Analytical Solution of Spectrum Changes in Simple Nonlinear Systems without Memory, Used in Digital Audio Signal Processing

JIRI SCHIMMEL, JIRI MISUREC
Department of Telecommunications FEEC
Brno University of Technology
Purkynova 118, 612 00 Brno
CZECH REPUBLIC

Abstract: This paper deals with an analytic solution of spectrum changes that appear in nonlinear discrete systems without memory whose transfer characteristics can be approximated via a piecewise-linear function. Such systems appear frequently in digital audio signal processing, for example in the case of signal saturation, overflow, elimination of denormal numbers, etc. The paper also deals with relations between the mutual ratio of amplitudes of the higher harmonics produced and the approximation parameters.

Key-Words: Nonlinear distortion, Nonlinear functions, Nonlinear systems, Piecewise-linear approximation, Saturation, Denormalized numbers

1 Introduction
We have to deal with aliasing when a digital signal is being processed by a nonlinear system. The aliasing is caused by bandwidth extension if the highest frequency component exceeds half the sampling frequency. To prevent that, we can either up-sample the input signal or approximate the transfer function of the system via the finite sum of terms of Taylor’s series and use nonlinear processing by the method of limiting the input signal bandwidth as published in [1].

That is why nonlinear systems are used with such a type of approximation whose response to an input signal with limited bandwidth has a limited bandwidth as well (e.g. polynomial approximation [2]). The polynomial and exponential approximations also have the advantage that approximation parameters can be evaluated by solving a linear equation system according to the required mutual ratio of harmonics [2].

The response of a system with transfer characteristics approximated using piecewise-linear function to a limited-bandwidth input signal has not a limited bandwidth. However, such nonlinear systems are used in digital audio signal processing frameworks themselves, for example signal for saturation, or by an algorithm developer due to performance enhancement, for example for an elimination of denormalized numbers. The reason for this is that the computing-power demands of these simple algorithms are extremely low. In order to reduce the aliasing distortion caused by such systems we need to found approximation parameters which guarantee that amplitude of higher harmonics decreases with their order. Then we can use the oversampling ratio derifed from psychoacoustical model to suppress this distortion.

2 Spectrum Changes in System with Piecewise-Linear Transfer Function
Approximation of a nonlinear transfer function \( \Psi(\cdot) \) that uses piecewise-linear function is defined using \( R \) linear sections, for which the following equation holds for \( i = 1, 2, ..., R \)

\[
y[n] = S_i(n[x[n]-x_{pr}]) \quad \text{for} \quad x_{i-1} \leq x[n] < x_i,
\]

where \( S_i \) are the slopes of straight lines in the area \( x_{i-1} \leq x[n] < x_i \), \( x_{pr} \) are the points in which the line cuts the x axis, and \( x_i \) are the lower boundaries of a particular section of the function (see Fig. 1).

2.1 Analytic Solution of Approximation of Discrete Fourier Transform Coefficients
In the discrete domain, the spectrum change can be evaluated in such systems by computing the approximation of coefficients of the discrete Fourier transform [3]. The following equation holds for coefficients \( Y[k] \) of the discrete Fourier transform of real discrete signal \( y[n] \) [4],

\[
Y[k] = \frac{1}{N} \sum_{n=0}^{N-1} (y[n]\cos(2\pi kn/N) - jy[n]\sin(2\pi kn/N)),
\]

where \( N \) is the number of samples of \( y[n] \). An approximation of the discrete Fourier transform using Fourier series coefficients can be used for evaluating the discrete Fourier series coefficients of discrete-time analytic signal. The following equation is derived in [4]
for the Fourier series in goniometric form associated
with function \( f(x) \), which is periodical with period \( \xi \),
\[
f(x) = \frac{A_0}{2} + \sum_{k=1}^{\infty} A_k \cos \left( \frac{2\pi k x}{\xi} \right) + \sum_{k=1}^{\infty} B_k \sin \left( \frac{2\pi k x}{\xi} \right)
\]  
(3)
where \( A_k \) and \( B_k \) coefficients are given by equations
\[
A_k = \frac{2}{\xi} \int_{-\frac{\xi}{2}}^{\frac{\xi}{2}} f(x) \cos \left( \frac{2\pi k x}{\xi} \right) \, dx \quad \text{for } k = 0, 1, 2, \ldots,
\]  
(4)
\[
B_k = \frac{2}{\xi} \int_{-\frac{\xi}{2}}^{\frac{\xi}{2}} f(x) \sin \left( \frac{2\pi k x}{\xi} \right) \, dx \quad \text{for } k = 0, 1, 2, \ldots
\]  
(5)

The even-function attribute of the Fourier series [3] can be utilized in the case of spectral component analysis of
the output signal of a nonlinear system with cosine input
signal. It can be seen from Fig. 1 that the period of the
harmonic signal
\[ \sum_{k=-\infty}^{\infty} \cos k \alpha \] 
can be represented by a piecewise-linear function.

If we substitute an analytic expression of a discrete
harmonic signal
\[ x[n] = X_1 \cos \left( \frac{2\pi n}{N} \right) \phi_i \]
(6)
where \( X_1 \) is the amplitude and \( \phi_i \) the initial phase of a
harmonic signal, into equation (1), we obtain, using the
substitution
\[ \alpha = \frac{2\pi n}{N} \]
(7)
the equation of output signal of the system for \( \phi_i = 0 \)
\[ y(\alpha) = S_i \left( X_1 \cos \alpha - x_{p_i} \right) \quad \text{for } \Theta_i < \alpha < \Theta_{i+1}, \]
(8)

where
\[ \cos \Theta_i = \frac{x_{i+1}}{X_1}, \quad \Theta_i > \Theta_{i+1} \quad \text{for } i = 0, 1, 2, \ldots, R. \]
(9)

The even-function attribute of the Fourier series [3] can be utilized in the case of spectral component analysis of
the output signal of a nonlinear system with cosine input
signal. It can be seen from Fig. 1 that the period of the
harmonic signal
\[ \sum_{k=-\infty}^{\infty} \cos k \alpha \]
can be approximated by a piecewise-linear function.

Using goniometrical identities we obtain the following
equation for the approximation of the discrete Fourier series coefficients of output signal
\[
Y(\alpha) = \frac{1}{\pi} \int_{\pi}^{\pi} y(\alpha) \cos \alpha \, d\alpha =
\]  
(10)
\[
= \frac{1}{\pi} \int_{\pi}^{\pi} y(\alpha) \cos \alpha \, d\alpha + \frac{1}{\pi} \int_{0}^{\pi} y(\alpha) \cos \alpha \, d\alpha.
\]

If we substitute equation (1) into equation (10) we obtain
the following function
\[
Y(\alpha) = \frac{X_1}{\pi} \sum_{i=0}^{\infty} S_i \left( \int_{\Theta_i}^{\Theta_{i+1}} P_{k_i}(\alpha) d\alpha + \int_{\Theta_{i+1}}^{\Theta_{i+1}} P_{k_i}(\alpha) d\alpha \right),
\]  
(11)
where \( \Theta_0 = \pi, \Theta_R = 0, \) and
\[ P_{k_i}(\alpha) = \cos \alpha \cos \alpha - \frac{x_{p_i}}{X_1} \cos \alpha. \]
(12)

Using goniometrical identities we obtain the following
equation for the amplitude of \( k \)th harmonics for \( k > 1 \) [2]
\[
Y_k(\alpha) = \frac{2X_1}{\pi} \sum_{i=0}^{\infty} \frac{S_i - S_{i+1}}{k^2 - 1} \left( k \sin \theta \cos \theta - \sin \theta \cos \theta \right) - \frac{S_{x_{p_i}} - S_{x_{p_{i+1}}}}{kX_1} \sin \theta.
\]  
(13)

One can see that equation (13) can be used only for
\( k > 1 \). The following equations for the amplitude of the
first harmonic and the direct component can be derived from
(11) and (12) using a different procedure [2]
\[
Y_1(\alpha) = \frac{X_1}{\pi} \sum_{i=0}^{\infty} \left( S_i - S_{i+1} \right) \left( \Theta_i + \cos \Theta_i \sin \Theta_i \right) - \frac{2}{X_1} \left( S_{x_{p_i}} - S_{x_{p_{i+1}}} \right) \sin \Theta_i,
\]  
(14)
\[
\frac{Y_0(\alpha)}{2} = \frac{X_1}{\pi} \sum_{i=0}^{\infty} \left( S_i - S_{i+1} \right) \sin \Theta_i - \left( S_{x_{p_i}} - S_{x_{p_{i+1}}} \right) \frac{\Theta_i}{X_1},
\]  
(15)
2.2 Relations between Piecewise-Linear Function Characteristics and Parameters

It can be seen from equation (13) that the output signal of a nonlinear system with transfer characteristics approximated by the piecewise-linear function has not a limited bandwidth if $R > 1$. Equation (13) is a sum of goniometric functions and the common period $\xi$ of spectral component amplitude repetition can be found for a finite number of limit angles $\Theta_i$. However, their amplitudes decrease very slowly [2]. When digital audio signal is processed, the highest harmonic of the output signal spectrum that is not masked by lower harmonics can be found using the psychoacoustical model. The upsampling ratio of the input signal can be chosen according to the order of this harmonic.

It can be also seen from equation (13) that the amplitudes of the harmonics depend on the difference of the slopes $S_i - S_{i-1}$ of subsequent linear sections and on the difference $S_{X_{f(i)}} - S_{X_{f(i)-1}}$. The amplitudes of higher harmonics thus increase with the difference of the right and left limits at the points of function discontinuities. The following equations hold for $i = 0, 1, ..., R - 1$ if the approximation function is continuous

$$S_{i-1}(x_i - x_{p_{i-1}}) = S_i(x_i - x_{p_i}),$$  \hspace{1cm} (16)

due to the $S_i$ and $x_{p_i}$ parameters are linearly dependent according to the equation

$$x_{p_i} = x_i + \frac{S_{i-1}}{S_i}(x_{p_{i-1}} - x_i).$$  \hspace{1cm} (17)

One could say from equation (13) that the amplitudes of output signal harmonics are linearly dependent. However, the substitutions $x_0 = -X_1$ and $x_R = X_1$ were used when equation (13) was derived (see Fig. 1). Change the input signal amplitude will influence all limit angles $\Theta_i$. The input signal will not span the $i^{th}$ section of the characteristic if $|\cos \Theta_i| > 1$.

Generally, the piecewise linear approximation of transfer characteristics has $R$ linear sections has $3R - 1$ parameters. The number of parameters decreases if the $S_i$ and $x_{p_i}$ parameters are linearly dependent according to (17). The following equations hold for the $x_{p_0}$ and $S_{R-1}$ parameters if the output signal range is $[y_{\min}, y_{\max}]$

$$x_{p_0} = -X_1 - y_{\min}/S_0,$$

$$S_{R-1} = y_{\max}/(X_1 - x_{p_{R-1}}).$$  \hspace{1cm} (18)

The total number of parameters of a piecewise linear continuous function with $R$ sections is $2R - 2$.  

2.3 Nonlinear Distortion Caused by Overflow

Arithmetic overflow occurs when an operation in arithmetical-logic unit produces a result that is greater than number that a given register or storage location can store. There are two common modes of arithmetical-logic unit that handle an arithmetic overflow: wrap-around and saturation mode.

2.3.1 Wrap-Around Mode

Wrap-around mode only sets the overflow flag and discards most significant bits of result that have exceeded the register or storage location range. Fig. 1 shows transfer characteristic of a nonlinear system representing two’s-complement wrap-around arithmetic. One can see from Fig. 1 that

$$S_i = 1 \text{ for } i = 0, 1, ..., R - 1,$$

$$x_{p_i} - x_{p_{i-1}} = 0,$$

$$R = 2.\text{floor}(X_1) + 1.\hspace{1cm} (19)$$

![Fig. 1. Transfer characteristic of a nonlinear system representing two’s-complement wrap-around arithmetic.](image)

We can obtain equations for computing the relative amplitudes $Y_i(\alpha)/Y_1(\alpha)$ of the output signal harmonics for such a type of system from equations (13) and (14)

$$Y_i(\alpha) \sim \frac{1}{k} \sum_{k=1}^{\infty} \sin k\alpha,$$

$$Y_1(\alpha) = \frac{1}{k} \sum_{k=1}^{\infty} \sin k\Theta_i \hspace{1cm} (20)$$

We have to realize that $\Theta_i$ depend on the harmonic input signal amplitude $X_i$. Fig. 2 shows dependence of relative levels $Y_1(X_1)/Y_1(X_i)$ of odd higher harmonics of output signal of a nonlinear system representing two’s-complement wrap-around arithmetic for signal that exceeds the arithmetic range four times. The amplitudes of even harmonics are zero because the transfer characteristic is odd function.

2.3.2 Saturation Mode

In the saturation mode, a result that is lower than a low boundary or higher than a high boundary is replaced by the appropriate boundary. The transfer characteristic of a...
nonlinear system representing two’s-complement saturation arithmetic is in Fig. 3. With this type of nonlinear system, some sections of the function are constant, i.e., \( S_i = 0 \), and for these sections the equation \( y[n] = S_{i-1}x_{P_i-1} \) has to be substituted into (10) (assuming continuous function).

\[
Y_k(X_1)/Y_1(X_1) = \frac{2S_X}{\pi(k^2 - 1)} \left( \left(1 + (-1)^{k}\right) k \sin k\Theta_2 \cos \Theta_2 - \sin \Theta_2 \cos k\Theta_2 \right) 
\]

\[ Y_2(\alpha) = \frac{2S_X}{\pi} (\Theta_2 + \sin \Theta_2 \cos \Theta_2), \quad Y_2(\alpha) = 0. \]  

One can see from equation (21) that the amplitudes of the even harmonics are zero. In fact, the nonlinear system representing two’s-complement saturation arithmetic has not symmetrical characteristic so even harmonics are generated as well. Equations for computing amplitudes of the harmonics of the output signal spectrum of a non-symmetrical limiter can be found in [2]. However, this asymmetry is so small that level of the even harmonics is more than 100 dB below level of the first harmonics for 16-bit numbers.

Fig. 4 shows dependence of relative levels \( Y_k(X_1)/Y_1(X_1) \) of odd higher harmonics of output signal of a nonlinear system representing two’s-complement saturation arithmetic for signal that exceeds the arithmetic range four times.

2.4 Elimination of Denormalized Numbers

Personal computers equipped architectures, which follows IEEE 754 standard for representing floating-point value, are widely used for digital audio processing. This standard specifies formats for representing floating-point values, namely single-precision and double-precision. Binary floating-point numbers are stored in a sign-magnitude form consisting of sign bit, biased exponent, and fraction. Non-zero, finite numbers are divided into two classes: normalized and denormalized [6]. When floating-point numbers become very close to zero, the normalized-number format can no longer be used to represent the numbers. This is because the range of the exponent is not large enough to compensate for shifting the binary point to the right to eliminate leading zeros [7]. When performing normalized floating-point computations, a processor operates on normalized numbers and produces normalized numbers as results. Denormalized numbers represent an underflow condition. A denormalized number is computed through a gradual underflow technique [7]. This causes that arithmetical operations with denormalized numbers are much slower than for normalized numbers [8].
In digital signal processing applications denormalization is often caused by the use of feedback structures [8]. There are several methods widely used by algorithm developers to prevent the denormalization of numbers and thus decreasing the computer performance. Beside a white noise addition, there are several methods that use nonlinear systems with odd transfer characteristic.

2.4.1 Zero Substitution
This method tests a floating point number and sets it to zero if it is denormalized. Addition and consequent subtraction of a small normalized number can be alternatively performed. Such an operation can be represented using nonlinear system with transfer characteristics

\[ y[n] = \begin{cases} x[n] & \text{for } |x[n]| > x_D \\ 0 & \text{for } |x[n]| \leq x_D, \end{cases} \]  

(23)

where \( x_D \) is so called anti-de-normal number. Although smallest normalized number is approx. \( \pm 1.18 \cdot 10^{-38} \) [6], larger numbers are usually used [8]. Fig. 5 and 6 show the transfer characteristic and dependence of relative levels of the higher harmonics \( k \) on amplitude \( X_1 \) of a nonlinear system representing algorithm using zero substitution with \( x_D = 1 \cdot 10^{-20} \).

![Fig. 5. Transfer characteristic of a nonlinear system representing algorithm using zero substitution.](image1)

![Fig. 6. Dependence of relative levels of the higher harmonics on amplitude \( X_1 \) of a nonlinear system representing algorithm using zero substitution.](image2)

2.4.2 Substitution of Constant Value
This method tests a floating point number and sets it to normalized constant value \( x_0 \) if it is denormalized. Such an operation can be represented using nonlinear system with transfer characteristics

\[ y[n] = \begin{cases} x[n] & \text{for } |x[n]| > x_D \\ x_0 & \text{for } |x[n]| \leq x_D. \end{cases} \]  

(24)

Fig. 7 and 8 show the transfer characteristic and dependence of relative levels of the higher harmonics \( k \) on amplitude \( X_1 \) of a nonlinear system representing denormalization-preventing algorithm using substitution of constant value \( x_D = 1 \cdot 10^{-20} \).

![Fig. 7. Transfer characteristic of a nonlinear system representing algorithm using constant value substitution.](image3)

![Fig. 8. Dependence of relative levels of the higher harmonics on amplitude \( X_1 \) of a nonlinear system representing algorithm using constant value substitution.](image4)
2.4.2 Addition of Constant Value

This method adds a small normalized number \( x_D \) without any testing. Such an operation can be represented using nonlinear system with transfer characteristics

\[
y[n] = \begin{cases} 
  x[n] + x_D & \text{for } x[n] > 0 \\
  x[n] - x_D & \text{for } x[n] < 0.
\end{cases}
\] (25)

Fig. 9 and 10 show the transfer characteristic and dependence of relative levels of the higher harmonics \( k \) on amplitude \( X_1 \) of a nonlinear system representing denormalization-preventing algorithm using addition of constant value \( x_D = 1 \cdot 10^{-20} \).

4 Conclusion

It is possible to affect amplitudes of higher harmonics produced by a nonlinear system without memory with system function approximated using piecewise-linear approximation, as demonstrated in section 2.2. The amplitudes of the harmonics depend on difference of the slopes and zero points of the subsequent linear sections. Therefore the saturation arithmetic is better than wrap-around one from the aliasing distortion point of view, which is well known fact. More suppression of higher harmonics caused by a signal overflow can be achieved using piecewise-linear compression which minimizes difference of slopes and zero points of the subsequent sections. However, compression decreases the dynamic range and higher harmonics are generated even for regular-range numbers.

As can be seen form Fig. 6, denormalization-preventing algorithm using zero substitution is the best one from the aliasing distortion point of view. This algorithm causes higher distortion than other algorithms for input signal amplitudes slightly above the antidenormal value \( x_D \) but amplitudes of higher harmonics are lower than –70 dB for amplitudes above \( 3x_D \). On the other hand, the amplitudes of higher harmonics are lower for input signal amplitudes slightly above \( x_D \) when algorithm using addition of constant value is used but the amplitudes decrease slowly. The algorithm that uses constant value substitution is the compromise between the two previous ones. It is important to realize that we have to deal with the aliasing distortion caused by denormalization-preventing algorithm only for small normalized numbers.

References: