Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 4º Congresso da AES Brasil
10ª Convenção Nacional da AES Brasil
08 a 10 de Maio de 2006, São Paulo, SP

Accurate and Efficient Fundamental Frequency Determination from Precise Partial Estimates

Adriano Mitre[1], Marcelo Queiroz[1], Regis R. A. Faria[2]

[1]Department of Computer Science,
Institute of Mathematics and Statistics, University of São Paulo

[2]Laboratory of Integrated Systems,
Polytechnic School, University of São Paulo

adriano@mitre.com.br, regis@lsi.usp.br, mqz@ime.usp.br

ABSTRACT

An algorithm is presented for the estimation of the fundamental frequency (F_0) of monophonic sounds. The method relies upon accurate partial estimates, obtained on a frame basis by means of enhanced Fourier analysis. The use of state-of-the-art sinusoidal estimators allows the proposed algorithm to work with frames of minimum length (i.e., about two fundamental periods). The accuracy of the proposed method does not degrade for high pitched sounds, making it suitable for musical sounds.

INTRODUCTION

Extracting the fundamental frequency (F_0) contour of a monophonic sound recording has a number of applications, such as audio coding, prosodic analysis, melodic transcription and onset detection.

Pitch determination in speech signals is a extensively studied topic, mostly motivated by immediate applications in telecommunications. Musical pitch estimation, however, has received considerably less attention.

Speech and musical pitch estimation pose different challenges for pitch determination algorithms (PDA). Fundamental frequency estimation in music signals is in many ways more challenging than that in speech signals. In music, the pitch range can be wide, comprising more than seven octaves, and the sounds produced by different musical instruments vary a lot in their spectral content. The inharmonicity phenomenon has to be taken into account.

On the other hand, the dynamic (time-varying) properties of speech signals are more complex than those of an average music signal. The F_0 values in music are temporally more stable than in speech.

Despite the aforementioned differences, it is occasionally possible to employ speech-tailored PDAs to monophonic musical recordings, with variable degree of success.

The human voice and most pitched musical instruments used in Western music produce quasi-harmonic sounds[1]. The reason for this is encountered in the physics of vibrating strings and tubes. As the pitch of a quasi-harmonic sounds is closely related to its fundamental frequency, both terms were used indistinctly in the present work.

PROPOSED METHOD

A number of techniques have been proposed for pitch estimation, mostly aiming at measuring periodicity in the time or frequency domain. Most funda-

[1] The mallet percussion family is a notable exception.

mental frequency estimation methods may be classified according to the domain on which they operate. The ones which operate directly on the signal waveform are termed time-domain methods. Methods which transform the waveform to a spectral representation are called frequency-domain methods. This transformation is usually carried out by means of constant Q or short-time Fourier transforms (STFT).

Although the proposed method employs the Fourier transform, it does not operate on the complete spectrum signal, but rather on a small set of partials. It requires frequency analysis, followed by extraction and estimation of partials. The list of partials in each frame is the input to the proposed algorithm.

The main steps of the proposed method are shown in Figure 1.

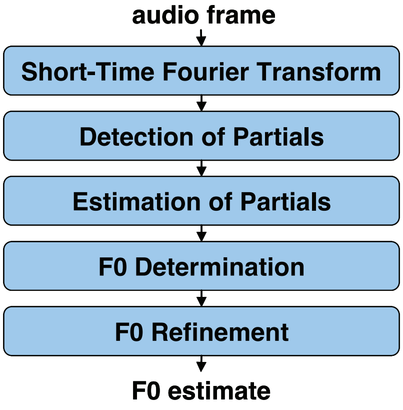

Figure 1: Flowchart of the proposed method.

Detection of Partials

The spectral analysis module produces, for each audio frame, its corresponding complex spectrum. Notwithstanding, we note that only prominent partials are relevant for fundamental frequency estimation.

Under reasonable assumptions, each partial in the input signal produces a local maximum in the magnitude spectrum; the converse is not true due to smearing effects and noise intrinsic to discrete analysis. Therefore several heuristics were proposed to discriminate local maxima induced by partials from those induced by noise. A popular strategy in analysis/resynthesis systems is partial tracking [1, 2], which does not operate on isolated frames and thus suggests an offline partial filtering strategy.

In the present study, the discrimination between genuine and spurious peaks is postponed to the subsequent module. In this approach every peak is estimated "as if it were" a partial. Then, the ones whose frequency estimate depart more than half bin from its original value are discarded as noise.

Estimation of Partials

In order to correctly estimate a 12-tone-equal-tempered pitch from a given fundamental frequency, an accuracy[2] of at least $\mathrm{F0}_{\mathrm{min}} \left(\sqrt[24]{2} - 1 \right)$ Hz is needed, where $\mathrm{F0}_{\mathrm{min}}$ denotes the lowest expected fundamental frequency in the input signal. In order to accurately follow expressive subtleties such as *vibrati* and *glissandi* a higher accuracy is needed.

Frequency accuracy of conventional STFT is half the inverse of frame length, represented by $\{2\tau\}^{-1}$ Hz. STFT's frequency resolution[3], although constrained by the frame length, depends also on the window shape. More precisely, it is determined by the 6 dB bandwidth of the window power spectrum main lobe and is given by $L_w \cdot \tau^{-1}$ Hz, where L_w depends on the window. For classic windows, such as Hann and Blackman, L_w lies between 1.2 and 3.1 [3, 4].

For instance, in order to discriminate between pitches of a 6-stringed guitar whose lowest pitch is an E corresponding to 82.4 Hz, one needs a frame of duration at least $\left[2 \times 82.4 \times \left(\sqrt[24]{2} - 1 \right) \right]^{-1} \simeq 207$ ms. Musical signals seldom exhibit quasi-periodic behaviour for so long. Large frames tend to lower temporal precision because of contamination from two or more succesive notes occurring in a single analysis frame. In addition, a temporal accuracy of 20 ms asks for an overlap factor of 90% and therefore raises the computational workload by a factor of ten.

In monophonic quasi-harmonic signals any two partials are at least $\mathrm{F0}_{\mathrm{min}}$ Hz apart and thus a frame length of $L_w \cdot \mathrm{F0}_{\mathrm{min}}{}^{-1}$ s is enough for them to be resolved (i.e., separated). This new bound is much tighter than the previous one. For the guitar example, a Hamming-windowed frame of $1.81 \times 82.4^{-1} \simeq 22$ ms is enough.

Fortunately, several techniques exist for improving the estimates of resolved partials. These generally fall into two categories, phase-based and interpolation-based.

Interpolation-based Techniques

One of the techniques for improving the estimates of sinusoidal components is spectral oversampling. It is usually attained by means of zero-padding, which consists in adding a sequence of zeros to the windowed frame before computing the STFT. The disadvantage of spectral oversampling is that the increase in the computational workload is proportional to the improvement in accuracy.

Another technique is quadratic (or parabolic) interpolation, whose estimates are computed using each local maximum of the spectrum and its adjacent bins. It benefits from the fact that the main lobe of the logarithmic power spectrum of several windows are

[2] In the present work, the term accuracy is used in the sense of exactness. An estimator is thus said to have accuracy ϵ if every estimate is within ϵ of its true value, i.e., $|\hat{f}_i - f_i| < \epsilon$ for all i.

[3] Throughout the text, frequency resolution will refer to how close two sinusoids may get while still being separable in the spectrum. A resolution of Δ means that two sinusoids with same amplitude and frequencies f_1 and f_2 may separated if and only if $|f_1 - f_2| \geq \Delta$ and $\min\{f_1, f_2\} \geq \Delta$. The second inequality is due odd-simmetry of the spectrum of real signals.

very close to a quadratic function. Purposefully designed windows are sometimes employed, which are obtained by taking the inverse transform of a perfect quadratic function. The parabolic interpolation technique is often combined with spectral oversampling.

For the special case of the Hann window, Grandke designed an interpolation technique which considers each peak and its greatest neighbour [5].

A number of interpolation techniques exist for the rectangular-windowed STFT[4], however spectral leakage problems prevent the use of rectangular window for musical signal analysis.

Phase-based Techniques

More sophisticated partial estimation techniques use the phase spectrum in addition to magnitude information. The Derivative Method [6] uses the spectra of the original signal and its derivative (aproximated by a low-pass filter) and the Spectral Reassignment Method [7, 8] associates energy content to the cells of a time-frequency representation in order to improve accuracy of the estimates. Thanks to a trigonometric interpretation of the Derivative Method, an improved estimator was derived in [9]. The new estimator is as precise for close-to-Nyquist frequencies as the Derivative is for low frequencies.

These techniques give better estimates at the expense of additional STFT computations. Comparative studies of these techniques with respect to mean error, variance and bias can be found in [10] and [11].

Amplitude Estimation

Except by quadratic interpolation and spectral oversampling, the aforementioned techniques only estimate the frequency of partials. Nevertheless, one can obtain precise amplitude estimates of partials by applying analytical knowledge about the window used.

Denoting by $\hat{f}_k$ the frequency estimate of the partial at the k-th bin, whose center frequency is f_k, and by W the frequency response of the window, the precise amplitude estimate for the partial is given by the formula

$$\hat{a}_k = \frac{a_k}{W\left(\left|\hat{f}_k - f_k\right|\right)} \tag{1}$$

Prior to fundamental frequency determination, described in the "Fundamental Frequency Determination" section, the magnitude of the partials must be normalized to absolute decibels. This is accomplished by the following formula.

$$\hat{a}_k^{\mathrm{dB-norm}} = \alpha + 20 \cdot \log_{10} \hat{a}_k \tag{2}$$

The term α is set to map the maximum possible amplitude to 70 dB. It is determined by the window size (in samples), the windowing function and the recording bit-depth.

Finally, non relevant partials are filtered prior to fundamental frequency determination. A partial is considered relevant if its frequency is within human hearing range (20−20, 000 Hz) and its magnitude is strictly positive.

Fundamental Frequency Determination

The proposed method assumes that the strongest partial belongs to the main harmonic series, thus its frequency is expected to be multiple of F_0. Letting $f_\star$ denote the frequency corresponding to the strongest partial, the set of candidates for F_0 is composed by submultiples of $f_\star$. Formally,

$$C = \left\{ c_n \overset{\mathrm{def}}{=} \frac{f_\star}{n} : 1 \le n \le \left\lfloor \frac{f_\star}{\mathrm{F0}_{\min}} \right\rfloor \right\} \tag{3}$$

The next step consists in collecting the harmonic series corresponding to each F_0 candidate. This is carried out by the following algorithm: firstly, partials are sorted in decreasing order of magnitude; then, each partial is sequentially assigned to the nearest (in a quarter tone vicinity) "empty slot" of the candidate's harmonic series.

As a result of the previous algorithm, the i-th harmonic of the n-th candidate is given by

$$H[n][i] = \arg\max_{p \in \Lambda_i^n} \left\{ p_{\mathrm{mag}} \right\} \tag{4}$$

where p denotes a partial with frequency p_{freq} and magnitude p_{mag}. In words, $H[n][i]$ is the partial with greatest magnitude among the set of potential i-th harmonic of the n-th candidate, given by

$$\Lambda_i^n = \left\{ p : l_i < \frac{p_{\mathrm{freq}}}{i c_n} < h_i \right\} \tag{5}$$

where l_i and h_i ensure smaller than quarter-tone deviation and, in the case of higher order harmonics, prevent single partials from being assigned to multiple adjacent harmonics "slots". Formally,

$$l_i = \max \left\{ \sqrt[24]{2^{-1}}, \sqrt{\frac{i-1}{i}} \right\} \tag{6}$$

$$h_i = \min \left\{ \sqrt[24]{2}, \sqrt{\frac{i+1}{i}} \right\} \tag{7}$$

In short, if the i-th harmonic of the n-th candidate belongs to the spectrum, it will be assigned to $H[n][i]$. Otherwise, it is agreed that $H[n][i]_{\mathrm{mag}} = 0$.

It is further necessary to quantify the prominence of each candidate according to its harmonic series. This takes into account psychoacoustic factors, particularly the critical band [12, §2.4 and §3.4]. The functions Φ and Ψ defined below are based on the harmonic sum model [13, §6.3.3]. The psychoacoustic motivation for these formulas can be found in the same reference.

[4] Rectangular-windowed STFT is often misleadingly referred to as unwindowed, instead of *unsmoothed*, STFT.

Formally stating, the prominence of the n-th candidate is given by

$$\Phi(n) = \sum_{i=1}^{I(n)} H[n][i]_{\text{mag}} \cdot \Psi(i) \tag{8}$$

$$I(n) = \max\left\{ j : H[n][j]_{\text{mag}} > 0 \right\} \tag{9}$$

and $\Psi(i)$ denotes the fraction of the critical band which corresponds to the i-th harmonic, given by

$$\Psi(i) = \begin{cases} 1, & \text{if } i \leq 4 \\ \Gamma(i) - \Gamma(i-1), & \text{otherwise} \end{cases} \tag{10}$$

$$\Gamma(n) = \log_{2^{1/3}} \left(n \cdot \sqrt{\frac{n+1}{n}} \right) \tag{11}$$

The fundamental frequency estimation is performed in three steps, given the prominence of the candidates as defined above. The first step selects those candidates with relative prominence of at least $\beta \in [0, 1]$ with respect to the maximal prominence:

$$C^{\Phi} = \left\{ c_n \in C : \Phi(n) \geq \beta \cdot \max_{m | c_m \in C} \{\Phi(m)\} \right\} \tag{12}$$

For each of these candidates the weighted average harmonic magnitude is computed as:

$$\chi(n) = \frac{\sum_{i=1}^{I(n)} H[n][i]_{\text{mag}} \cdot \Psi(i)}{\sum_{i=1}^{I(n)} \Psi(i)} \tag{13}$$

Then the one with the highest value of χ is selected as F_0, whose index is

$$\varphi = \arg\max_{n : c_n \in C^{\Phi}} \{\chi(n)\} \tag{14}$$

Fundamental Frequency Refinement

The exact value of the estimated F_0 was based on the frequency estimate of a single partial: the strongest one. However, the F_0 estimate may be improved by considering frequency estimates of all partials in the harmonic series of the winner candidate. Since partial estimates are expected to be non-biased, individual errors should cancel each other out by averaging.

The realiability of a partial estimate is affected by its signal-to-noise ratio (SNR) and the stability of its absolute frequency. Therefore strong and small indexed harmonics should be privileged, since they have the higher SNR and smallest absolute frequency modulations.

Taking these facts into account, we propose the following formula for further refining the initial fundamental frequency:

$$\hat{F}_0 = \frac{\sum_{i=1}^{I(n)} H[i]_{\text{freq}}/i \cdot H[i]_{\text{mag}} \cdot \Psi(i)}{\sum_{i=1}^{I(n)} H[i]_{\text{mag}} \cdot \Psi(i)} \tag{15}$$

where $H[i]$ denotes the i-th partial of the harmonic series of c_{φ}, which is, $H[i] \overset{\text{def}}{=} H[\varphi][i]$.

The F_0 refinement might be thought as an weighted average of local F_0 estimates. Local estimates should be understood regarding the harmonic indice, i.e., the local F_0 estimate for the i-th harmonic is $H[i]_{\text{freq}}/n$.

ADVANTAGES AND DRAWBACKS

It is well known that spectral and temporal resolutions are reciprocals and thus detecting F_0 as low as f Hz requires a window whose length is at least $K \cdot f^{-1}$ s, where K is independent of f. In the case of Fourier spectrum based methods, K is mainly determined by the window [3].

On the one hand, all short-time F_0 estimators suffer from this limitation. On the other hand, while waveform-based PDAs have their precision determined (i.e., fixed) by the signal's sample rate, the precision of F_0 estimates produced by spectrum-based PDAs might be increased by employing longer windows. Notwithstanding, the use of interpolation may be helpful for methods on either domain.

The precision of the proposed method has the same order of magnitude as that of the sinusoid estimator employed, occasionally surpassing it due to the refinement procedure. It must be noted, however, that if spurious peaks in the magnitude spectrum are incorrectly classified as partials and collected to the harmonic series of the winner F_0 candidate, the refinement stage may degrade, instead of enhance, the initial F_0 estimate.

The method is timbre-independent, being robust to the following phenomena:

- weak or absent fundamental
- incomplete series (e.g., only odd harmonics)
- sinusoidal-like sounds
- moderate levels of inharmonicity (as found in acoustic instruments)

It must be noted that although inharmonicity is not explicitly modelled, the tolerance of the harmonic series collector allows for moderately inharmonic low order partials.

Experiments conducted with severely bandlimited (e.g. telephone-like bandpass filtered) versions of musical recordings have shown that the method is robust against bandlimiting. In some sense this is expected, since the method is partially derived from a bandwise multiple-F_0 estimator [14].

IMPLEMENTATION ISSUES

Profiling revealed that the most processing-intense step of the proposed method is the calculation of the STFT, which can be carried out by the Fast Fourier Transform algorithm.

The memory required by the method, excluding the STFT, is proportional to $|C|$, the number of candidates. It can be seen from Equation 3 that $|C|$ is indirectly dependant on the window length, as $\mathrm{F0}_{\min}$ should never be lower than $L_w \cdot \tau^{-1}$. Notwithstanding, the number of candidates can be safely assumed to be smaller than 200, as in musical sounds it is usually the case that $f_\star < 5$ kHz and $\mathrm{F0}_{\min} > 27.5$ Hz.

Thus, not only the processing, but also the memory requirements of the proposed method are dominated by the STFT.

EXPERIMENTS AND RESULTS

By the writing of this article, only informal (although extensive) evaluation was conducted. The results were, in general, very encouraging. Figures 2 and 3 show F_0 contours produced by the proposed method with expressive recordings of acoustic instruments.

There were two main reasons that retarded formal evaluation. The first reason is that there is no standardized *musical* database available for the task of PDA evaluation, i.e., one which provides reference F_0 tracks along with the audio recordings. The second reason is that, to the best of authors knowledge, there is no tool available for automatic generating statistics from reference and estimated F_0 tracks.

In an effort to remedy the situation, an automatic PDA evaluation tool was developed and musical monophonic recordings were collected, comprising most acoustic, electric and electronic instruments. In spite of this, manually obtaining reference F_0 tracks for the recordings is a laborious process which could not be concluded until the article's submission deadline.

It must be stressed that formal evaluation will be carried out. As soon as the work is done, the recordings, reference F_0 tracks, evaluation tool and results will be made available at http://www.mitre.com.br/pda.

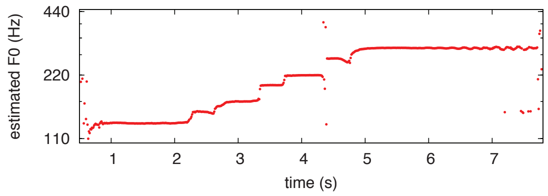

Figure 2: Expressive saxophone performance of the initial notes of a jazz standard.

CONCLUSION

A new algorithm was proposed for monophonic F_0 estimation. The method benefits from state-of-the-art partial estimators to reduce the required analysis frame length to a minimum (i.e., about two fundamental periods). This accounts for increased time resolution and reduced computational workload. The reduced number of configuration parameters makes it easier to fine-tune the method. Furthermore, informal evaluation suggests that the method is very robust for musical sounds.

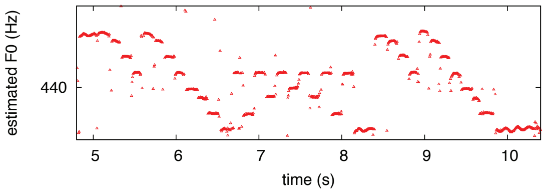

Figure 3: Expressive violin performance of an excerpt from a classical piece.

References

[1] Robert J. McAulay and Thomas F. Quatieri. Speech Analysis/Synthesis Based on a Sinusoidal Representation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 34(4):744–754, 1986.

[2] Mathieu Lagrange, Sylvain Marchand, Martin Raspaud, and Jean-Bernard Rault. Enhanced Partial Tracking Using Linear Prediction. In *Proceedings of the 6th International Conference on Digital Audio Effects(DAFx-03)*, Londres, Reino Unido, 2003.

[3] Fredric J. Harris. On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform. *Proceedings of the IEEE*, 66(1), January 1978.

[4] Albert H. Nuttall. Some Windows with Very Good Sidelobe Behavior. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(1):84–91, February 1981.

[5] Thomas Grandke. Interpolation algorithms for discrete Fourier transforms of weighted signals. *IEEE Transactions on Instrumentation and Measurments*, 32(2):350–355, June 1983. 1983.

[6] Myriam Desainte-Catherine and Sylvain Marchand. High Precision Fourier Analysis of Sounds Using Signal Derivatives. *Journal of the Audio Engineering Society*, 48(7/8):654–667, July/August 2000.

[7] Kunihiko Kodera, Roger Gendrin, and Claude de Villedary. Analysis of time-varying signals with small BT values. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):64–76, February 1978.

[8] Francçois Auger and Patrick Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5):1068–1089, May 1995.

[9] Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault. Improving sinusoidal frequency estimation using a trigonometric approach. In *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx-05)*, Madrid, Spain, September 20-22 2005.

[10] Florian Keiler and Sylvain Marchand. Survey On Extraction of Sinusoids in Stationary Sounds. In *Proceedings of the 5th International Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, September 2002.

[11] Stephen Hainsworth and Malcolm Macleod. On Sinusoidal Parameter Estimation. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, London, United Kingdom, September 2003.

[12] Juan G. Roederer. *The Physics and Psychophysics of Music: An Introduction*. Springer-Verlag Telos, 3rd edition, 1995.

[13] Anssi Klapuri. *Signal Processing Methods for the Automatic Transcription of Music*. PhD thesis, Tampere University of Technology, March 2004.

[14] Anssi P. Klapuri. Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6):804–816, November 2003.