Automatic Amplitude Estimation Strategies for CBM Applications

Thomas L LAGÖ
Tech Fuzion, P.O. Box 971, Fayetteville, AR, 72701, USA; Phone: +1 479 571 0035; e-mail: thomas.lago@techfuzion.com

Abstract
In many CBM applications, the basic assumptions for the analysis used is forgotten or “ignored.” This can lead to large amplitude errors and completely mislead the analysis. This will most likely be devastating for prognosis since the amplitude errors are likely to be very large. This paper describes some of these errors and how this can lead to wrong conclusions and or results. In this paper, an example of 100,000 per cent amplitude error is presented and this actually happens in real life. By understanding these errors and why they exist, it is possible to mitigate and/or minimize their impact on the end results. The challenge related to that signals analysed do not obey the proper properties and hence large amplitude errors could become the result can be handled automatically. The background to the challenges is discussed and proper solutions and strategies to avoid such errors are presented. The automatic method will tell which of the frequency lines can be trusted from an amplitude accuracy point of view.

Keywords: Signal Processing, sampling, FFT, estimation, CBM.

1. Introduction

Estimation of the amplitude of an unknown signal is one of the basic requirements in measuring techniques. A-priori knowledge about the signal is often given before the measurement takes place. One example is when using a Digital Voltmeter (DVM), and the signal is assumed to be sinusoidal. Another could be when the signal is “broadband” or “noise like” and the power is measured using an octave filter. These assumptions about the signal bandwidth have to be fulfilled if the correct amplitude is to be estimated. This is a complex problem and a good understanding of the measurement situation, is therefore essential.

In the literature for frequency analysis and measurement techniques, a simplification of the signal type is made to avoid complexity, [1][2][3]. It is common to use either the sinusoidal approach or the broadband noise approach, [4]. If the signal belongs to one of these classes, several difficulties disappear since a-priori information about the signal is inherent. If the signal does not belong to one of these classes, serious measurement errors can result. If relative measurements are made, the errors are usually small, since no changes are made on the instrument side. In general, it is impossible to correctly measure a completely unknown signal without a risk of severe measurement errors. In theory, we can make sure we have one type but in real life we often have a combination of multiple signal types. It is common that this challenge is ignored and a PSD or Power Spectra is used. This will most likely lead to amplitude errors.

2. Frequency Domain Analysis

When performing FFT analysis, it is not always fully understood that the base assumption for FFT analysis is based on a deterministic signal assumption as described by Figure 1 below. Fourier analysis can only be used on deterministic and repetitive signals. If signals belong to another class than the deterministic and repetitive one (discrete components), large amplitude scaling errors are likely to happen. If we know that the signal is “broad band” like e.g. noise,
we can scale the data using power spectral density and hence compensate for this error. The same applies if we know that we have e.g. ergodic data or transients. However, it is rarely the case. In CBM applications, a combination of signals is the most common case and all levels (different peaks in the data) cannot be properly scaled unless they belong to the same category. Hence, some of the peaks in the data can have rather larger amplitude errors.

![Diagram of signal classes]

Figure 1. Illustration of the main signal classes. The “green categories” can be analysed by an FFT and described using a Fourier methodology. The “red” categories are very challenging.

3. Measurement Setup

The measurement setup that has been used in the tests is given in figure 2 below. The Agilent Technology’s (former Hewlett-Packard) Dynamic Signal Analyser HP35670A is a common instrument in sound and vibration work worldwide, and the used setup is typical and many engineers use it every day. A completely unknown signal is fed to the analyser through a coaxial cable. Despite the fact that the signal is well within the analyser’s frequency and dynamic range, negligible aliasing and very good signal conditioning, it is not possible to accurately estimate the amplitude of the unknown signal/signals in the cable. This is difficult for many engineers to understand, but this is a very important fact. Some a-priori information has to be given in order to succeed. It is possible to be “lucky” and manage an accurate measurement, but it is more luck than anything else, given that we have no information whatsoever in regards to the signal.

![Illustration of measurement setup]

Figure 2. Illustration of the measurement setup. An HP35670 DSA has been used to collect time data and perform the frequency analysis using an FFT.
When analysing the unknown signal, the default instrument setup in the analyser has been used: 400 frequency lines, DC-51.2 kHz frequency span, a Hanning window and one FFT calculation (no averaging). The results from the measurement using this setup are depicted in figure 3 and figure 4. At first sight, it seems like the signal consists of a sinusoid at 15 kHz with a distortion component at 45 kHz. However, on analysing the picture in more detail using the markers, it is found that the second peak is located at 46 kHz, not at 45 kHz. Therefore, this peak cannot be a distortion component. Are the first and second peaks really sinusoids?

![Figure 3](image.png)

Figure 3. A frequency analysis on an unknown signal. Two peaks are visible, one at 15 kHz and one at about 45 kHz. There is a 12 dB difference in signal level according to the display markers.

In Figure 3, two tones are visible but only one exists. The “tone” to the right is about -25 dB according to the display. However, the one to the right is really not a tone. If we try to zoom in on the signal, it will “disappear” despite clearly being there in this figure. Since we often base diagnosis on signal and/or component amplitudes, it is essential to scale them properly.

In Figure 4, a zoom has been performed covering the “peak” frequency range that was visible before. The peak is “gone” and levels have decreased to almost -51 dB. This compares to a large amplitude error and in this case it is about 100,000 per cent. If we make 100,000 per cent amplitude errors, this is bad, especially if we use this as the foundation for prognosis.

![Figure 4](image.png)

Figure 4. A zoom analysis on the unknown signal’s second peak. Now, the amplitude level is -51 dB and the signal seems to have disappeared since there is no peak.
By changing the time window for the frequency analysis, the second peak changes its amplitude by more than 30%, but not the first peak. This is puzzling. The situation is not satisfactory but realistic. One could of course avoid this situation by performing only one measurement, reading the markers and be happy. However, if an accurate measurement is the aim, a better understanding of the principles of measurement is necessary.

What is right? Is there a peak to the right or not? This is very difficult to tell at this stage even though it looks like it on the screen. Therefore, most engineers would have assumed that they are sinusoids. One idea could be to use time domain to see if this gives a clue. The time series will show one sinusoid not two. However, the frequency analysis views two tones not one. By changing window for the frequency analysis the second peak changes its amplitude with more than 30% but not the first peak. This is puzzling. The situation is not good, but real. One could of course avoid this situation by performing only one measurement, reading the markers and be happy. However, if an accurate measurement is the goal, a better understanding of the measurement principles has to be considered.

Another important tool that usually gives a good indication of signal type is the histogram. Figure 5 illustrates a histogram analysis. This result indicates that the signal consists mainly of one sinusoid. However, the frequency analysis clearly shows two peaks. Fig. 5 illustrates a Matlab simulation of two sinusoids, with both the time signal as well as the histogram presented.

![Figure 5. Histogram representation of some typical signals.](image)

4. **Correct Amplitude Scaling**

The main reason for the above problem is that the signal at 46 kHz is not a sinusoid, but the signal at 15 kHz is. The 46 kHz component consists of narrowband noise at a level of almost 30 dB lower than the analyser shows, given the analysis bandwidth used. It is important to note that the problem lies not with the analyser but the user. The measurement error in this case is 27 dB, equivalent to a measurement error of almost 100,000%! This problem is due to the fact that the analyser must be set using the correct scaling method according to the input signal type. The following scaling methods must be used:

- Tonal components (sinusoids): Use PPS scaling (Power Spectra)
- Broadband signals: Use PPSD scaling (Power Density Spectra)
- Transients: Use PESD scaling (Energy Density Spectra)
If these rules are not followed serious amplitude scaling errors will result, unless an analysis bandwidth of 1 Hz is used. This is how several text books avoid the problem, [4][5][6]. On the first page they assume that the rest of the book is based on a 1 Hz bandwidth. In this case these scaling problems can be neglected. In real life, it is very rare for the analyser to use 1 Hz bandwidth. Therefore, the scaling problem must be addressed properly. Some text books state that all signals are “noise like”, narrowband or broadband, [1][2][3]. The histogram for a sinusoid will be different from that for a narrowband noise signal. Therefore, the histogram could be a good indication when determining what signal type the signal belongs to. However, most real life signals are composed of several types making it difficult to classify the signal as belonging to one type only. The dynamic signal analyser has a facility for performing a histogram calculation, a tool underestimated by many engineers.

It can be helpful to take a look at the histogram for the signal. On the above signal, a small increase in the middle of the histogram would hint us on the fact that the signal does contain more than just sinusoids. If this is not the case, frequency information may leak from one frequency line to another. The window will thus reduce leakage, but also change the analysis bandwidth. There is a trade-off between time signal energy, analysis bandwidth, picket fence effect (amplitude ripple), side lobes and spectral leakage. There are several windows available in most commercial Dynamic Signal Analysers, but the most commonly used in industrial measurements are: No Window (Rectangular), Hanning, Flat top and Force-Exponential. It is however important to note that there are several flat top windows. The P401 developed by Hewlett Packard (today Agilent Technologies) has lower side lobes than its predecessor P301, which is the flat top window most often used in general measurement equipment. Different key windows are presented with their key parameters in Table 1 below, [7].

<table>
<thead>
<tr>
<th>Window</th>
<th>Amplitude error</th>
<th>Bandwidth, BW</th>
<th>First Sidelobe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hanning</td>
<td>1.43 dB</td>
<td>6 Hz</td>
<td>-31.5 dB</td>
</tr>
<tr>
<td>Hamming</td>
<td>1.75 dB</td>
<td>5.5 Hz</td>
<td>-43.2 dB</td>
</tr>
<tr>
<td>Flat top, P301</td>
<td>0.01 dB</td>
<td>13.7 Hz</td>
<td>-70.4 dB</td>
</tr>
<tr>
<td>Flat top, P401</td>
<td>0.01 dB</td>
<td>15.3 Hz</td>
<td>-82.1 dB</td>
</tr>
<tr>
<td>Rect, no window</td>
<td>3.94 dB</td>
<td>4 Hz</td>
<td>-13.2 dB</td>
</tr>
</tbody>
</table>

It is very clear that the spectral resolution is good with a Hanning window, and the amplitude accuracy is best for the flat top window. It does not make a big difference which type of Flattop windows is used in terms of amplitude accuracy or ripple. The big difference in between them has to do with the side lobes. This will indirectly affect the amplitude accuracy due to possible coupling to nearby frequency components. In this exercise, such effects have been assumed to be negligible. The frequency domain signal consists of a real and an imaginary part, and has negative frequencies. In real life, there are no negative frequencies and thus, a Power Spectrum with only positive frequencies is required. In the above equations, the voltage dimension is included. There are three scaling methods to choose from. This is a difficult choice, since it requires a-priori information about the signal before
choosing the right type of scaling. Without this information it is impossible to be sure that the amplitude is correct, [8][9][10]. In [4] it is stated that $P_{PS}$ and $P_{PSD}$ are the same. This is a correct statement given the assumption as in the book, of a 1 Hz analysis bandwidth. In most practical cases the analysis bandwidth is not 1 Hz. An amplitude scaling error will thus be the consequence, often of several 1000%. The above discussion shows that it is important to have a-priori information about the signal in order to evaluate absolute amplitude levels. It is not possible, in general, to find absolute amplitude levels without some knowledge about the signal. This information can be achieved by using a set of measurements and then using that information as the basis for further action.

5. TRANSIENTS

When handling transients, another complication with the FFT calculation comes into play. The FFT assumes stationarity and in real life, no signal is stationary. Thus we often redefine the assumption and assume that the signal obeys quasi-stationary properties. This means that the signal is not allowed to vary during the collection of the time record. If that happens, the absolute amplitude level will be wrong. Also, if the frequency changes during collection of the time block, a smearing of the frequency content will be the result. Many signals that are collected in vehicle applications do not fulfil the needed assumption and special care must be taken when deciding on the amplitude level. Since the record length of the FFT analyser changes with frequency range, the smaller frequency range to be analysed, the more stationarity restrictions on the signal. In figure 6, a sinusoidal burst signal with duration 10% of the time record has been measured. The amplitude will thus be off by a factor of 10, since the FFT divides the amplitude with the length of the time record, [5]. The amplitude error when using PS scaling is thus 20 dB, a major error. If the frequency range is changed, the time record will change and the error change since the signal duration will change from 10% to some other value, up or down, depending on change in frequency range. A common practice to at least avoid the change with frequency range is to use the ESD, Energy Spectral Density scaling. In this case the time record is multiplied with [s] or time, and thus the record length change is normalized. That does not mean that the amplitude for a transient is known in absolute terms, but the amplitude level will be independent of then analyser settings. This is one step in the right direction. When analysing transients it is of utmost importance to really start thinking and asking yourself what the amplitude levels presented by the FFT analyser really means. There is also a misconception that, especially for transients, the sparse sampling in an FFT analyser will make the amplitude results in time domain incorrect. That is not fully correct. However, due to the sparse sampling, all amplitude information is not visible but can be made visible. Common techniques would be using time domain interpolation via a zero-padded FFT or a straight digital filter interpolation algorithm.

6. RECOMMENDED MEASUREMENT PROCEDURE

When absolute amplitude levels are important and the signal is unknown, it is important to start with a measurement, enabling the classification process of the signal. The FFT process and the analyser is deceiving, but it is not really possible to determine the signal type from one measurement alone. Thus, several measurements must be performed and a proper classification will be based on at least two measurements. Start the analyser with a Power Spectrum scaling method (often default). Then,

1. Measure with one frequency range. Read all amplitude levels for all peaks.
2. Measure the same signal again, but with a factor of two decrease in frequency span. This gives twice the measurement time and consequently half the measurement bandwidth BW. If some amplitude levels (peaks) change levels when compared to the previous measurement, the signal cannot be scaled correctly using P<sub>PS</sub>.

3. Continue to change the frequency range until the amplitude levels are stable and do not change when the measurement settings are changed. When this happens, the amplitude values can be read, and they have the correct amplitude scaling.

![Power Spectrum and Power Spectral Density](image)

Figure 6. Illustration of the amplitude situation for a P<sub>PS</sub> and a P<sub>PSD</sub> scaling. In the left figure the filter should give the true value, irrespective of the width of the analysis filter. In the right figure, it is necessary to compensate for the width of the analysis filter since the output will be the sum of all components within that filter.

If the signal keeps changing for each change of frequency range, try using P<sub>PSD</sub> scaling instead. If the levels do not change when using a P<sub>PSD</sub> scaling, the amplitude levels are correctly scaled. Observe that there are signals where it is not possible to reach a solution for either P<sub>PS</sub> or P<sub>PSD</sub>. In such cases, it is difficult to rely on the amplitude values. A rule of thumb when determining the amplitude scaling method is:

- If the analysis bandwidth ≪ the signal bandwidth: Use P<sub>PSD</sub> scaling (=broadband signals).
- If the analysis bandwidth ≫ the signal bandwidth: Use P<sub>PS</sub> scaling (=tonal components).
- If the input signal is transient: Use P<sub>ESD</sub> scaling (Energy Density) (=transients).

**ALWAYS perform at least two measurements with different analysis bandwidths and compare. If they are equal, the right amplitude scaling is being used!**

The figure to the left in figure 6 illustrates how the scaling is correct using P<sub>PS</sub>, since there is no compensation for the width of the filter. If there is more than one signal component in the left marked filter, then the amplitude is wrong. The inverse is applicable for the figure to the right in figure 6. In this case, it is necessary to compensate for the width of the filter. If no compensation is made, the amplitude will be scaled incorrectly. That is why the signal would be overestimated if the analysis bandwidth BW is larger than 1 Hz, otherwise underestimated.
7. Automated Method

Since individual components can have different signal characteristics, it is helpful if we can see this directly. There are methods, where it is possible to quickly determine if signal components belong to the proper category. This is a typical example:

- Calculate three FFTs with resolutions: 400, 2400, 9600 lines.
- Overlay the data in the same plot using different colours for the three plots.
- Assume “tonal” or deterministic ($P_{PS}$ scaling). If all three overlay properly on all components of interest, the assumption was right.
- If the signal instead is “random” or ergodic, it will drop in amplitude every time the resolution is increased.

In Figure 7, an example of such “multiple analysis” is performed. It is clear that the first component has the same value every time and the second “peak” has not. Hence, the second peak is not a peak and we need to investigate further before we can tell its real amplitude.

![Figure 7. Example of three FFT analysis overlaid in the same plot, showing which components are correctly estimated and which ones are not.](image)

The above methodology is easy and gives a good handle on if my assumptions are right – deterministic signals. However, it is most likely that several of the peaks of interest will change their amplitude values when I change resolution. In that case, the amplitude value has not been properly described or estimated. This approach can be fully automated and all lines that are correct are plotted using a “green” colour and the wrongly estimated are plotted using a “red” colour.

8. Conclusions

Spectral estimation using FFT analysers such as a Dynamic Signal Analyser is very common. If these analysers are used without proper knowledge of the analysed signal, serious amplitude scaling errors may be the result. This has nothing to do with the instrument. Depending on whether the signal is sinusoidal, noise or transient, different scaling methods must be chosen. This requires the user to push a key: Power Spectra, Power Spectral Density or Energy Spectral Density. With the wrong scaling method, errors of several thousand per
cent can occur. There are also real life signals that are sinusoidal for one component, but broadband for others. In these cases, it is very important to continue adjusting the resolution until the amplitude levels are steady. If not, some components will be estimated with large errors. Several examples from real life measurements illustrate that this is far from being an academic problem. Most text books avoid the problem by assuming that the analysis bandwidth is 1 Hz whereby the scaling problem does not exist. A recommended analysis technique when performing spectral estimation where amplitude is of importance has been proposed. If this approach is followed, the amplitude errors will be controlled.

References