# Fundamental Frequency Estimation in the SMS Analysis

Pedro Cano
Audiovisual Institute, Pompeu Fabra University
Rambla 31, 08002 Barcelona, Spain
pcano@iua.upf.es http://www.iua.upf.es

**Abstract**

This paper deals with the fundamental frequency estimation for monophonic sounds in the SMS analysis environment. The importance of the fundamental frequency as well as some uses in SMS is commented. The particular method of $F_0$ estimation based on a two-way mismatched measure is described as well as some modifications. Finally we explain how pitch-unpitched decision is performed.

## 1 Introduction

The particular approach of SMS is based on modeling sounds as stable sinusoids (partials) plus noise (residual component), therefore analyzing sounds with this model and generating new sounds from the analyzed data. The analysis procedure detects partials by studying the time-varying spectral characteristics of a sound and represents them with time-varying sinusoids. These partials are then subtracted from the original sound and the remaining "residual" is represented as a time-varying filtered white noise component [1].

This article is part of the current work at the Audiovisual Institute of the Pompeu Fabra University by the Music Technology Group to improve the original SMS system developed by X. Serra.

## 2 Importance of the fundamental frequency in SMS

In SMS, an input sound, without any harmonic assumption, is modeled by

$$s(t) = \sum_{r=1}^{R} A_r(t)\cos[\theta(t)] + e(t)$$

where $A_r(t)$ and $\theta_r(t)$ are the instantaneous amplitude and phase of the $r^{th}$ sinusoid, respectively, and e(t) is the noise component at time t.

However, when the input sound is pseudo-harmonic, it is useful to take advantage of this and group the partials related harmonically and leave to the residual both non-harmonic and noisy contributions to the sound.

The sound is modeled then by:

$$s(t) = \sum_{k=1}^{K} A_k(t)\cos\left(2\pi k \cdot \left(f_0(t) + \Delta_k f_0(t)\right)\cdot t + \psi_k(t)\right) + e(t)$$

where $A_k(t)$ and $\psi_k(t)$ are the instantaneous amplitude and phase of the $k^{th}$ sinusoid, respectively, $\Delta_k$ is the frequency deviation of the $k^{th}$ sinusoid to a perfect harmonic series normalized to the fundamental frequency $f_0$, and e(t) is the noise component at time t.

This representation is far more flexible and allows for more and better transformations but it raises the question of what to do if a portion of the sound becomes inharmonic or noise-like. Our choice has been to perform time segmentation on the sound with a pitch-unpitch detector. If a frame is classified as unpitched, the whole spectrum is considered residual and is treated with a different model. If, on the other hand, a frame is considered pitched the separation between harmonic and noisy components relies, besides the general peak detection and continuation routines, on the fundamental frequency estimation. The selection of sinusoidal components is heavily based on the fundamental frequency estimation.

A part from allowing for a more musically meaningful representation, dealing with harmonic signals has more advantages; a frequent problem in sinusoidal representation is the selection of a window for a good frequency trade-off. With the knowledge of the fundamental frequency, an approach to pitch-synchronous analysis can be performed by adapting the window size to a specific number of pitch periods.

The fundamental frequency can also be used to transform the residual component. If for instance, a pitch scaling is performed in a sound, unless we do something on the residual, we will find that harmonic and residual do not mix well. A possible way to solve this is by comb-filtering the residual, subtracting the old tonality and adding in the new one so that the sounds mix better.

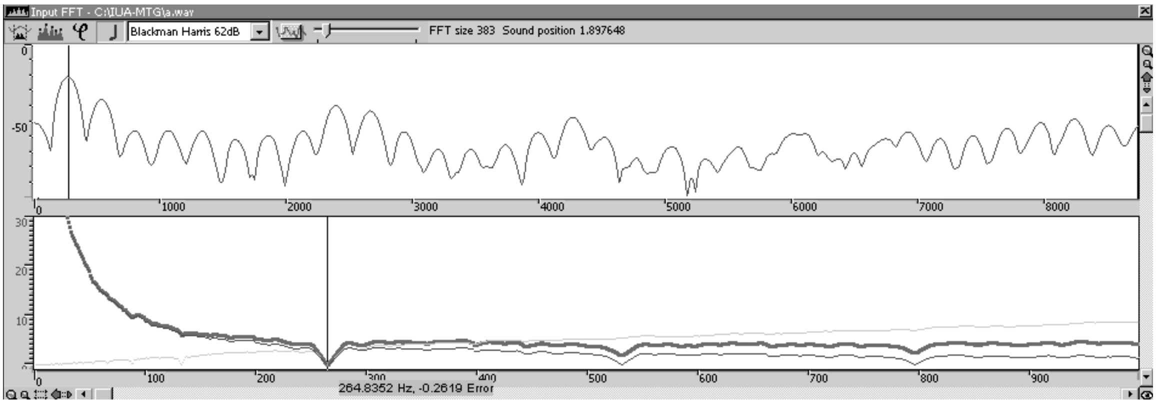


Figure 1. In this figure it is easy to see how the two errors, measured to predicted and predicted to measure, complement each other to avoid octave errors and to find the correct $F_0$. Note that the predicted to measured presents a minimum in the octaves superior to the correct $F_0$ (around 530 Hz) and the measured to predicted does it in the lower octaves (around 130Hz).

There is not a unique solution when determining a period of a quasi-periodic signal. In the context of SMS, we look for the $F_0$ of the pure periodic sound that best resembles locally to the input sound. A possible search for $F_0$ could be accomplished by minimizing the residual energy of the signal once the harmonic partials are subtracted. This is the common approach of analysis by synthesis systems [2] in the frequency domain or the Least-square algorithm proposed by Choi [3] in the time domain. For our purposes, we wanted a technique with a good trade-off between accuracy and computation time. Thus, what we do is the following: given the set of spectral peaks calculated as part of the SMS analysis, with magnitude and frequency values for each one, we estimate the fundamental frequency by measuring the "goodness" of the possible harmonic series with the actual spectral peaks, in a way similar to the method proposed by Maher and Beauchamp as the Two-Way Mismatch (TWM) procedure [4].

# 3 The TWM Procedure

In the Two-Way Mismatch procedure the estimated $F_0$ is chosen as 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.

In the original algorithm they propose using a short-time Fourier transform analysis of an acoustic signal input using fixed window size (typically, 46 ms). For each time frame (typically, 5.8 ms), they obtain the magnitude spectral peaks and apply the TWM procedure.

The predicted to measured error is defined as:

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

where $\Delta f_n$ is the difference between a predicted and its closest measured peak, $f_n$ and are the frequency and magnitude of the predicted peaks, and $A_{max}$ is maximum peak magnitude.

The measured to predicted error is defined as:

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

where $\Delta f_k$ is the difference between a measured and its closest predicted peak, $f_k$ and $a_k$ are the frequency and magnitude of the measured peaks, and $A_{max}$ is maximum peak magnitude.

The total error is:

$$
Err_{total} = Err_{p\to m} / N + \rho Err_{m\to p} / K
$$

The values they propose are p=0.5, q=1.4, r=0.5 and ρ=0.33. We have validated this values experimentally and tested that they work optimally for most sounds.

As a simple example of the TWM calculation consider the measured sequence of partials {200, 300, 500, 600, 700, 800}. In this example we would like to determine whether 50, 100, or 200 Hz is the best $F_0$

assuming all the measured partials are approximately equal in amplitude. In the following table (*Table 1*) we show the formula results.

| | $Err_{pm}$ | $Err_{mp}$ | Err |
|---|---|---|---|
| 50 Hz | 122.58 | -3.0 | 7.49 |
| 100 Hz | 32.0 | -3.0 | 3.83 |
| 200 Hz | 10.0 | 30.66 | 4.2 |

Table 1. Example of error measures based on the TWM calculation.

Note that neither of the errors acting alone can achieve an unambiguous $F_0$ choice. This can also be checked in the graph of *Figure 1*, where the $Err_{p\to m}$ (in darker gray), $Err_{m\to p}$ (in lighter gray) add up to a total error (in black) with the correct solution.

# 4 Modifications

Although the method described works well for a general purpose application, in the context of SMS we have made some changes to better suit our needs.

## 4.1 Pitch dependent analysis window

Some of the analysis steps in SMS help to improve the performance of the fundamental frequency estimation as a side effect. The analysis window, for instance, adapts its size and type depending on the fundamental frequency. Doing so, the accuracy of the result is less depending on pitch than using a fixed window.

In order to change the window to get a good time-frequency trade off, the algorithm has to be sure that the $F_0$ detected is a reliable one. In our case, we consider a good $F_0$ when the pitch error is sufficiently low and the pitch estimate has been relatively constant for a few frames. There is also a hystheresis cycle to avoid changing windows back and forth if $F_0$ is around a changing point.

In our current implementation of SMS, the way that the window size and the window type change as a function of the period detected is selectable by the user. The values in *Figure 2* have proven to work correctly and result in a minimum latency analysis and good time resolution. There are reasons why window size is not a constant number of periods of the signal. For low fundamental frequency sounds, in order to avoid the delay caused by using a big window and also, to improve time resolution, the window size is chosen to be less than 3 periods. As fundamental frequency rises, each pitch period is represented by fewer samples and to keep a sufficient frequency resolution we have to increase the size of the window, which reaches 3.5 periods at 500Hz. For the above configuration to work is necessary to change the window type as well. We thus select a window with a narrow main lobe, for example the Kaiser-Bessel 1.8. for low pitches while we can afford a smoother window with a wider main lobe, like the Kaiser-Bessel 2.5. for the higher pitches.

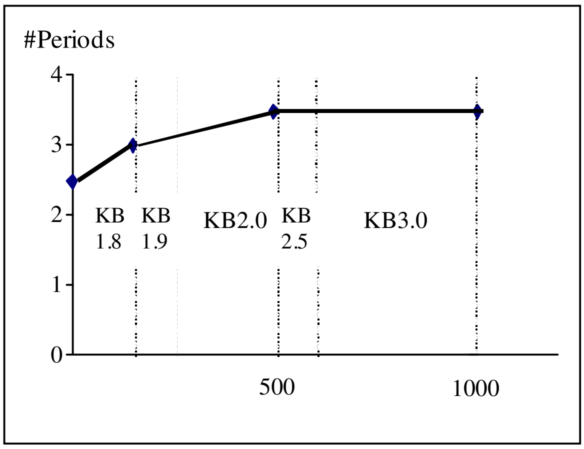


Figure 2. Windows size and window type as a function of fundamental frequency. KB stands for Kaiser-Bessel and the number refers to the $\alpha$.

## 4.2 Optimization in the search for candidates

The original TWM method is computationally quite costly because it sweeps through all the possible fundamental frequency range with very small frequency increments. It is faster to first find possible candidates and apply the algorithm to these ones only. By only considering as candidates the frequencies of the measured spectral peaks, frequencies related to them by simple integer ratios (e.g., 1/2, 1/3, 1/4), and the distances between well defined consecutive peaks, the search is simplified enormously. From the best candidates we can also refine the measure by searching around them with small frequency increments for a better fundamental frequency.

## 4.3 Peak selection

The TWM error, which works optimally when all the peaks are partials, is corrupted by side lobes and noise peaks. It is useful then to improve the peak selection step, discarding peaks that are clearly not partials, thus not using all the spectral peaks. A good approach to perform this selection is correlating the spectrum with the transform of the window [5], but that is time consuming and it also implies losing frequency resolution [2]. For monophonic sounds from a good recording is enough using some simple rules. In our case, in order for a peak to be accepted it has to be less than 40 dB below the highest peak and it has to have a minimum bandwidth. Another criteria that is useful in some cases is to accept peaks whose phase slope around the peak is close to zero.

### 4.4 Fundamental frequency tracking

$F_0$ estimators and TWM algorithm as well work with the analysis based on a narrow time segment. Since it works only from local details, it may myopically track irrelevant details or event mistake the estimation. The musical signals we use, on the other hand, have smooth $F_0$ trajectories. With the purpose of avoiding gross errors or simply to have smoother track, it is useful to move to higher levels of processing and consider neighbor frames.

We use the history of the pitch to choose between candidates with similar TWM Error. There is the option to take into account future frames to help us decide the best $F_0$, both in describing the signal at the local frame, and having a smooth evolution with neighbor frames. As additional information to modify the search among $F_0$ candidates we have the phase: The second derivative of the phase in a peak is positive if a partial is decreasing its frequency and negative if the partial is increasing. This information is used as well when deciding best $F_0$ trajectory to use. Note that the pitch contour is not smoothed independently of the knowledge of the sound as median techniques do.

### 4.5 Knowledge-based rules

The general method described so far is independent of the type of sound. However context specific optimization can be done when knowledge on the signal is available. Knowing, for instance, the spectral range of the $F_0$ in a particular sound helps both the accuracy and the computational cost. Then, there are sounds with specific characteristics, like in a clarinet where the even partials are softer than the odd ones. From this information we can define a set of rules that will improve the performance of the estimator.

## 5 Pitched-Unpitched decision

As explained above, the portions of a sound that cannot be well represented with the harmonic model are considered as residual. There is then a strict segmentation in time with routines that use the error measure of the TWM along with other measures that are easily computed as part of the SMS analysis. In particular we have used: Zero Crossing, Energy, Noisiness and Harmonic Distortion [9].

## 6 Conclusions

Fundamental frequency estimation is an old problem with many approaches and solutions depending on the particular target application. In our case, the main goal is to estimate parameters for the harmonic model in SMS while describing a reliable and smooth track that allows us to use $F_0$ for many other applications. There is room for many improvements, specially incorporating the knowledge of particular sound families and the use of the phase information.

## 7 Acknowledgments

This work has been supervised by Xavier Serra with many contributions by Jordi Bonada and other members of the Music Technology Group.

## References

[1] X. Serra. “Musical Sound Modeling with Sinusoids Plus Noise”. C. Roads, S. Pope, A. Picialli, G. De Poli, editors. *Musical Signal Processing*. Swets & Zeitlinger Publishers, 1997.

[2] P. Fernández-Cid. *Transcripción automática de señales polifónicas*, Ph. D. Thesis. Polytechnic University of Madrid, 1998.

[3] A. Choi.. “RealTime Fundamental Frequency Estimation by Least-Square Fitting”, *IEEE Trans. Speech & Audio Processing* 5 (2) pp. 201-205, 1997.

[4] R. C. Maher and J. W. Beauchamp. “Fundamental Frequency Estimation of Musical Signals using a two-way Mismatch Procedure”, *Journal of the Acoustical Society of America*. (4):2254—2263, 1994.

[5] B. Doval and X. Rodet. “Fundamental Frequency Estimation and Tracking using Maximum Likelihood Harmonic Matching and HMMs”, *Proceedings of ICASSP*, 1993.

[6] R. J. McAulay and T. F. Quatieri. “Pitch Estimation and Voicing Detection based on a Sinusoidal Speech Model”. *Proceedings of ICASSP*, 1990.

[7] W. Hess. Pitch Determination of Speech Signals. Springer-Verlag, Berlin, 1983.

[8] M. M. Goodwin. *Adaptive Signal Models: Theory, Algorithms, and Audio Applications*. Ph.D. Thesis, University of California at Berkeley, 1997.

[9] X. Serra and J. Bonada. “Sound Transformations based on the SMS High Level Attributes”. *Proceedings of the Digital Audio Effects Workshop*, 1998.