ACCURATE AND ROBUST FREQUENCY ESTIMATION IN THE ODFT DOMAIN

Aníbal Ferreira

University of Porto, Porto, Portugal
ATC Labs, New Jersey, USA
ajf@atc-labs.com

Deepen Sinha

ATC Labs, New Jersey, USA
sinha@atc-labs.com

ABSTRACT

This paper presents new results improving by a factor of 10 the accuracy of an ODFT-based frequency estimation algorithm. These results are shown to be robust to the influence of additive noise and compare favorably to other non-iterative frequency domain estimation algorithms. A perspective is given on possible application areas, namely those involving real-time constraints.

1. INTRODUCTION

Accurate frequency estimation is required in many different application scenarios including speech coding [1], speech recognition [2], automatic transcription of music, special effects in audio, multimedia indexing and real-time interactive multimedia systems. In addition, in recent years, a number of audio coding technologies have been developed that perform a decomposition of the audio signal into sinusoids and noise, prior to coding, and therefore rely on accurate frequency estimation [3, 4, 5, 6].

Many frequency estimation algorithms have been proposed in the literature and can be broadly classified as time-domain algorithms and as frequency-domain algorithms. Time-domain algorithms are mainly based on auto-correlation measures (*e.g.,* [7]) and are frequently used in the context of pitch estimation of speech signals [2]. These methods have an inherent difficulty in dealing with polyphonic signals (*i.e.,* when different pitches occur simultaneously) and therefore in this case frequency-domain algorithms are preferred. In turn, frequency domain algorithms can be further classified into different categories as a function of the main underlying approach, including cepstrum analysis, peak picking and analysis, or phase analysis (*e.g.,* time derivative of the phase) of a Fourier representation of the signal [8].

The context of our research presumes polyphony and presumes that the frequency estimation must be carried out using a single audio frame and associated Fourier representation, which means time derivative of the phase can not be implemented. As a consequence, we focus on spectral peak picking and analysis techniques. As in [9], we presume in this paper the Odd-DFT (ODFT) transformation and time windowing by the square root of a shifted Hanning window [4].

Peak picking and analysis algorithms can be window agnostic (such as the parabolic interpolation or the triangle algorithms [10]) or may take into consideration (and advantage of) the frequency response of the time analysis window (FRTAW). These algorithms are likely to deliver more accurate frequency estimates as it was shown in [9]: for similar conditions, the frequency estimation error of a window-aware algorithm can be about six times smaller than the estimation error of a window-agnostic algorithm (the parabolic algorithm [11]).

This paper presents new results improving the technique presented in [9] in two important ways: accuracy and robustness to noise. In fact, it has been shown in [9] that the frequency of a sinusoid can be estimated with an error less than 1% of the bin width, in the absence of noise, and using a single closed form expression. In this paper we show how to reduce this estimation error to less than **0.1%** of the bin width, while improving simultaneously the robustness to noise.

This paper is organized as follows. In section 2 we present the basic signal analysis framework and detail the main approximation criteria. In section 3 we review the single-rule frequency estimation model that has been presented in [9]. In section 4 we derive two new rules and explain the construction of an improved three-rule frequency estimation model. In section 5 we characterize the performance of the new frequency estimation model and evaluate its robustness to noise in section 6. Section 7 presents the main conclusions of this paper.

2. ANALYSIS FRAMEWORK

Given a discrete signal represented by

$$x(n) = A \sin\left[\frac{2\pi}{N}(\ell + \Delta\ell)n + \phi\right] + r(n)\,, \tag{1}$$

where A represents the magnitude of a sinusoid, ℓ and $\Delta\ell$ represent respectively the integer part and the fractional part of the frequency of the sinusoid on a DFT-type frequency bin scale whose periodicity is N, ϕ represents the phase of the sinusoid, and $r(n)$ represents additive noise; the frequency estimation problem consists in finding the values of ℓ and $\Delta\ell$ after $x(n)$ has been windowed by $h(n)$ and transformed to the frequency domain by means of an ODFT of size N:

$$X_O(k) = \sum_{n=0}^{N-1} h(n)x(n)e^{-j\frac{2\pi}{N}(k+\frac{1}{2})n}. \tag{2}$$

When $x(n)$ and $h(n)$ are real, $X_O(k) = X_O^*(N-1-k)$, where * denotes complex conjugation. As explained in [9], in our framework the ODFT is used in a 50% overlap-add scheme achieving perfect reconstruction, which requires a special time window such as the square root of a shifted Hanning window [4]:

$$h(n) = \sin\frac{\pi}{N}(n+\frac{1}{2})\ ,\ 0 \le n \le N-1, \tag{3}$$

which we presume in this paper.

As shown in [9], the integer ℓ is readily extracted from the ODFT magnitude spectrum since it corresponds to the bin index of a local maximum. Ignoring the influence of the negative part of the spectrum, which is a fair assumption unless ℓ is too close to 0 or to $\frac{N}{2}-1$, the fractional part of the sinusoid frequency, $\Delta\ell$, can be estimated from the ODFT spectral coefficients surrounding the spectral coefficient with index ℓ and corresponding to a local maximum. In fact, given that each channel of the ODFT filter bank is obtained by modulating the FRTAW, $H(\omega)$, to the discrete center frequencies $\omega=\left(k+\frac{1}{2}\right)\frac{2\pi}{N}$, with $k=0,1,\ldots,N-1$; the magnitude of each ODFT spectral coefficient (or bin) depends on $H(\omega)$. Thus, $\Delta\ell$ can be estimated by an appropriate relation of the ODFT bins surrounding a local maximum. In this context we take $H(\omega)=\sum_{n=0}^{N-1}h(n)e^{-j\omega n}$, where $h(n)$ is given by equation (3).

The normalized magnitude of $H(\omega)$, which we represent as $|\widehat{H(\omega)}|$, can be obtained as [9]:

$$|\widehat{H(\omega)}|=\frac{\left|\cos N\frac{\omega}{2}\right|}{2\sin\frac{\pi}{2N}}\left|\frac{1}{\sin\frac{1}{2}\left(\frac{\pi}{N}-\omega\right)}+\frac{1}{\sin\frac{1}{2}\left(\frac{\pi}{N}+\omega\right)}\right|. \quad (4)$$

This equation is not directly tractable because of the pole-zero cancellation at frequencies $\omega=\frac{\pi}{N}$ and $\omega=-\frac{\pi}{N}$. However this problem can be circumvented by considering an approximation to the main lobe of the FRTAW:

$$|\widehat{H(\omega)}|\simeq\left[\cos\frac{N}{6}(\omega)\right]^{G}, \quad |\omega|<\frac{3\pi}{N}, \quad (5)$$

where G is a real constant. Equation (5) can therefore be used to relate the ODFT bins surrounding a local maximum and extract, as a result, an estimate of $\Delta\ell$.

3. SINGLE-RULE ESTIMATION MODEL

If ℓ is the index of a local maximum in the ODFT magnitude spectrum, the ratio of the magnitudes of the ODFT spectral coefficients in subbands $k=\ell-1$ and $k=\ell+1$ are obtained as [9]:

$$R=\frac{|X_O(\ell-1)|}{|X_O(\ell+1)|}=\frac{\left|H\left(\frac{2\pi}{N}(\Delta\ell+\frac{1}{2})\right)\right|}{\left|H\left(\frac{2\pi}{N}(\Delta\ell-\frac{3}{2})\right)\right|}. \quad (6)$$

Using the approximation (5), it can be shown that equation (6) leads to [9]:

$$\Delta\ell\simeq\frac{3}{\pi}\arctan\frac{\sqrt{3}}{1+2R^{1/G}}\ , \quad 0.0\leq\Delta\ell<1.0\ . \quad (7)$$

The value of G has been adjusted to $27.4/20.0$ ir order to optimize the estimation error in the *minmax* sense, when $0.0\leq\Delta\ell<1.0$. In this case and in the absence of noise, the maximum estimation error has been found to be less than 1% of the bin width, and to be essentially independent of N, ℓ, A and ϕ [9].

In practice it has been verified that equation (7) delivers very accurate frequency estimates, except when there is an influence of noise such that $|X_O(\ell+1)|$ does not approach 0.0 as it should when $\Delta\ell$ approaches 0.0, and $|X_O(\ell-1)|$ does not approach 0.0 as it should when $\Delta\ell$ approaches 1.0.

4. THREE-RULE ESTIMATION MODEL

In order to make the estimation process less vulnerable to the influence of noise, the (three) strongest spectral lines in the ODFT magnitude spectrum and falling within the main lobe of the FRTAW (whose width is $6\pi/N$, [9]), should be used. Taking into consideration the relative magnitudes of the three ODFT spectral lines (falling within the main lobe of the FRTAW) as a function of $\Delta\ell$ [4, 9], it becomes clear that a new formula should relate $|X_O(\ell-1)|$ and $|X_O(\ell)|$ when $\Delta\ell\leq 0.5$, and that another new formula should relate $|X_O(\ell)|$ and $|X_O(\ell+1)|$ when $\Delta\ell\geq 0.5$. Following an analytical work similar to that reviewed in section 3, it can be concluded that in the former case, if we represent the associated relation of magnitudes as:

$$Q=\frac{|X_O(\ell-1)|}{|X_O(\ell)|}=\frac{\left|H\left(\frac{2\pi}{N}(\Delta\ell+\frac{1}{2})\right)\right|}{\left|H\left(\frac{2\pi}{N}(\Delta\ell-\frac{1}{2})\right)\right|}\ , \quad (8)$$

the second frequency estimation formula is obtained as:

$$\Delta\ell\simeq\frac{3}{\pi}\arctan\sqrt{3}\frac{1-Q^{1/F}}{1+Q^{1/F}}\ , \quad 0.0\leq\Delta\ell\leq 0.5\ , \quad (9)$$

and in the latter case, if we represent the associated relation of magnitudes as:

$$S=\frac{|X_O(\ell+1)|}{|X_O(\ell)|}=\frac{\left|H\left(\frac{2\pi}{N}(\Delta\ell-\frac{3}{2})\right)\right|}{\left|H\left(\frac{2\pi}{N}(\Delta\ell-\frac{1}{2})\right)\right|}\ , \quad (10)$$

then the third frequency estimation formula is obtained as:

$$\Delta\ell\simeq\frac{3}{\pi}\arctan\sqrt{3}\frac{S^{1/H}}{1+Q^{1/H}}\ , \quad 0.5\leq\Delta\ell<1.0\ . \quad (11)$$

The constants F and H in these new formulas must be optimized so as to minimize the estimation error. It should be noted that these three formulas represent all possible combinations of two magnitudes out of the three magnitudes describing the (ODFT) sampled main lobe of the FRTAW.

Attempts to build a two-rule estimation model using equation (9) if $\Delta\ell\leq 0.5$, and equation (11) if $\Delta\ell\geq 0.5$, have revealed that the optimal estimation error in the least squares (LS) sense is obtained as 0.72% of the bin width, and that the optimal estimation error in the *minmax* sense is obtained as 0.46% of the bin width. Either case represents already an improvement over the single-rule estimation model reviewed in the previous section.

These new results have also revealed that the largest estimation error resulting from the two-rule estimation model occurs in the vicinity of $\Delta\ell=0.5$. Since this is the region where the estimation error resulting from the single-rule estimation model (equation (7)) is smaller, it makes sense to build a three-rule estimation model involving equation (9) when $0.0\leq\Delta\ell\leq 0.5-\gamma/2$, equation (7) when $0.5-\gamma/2<\Delta\ell<0.5+\gamma/2$, and equation (11) when $0.5+\gamma/2\leq\Delta\ell<1.0$. A study was conducted on the optimization (in the LS sense) of the estimation error as a function of γ, which represents the width of the rule centered on $\Delta\ell=0.5$. The results of this study are depicted in Fig. 1 when γ varies in the range $[0.0,\ 1.0]$. It should be noted that when $\gamma=0.0$ our model reduces to the two-rule model (equations (9) and (11)), and that when $\gamma=1.0$ our model reduces to the single-rule model (equation (7)).

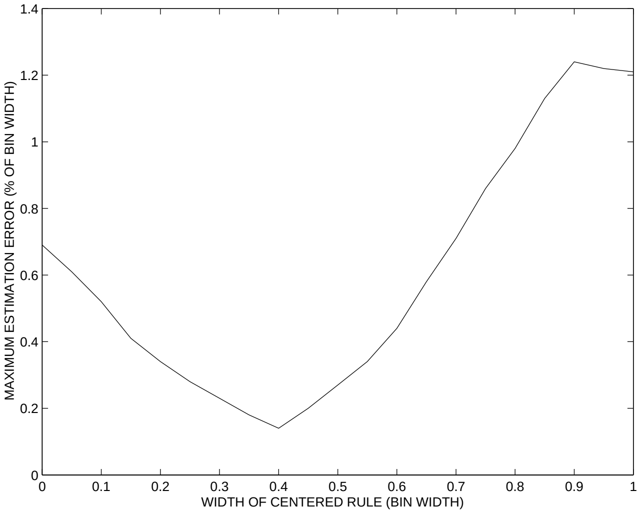

Figure 1: Maximum frequency estimation error (using LS optimization) as a function of the width of rule centered on $\Delta\ell = 0.5$.

5. PERFORMANCE OPTIMIZATION AND EVALUATION

Fig. 1 reveals that the best performance of the three-rule model is achieved when γ is near 0.4, specifically $\gamma = 0.42$, as can be concluded after a detailed analysis of the problem. Using this value, the optimization of the (global) estimation error, in the LS sense, leads to a maximum frequency estimation error of 0.14% of the bin width, when $\Delta\ell$ varies in the range $[0.0,\ 1.0[$. Using the *minmax* optimization criterion, the maximum frequency estimation error reduces even further to 0.096% of the bin width (about 0.1%), as can be seen in Fig. 2. Table 1 indicates the optimal value found for parameters G, F and H, according to the optimization criterion.

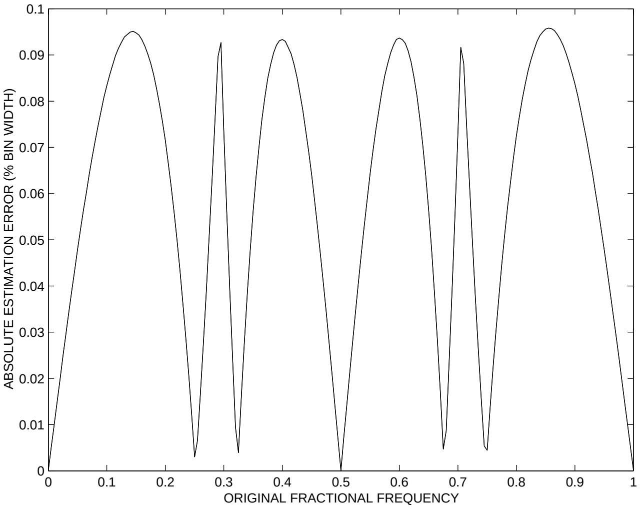

Figure 2: Frequency estimation error of the three-rule estimation model after *minmax* optimization.

The ideal and actual fractional frequency estimation functions as a function of $\Delta\ell$, are represented in Fig. 3. We have considered in this figure the optimal G, F, and H parameters resulting from the *minmax* optimization criterion. This figure reflects a significant improvement over the results presented in [9]. These results are fairly independent on the values of N, ℓ, A and ϕ, in equation (1).

Table 1: *Optimal values of the G, F, and H parameters as a function of the optimization criterion (least-squares or* minmax*).*

	LS	***minmax***
20G	29.08	29.00
20F	32.82	32.75
20H	32.82	32.75

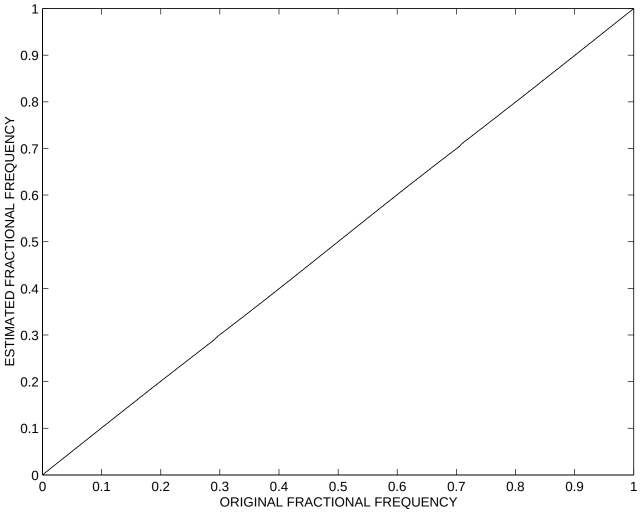

Figure 3: Ideal (dashed line) and actual (solid line) frequency estimation functions as a function of $\Delta\ell$.

In order to compare these results with other (non-iterative) peak picking and analysis algorithms, we have implemented the parabolic interpolation and triangle algorithms as described in [10]. Under test conditions similar to those used in this paper, we have concluded that the maximum fractional frequency estimation error of the parabolic interpolation algorithm is in the order of 4.4% of the bin width (which is a little better than the 5.7% error reported in [11]), while that of the triangle algorithm is in the order of 20% of the bin width. These results just confirm that the performance of a window-aware algorithm are far better than the performance of window-agnostic algorithms.

6. ROBUSTNESS TO NOISE

The influence of additive white gaussian noise ($r(n)$ in equation (1)), on the performance of the frequency estimation, has been evaluated in two cases: using the single-rule estimation model, and using the three-rule estimation model. The Signal-to-Noise ratio (SNR) relating the average power of a single sinusoid and the average power of the noise, was forced to vary between 70 dB and -10 dB. These two limits correspond to extreme situations: in the former case the influence of the noise is negligible, and in the latter case the influence of the noise is very strong. The results are represented in Fig. 4. This figure shows that, as expected, the performance of the three-rule estimation model is always superior

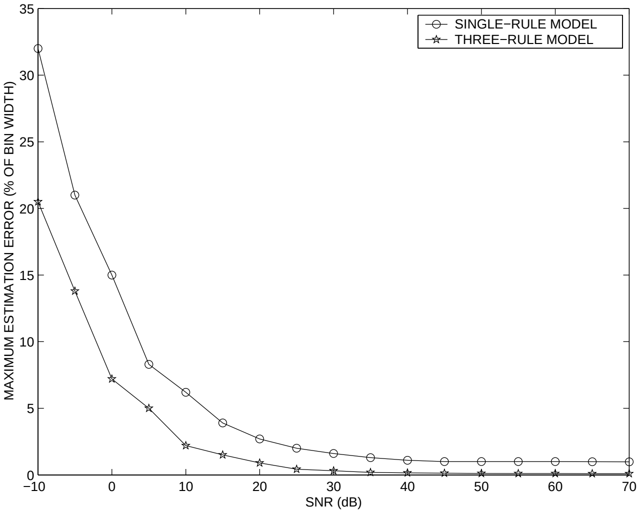

Figure 4: Performance of the frequency estimation process as a function of the SNR.

to that of the single-model estimation rule, notably when the noise is higher. In order to understand better this conclusion, Fig. 5 represents the ideal (dashed line), the actual single-rule frequency estimation function (dotted line), and the actual three-rule frequency estimation function (solid line), when the SNR is 0 (zero) dB. In can be concluded from this figure that the single-rule estimation model performs especially poorly when $\Delta\ell$ is close to 0.0 or to 1.0, while the three-model estimation rule performs without such a bias when $\Delta\ell$ varies in the range $[0.0,\ 1.0[$. This conclusion is valid for all SNR values and confirms that the three-rule estimation model is extremely robust to noise.

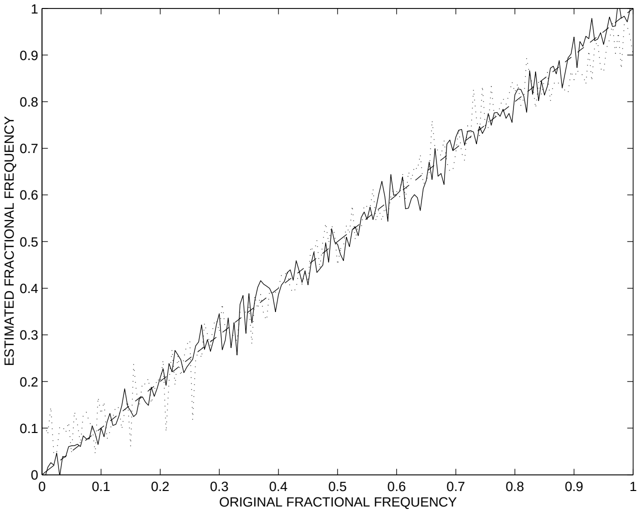

Figure 5: Ideal (dashed line), single-rule (dotted line) and three-rule (solid line) frequency estimation functions when the SNR is zero dB.

7. CONCLUSION

In this paper we have presented new results allowing to improve by an order of magnitude the performance of a previously published frequency estimation algorithm that operates in the frequency domain, that presumes a time window corresponding to the square root a shifted Hanning window, and that is non-iterative. The new algorithm was shown to be robust to the influence of additive noise, and is suited to real-time applications, notably in the context of high-quality audio analysis, coding, modification, and bandwidth extension [12].

8. REFERENCES

[1] I. Trancoso, L. Almeida, J. Rodrigues, J. Marques, and J. Tribolet, “Harmonic Coding: State-of-the-Art and Future Trends,” *Speeech Communication*, vol. 7, no. 2, pp. 239–245, July 1988.

[2] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., 1993.

[3] S. N. Levine, “Audio Representations for Data Compression and Compressed Domain Processing,” Ph.D. dissertation, Stanford University, 1998.

[4] A. J. S. Ferreira, “Spectral Coding and Post-Processing of High Quality Audio,” Ph.D. dissertation, Faculdade de Engenharia da Universidade do Porto-Portugal, 1998, http://telecom.inescn.pt/doc/phd_en.html.

[5] ——, “Efficient Intraframe Coding of Monophonic Audio,” *116th Convention of the Audio Engineering Society*, May 2004, paper 6166.

[6] K. Hamdy, M. Ali, and A. Tewfik, “Low Bit Rate High Quality Audio Coding with Combined Harmonic and Wavelet Representations,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[7] A. de Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *J. Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, April 2002.

[8] T. Nakatani and T. Irino, “Robust and accurate fundamental frequency estimation based on dominant harmonic components,” *J. Acoustical Society of America*, vol. 116, no. 6, pp. 3690–3700, December 2004.

[9] A. J. S. Ferreira, “Accurate Estimation in the ODFT Domain of the Frequency, Phase and Magnitude of Stationary Sinusoids,” in *2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 21-24 2001, pp. 47–50.

[10] F. Keiler and S. Marchand, “Survey on extraction of sinusoids in stationary sounds,” in *5th Int. Conf. on Digital Audio Effects (DAFx-02)*, 2002, pp. 51–58.

[11] J. C. Brown, “Frequency Ratios of Spectral Components of Musical Sounds,” *Journal of the Acoustical Society of America*, vol. 99, no. 2, pp. 1210–1218, February 1996.

[12] A. J. S. Ferreira and D. Sinha, “Accurate Spectral Replacement,” *118th Convention of the Audio Engineering Society*, May 2005, submitted.