands Organisation for Scientific Research (NWO).

References

Boashash, B. (1992). Estimating and interpreting the instantaneous frequency of a signal. In *Proceedings of the IEEE*, Volume 80, no.4, pp. 539 – 568. IEEE.

Brown, J. C. and M. S. Puckette (1993). A high resolution fundamental frequency determination based on phase changes of the Fourier transform. *Journal of the Acoustical Society of America 94, no. 2*, 662 – 667.

Desain, P. and H. Honing (1996). Modeling continuous aspects of music performance: Vibrato and portamento. In *Proceedings of the International Music Perception and Cognition Conference.* B. Pennycook & E. Costa-Giomi, CD-ROM.

Desain, P., H. Honing, R. Aarts, and R. Timmers (2000). *Rhythmic Aspects of Vibrato* (In P. Desain and W. L. Windsor, Rhythm Perception and Production), pp. 203–216. Swets & Zeitlinger.

Hess, W. (1983). *Pitch determination of speech signals.* Springer-Verlag.

Maragos, P. and J. K. Kaiser (1993). Energy separation in signal modulations with application to speech analysis. In *IEEE Transaction on Signal Processing*, Volume 41, no. 10, pp. 3024 – 3050. IEEE.

Timmers, R. and P. Desain (2000). Vibrato: the questions and answers from musicians and science. In *Proceedings of the International Conference on Music Perception and Cognition.* UK, Keele University, Department of Psychology, CD-ROM.

Vakman, D. (1996). On the AS, the TK energy algorithm and other methods for defining amplitude and frequency. In *IEEE Transaction on Signal Processing*, Volume 44, no. 4, pp. 791 – 797. IEEE.

Wang, A. L.-C. (1994). *Instantaneous and frequency-warped signal processing techniques for auditory source separation.* Ph. D. thesis, Stanford University.