HAL open science

Fourier-based Methods for the Spectral Analysis of Musical Sounds

Sylvain Marchand

To cite this version:

Sylvain Marchand. Fourier-based Methods for the Spectral Analysis of Musical Sounds. European Signal Processing Conference (EUSIPCO), Sep 2013, Marrakech, Morocco. pp.1569744487. hal-00950121

HAL Id: hal-00950121

https://hal.science/hal-00950121v1

Submitted on 20 Feb 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FOURIER-BASED METHODS FOR THE SPECTRAL ANALYSIS OF MUSICAL SOUNDS

Sylvain Marchand

University of Brest, Lab-STICC – CNRS UMR 6285, 29200 Brest, Brittany, France

ABSTRACT

When dealing with musical sounds, the short-time Fourier transform prevails and sinusoids play a key role, according to both acoustics (vibrating modes) and psychoacoustics (pure tones). The values obtained when decomposing the signal on the time-frequency atoms are usually assigned to their geometrical center, leading to estimation errors for the sinusoidal parameters. To correct this, one can exploit the amplitude or phase information, use the derivatives of the analysis window, or those of the audio signal. This leads to three methods (phase vocoder, spectral reassignment, derivative algorithm) equally efficient: they are in fact different formulations of the best analysis method based on the Fourier spectrum.

***Index Terms*—** Sound analysis, sinusoidal modeling, Fourier transform, spectral reassignment

1. INTRODUCTION

Additive synthesis can be considered as a spectrum modeling technique. It is originally rooted in Fourier's theorem, which states that any periodic function can be modeled as a sum of sinusoids at various amplitudes and harmonically related frequencies. Sinusoidal modeling consists in considering the trajectories in time of the amplitude and frequency parameters of each sinusoid present in the sound. This was proposed by McAulay and Quatieri [1] for speech signals and by Smith and Serra [2] for musical sounds. Sinusoidal modeling leads to meaningful sound representations, suitable *e.g.* for audio effects (time stretching, pitch shifting, etc.), audio coding, source separation, or music transcription.

The sinusoidal model being parametric, an important problem is to be able to estimate the model parameters as accurately as possible, to get high quality sounds. In this paper, we focus on estimators based on the Fourier spectrum and well-suited for musical sounds, although other approaches exist (*e.g.* see [3]) and other signals could also be considered.

After a presentation of sinusoidal modeling in Section 2, Section 3 describes three methods for the estimation of the sinusoidal parameters, and Section 4 shows that they are all efficient in practice and in fact equivalent in theory: they are different formulations of the best Fourier-based analysis method.

Together with Lagrange we studied in [4] the equivalence of these estimators in the stationary case and only for one sinusoidal parameter (the frequency). This paper can be regarded as an extension, where spectral reassignment plays a central role.

This research was partly supported by the ANR agency, DReaM project (ANR-09-CORD-006).

2. SPECTRAL SOUND MODELING

2.1. Sinusoidal Modeling

Let us consider here the sinusoidal model under its most general expression, which is a sum of complex sinusoids / exponentials (the *partials*) with slow time-varying amplitudes a_p and non-harmonically related frequencies ω_p (defined as the first derivative of the phases ϕ_p). The resulting signal s is thus

$$s(t) = \sum_{p=1}^{P} a_p(t) \exp(j\phi_p(t)) \tag{1}$$

where P is the number of partials. Since this paper focuses on the statistical quality of the parameters' estimators rather than their frequency resolution, the signal model is reduced to only one partial ($P = 1$). The subscript notation for the partials is then useless. Let us also define Π_0 as being the value of a given parameter Π at time 0, corresponding to the center of the analysis frame. The signal s is then

$$s(t) = \exp\left(\underbrace{(\lambda_0 + \mu_0 t)}_{\lambda(t)=\log(a(t))} + j \underbrace{(\phi_0 + \omega_0 t)}_{\phi(t)} \right) \tag{2}$$

where μ (the amplitude modulation) is the derivative of λ (the log-amplitude), and ω (the frequency) is the derivative of ϕ (the phase). Thus, the log-amplitude and the phase are modeled by polynomials of degree 1, which can be viewed either as truncated Taylor expansions of more complicated amplitude and frequency modulations (*e.g.* tremolo / vibrato), or either as an extension of the stationary case where $\mu_0 = 0$.

2.2. Spectral Analysis

The main problem we have to tackle now is the estimation of the model parameters. This can be achieved, as in the stationary case, by using the short-time Fourier transform (STFT):

$$S_w(t, \omega) = \int_{-\infty}^{+\infty} s(\tau) w(\tau - t) \exp\left(-j\omega(\tau - t)\right) \, d\tau \tag{3}$$

where S_w is the short-time spectrum of the signal s.

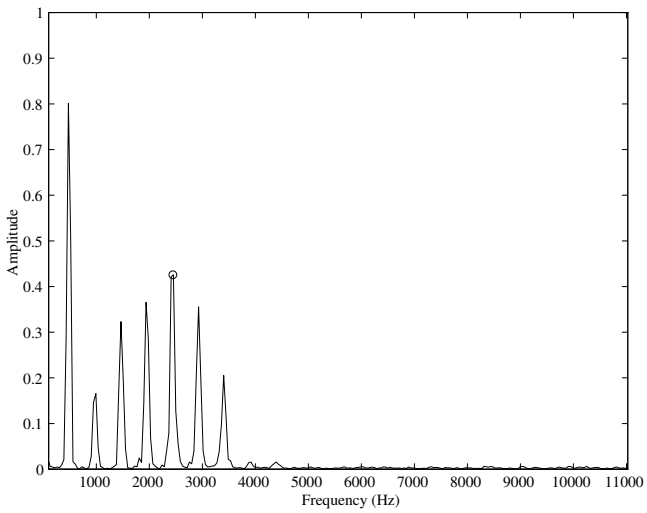

Fig. 1. Magnitude spectrum of a harmonic sound, resulting from a fast (discrete) Fourier transform (FFT). Each partial causes a peak (*e.g.* see ∘) and will be considered individually.

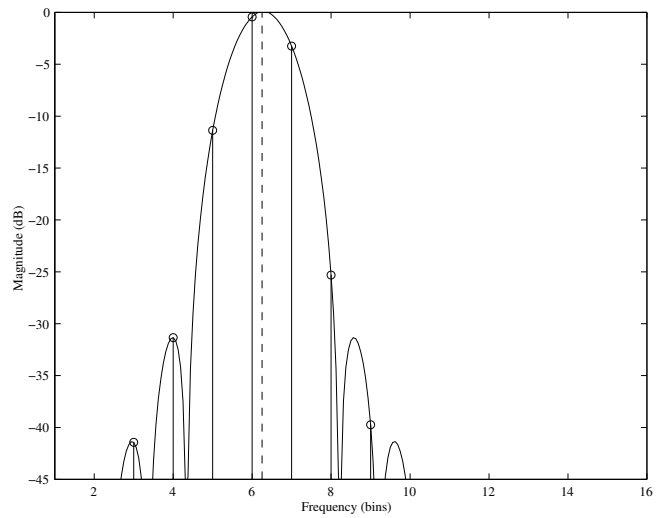

Fig. 2. Zoom on one peak: The analyzed sinusoid (dashed line) is observed from the spectrum of the analysis window through the discrete frequencies (solid lines) of the Fourier transform, leading to several bins with significant energy (∘).

Note that, as in [5], we use here a slightly modified definition of the STFT. Indeed we let the time reference slide with the window, which is also the case in practice when the STFT is implemented using a sliding fast Fourier transform (FFT).

For the sake of simplicity, all the mathematical derivations will be done in the continuous domain. However, in practice the signals are discrete (with some sampling frequency F_s). The Fourier transform (FT) will then be replaced by its discrete version (DFT) of size N, and the time will be expressed in samples (sample n being at time n/F_s) and the frequency in bins (bin m being at frequency mF_s/N).

Many instrumental sounds are harmonic, meaning that the frequencies of the partials are multiple of some fundamental frequency (related to our perception of pitch). The magnitude of the Fourier spectrum exhibits then a series of peaks (see Fig. 1). Each peak is a local maximum m in the magnitude spectrum and corresponds to some partial.

S_w involves an analysis window w, usually symmetric and band-limited in such a way that for any frequency corresponding to one specific partial, the influence of the other partials can be neglected (in the general case when $P > 1$).

In the stationary case ($\mu_0 = 0$), the spectrum of the analysis window gets simply centered on the frequency ω_0 and multiplied by the complex amplitude

$$s_0 = a_0 \exp(j\phi_0) = \exp(\lambda_0 + j\phi_0), \tag{4}$$

as shown in Fig. 2, which can be regarded as a zoom on one of the peaks of the preceding figure. In the non-stationary case however, considering Equation (3) at estimation time 0, we see that s_0 gets multiplied by $\Gamma_w(\omega_0 - \omega, \mu_0)$ where

$$\Gamma_w(\omega, \mu_0) = \int_{-\infty}^{+\infty} w(t) \exp\left(\mu_0 t + j\omega t\right)\, dt. \tag{5}$$

In the special case of using a Gaussian window for w, an analytic formula can be derived. Else, it is always possible to compute Γ_w directly from Equation (5).

Once the estimated amplitude modulation $\hat{\mu}_0$ and frequency $\hat{\omega}_0$ are known, the amplitude and phase parameters can eventually be estimated since

$$\hat{s}_0 = \frac{S_w(0, \omega_m)}{\Gamma_w(\hat{\omega}_0 - \omega_m, \hat{\mu}_0)}, \tag{6}$$

where ω_m is the (discrete) frequency of the local maximum of the magnitude spectrum where the partial is detected.

3. SINUSOIDAL ESTIMATION

The problem is yet to estimate the amplitude modulation and frequency parameters. In this section, we present 3 methods providing estimation functions $\hat{\mu}$ and $\hat{\omega}$ for these parameters. In practice, for each detected partial at time 0 (center of the analysis frame) and (discrete) frequency ω_m, the estimates of its parameters are given by:

$$\hat{\mu}_0 = \hat{\mu}(0, \omega_m), \tag{7}$$
$$\hat{\omega}_0 = \hat{\omega}(0, \omega_m). \tag{8}$$

3.1. Difference Method (Phase Vocoder)

The Fourier-based approach started together with computer music, about 50 years ago. The phase vocoder introduced by Flanagan and Golden [6] was already using the phase of the Fourier spectrum to estimate the frequency of the partials, and more precisely the phase difference of consecutive spectra [7]. This simple yet efficient difference approach was generalized recently to the non-stationary case [8].

Thus $S_w(t,\omega)$ is the spectrum of the frame centered at the desired (discrete) estimation time, and let $S_w^{\mp}(\omega) = S_w(t \mp 1/F_s, \omega)$ be its left (previous, *i.e.* one sample before) and right (next, *i.e.* one sample after) neighboring spectra, respectively (F_s denoting the sampling frequency).

Since the log-amplitude and phase differences correspond to the real and imaginary parts of the logarithm of spectral ratios, respectively, let us define:

$$\begin{aligned} \Delta_\lambda(S_1,S_2) &= \log|S_1| - \log|S_2| \\ &= \Re\left(\log\left(S_1/S_2\right)\right), \qquad (9) \\ \Delta_\phi(S_1,S_2) &= \angle S_1 - \angle S_2 \\ &= \Im\left(\log\left(S_1/S_2\right)\right) \qquad (10) \end{aligned}$$

(S_1 and S_2 denoting two arbitrary complex spectra).

Since we can measure the amplitude of the spectra, we can compute the left and right estimates of the amplitude modulation, and retain their mean as the final estimation:

$$\begin{aligned} \mu_- &= \Delta_\lambda(S_w, S_w^-)\cdot F_s, \\ \mu_+ &= \Delta_\lambda(S_w^+, S_w)\cdot F_s, \\ \hat{\mu} &= (\mu_- + \mu_+)/2. \qquad (11) \end{aligned}$$

Similarly, with the measured phase of the spectra, we can compute an estimation of the instantaneous frequency:

$$\begin{aligned} \omega_- &= \Delta_\phi(S_w, S_w^-)\cdot F_s, \\ \omega_+ &= \Delta_\phi(S_w^+, S_w)\cdot F_s, \\ \hat{\omega} &= (\omega_- + \omega_+)/2. \qquad (12) \end{aligned}$$

In practice, we are looking for positive frequencies and since the phase is measured modulo 2π, after each call to the Δ_ϕ function we must apply the phase unwrapping procedure of the phase vocoder, *i.e.* adding 2π to the result if lower than 0.

3.2. Spectral Reassignment

Reassignment was first proposed by Kodera *et al.* [9] and was generalized by Auger and Flandrin [10] to improve time-frequency representations. Usually, the values obtained when decomposing the signal on the time-frequency atoms are assigned to the geometrical center of the cells (center of the analysis window and bins of the Fourier transform). The reassignment method assigns each value to the center of gravity of the cell's energy. The method uses the knowledge of the first derivative w' – obtained by analytic differentiation – of the analysis window w in order to adjust the frequency inside the Fourier transform bin. This approach was generalized for the amplitude modulation in the non-stationary case (see [5]).

The complex short-time spectrum – resulting from the STFT – is, in the polar form:

$$S_w(t,\omega) = a(t,\omega)\exp\left(j\phi(t,\omega)\right) \qquad (13)$$

where the instantaneous amplitude a and phase ϕ are real-valued functions of time t and frequency ω. By considering Equation (3), we can easily derive:

$$\frac{\partial}{\partial t}\log\left(S_w(t,\omega)\right) = j\omega - \frac{S_{w'}(t,\omega)}{S_w(t,\omega)} \qquad (14)$$

where w' denotes the derivative of w. Then, since the amplitude modulation (resp. frequency) is the derivative of the amplitude (resp. phase), from Equations (13) and (14), we obtain the reassigned parameters:

$$\begin{aligned} \hat{\mu}(t,\omega) &= \frac{\partial}{\partial t}\Re\left(\log\left(S_w(t,\omega)\right)\right) \\ &= -\Re\left(\frac{S_{w'}(t,\omega)}{S_w(t,\omega)}\right), \qquad (15) \\ \hat{\omega}(t,\omega) &= \frac{\partial}{\partial t}\Im\left(\log\left(S_w(t,\omega)\right)\right) \\ &= \omega - \Im\left(\frac{S_{w'}(t,\omega)}{S_w(t,\omega)}\right). \qquad (16) \end{aligned}$$

3.3. Derivative Method

Together with Desainte-Catherine in [11], we proposed to use the signal derivatives to estimate the sinusoidal parameters in the stationary case; and with Depalle in [5], we generalized this derivative method to the non-stationary case. Indeed, considering Equation (2), since the derivative of an exponential is an exponential, we have:

$$s'(t) = (\mu_0 + j\omega_0)\cdot s(t) \quad \text{and thus}$$
$$\Re\left(\frac{s'}{s}\right) = \mu_0 \quad \text{and} \quad \Im\left(\frac{s'}{s}\right) = \omega_0.$$

For this method to work in the case of a signal made of several partials, we have to switch to the spectral domain and define:

$$\hat{\mu} = \Re\left(\frac{S'_w}{S_w}\right), \qquad (17)$$

$$\hat{\omega} = \Im\left(\frac{S'_w}{S_w}\right), \qquad (18)$$

where S'_w is the short-time spectrum of the signal derivative s'. As shown in [5], in practice this (discrete) derivative s' can be obtained by convolving the discrete signal s by the following differentiator filter:

$$h[n] = F_s\frac{(-1)^n}{n} \text{ for } n \neq 0, \text{ and } h(0) = 0 \qquad (19)$$

of infinite time support. Thus, in practice, we multiply h by a (finite-length) Hann window. This results in a high-pass filter, and can lead to estimation problems in the high frequencies (above approx. $3/4$ of the Nyquist frequency), fortunately above the audible limit (16kHz).

So, for each partial m detected in the (discrete) Fourier spectrum at time t and frequency ω_m, together with Equations (7) and (8), we have now 3 ways to estimate the amplitude modulation and frequency parameters of the partial:

- the difference approach with Equations (11)-(12),
- the reassignment approach with Equations (15)-(16),
- the derivative approach with Equations (17)-(18).

Once these parameters are known, the others (amplitude and phase) can be estimated in turn using Equation (6).

4. COMPARING THE ESTIMATORS

Now the question is: Which is the best approach?

4.1. Experimental Results

To quantitatively evaluate the precision of theses approaches for the estimation of all the model parameters, we ran the same experiments as in [5, 8].

We consider discrete-time signals s, with sampling rate F_s, each consisting of 1 complex exponential generated according to Equation (2) with an initial amplitude $a_0 = 1$, and mixed with a Gaussian white noise. In our experiments, we set the sampling frequency $F_s = 44100$Hz, the FFT size $N = 511$, and the signal-to-noise ratio (SNR) goes from -20dB to $+100$dB by steps of 5dB. For each SNR and for each analysis method, we test several parameter combinations: 99 frequencies (ω_0) linearly distributed in the $(0, 3F_s/8)$ interval, 9 phases (ϕ_0) linearly distributed in $(-\pi, +\pi)$, and 5 amplitude modulations (μ_0) linearly distributed in $[-100, +100]$. For the analysis window w, we use the symmetric Hann window. We focus on the variance of the estimation error over this test set (the mean being zero for unbiased estimators). We then compare the difference method (D), the reassignment method (R), and two variants of the derivative method: The estimated derivative method (ED), where the derivative s' is estimated using the differentiator filter of Equation (19) of size 1023; and the theoretic derivative method (TD), where the exact derivative is used for s' – since it is analytically known for the test signals. The results of the TD method can be regarded as the best performance the ED method could achieve, at the expense of a longer differentiator filter though.

When looking at the results of these experiments (see Fig. 3), we see that all these methods are very efficient, close to the the Cramér-Rao bounds (CRB), which are the limit to the best possible performance achievable by an unbiased estimator given a data set (see [5]).

In the high SNRs, the ED method is biased because of the approximation of the derivative by the finite-length differentiator filter. Applying the spectral reassignment on the discrete spectrum causes a bias, as noticed by Hainsworth [12], degrading the performances of the R method. Perhaps surprisingly, the simplest D method is the most efficient.

4.2. Theoretical Equivalences

4.2.1. Reassignment and Derivative

In [4], the reassignment (Section 3.2) and derivative (Section 3.3) methods were proven to be theoretically equivalent, at least as regards the estimation of the frequency in the stationary case. In [5], we generalized the proof of the equivalence to the non-stationary case, and for the estimation of both the frequency and the amplitude modulation. More precisely, we introduce $\rho = \tau - t$ which gives another (equivalent) expression for the STFT (see Equation (3)):

$$S_w(t,\omega) = \int_{-\infty}^{+\infty} s(t+\rho)w(\rho)\exp\left(-j\omega\rho\right)\, d\rho \qquad (20)$$

from which we can derive

$$\frac{\partial}{\partial t}\log\left(S_w(t,\omega)\right) = \frac{S'_w(t,\omega)}{S_w(t,\omega)}. \qquad (21)$$

By considering Equation (21) instead of Equation (14) in Section 3.2 (reassignment approach), we would have obtained the equations of Section 3.3 (derivative approach). Thus the reassignment approach is equivalent to the derivative approach, at least in the continuous case. A different proof, based on integration by parts, can be found in [13].

4.2.2. Reassignment and Difference

It turns out that the reassignment approach is also equivalent to the difference approach in the discrete case. Indeed, in the phase vocoder approach the parameters are estimated by first-order differences, approximating the differentiation of the spectrum of Equations (15)-(16) in the discrete-time case.

5. CONCLUSIONS

Although the three approaches we presented in Section 3 are equivalent in theory, the small differences observed in practice in Section 4.1 are due to a bias of the reassignment method in the discrete case, and of the derivative method when using a finite-length differentiator filter. The most efficient approach turns out to be the simplest one, based on first-order differences, *i.e.* a rather crude approximation of differentiation... But further investigations are necessary, since in theory the reassignment method (resp. the derivative method) requires the analysis window (resp. time signal) to be differentiable, which are *a priori* different conditions.

6. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. on Acous., Speech, and Sig. Proc.*, vol. 34, no. 4, pp. 744–754, 1986.

[2] J. O. Smith III and X. Serra, "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds based on a Sinusoidal Representation," in *Proc. Int. Computer Music Conf.*, 1987, pp. 290–297.

[3] R. Badeau, G. Richard, and B. David, "Performance of ESPRIT for Estimating Mixtures of Complex Exponentials Modulated by Polynomials," *IEEE Trans. on Sig. Proc.*, vol. 56, no. 2, pp. 492–504, 2008.

[4] S. Marchand and M. Lagrange, "On the Equivalence of Phase-Based Methods for the Estimation of Instantaneous Frequency," in *Proc. 14th European Conf. on Sig. Proc.*, 2006.

[5] S. Marchand and Ph. Depalle, "Generalization of the Derivative Analysis Method to Non-Stationary Sinusoidal Modeling," in *Proc. Int. Conf. on Digital Audio Effects*, 2008, pp. 281–288.

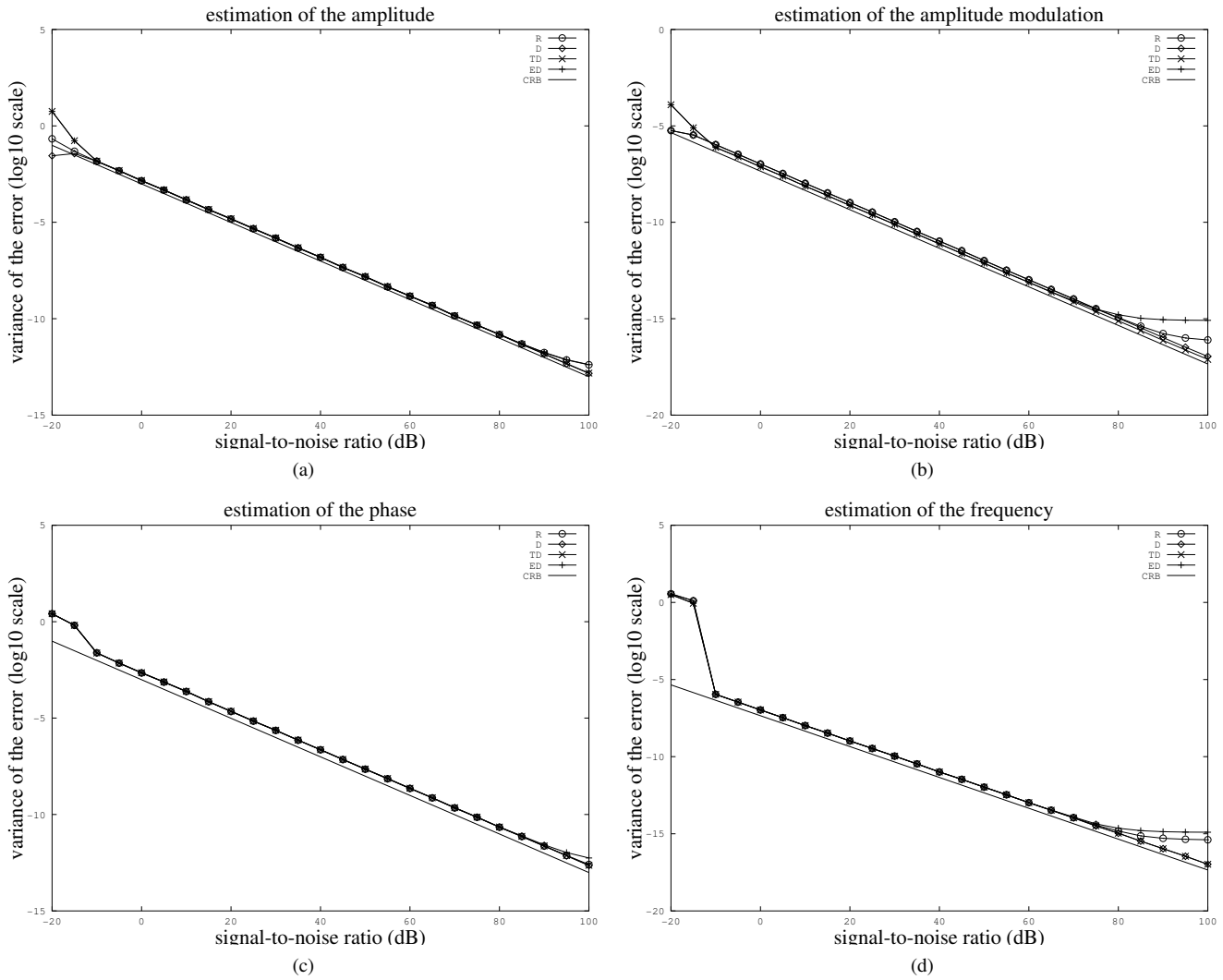

Fig. 3. *Estimation errors as functions of the SNR for the amplitude (a), amplitude modulation (b), phase (c), and frequency (d), with the reassignment (R), difference (D), theoretical derivative (TD), and estimated derivative (ED) methods, and with comparison to the Cramér-Rao Bounds (CRB).*

[6] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Tech. Journal*, vol. 45, pp. 1493–1509, 1966.

[7] M. B. Dolson, "The Phase Vocoder: A Tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[8] S. Marchand, "The Simplest Analysis Method for Non-Stationary Sinusoidal Modeling," in *Proc. Int. Conf. on Digital Audio Effects*, 2012, pp. 23–26.

[9] K. Kodera, R. Gendrin, and C. de Villedary, "Analysis of Time-Varying Signals with Small BT Values," *IEEE Trans. on Acous., Speech, and Sig. Proc.*, vol. 26, no. 1, pp. 64–76, 1978.

[10] F. Auger and P. Flandrin, "Improving the Readibility of Time-Frequency and Time-Scale Representations by the Reassignment Method," *IEEE Trans. on Sig. Proc.*, vol. 43, no. 5, pp. 1068–1089, 1995.

[11] M. Desainte-Catherine and S. Marchand, "High Precision Fourier Analysis of Sounds Using Signal Derivatives," *Journal of the AES*, vol. 48, no. 7/8, pp. 654–667, 2000.

[12] S. W. Hainsworth, *Techniques for the Automated Analysis of Musical Audio*, Ph.D. thesis, University of Cambridge, United Kingdom, 2003.

[13] X. Wen and M. Sandler, "Notes on Model-Based Non-Stationary Sinusoid Estimation Methods Using Derivatives," in *Proc. Int. Conf. on Digital Audio Effects*, 2009.