InnoSpaceTool Unit 8 complementary material - Modulation

Endurosat Team

This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 768049

In this unit we will study modulation in more detail. We will begin by introducing the main types of modulation - Frequency Modulation (FM), Amplitude Modulation (AM) and Phase Modulation (PM). Historically, AM was the first method used to transfer voice by radio and it remains in use today for many applications, but FM has the benefit of larger signal-to-noise ratio and it is the method we will consider in depth after introducing both in the first section. This unit can be seen as a continuation of the previous one, since we will consider modulation of bits, therefore the signal must have been digitalized prior to the process.

1 FM, AM and PM - Introduction

We already mentioned that to use higher frequency waves for the transmission of signals with lower frequency, we will need to perform a process called modulation, which somehow "embeds" the latter into the former. From this point on, we will refer to the higher-frequency signal which we use for the transmission as the "carrier signal" and to the lower-frequency signal which carries the information as the "modulation signal". To perform modulation, we must change some measurable property (properties) of the carrier and use it to encode the information. Let us consider a sinusoidal carrier signal - one of the form $U(t) = Asin(\omega t + \varphi)$. The three main properties such a signal has are its amplitude A, its frequency ν (or angular frequency ω respectively) and its phase φ. These three are exactly the main parameters which are used for modulation, each type carrying the respective name of the parameter modified: Amplitude Modulation (AM), Frequency Modulation (FM) and Phase Modulation (PM).
To illustrate the difference between these three, we will first consider the simple example of representing a lower-frequency sine wave with a higher-frequency such waves in each respective way (carrying no information). Let our modulation and carrier signals be $u_{mod} = U_M \sin(\Omega t + \varphi_M)$ and $u_c = U_c \sin(\omega_c t + \varphi_c)$ respectively, where $\Omega < \omega_c$. Then we can express the AM, FM and PM signals as:

$$u_{AM}(t) = U_c[1 + m_A \sin(\Omega t + \varphi_M)] \sin(\omega_c t + \varphi_c) \tag{1.1}$$

$$u_{FM}(t) = U_c \sin[(\omega_c + \Delta\omega \sin(\Omega t + \varphi_M))t + \varphi_c] \tag{1.2}$$

$$u_{PM}(t) = U_c \sin[\omega_c t + \varphi_c + m_\varphi \sin(\Omega t + \varphi_M)] \tag{1.3}$$

Where $m_A = U_M/U_c$ is the "depth" of amplitude modulation, $\Delta\omega$ is the angular frequency deviation (the largest frequency difference with respect to the carrier's) for frequency modulation and m_φ is the so called phase modulation index (the largest phase difference $\Delta\varphi_c$). The plots of these functions can be seen on the next page (Fig.8.1). One can clearly see how the amplitude of the carrier follows the modulation signal in the first case, it is a bit harder to see the effect for the FM and PM, but there are clear changes in the carrier, corresponding to changes in the modulation signal, and so given any of the three, we can reconstruct the modulation signal.

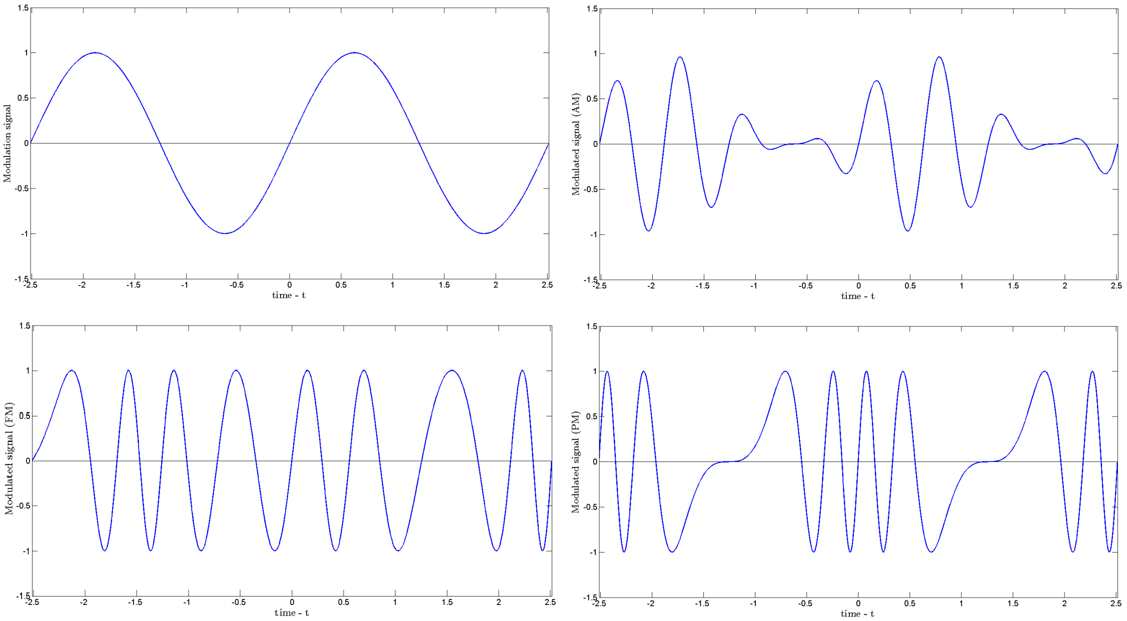

Fig.8.1 - Sine signal modulated by AM (top right), FM(bottom left) and PM(bottom right).

Now that we have a basic idea of what the difference between different modulations is, note that a general modulation signal would not be sinusoidal, but composed of N harmonic components with different amplitudes and frequencies:

$$u_M(t) = \sum_{n=1}^{N} U_{M_i} \sin(\omega_i t + \varphi_{M_i}) \tag{1.4}$$

This looks very familiar to a Fourier series, except that it has only a finite number of components. In fact, 1.4 represents a general radio signal (which as mentioned before, consists of only a finite number of components). For our signal, using Fourier series, we can see that the spectre in the AM case consists of two sections (symetrical about the carrier frequency), which correspond to the frequencies $\omega_c - \omega_i$ and $\omega_c + \omega_i$. The spectre is easiest to find in the case of AM, since $\sin\alpha \sin\beta = 1/2[\cos(\alpha-\beta) - \cos(\alpha+\beta)]$ and we don't need to evaluate any integrals. The effective bandwidth of the AM signal is then:

$$B_{AM} = 2\nu_{max} \tag{1.5}$$

where ν_{max} corresponds to the largest frequency component in the sum $\omega_{max} = 2\pi\nu_{max}$. The spectre of the same signal in the FM case consists of infinitely many frequencies - the frequency of the carrier ω_c and infinitely many symmetrically positioned frequencies $\omega_{\pm k} = \omega_c \pm k\Delta\omega$. The amplitudes of the spectral components depend on the ratio $\Delta\omega/\Omega$ and can be determined using Bessel's functions (a well tabulated set of infinitely many functions). We will not go into the details of Bessel's functions here (the reader can find information about their representations and tables of values in many books). The effective bandwidth of the FM modulated signal is determined by neglecting spectral components with amplitude less or equal to 1% of the maximum value. It can be shown that these are the spectral components with k in the range between 1 and $\Delta\omega/\Omega + 1$. Taking into account the fact that these are all equally-spaced in the frequency domain (at a distances $F = \Omega/2\pi$ from each-other), we can give the effective bandwidth for an FM signal as:

$$B_{FM} = 2F\left(\frac{\Delta\omega}{\Omega} + 1\right) = 2F\left(\frac{\Delta\nu_c}{F} + 1\right) = 2(\Delta\nu_c + F) \tag{1.6}$$

where $\Delta\nu_c$ is simply the frequency difference, corresponding to $\Delta\omega$ ($\Delta\omega = 2\pi\Delta\nu_c$). In the case when the modulation signal is not sinusoidal but of the form 1.4, we simply need to replace F with the highest-frequency component F_{max}, so the bandwidth becomes approximately:

$$B_{FM} = 2(\Delta\nu_c + F_{max}) \tag{1.7}$$

The spectre of the PM signal is the same as that of the FM signal, and so its effective bandwidth is given with a similar formula:

$$B_{PM} = 2F_{max}(\Delta\varphi_c + 1) \tag{1.8}$$

In the cases when $\Delta\nu_c >> F$ for FM or $\Delta\varphi_c >> 1$ for PM, 1.7 and 1.8 become simply:

$$B_{FM} = 2\Delta\nu_c \tag{1.9}$$

$$B_{PM} = 2F_{max}\Delta\varphi_c \tag{1.10}$$

It is important to note here that the AM does not use the power of the transmitter as efficiently as FM and PM, but from the expressions above we can see that it requires much smaller bandwidth in general! Furthermore, AM is far less resistive to external noises, and as a result it is not widely used for communication purposes today. From 1.9 and 1.10 we can also see that the bandwidth for PM signals depends on the frequency of the modulation signal, while that for FM signals remains unchanged for practically any spectral width of the modulation signal, this is why the use of frequency modulation is preferred in radio communications!

2 Digital methods for forming signals

Let us now consider the problem of transmitting digital information through a radio channel. Once again, we must encode using some parameter of the carrier wave, but this time we need only two states, which will allow us to differentiate between 0 and 1. The main concern with digital signal transmission is the transmission speed. Since the channel bandwidth is fixed, one must take into account the effective bandwidth of the manipulated signal when choosing the type of encoding. In practice the signal may be required to pass through a filter so that its effective bandwidth can be reduced, but this leads to unwanted noises which we will explore in the next section. When comparing the signal manipulation methods, one usually uses the parameter link spectral efficiency - BR, which gives the realized transmission speed per unit bandwidth:

$$BR = \frac{R_b}{B_{SK}} \tag{2.1}$$

where R_b is the signal transmission speed in bits per second and B_{SK} is the minimum bandwidth needed in the frequency domain for complete recovery of the signal in the receiver. The simplest digital modulation methods are the Amplitude Shift Keying (ASK), Frequency Shift Keying (FSK) and Binary Phase Shift Keying (BPSK) manipulations. We will review each separately now.

1. ASK - Realized by jump-changing the amplitude of the carrier signal in accordance with the digital modulation signal. An ASK modulated signal can be described as:

   $$u_{ASK}(t) = L(t)U_c \sin(\omega_c t) \tag{2.2}$$

   where $L(t)$ is a logical function of the input bit (so $L(t) = 1$ for bit 1 and $L(t) = 0$ for bit 0).

2. FSK - Realized by jump-changing the frequency of the carrier signal in accordance with the digital modulation signal. An FSK modulated signal can be described as:

   $$u_{FSK}(t) = U_c \sin(\omega_c + L(t)\Delta\omega_c)t] \tag{2.3}$$

   Where $L(t)$ is the same logical function, so the frequency is higher for bit 1 and equal to the carrier frequency for bit 0.

3. BPSK - Realized by switching the carrier's phase by $\pm 180^o$ every time a logical 1 becomes 0 or a logical 0 becomes 1 respectively. A BPSK modulated signal can be described as:

   $$u_{BPSK}(t) = U_c \sin(\omega_c t + L(t)\pi) \tag{2.4}$$

   where $L(t)$ is a logical function again, but having the values 1 and -1 this time.

Note that we can define the logical functions for the ASK and FSK in the same way as that for the BPSK. This can improve the ASK manipulation, since for the function described in 2.2, the receiver will not be receiving any signal whenever there is a long sequence of zeros. This may lead to loss of information for the carrier's frequency and introduce additional complications. To see how each of these works, let us consider the binary sequence 01001101. Modulating a carrier signal with it in each of the three listed ways is shown on the next page (Fig.8.2):

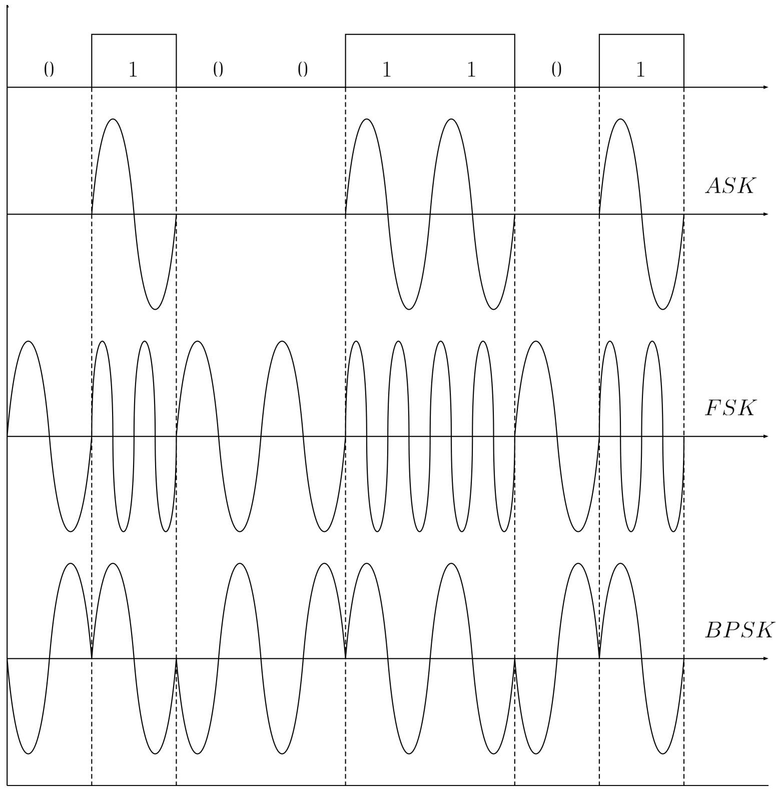

Fig.8.2 - ASK, FSK and BPSK modulation for the sequence 010011001.

We showed in the last unit that for the correct reproduction of a digital signal, one needs to transmit the part of its spectre within the Nyquist bandwidth B_N. Since we are dealing with 2-state bits, the required bandwidth for the ASK signal is:

$$B_{ASK} = 2B_N = f_T \tag{2.5}$$

To guarantee better transmission, a 1.4 times larger bandwidth is used in practice $(1.4f_T)$. Substituting B_N and using our link spectral efficiency definition 2.1, we obtain it to be 1bit/s/Hz. The same band efficiency is obtained in the other two cases as well (FSK and BPSK). One can obtain larger value by manipulating parameters which can take more than two states. The most widely used such modulation is the Quadrature Phase Shift Keying (QPSK). Unlike the previous considered examples, QPSK uses four states (embedded as changes in the carrier's phase), each of which carries information for two bits. Similarly to the BPSK, a switch of the carrier's phase is performed, but unlike the BPSK (where each change was by 180^o), different jumps are used to indicate different combinations of two bits in the QPSK (each corresponding to an angle in one of the 4 quadrants). The bits corresponding to each quadrant (and so each phase change) can be seen on Fig.8.3:

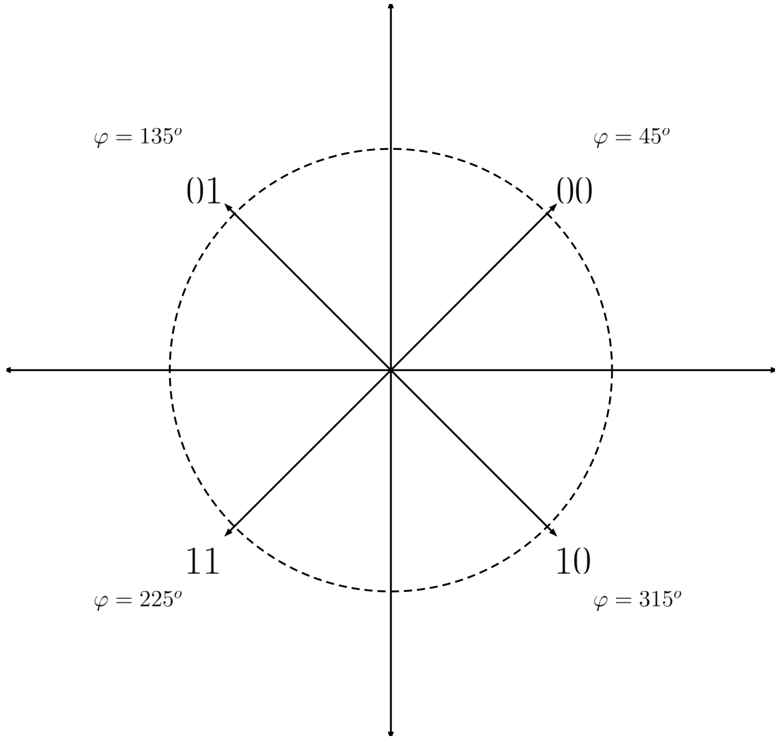

Fig.8.3 - QPSK bits to phase change correspondence.

The signal vector can be obtained as a sum of two components (one of the type $\cos(\omega_c t)$ pointing in the x direction and one of the type $\sin(\omega_c t)$ pointing in the y direction of the graph). Quadrature representation relies on the trigonometric identity:

$$\sin(\omega_c t + \varphi) = \sin(\omega_c t)\cos(\varphi) + \cos(\omega_c t)\sin(\varphi) \tag{2.6}$$

Since the allowed phases are $\varphi = \pi/4 + n\pi/2$, the functions $\sin\varphi$ and $\cos\varphi$ have values of $\pm\sqrt{2}/2$ and so we can represent a carrier with arbitrary phase φ as a superposition in the form:

$$u_{QPSK}(t) = \frac{\sqrt{2}}{2} U_c[L_x(t)\sin(\omega_c t) + L_y(t)\cos(\omega_c t)] \tag{2.7}$$

where L_x and L_y are logical functions, each of which can have the values ± 1 depending on the bits as follows: for 00 we have $L_x = L_y = 1$ (corresponding to $\varphi = 45^o$); for 01 we have $L_x = -1$, $L_y = 1$ (corresponding to $\varphi = 135^o$; for 10 we have $L_x = 1$, $L_y = -1$ (corresponding to $\varphi = 315^o$); for 11 we have $L_x = L_y = -1$ (corresponding to $\varphi = 225^o$). As an example, the same sequence (01001101) has been shown on Fig.8.4, this time modulated using QPSK:

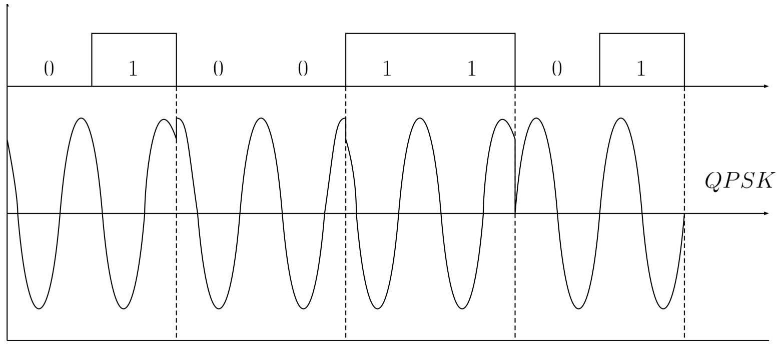

Fig.8.4 - QPSK modulation for the sequence 010011001.

Note that unlike the modulations on Fig.8.2 where each bit change corresponded to some amplitude, frequency or phase change, now the change is responsible for the next two bits in the sequence. As a result, the required bandwidth for the QPSK signal is:

$$B_{QPSK} = 2B_N = \frac{f_T}{2} \tag{2.8}$$

This means that reducing the manipulation speed for forming the QPSK signal twice (compared to that for the BPSK signal) leads to a double reduction of its spectre also. More possible states can be included as well by dividing the phases in more intervals. Among such modulations are the 8PSK, 16PSK and so on (with 8 phase states, 16 phase states and etc. respectively). Combinations of the manipulated quantities are also possible - for example the 16QAM uses amplitude and phase modulation with a total of 16 states again. In it, there are 4 possible states in each quadrant, but phase alone can differentiate between just 3 of them, so amplitude modulation is required also. We will not dive deeper into these methods, since the main ones used (especially for the applications this book is dedicated to) are FSK and QPSK.

3 Restriction and reduction of the spectre - filters

In all of the considered above methods of modulation, the spectre of the signal is theoretically infinite. As stated before, no machine can truly produce a signal with infinite spectre, and so one needs to cut off the higher frequencies to allow transmission. As we've already seen, the sufficient requirement to correctly restore the signal after transmission is that the first harmonic is transmitted. It defines the minimal required frequency bandwidth f_T, which corresponds to twice the Nyquist band. One can cut the higher frequencies by applying a filter to the signal, but there are downsides to this, as filtering PSK signals leads to additional (unwanted) amplitude modulation. The appearance of this additional unwanted modulation leads to reduction of noise resistivity and so it is unwanted.

The simplest and most effective filtering method is filtering of the modulating signal. The lower frequency of the modulating signal allows the use of digital filters and the suppression of the higher-frequency harmonics of the rectangular signal in the source, which leads to lower intermodulation products in the next steps of the transmitter.

Another unwanted effect due to filtering is the inter-symbol interference, which leads to errors in the digital signal after filtration. Since the filter cuts the higher frequencies of a rectangular impulse, it changes its shape in a similar manner to that observed on Fig.7.4 and so nearby impulses will overlap (the signals are interfering). One can avoid such errors by the use of ideal low-frequency filter, whose transmission and impulse characteristics are shown on the figure bellow:

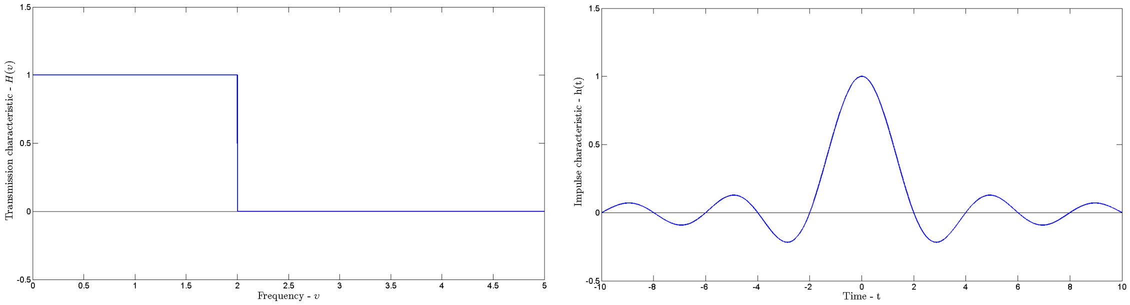

Fig.8.5 - Transmission and impulse characteristics of an ideal filter for T=2.

By impulse characteristic we mean the shape which a rectangular impulse gets after it passes through the filter. For digital signal transmission speed of 1bit/T, the bandwidth allowed by the filter is equal to the Nyquist band $B_N = 1/2T = f_T/2$. Since the impulse characteristic of this filter (seen on Fig.8.5) is described by the SI function (sine integral function), the impulse reaction of neighbouring rectangular impulses will also have the same shape, but displaced by a certain number of periods. As a consequence, in any given moment $t_n = nT$, the impulse reaction of the n^{th} impulse reaches its maximum while the impulse reactions of the previous and consequent impulses interfere and amount to zero. In this way, one eliminates the influence of the neighbouring bits on the one currently being transmitted.

In practice, however, one cannot realize ideal low-frequency filter. Furthermore, its very steep cut-off would cause even more interference if the filtration was not exact in periodicity (acting on each rectangular impulse exactly). For these reasons, more smooth filters are used in practice. The most important such are the $\cos^2$ and the Gauss filter. We will now consider the second one in some more depth.

The transmission characteristic of a Gauss filter used for limiting the spectre of digital signals is described by:

$$H(\nu) = e^{-\frac{\pi^2}{k}(\nu T)^2} \tag{3.1}$$

Where the parameter k determines how steep the Gaussian is (usually it takes its values in the interval $[3, 4]$). Due to the Fourier transform properties of a Gaussian, the impulse characteristic of such a filter is once again a Gaussian with amplitude falling off rapidly for $t > T$ as shown on Fig.8.6. The Gaussian filter is characterized by 3 dB admission bandwidth normed about the bit speed (the frequency $f_T = 1/T$). The value is chosen less than 0.5 which means that for the Gaussian filter, the bandwidth at level 3 dB is less than the Nyquist bandwidth.

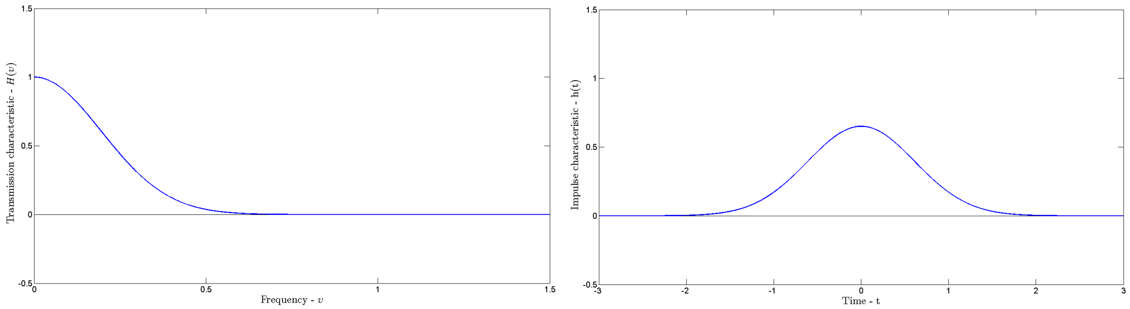

Fig.8.6 - Transmission and impulse characteristics of a Gaussian filter for T=2.

Apart from filters, reduction of the bandwidth of the spectre can be obtained by the so called continuous phase manipulation methods, which (unlike the phase modulation methods considered previously) changes the phase in a smooth way. We will not consider these methods in depth, but an example for such a method is frequency modulation, for which a single generator with smoothly varying frequency is used (unlike the jump-change between frequencies, considered before).