Class Notes, 31415 RF-Communication Circuits

Chapter I

MODULATION, TRANSMISSION, and DEMODULATION

Jens Vidkjær

NB 229

Contents

I Modulation, Transmission, and Demodulation

Modulation is the process of transforming a baseband message to a form suitable for transmission through the channel in consideration. Demodulation is the reverse process of again recovering the original message. The scope of the demodulation depends on the type of data being send. In a radio telephony channel it may suffice at the receiver site to get an output with a power spectrum that contains the dominant part of the input power spectrum, in a television video channel it is important to reconstruct in time-domain the shape of the signal being send. In digital transmissions, the goal is to rebuild a logical bitstream represen-

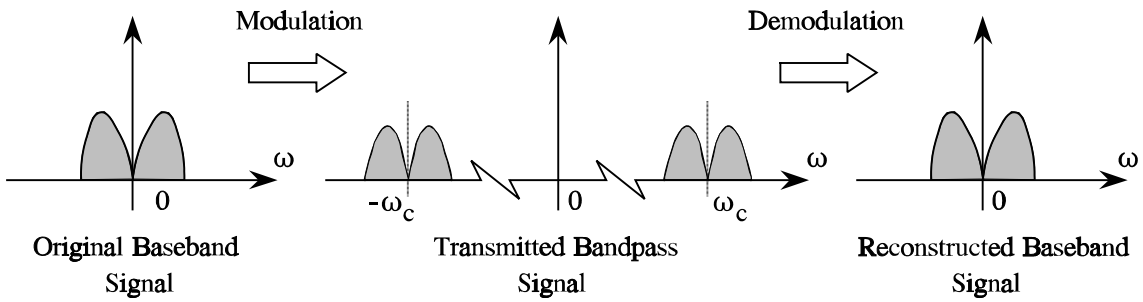


Fig.1 Power spectra for signals in the modulation and demodulation processes. It depends on the actual modulation type whether or not the spectra become similar with respect to shapes, symmetries, and bandwidths.

tation equivalent to the input stream. These and more distinct requirements are the background for the variety of modulating methods and modulator/demodulator circuits in use. However, all the types we shall consider below for RF communications have in common, that the modulating process transforms the low frequency baseband signal to a bandpass signal around a carrier frequency as sketched in Fig.1. The bandpass signal is the one actually transmitted to the receiver where the demodulator reconstruct the low-frequency baseband message. Obvious reasons for the procedure are that more baseband signals may be transmitted simultaneously through the same channel at different carrier frequencies. In radio-communication, moreover, efficient radiation and reception of signals through antennas require that the wavelength is comparable to their physical dimensions, so a move towards high frequencies makes the equipment manageable in size.

The scope of the presentation in this chapter is to provide a background for designing circuits and sub-systems that operate in the RF-frequency range. As will be apparent, there are still holes to be filled even to accomplish this limited goal. To get coverage of omissions and especially all the further aspects that are important to understand, plan, and design complete RF-communication systems, the reader should consult the comprehensive literature about the system aspects of communication. A few examples are given in the reference list where ref's [1] and [2] encompass all types of modulations, [3] and [4] concentrate on digital communication, and [5] includes a detailed account on system design aspects.

I-1 Basic Modulation Types and Concepts

The types of modulations suited for RF communication we consider are called continuous wave modulations, CW, where the baseband information is impressed upon a sinusoidal carrier wave of amplitude A_{c0} and angular frequency ω_c.

$$\begin{aligned} \textit{carrier} \;:\; & y(t) = A_{c0} \cos \omega_c t, & (a) \\ \textit{modulated carrier} \;:\; & y(t) = A(t) \cos(\, \omega_c t + \varphi(t) + \varphi_0 \,) \,. & (b) \end{aligned} \qquad (1)$$

The time dependencies of A(t) and φ(t) in the last equation contain the baseband message and the angle φ_0 represents an offset phase for the carrier compared to the timing of the baseband message. If there is no synchronism between the two, the offset may be set to zero without loss of generality. Eq.(1)(b) is called the envelope-phase representation of a modulated signal.

One major distinction between different modulation types is on how a baseband message contained in the signal x(t) is impressed upon the modulated output y(t). Here the basic extremes are amplitude modulation, where the phase does not depart from the carrier phase, and angle modulation, where the amplitude is kept constant.

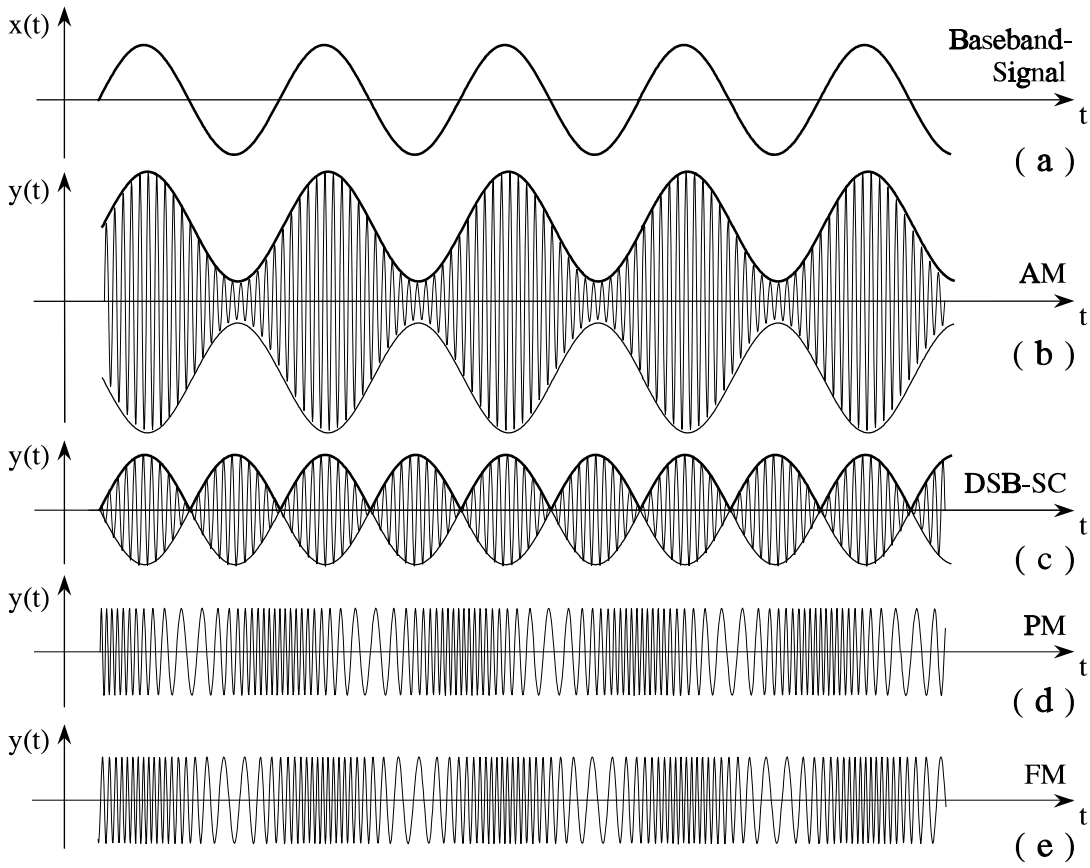


Fig.2 Examples of modulation waveshapes from a sinusoidal baseband signal x(t). The modulation types shown are AM, amplitude, DSB(-SC) double sideband (suppressed carrier), PM, phase, and FM, frequency modulation.

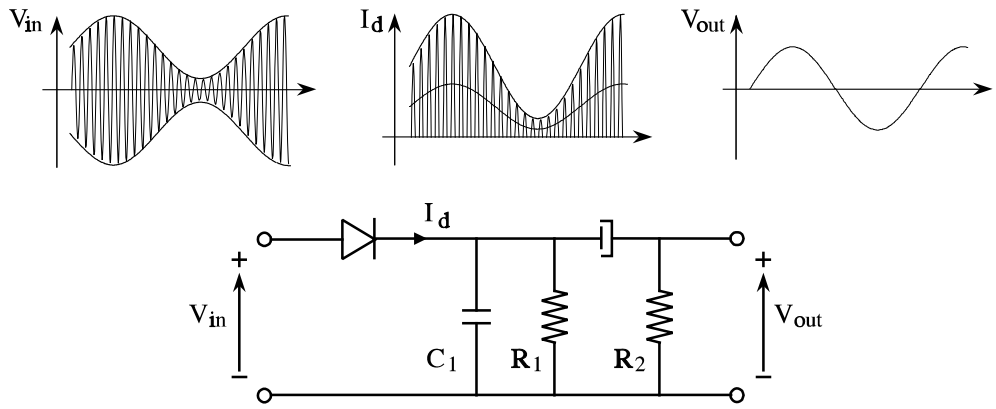


Fig.3 Envelope detector. The short-term mean-value of the diode current holds the signal envelope. LP filtering by C_1, $R_1||R_2$ removes the carrier component.

Amplitude Modulations, AM and DSB-SC

$$\underline{\textit{Examples of Amplitude Modulation}, \ \varphi(t) = 0.}$$

$$\textit{amplitude modulation, AM} \ : \ y(t) = A_{c0}(1 + m\,x(t))\cos\omega_c t\,, \qquad (a) \quad (2)$$

$$\textit{double-sideband, (suppressed carrier), DSB or DSB-SC} \ : \ y(t) = A_{c0}x(t)\cos\omega_c t\,. \qquad (c)$$

Two common amplitude modulation methods are given above. Their waveshapes for a sinusoidal baseband signal are exemplified in Fig.2. The AM modulation in Eq.(2)(a) is intended to transfer the waveshape of the modulating baseband signal x(t) to the envelope of the carrier. Scaling of the signal levels is here quantified by the modulation index m. With a normalized baseband signal $|x(t)|\leq 1$, the condition $m\leq 1$ (or 100%) implies undistorted reproduction of the baseband signal to the carrier envelope. It is easy to reconstruct the baseband signal from an AM modulated wave in a receiver by the simple envelope detector circuit in Fig.3. Its detailed function will be considered later, but it should be realized that the low-pass filter bandwidth must exceed the envelope frequency. An AM modulated wave gets spectral components stemming from the baseband signal below and above the carrier. This is demonstrated for a sinusoidal baseband signal $x(t)= A_x\cos\omega_x t$, $A_x\leq 1$, and the trigonometric identity

$$\cos a\,\cos b = \tfrac{1}{2}\cos(a+b) + \tfrac{1}{2}\cos(a-b) \qquad (3)$$

that yields

$$\begin{aligned} y(t)\,|_{AM} &= A_{c0}\,(1 + m\,A_x\cos\omega_x t)\cos\omega_c t \\ &= A_{c0}\cos\omega_c t + \frac{A_{c0}}{2}\,m\,A_x\cos(\omega_c - \omega_x)t + \frac{A_{c0}}{2}\,m\,A_x\cos(\omega_c + \omega_x)t. \end{aligned} \qquad (4)$$

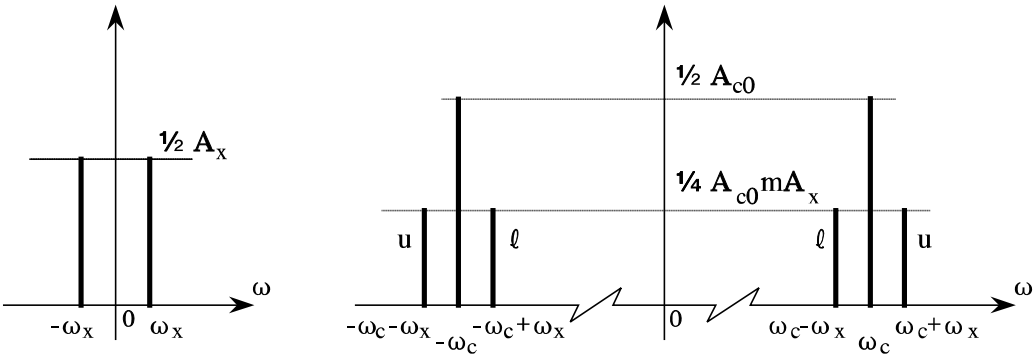


Fig.4 Spectral components in a double-sided amplitude spectrum of the sinusoidal baseband signal and the AM modulated waveform from Eq.(4). u and ℓ are the upper and lower sideband components.

Besides the distinct carrier term of frequency ω_c we get two so-called sidebands caused by the baseband signal, a lower sideband at the difference frequency ω_c-ω_x and a corresponding upper sideband at ω_c+ω_x. A doublesided amplitude spectrum portray of this situation is shown in Fig.4, where each term contributes a positive and a negative frequency component according to the Euler identities,

$$\cos a = \frac{1}{2}\left(e^{ja} + e^{-ja} \right), \qquad \sin a = \frac{1}{j2}\left(e^{ja} - e^{-ja} \right). \tag{5}$$

With maximum undistorted modulation, i.e. m=1 and A_x = 1, the power of the AM modulated wave, say it is a voltage across a 1Ω resistor, becomes,

$$P_{AM} = 2\,\frac{A_{c0}^2}{4} + 4\,\frac{A_{c0}^2\, m^2 A_x^2}{16}\Bigg|_{m=1,\,A_x=1} = \underbrace{\frac{1}{2}A_{c0}^2}_{carrier} + \underbrace{\frac{1}{4}A_{c0}^2}_{envelope}, \tag{6}$$

so at most 33% of the transmitted power contains the message from the baseband signal. However, the carrier power is required if simple envelope detection should be used in a receiver for the signal as suggested by the AM modulator and demodulator pair in the block scheme of Fig.5 a,b. The ideal multiplying element is sometimes called a balanced modulator while its practical realization is called a balanced mixer. Note that the insertion of the carrier corresponds to adding a DC component to the baseband signal. To get explicit control of the carrier it is generally supposed that the baseband signal itself, x(t), holds no DC term.

Compared to the result above, the DSB modulation from Eq.(2)(b) provides no distinct carrier component. This fact is often emphasized by adding the phrase "suppressed carrier" or SC. We have,

$$
\begin{aligned}
y(t)\,|_{DSB-SC} &= A_{c0}\,A_x\,\cos\omega_x t\,\cos\omega_c t \\
&= \frac{A_{c0}A_x}{2}\cos(\omega_c - \omega_x)t + \frac{A_{c0}A_x}{2}\cos(\omega_c + \omega_x)t\ ,
\end{aligned}
\tag{7}
$$

and the doublesided spectrum of this signal corresponds to the spectrum in Fig.4 with m=1 but without the two carrier components at angular frequencies $\pm\omega_c$. We avoid assigning power to carrier components that bear no information, but the price paid is, that it becomes more complicated to get the baseband signal back in demodulation. As seen in the sinusoidal example of Fig.2, the envelope that would be sensed by an enveloped detector (shown in heavy line) is no longer the baseband signal. To detect the baseband from a DSB modulated signal it should again be multiplied by a carrier as shown in the DSB-SC modulator/demodulator pair in Fig.5 c,d. Although this at a first glance looks as simple as the AM pair, a prerequisite for proper operation is, that the two carrier oscillators are synchronized to run with equal phases. Let $\theta(t)$ represents a possible difference in phase between the two - it will grow linearly with time if the frequencies differ - then the output from the demodulator multiplier becomes,

$$
\begin{aligned}
u(t) &= x(t)\,\cos\omega_c t\,\cos(\omega_c t + \theta(t)) \\
&= \frac{x(t)}{2}\cos\theta(t) + \frac{x(t)}{2}\cos(2\omega_c t + \theta(t))\ .
\end{aligned}
\tag{8}
$$

The last resultant term holds high frequency components that are removed by low-pass filtering. The first term holds the recovered baseband signal, but if $\theta(t)$ differs from zero the cosine factor either reduces or distorts it, so to get a predictable result the oscillator in the demodulator must be synchronized to the carrier of the received signal. A simple method is to let a fragment of the full carrier - a pilot carrier - follow the signal. This is done by adding

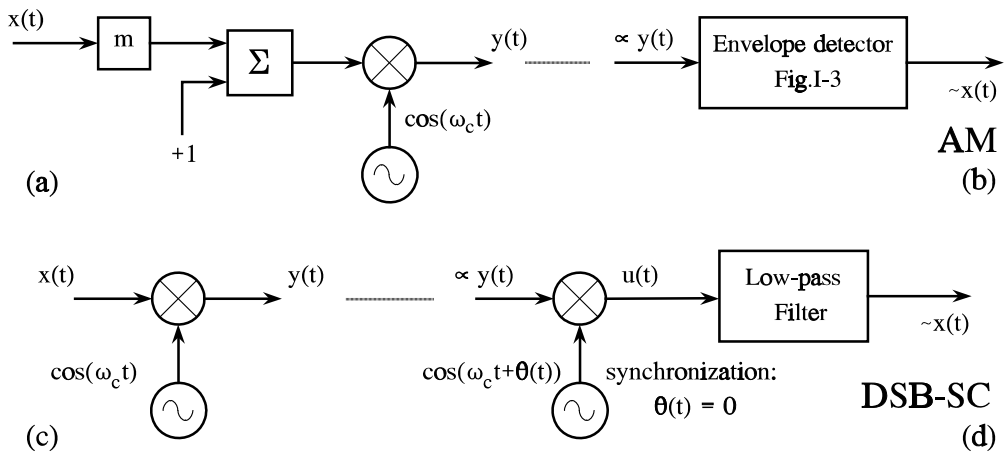


Fig.5 Block schemes for simple AM modulation (a), demodulation (b), and DSB-SC modulation (c) with the required synchronous demodulation (d).

a constant ≪1 instead of 1 in Fig.5a. The receiver extracts the carrier for demodulation through a narrow bandpass filter as sketched in Fig.6.

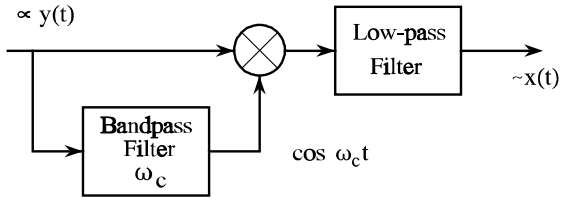


Fig.6 Demodulator principle for AM modulation with pilot carrier.
It is extracted and amplified by a band-pass filter.

Demodulation methods that require synchronization to the carrier are called coherent or synchronous. They are more fundamental than envelope detection. For instance, the AM signal may be coherently demodulated if the oscillators in Fig.5 a and d are synchronized.

Angle Modulations, PM and FM

Examples of Angle Modulation, $A(t) = A_{c0}$ ~ *constant*

$$\textit{phase modulation, PM} \;:\; y(t) = A_{c0}\cos(\omega_c t + \beta\, x(t)) \qquad (a)$$

$$\textit{frequency modulation, FM} \;:\; y(t) = A_{c0}\cos(\omega_c t + \varphi(t)), \quad \textit{where} \qquad (b) \tag{9}$$

$$\frac{d\varphi}{dt} = \Delta\omega(t) = \Delta\omega_{max}x(t) = 2\pi\Delta f_{max}x(t)\,. \qquad (c)$$

Angle modulated signals hold no information in the amplitude and may take the form of phase modulation, PM, where the phase of the modulated signal deviates from the phase of the carrier in proportion to the baseband signal. With a normalized baseband signal, $|x(t)|\leq 1$, the scaling constant β in Eq.(9)(a) is called the modulation index in angle modulations[1]. It specifies the maximum phase deviation from the carrier phase in either radians or degrees. The instantaneous frequency of the modulated signal is the time derivative of the total phase argument to the cosine factor in Eq.(1)(b). By frequency modulation, FM, the variation in instantaneous frequency from the carrier frequency is directly controlled by the baseband signal as shown by Eqs.(9)(b-c). The peak frequency deviation Δf_{max} replaces here β as the baseband signal scaling specification, but for sinusoidal modulation of a given frequency ω_x, the two scale factors are related,

1) This is the common definition when modulating with analog signals. With digital signals the modulation index is often given as the maximum phase excursion over a bit period in units of π, cf. Eq.(36).

$$x(t) = \cos\omega_x t \quad \Rightarrow \quad \varphi = \Delta\omega_{max} \int \cos\omega_x t \, dt = \frac{\Delta\omega_{max}}{\omega_x} \sin\omega_x t \, , \tag{10}$$

$$\beta = \frac{\Delta\omega_{max}}{\omega_x} = \frac{\Delta f_{max}}{f_x} \, . \tag{11}$$

The generation of angle modulation may be based on a phase modulating principle as sketched in Fig.7. In FM the baseband signal has to be integrated before it is applied to the modulator. This method is called indirect FM in contrast to direct FM that is shown in Fig.8. Here the basic building block is a voltage controlled oscillator, VCO. It has an output signal of constant amplitude and a instantaneous frequency, which is controlled linearly around the center frequency ω_c by an input voltage. The scaling factor K_V [Hz/Volt] is called the frequency gain in VCO terminology but it equals the peak frequency deviation Δf_{max} if the input x(t) is taken to be a voltage constrained to the interval ±1V.

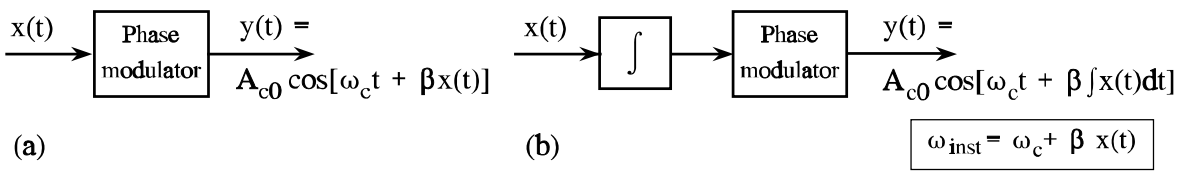


Fig.7 Generation of (a) phase modulation and (b) frequency modulation (indirect FM) with a phase modulator.

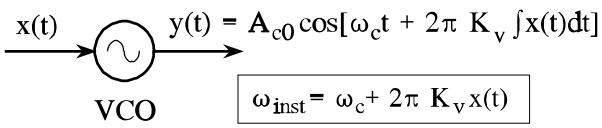


Fig.8 Principle of direct frequency modulation by a VCO, a voltage controlled oscillator.

Phase modulation concepts are mostly used in conjunction with transmission of digital signals and we shall postpone exemplifications to the discussion of that topic. One method of demodulating[2] FM signals in an asynchronous system is illustrated by Fig.9. The constant delay gives a phaseshift of $\pi/2$ (quadrature) at the carrier frequency. The difference term from the multiplication, cf.Eq.(3), transforms to a sine term through the identity

$$\cos(\, a - \pi/2 \,) = \sin a \, . \tag{12}$$

2) In angle modulation, demodulation is often called phase or frequency discrimination.

This component passes through the low-pass filter and approximates the original baseband messages according to Eq.(9)(c). The replacement of the sine term by its argument requires a small variation in phase over the period τ, i.e. the baseband message held in $\varphi(t)$ contains low-frequency components compared to the carrier frequency. The sum term from the multiplication gets twice the carrier frequency and it is removed by the low-pass filtering.

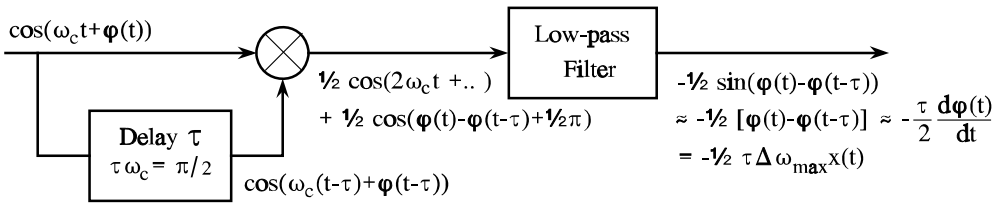


Fig.9 Quadrature FM demodulator block scheme. With lumped element approximation of the delay, cf. example II-7-2 in chapter II, the principle is common in receivers for FM broadcast.

With sinusoidal baseband signals PM and FM waveshapes get similar appearances as seen in the two lower curves of Fig.2. Although they look completely different from amplitude modulated signals, angle modulated signals imply also lower and upper sidebands around the carrier in their spectra. Due to the nonlinear relationship through the cosine factors in Eqs.(9)(a-b), this is most easily demonstrated if β is small, and we have the so-called narrowband PM or FM modulations. For $x(t)=A_x\cos\omega_x t$, $A_x \le 1$, and using the identity,

$$\cos(a+b) = \cos a \cos b - \sin a \sin b \ , \tag{13}$$

the FM modulated wave is written,

$$\begin{aligned} y(t)\,|_{FM} &= A_{c0}\cos(\omega_c t + \beta A_x \sin\omega_x t) \\ &= A_{c0}\cos\omega_c t\,\cos(\beta A_x \sin\omega_x t) - A_{c0}\sin\omega_c t\,\sin(\beta A_x \sin\omega_x t) \end{aligned} \tag{14}$$

By the estimations $\cos(a)\approx 1$ and $\sin(a)\approx a$ for $a \ll 1$ that are implied by the narrowband condition $\beta=\Delta\omega_{max}/\omega_x \ll 1$, and using the identity,

$$\sin a \sin b = \tfrac{1}{2}\cos(a-b) - \tfrac{1}{2}\cos(a+b), \tag{15}$$

the FM waveform under narrowband conditions, NBFM, may be approximated,

$$y(t)|_{NBFM} \approx A_{c0}\cos\omega_c t - \frac{A_{c0}}{2}\beta A_x \cos(\omega_c - \omega_x)t + \frac{A_{c0}}{2}\beta A_x \cos(\omega_c + \omega_x)t \ . \tag{16}$$

Comparing narrowband FM with the AM modulated wave from Eq.(4) it is seen, that the modulation indexes β and m play the same role. Both types of modulations have an upper and lower sideband at frequencies ω_x apart from the carrier and the only principal difference is the sign of the lower sideband component, so the amplitude spectrum of the narrowband

sinusoidally modulated FM signal equals the AM spectrum of Fig.4 setting m = β. It should also be realized that had we chosen baseband signal $x(t)=A_x \sin\omega_x t$, the results above transfer directly to so-called narrowband PM. The only difference to FM is the interpretation of the modulation index β according to Eq.(11). Without the narrowband assumptions, PM and FM modulations show more spectral components than the two $\omega_c-\omega_x$ and $\omega_c+\omega_x$ terms above, even if the baseband still holds only a single tone. We shall see this in the following example.

Example I-1-1 (Wideband FM and PM)

A single tone baseband signal $A_x \cos\omega_x t$ with scaling $|A_x|\leq 1$ is considered. We shall define an effective modulation index by

$$\beta_{eff} \equiv \beta A_x \ . \tag{17}$$

If this index becomes so large that the assumption for Eq.(16) is no longer valid, expressions for the sinusoidally modulated waveshape must be based on series expansions for the baseband factors in the two terms of Eq.(14). We have, cf. ref.[6] Eqs.9.1.42 and 43,

$$\begin{aligned} \cos(\beta_{eff} \sin\omega_x t) &= J_0(\beta_{eff}) + \sum_{n=2,even}^{\infty} 2J_n(\beta_{eff}) \cos n\omega_x t \,, \\ \sin(\beta_{eff} \sin\omega_x t) &= \sum_{n=1,odd}^{\infty} 2J_n(\beta_{eff}) \sin n\omega_x t \,. \end{aligned} \tag{18}$$

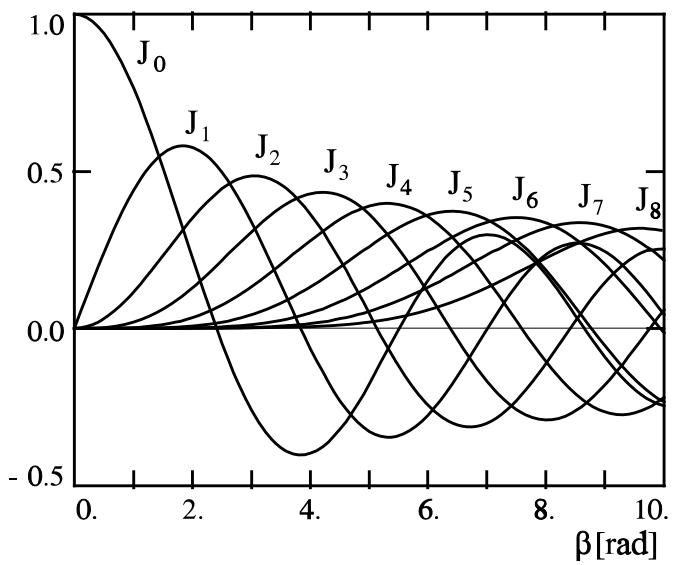


Fig.10 Bessel functions of the first kind and integer order. Tables may be found in ref.[6] but the functions are common in spreadsheets and mathematical programs like QuattroPro, Matlab, and Maple.

Table I $J_n(\beta_{eff})$, significant expansion coefficients from Bessel functions.

β_{eff}=0	1	2	3	4	5	6	7	8	9	10	n
1.0000	0.7652	0.2239	-0.2601	-0.3971	-0.1776	0.1506	0.3001	0.1717	-0.0903	-0.2459	0
	0.4401	0.5767	0.3391	-0.0660	-0.3276	-0.2767	-0.0047	0.2346	0.2453	0.0435	1
	0.1149	0.3528	0.4861	0.3641	0.0466	-0.2429	-0.3014	-0.1130	0.1448	0.2546	2
	0.0196	0.1289	0.3091	0.4302	0.3648	0.1148	-0.1676	-0.2911	-0.1809	0.0584	3
	0.0025	0.0340	0.1320	0.2811	0.3912	0.3576	0.1578	-0.1054	-0.2655	-0.2196	4
	0.0002	0.0070	0.0430	0.1321	0.2611	0.3621	0.3479	0.1858	-0.0550	-0.2341	5
		0.0012	0.0114	0.0491	0.1310	0.2458	0.3392	0.3376	0.2043	-0.0145	6
		0.0002	0.0025	0.0152	0.0534	0.1296	0.2336	0.3206	0.3275	0.2167	7
			0.0005	0.0040	0.0184	0.0565	0.1280	0.2235	0.3051	0.3179	8
			0.0001	0.0009	0.0055	0.0212	0.0589	0.1263	0.2149	0.2919	9
				0.0002	0.0015	0.0070	0.0235	0.0608	0.1247	0.2075	10
					0.0004	0.0020	0.0083	0.0256	0.0622	0.1231	11
					0.0001	0.0005	0.0027	0.0096	0.0274	0.0634	12
						0.0001	0.0008	0.0033	0.0108	0.0290	13
							0.0002	0.0010	0.0039	0.0120	14
							0.0001	0.0003	0.0013	0.0045	15
								0.0001	0.0004	0.0016	16
									0.0001	0.0005	17
										0.0002	18

The expansion coefficients $J_n(\beta_{eff})$ are Bessel functions of the first kind and integer order. They depend on argument β_{eff} as shown in Fig.10 and Table I. The outset from the narrow-band case is easily observed in the figure since, for small β_{eff}, J_0 equals one, J_1 is proportional to β_{eff} while all higher order functions approximate zero. With raising β_{eff}, terms of higher order than one get significance and more sidebands appear compared to the narrowband case in Eq.(16). Using the trigonometric identities for cosines and sine products, the FM signal for a cosine baseband tone - or the PM signal for a sine baseband tone - is obtained by inserting the expansions from Eq.(18) into Eq.(14) to yield

$$\begin{aligned} y(t)|_{FM\,or\,PM} &= A_{c0} J_0(\beta_{eff}) \cos(\omega_c t) \\ &+ A_{c0} \sum_{n=2,\,even}^{\infty} J_n(\beta_{eff}) \left[\cos(\omega_c t + n\,\omega_x t) + \cos(\omega_c t - n\,\omega_x t) \right] \\ &+ A_{c0} \sum_{n=1,\,odd}^{\infty} J_n(\beta_{eff}) \left[\cos(\omega_c t + n\,\omega_x t) - \cos(\omega_c t - n\,\omega_x t) \right] \\ &= A_{c0} \sum_{n=-\infty}^{\infty} J_n(\beta_{eff}) \cos(\omega_c t + n\,\omega_x t) \ . \end{aligned} \tag{19}$$

The last compaction includes negative order Bessel functions, where $J_{-n}(\beta)=(-1)^n J_n(\beta)$, cf.[6] Eq.9.1.5. Since the sums run to infinity, the bandwidth of the signal is in principle unlimited. In practice, however, an effective bandwidth may be defined as the one containing frequency components up to an order where 99% of the theoretical total power is included. Since the power contents of a sinewave is independent of frequency and the FM or PM modulated signal has constant envelope, the total power is simply half the squared carrier amplitude A_{c0}.

If $P_{ac,n}$ denotes the accumulated power of frequency components to order n in size, the relative power accumulation is a function of the effective modulation index only,

$$p_{ac,n}(\beta_{eff}) \equiv \frac{P_{ac,n}}{A_{c0}^2/2} = J_0^2(\beta_{eff}) + \sum_{i=1}^{n} 2J_i^2(\beta_{eff}) \ . \tag{20}$$

Table II shows these relative powers for various integer modulation indexes. For each index, the first figure where power accumulation exceeds 99% is printed in bold. Calling the corresponding order $n_{99\%}$ we have the following rule

$$n_{99\%} \approx \beta_{eff} + 1 \tag{21}$$

The spacing between the components in the FM or PM signal is that of the baseband sinusoid, ω_x, so the bandwidth required to transmit 99% of the theoretical power becomes

$$W_{99\%} = 2\,\omega_x\left(\beta_{eff} + 1\right) = \begin{cases} 2\left[\Delta\omega_{max} A_x + \omega_x\right] & \textit{FM modulation}, \\ 2\,\omega_x\left[\beta A_x + 1\right] & \textit{PM modulation}. \end{cases} \tag{22}$$

The splitting on modulation types here is based on Eq.(11). In PM it is the maximum phase excursion β, corresponding to the maximum effective modulation index in the limit of A_x=1, which characterizes the modulation and is assumed constant. Therefore, the bandwidth in PM is proportional to the baseband frequency ω_x. In the wideband case with β>1, the bandwidth also tends to become proportional with the amplitude A_x. In FM, the modulation is characterized by the maximum frequency deviation $\Delta\omega_{max}$ corresponding to a maximum amplitude of

Table II $p_{ac,n}(\beta_{eff})$, accumulated relative power of frequency components to and including order n in size. Figures in bold indicate the first 99% bound passing.

β_{eff}=0	1	2	3	4	5	6	7	8	9	10	n
1.0000	0.5855	0.0501	0.0676	0.1577	0.0315	0.0227	0.0901	0.0295	0.0082	0.0605	0
	0.9728	0.7154	0.2975	0.1665	0.2462	0.1758	0.0901	0.1396	0.1285	0.0643	1
	0.9992	0.9643	0.7701	0.4316	0.2505	0.2938	0.2718	0.1651	0.1705	0.1939	2
	1.0000	**0.9976**	0.9612	0.8017	0.5167	0.3201	0.3279	0.3346	0.2360	0.2008	3
		0.9999	**0.9960**	0.9598	0.8228	0.5759	0.3777	0.3568	0.3769	0.2972	4
		1.0000	0.9997	**0.9947**	0.9592	0.8381	0.6198	0.4258	0.3830	0.4068	5
			1.0000	0.9995	**0.9936**	0.9590	0.8499	0.6538	0.4665	0.4072	6
				1.0000	0.9993	**0.9926**	0.9590	0.8593	0.6809	0.5011	7
					0.9999	0.9990	**0.9918**	0.9592	0.8670	0.7032	8
					1.0000	0.9999	0.9987	**0.9911**	0.9594	0.8735	9
						1.0000	0.9998	0.9985	**0.9905**	0.9596	10
							1.0000	0.9998	0.9982	**0.9900**	11
								1.0000	0.9997	0.9980	12
									1.0000	0.9997	13
										1.0000	14

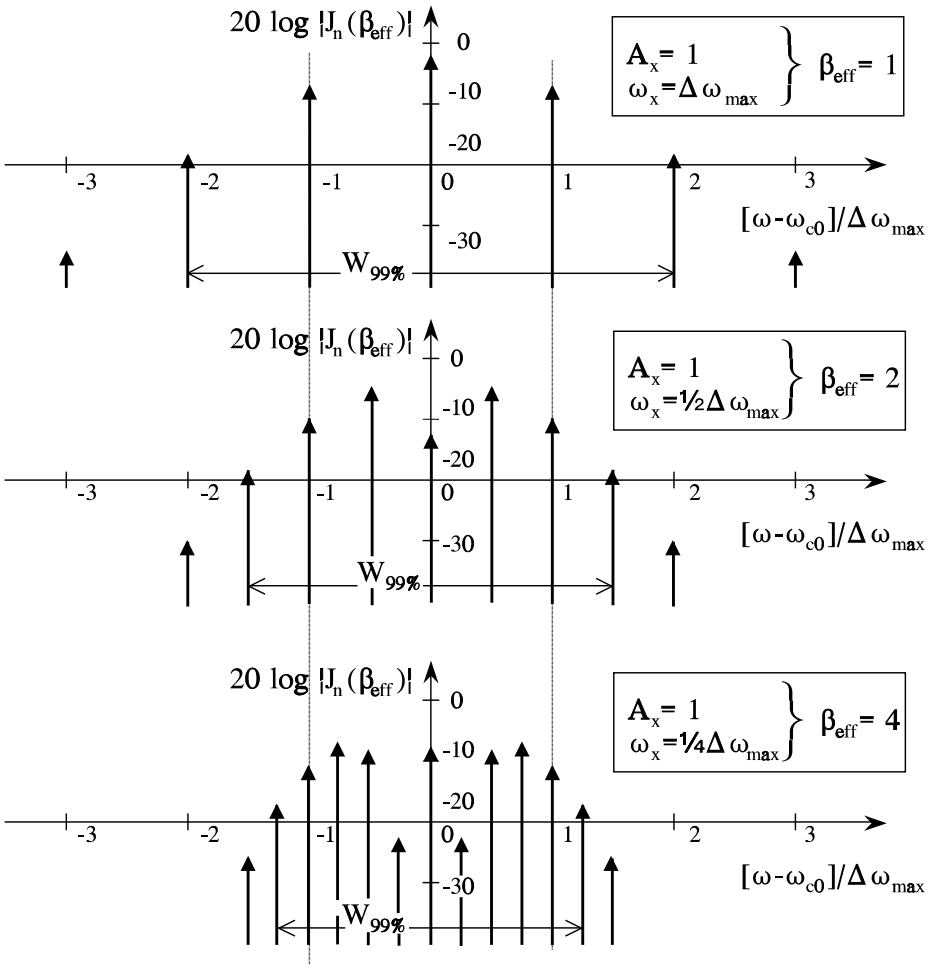


Fig.11 dB scaled FM spectra for single tone baseband signal of frequency equal to the whole, one half, and one fourth of the maximum frequency deviation, $\Delta\omega_{max}$

one. Keeping $\Delta\omega_{max}$ fixed, it is demonstrated by Fig.11, that the bandwidth approaches twice this parameter when the baseband frequency becomes a smaller and smaller fraction of the maximum deviation. In the limit where $\Delta\omega_{max} \gg \omega_x$, so the effective modulation index is large, Eq(22) shows that the bandwidth also in FM becomes proportional to amplitude A_x.

So far, we have considered a single tone baseband signal. With composite signals matters become more involved. For a periodic baseband signal, which may be described by a Fourier series, the resultant modulated waveform is not a sum of terms of the type above, one set for each baseband component. The baseband signal enters phase terms in either sine or cosine functions, so the modulation process is nonlinear. Therefore, a correct calculation of the FM and PM signal spectra in that case should start with series expansions of the type in Eq.(18), which, in the particular case of a sinusoidal baseband signal, gave the Bessel function coefficient. Only few other cases provide tractable analytical solutions and we shall not discuss them here but refer to [1], sec. 5.3 or [2] sec.5.2. It may be argued, however, that with an arbitrary baseband signal x(t), we may introduce modulation indexed according to

$$\textit{PM}: \quad \beta_{eff} \equiv \beta \ \max\{\,|\,x(t)\,|\,\}, \qquad \textit{FM}: \quad \beta_{eff} \equiv \Delta\,\omega_{max} \max\{\,|\,x(t)\,|\,\}/\,W_x\,, \qquad (23)$$

where W_x is the bandwidth of x(t). Using these definitions, Eq.(22) is known as Carson's rule, and it is still useful for estimating the bandwidth of the modulated signal.

Example I-1-1 end

Phasor Representation of Modulated Waveforms

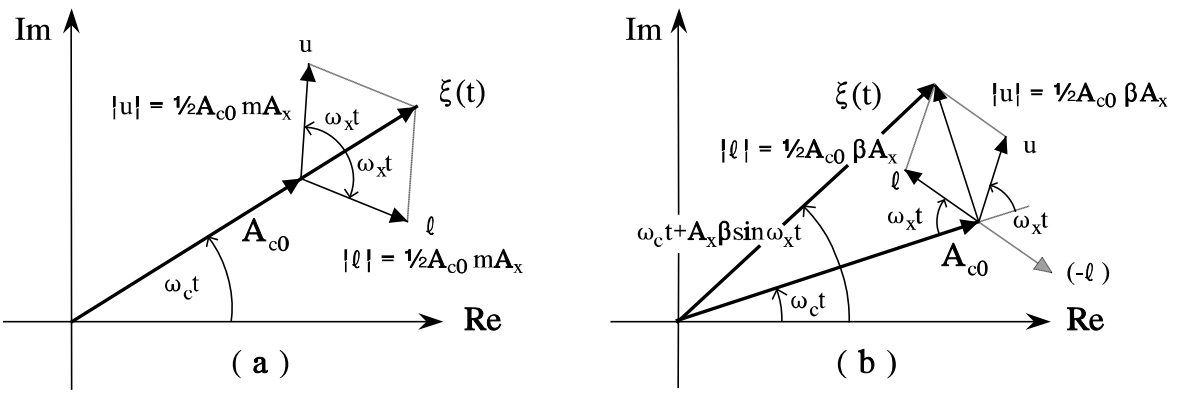


Fig.12 Phasor representation showing how the lower and upper sideband components, ℓ and u, add to the carrier A_{c0} in (a) AM modulation, and (b) narrowband FM. The modulated wave becomes y(t)=Re(ξ).

The contrast between the sideband terms in AM and narrowband FM is illuminated by the phasor representations in Fig.12. All terms in either Eq.(4) or Eq.(16) are taken as real valued projections of complex phasors at different angular frequencies. With AM the complex sum of the sideband components lie along the carrier. Due to the sign shift in the lower sideband term of narrowband FM, the sideband components here add to produce a component perpendicular to the carrier. The approximation in narrowband FM implies that the resultant length of the vector operations in Fig.12b stays close to the length of the carrier, i.e. $|\xi(t)| \approx A_{c0}$.

The phasor or complex number[3] representation above is a quite general tool when dealing with more complicated modulation types. The common formalism begins with Eq.(1)(b), which by Eq.(13) may be rewritten,

$$\begin{aligned} y(t) &= A(t)\cos(\varphi(t)+\varphi_0)\cos\omega_c t - A(t)\sin(\varphi(t)+\varphi_0)\sin\omega_c t \\ &= x_i(t)\cos\omega_c t - x_q(t)\sin\omega_c t\ , \qquad \textit{where} \end{aligned} \tag{24}$$

$$x_i(t) = A(t)\cos(\varphi(t)+\varphi_0), \qquad x_q(t) = A(t)\sin(\varphi(t)+\varphi_0)\ . \tag{25}$$

3) The terms phasor, complex number, and vector representations of modulated signals are used synonymously in the literature.

Eqs.(24) and (25) are called the quadrature-carrier or the I-Q representation of a modulated waveform. The naming refers to a coordinate system that rotates with the carrier, i.e. with the angular frequency ω_c. The "in phase" $x_i(t)$ and "quadrature" $x_q(t)$ components are the instant projections of the modulated signal vector $\xi(t)$ on the axis along the carrier and the axis perpendicular (in quadrature) to the carrier. The two examples of Fig.12 have the components

$$\begin{aligned} AM: \quad & x_i(t) = A_{c0}[1 + m x(t)], & x_q(t) &= 0. \\ & \quad\;\; = A_{c0}[1 + m A_x \cos\omega_x t] & & \\ narrowband\ FM: \quad & x_i(t) \approx A_{c0}, & x_q(t) &\approx A_{c0}\Delta\omega_{max}\int x(t)dt, \\ & & &\approx A_{c0}\beta\, A_x \sin\omega_x t. \end{aligned} \tag{26}$$

In a later context we shall benefit from a complex formalism for handling modulated signals, so it is worth already at this stage to observe that Eqs.(24) is equivalent to

$$y(t) = Re\left\{[x_i(t) + jx_q(t)]\, e^{j\omega_c t}\right\} = Re\left\{\zeta(t)\, e^{j\omega_c t}\right\} = Re\{\xi(t)\}, \tag{27}$$

where $\zeta(t)=x_i(t)+jx_q(t)$ is called the complex envelope of the modulated signal.

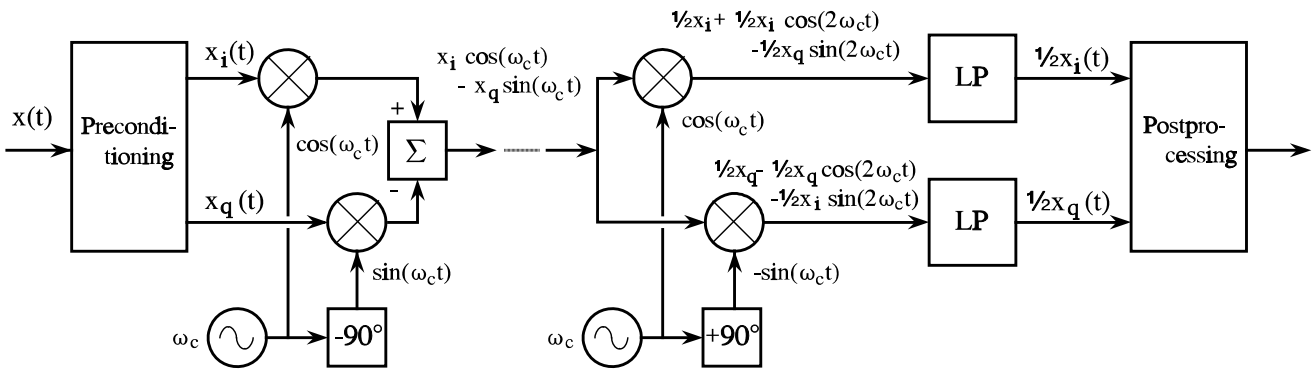


Fig.13 Quadrature modulator/demodulator pair. Note the sign shift in the demodulator sine carrier generation that corresponds to the subtraction in the modulator output.

Once the baseband signal is transformed into the components $x_i(t)$ and $x_q(t)$, the process of composing an arbitrarily modulated waveform and recover the two components again after undisturbed transmission may formally be made by the block scheme in Fig.13. The composite signals after the multipliers at the demodulator side include components based on the identity,

$$\sin a \cos b = \tfrac{1}{2}\sin(a - b) + \tfrac{1}{2}\sin(a + b)\ , \tag{28}$$

where the first sine term disappears if a = b. Besides being a vehicle for analysis of CW modulated communication systems, practical circuits may be build directly on basis of the quadrature-carrier representation. They are called vector or I/Q modulators and an example is given below in Example I-1-2. Fig.14 shows a realization of the phase modulation from Eq.(9)(a) where preconditioning of the baseband signals is the mapping through nonlinear sine and cosine characteristics. A subset of this scheme is utilized in Fig.15 to make an indirect narrowband FM modulator of the type outlined by Fig.7.

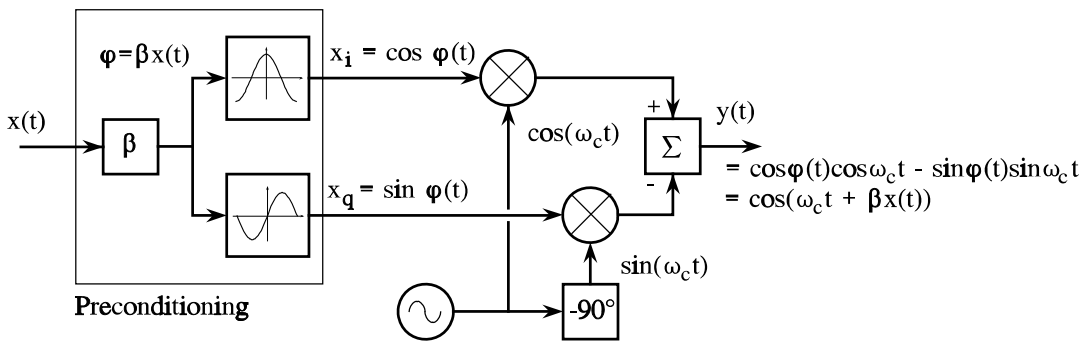


Fig.14 Phase modulator block scheme based on the quadrature-carrier resolution in Eqs.(24),(25). Note the nonlinear blocks holding cosine and sine functions.

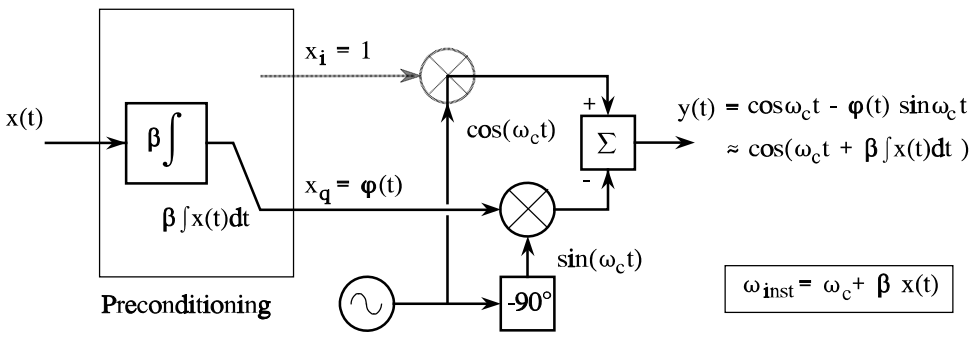


Fig.15 Indirect narrowband FM modulation scheme derived from Fig.14 by imposing the assumptions $\cos(\varphi)\approx 1$, $\sin(\varphi)\approx\varphi$ for $\varphi \ll 1$.

In case of amplitude modulated signals, the general quadrature-carrier representation may seem overly complicated. With no phase excursions from the carrier, there is no need to map the original baseband signal x(t) onto in-phase and qaudrature components. For instance, the DSB-SC scheme in Fig.5 could be either the transmission of $x_i(t)$ or $x_q(t)$ in Fig.13 setting the other one to zero. However, we could also use the quadrature principle to transmit two independent AM modulated signals simultaneously in the $x_i(t)$ or $x_q(t)$ branches, a technique that is called qaudrature AM or QAM. If the originating baseband signals in this case have equivalent bandwidths, so do their modulated counterparts. Within the same frequency band of transmission two messages may now be send and demodulated.

Example I-1-2 (SSB, single sideband modulation)

Another way to obtain efficient use of a given frequency band is to transmit only one of the sidebands from a AM-DSB modulation, which is the so-called single-sideband modulation, SSB. To generate this signal, the modulator may again be constructed to manipulate the in-phase and quadrature components of the signal to the desired property. The principle is shown by Fig.16 in the simple case of a sinusoidal baseband signal $x(t)=A_x\cos\omega_x t$. Preprocessing is here an introduction of a 90° phaseshift between the baseband components of the in-phase and quadrature branches. The sign of the shift determines whether it becomes the upper or the lower sideband that is produced.

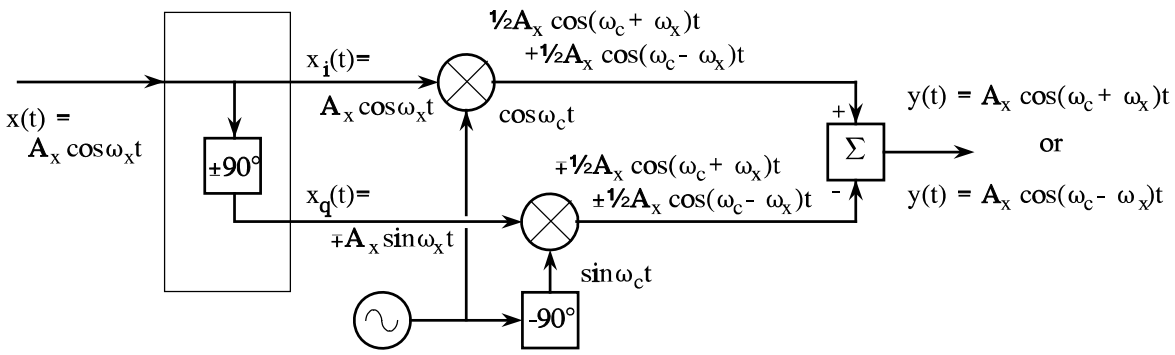


Fig.16 Phasing method of single sideband, SSB, modulation. The sign of the 90° phase-shifter determines whether the upper or lower sideband is produced.

Single sideband signals may be demodulated by the synchronous detector from Fig.5d, if it is possible to reconstruct the carrier. With a reconstructed carrier of $\cos(\omega_c t+\theta)$, θ represents a possible synchronization error, the detected signal before low-pass filtering in case of upper sideband transmission becomes

$$\begin{aligned} u(t) &= A_x\cos(\omega_c t+\omega_x t)\cos(\omega_c t+\theta) \\ &= \frac{A_x}{2}\cos\omega_x t\cos\theta + \frac{A_x}{2}\sin\omega_x t\sin\theta + \frac{A_x}{2}\cos(2\omega_c t+\ldots)\ . \end{aligned} \tag{29}$$

The first and the last terms here are equivalent to the two terms in the DSB-SC case from Eq.(8), where the last one is removed by filtering. The middle term is a new type of error - a distortion - that occur in SSB demodulation in addition to the baseband signal attenuation in the first term if θ differs from zero due to inaccurate synchronization.

Synchronous demodulation of SSB signals has the drawback that any signal component in the frequency band corresponding to the removed sideband must be absent from the input. Suppose the input signal in addition to the desired upperband A_x component includes a foreign component A_{lo} in the frequency of the removed sideband. In that case the output from error free synchronous detection corresponding to Eq.(29) becomes

$$\begin{aligned} u(t) &= A_x \cos(\omega_c t + \omega_x t)\cos\omega_c t + A_{lo}\cos(\omega_c t - \omega_x t)\cos\omega_c t \\ &= \frac{A_x + A_{lo}}{2}\cos\omega_x t + \frac{A_x}{2}\cos(2\omega_c t + \ldots) + \frac{A_{lo}}{2}\cos(2\omega_c t + \ldots)\ . \end{aligned} \tag{30}$$

Clearly, the desired output is now contaminated by the foreign component. It should be removed by filtering before the synchronous demodulator in order improve the situation. In the limit, however, the undesired component A_{lo} may represent unavoidable noise, so influence from the frequency range corresponding to the removed sideband cannot be totally neglected with this, in principle, rather simple demodulation scheme. To overcome that type of problems, i.e. to suppress responses in removed sideband, we must apply a demodulation scheme of the type that more closely follows the general demodulator pattern from Fig.13, for instance the SSB demodulator in Fig.17. As indicated by the figure, the quadrature demodulation allows for transmission of either upper or lower sidebands while the corresponding undesired frequency band, lower or upper sideband, cancels. It is the sign of the 90° phase shifter that determines the selection among the two possibilities.

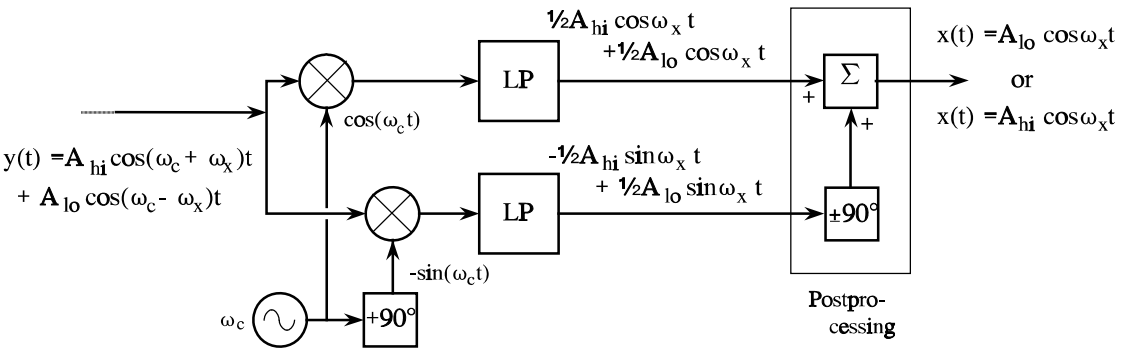


Fig.17 Single sideband demodulator scheme. The sign of the 90° phaseshifter determines whether the signal above, A_{hi} or below, A_{lo} the carrier is demodulated.

SSB modulation is used extensively to transfer analog voice signals in radio telephony. It is less suited for digital data communications. To see this we consider a general input x(t) instead of the sinusoid that was used to illuminate the modulator operation in Fig.16. A 90° phase shifting that applies to all frequencies of x(t) is called the Hilbert transformation, sometimes denoted $\hat{x}(t)$. It has transfer function $H(\omega)=-j$ for $\omega>0$ and $H(\omega)=j$ for $\omega<0$. The corresponding impulse response h(t) follows from the transform pair

$$H(\omega) = -j\,sgn(\omega) \quad \leftrightarrow \quad h(t) = \frac{1}{\pi t}\ , \tag{31}$$

where sgn(ω) is the signum function. If the input signal is a pulse

$$x(t) = x_i(t) = \begin{cases} 1, & |t| \le T/2 , \\ 0, & otherwise , \end{cases} \tag{32}$$

the quadrature component of the baseband signal is expressed by the convolution

$$x_q(t) = \hat{x}(t) \equiv x(t) \star h(t) = \frac{1}{\pi} \int_{-T/2}^{T/2} \frac{1}{t-\tau} \, d\tau = \frac{1}{\pi} \log \left| \frac{T/2 - t}{T/2 + t} \right| . \tag{33}$$

With pulse input the quadrature component $x_q(t)=\hat{x}(t)$ gets singularities at the pulse boundaries as shown in Fig.18. This means that the mixer in the quadrature branch of the modulator, the summing component and any subsequent power amplifiers momentarily should be capable of delivering infinitely large output signals, which clearly is an unrealistic requirement. Restricting output from the modulator distorts the resultant pulsed signals, and this is one of the reasons why SSB is not used in data communications. Another reason is that we get spectral efficiency corresponding to SSB with simpler means in the quaternary digital modulations that are presented in section I-3.

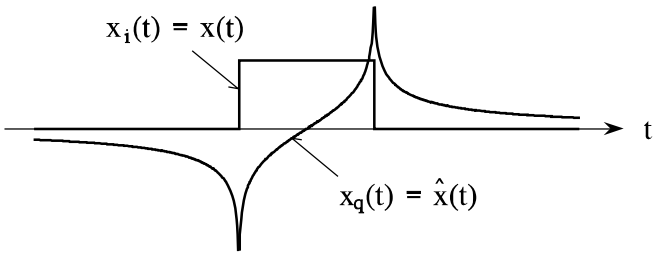


Fig.18 Baseband components for a pulse input to the SSB modulator in Fig.16. In-phase $x_i(t)$ duplicates the pulse. Quadrature component $x_q(t)$ approaches infinity at pulse boundaries.

Example I-1-2 end

Example I-1-3 (vector modulator IC)

Fig.19 shows an example integrated circuit implementation of a vector modulator, which includes an output amplifier after the summing point. RF integrated circuits have often differential or balanced signal paths, i.e. the signal are presented as the voltage between two dedicated conductors instead of the single ended voltage difference between one signal lead and a common ground. In integrated circuits signals are in most cases processed differentially, as it is difficult to establish a reference ground inside the circuits.

HEWLETT PACKARD

Silicon Bipolar Monolithic QPSK Modulator

Technical Data

HPMX-2001

Features

- **I/Q Bandwidth: DC - 700 MHz**
- **LO Frequency Range: DC-2000 MHz**
- **Standby Mode - P_d 10 mW**
- **Low Cost Plastic Surface Mount Package**

Applications

- **Digital Cellular Radio (eg., GSM, ADC)**
- **RF Data Links**
- **Vector Generators**
- **AM Modulator**
- **Single Sideband Mixer**

Description

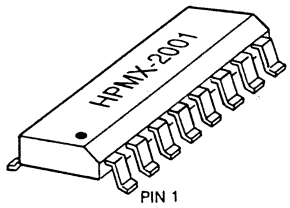
The HPMX-2001 is a silicon monolithic quadrature phase shift keyed modulator in a plastic surface mount SO-16 package. It is designed for wide or narrow band applications and has a typical LO operating frequency range of DC-2000 MHz and typical I/Q bandwidth of DC-700 MHz.

The modulator may be operated in any combination of single or differential input/output configurations. The LO inputs are self biasing and have 50 Ohm termination impedances. The I & Q inputs are open base and require 2.5 V DC bias. This allows direct coupling to the I & Q inputs. The RF output is designed to drive a 50 Ohm load and has a saturated output power level of -5.5 dBm.

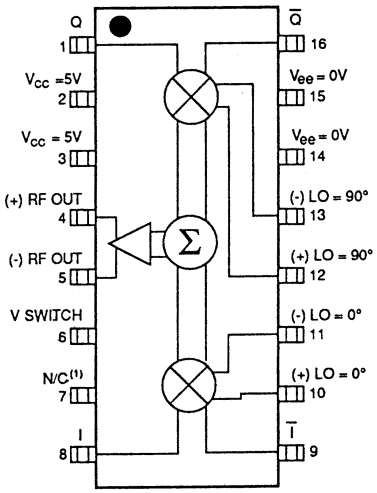
The internal signal paths of the modulator are fully differential to help reduce susceptibility to common mode noise. The small amplitude and phase imbalance of the device, 0.5 dB and 1 degree respectively, makes the modulator suitable for many communication applications. The HPMX-2001 also features a standby mode in which the device consumes only 10 mW of power.

The HPMX-2001 is manufactured using Hewlett-Packard's 13 GHz F_t, 25 GHz F_{max} silicon bipolar integrated circuit process.

Plastic SO-16 Package



Functional Block Diagram and Pin Configuration



NOTE 1: DO NOT GROUND PIN 7.

Fig.19 Vector modulator IC example, HPMX-2001 from Hewlett Packard, Communications Components, Designer´s Catalog 1993.

Example I-1-3 end

I-2 Binary Digital Modulations, ASK, PSK, FSK

The term analog modulation is used if the purpose of a modulation, transmission, and demodulation process is to reconstruct the baseband signal x(t) in the receiver. Typical examples are x(t) holding speech, music, or television signals in analog form. If the message to be impressed upon a modulated waveform originates from a digital signal, the goal at the receiver side is to minimize the probability of making wrong detections of the digits being send using initial knowledge about the waveforms to choose among.

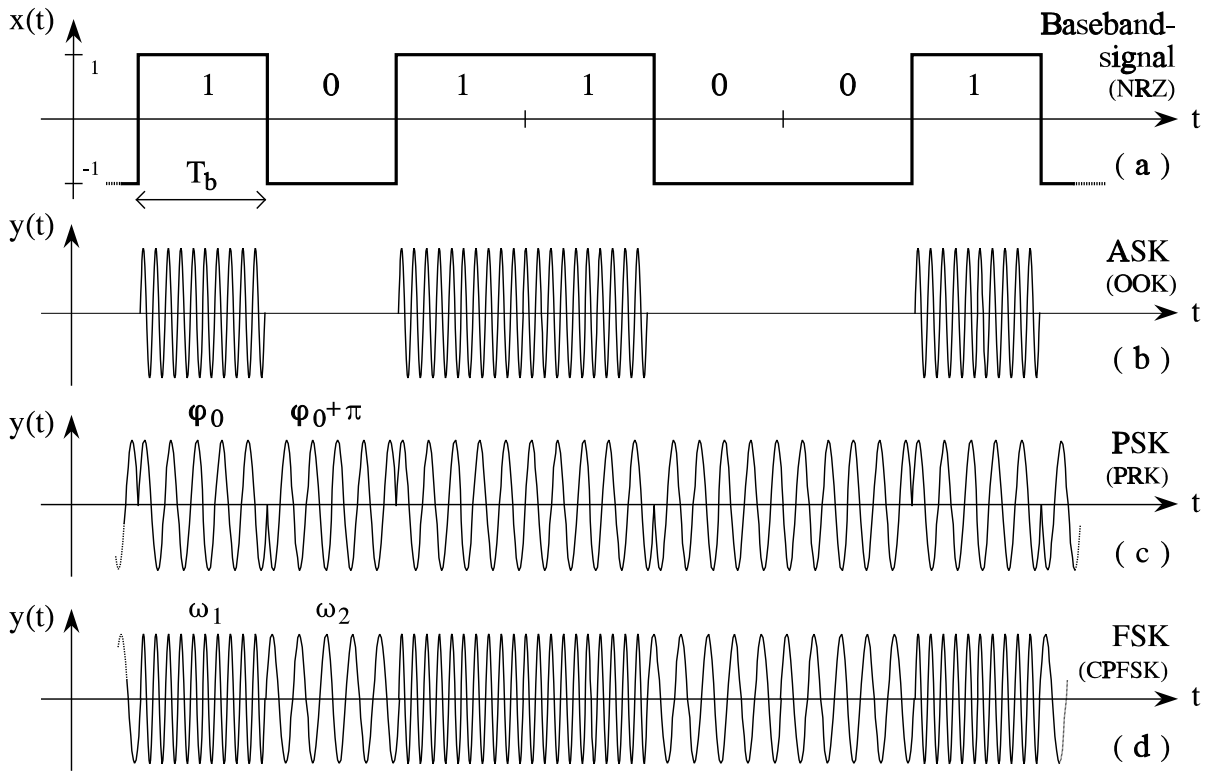


Fig.20 Examples of digitally modulated waveforms. ASK, amplitude shift keying (OOK, on-off keying). PSK, phase shift keying (PRK, phase reversal keying). FSK, frequency shift keying (CPFSK continuous phase FSK variant).

If a digital message modulates a carrier bit by bit using two distinct waveforms, the process is called binary digital modulation. Examples are shown in Fig.20. They are the simple digital counterparts to the analog modulation types from Fig.2. The upper curve shows the baseband signal that imposes the bitstream to the modulators. This function is an important design objective, which is chosen here to let logical 1 translates to +1 and logical 0 to -1 throughout the duration of each bit. A baseband signal that this way stays constant in the bit period is termed NRZ, non return to zero.

Applying the digital baseband signal to the AM modulator from Fig.5a produces an amplitude shift keying signal, ASK. The particular example in Fig.20b has an AM modulation index m=1, so no signal is transmitted with logical zero. This is called on-off keying, OOK.

Phase modulation by a digital signal is called phase-shift keying, PSK. The case shown in Fig.20c corresponds to a phase difference of π, which for obvious reasons also is called phase-reversal keying, PRK, but it should be mentioned, that the term binary phase-shift keying, BPSK, often is associated with this particular choice of the phase difference between the two bit signals. Observe also that the same modulated wave results if the baseband modulates a carrier in DSB-SC amplitude modulation.

By frequency-shift keying, FSK, the instant frequency of the transmitted wave follows the baseband signal. A digitally controlled FSK wave may be produced by switching between two oscillators at fixed frequencies ω_1, ω_2. Applying the digital baseband to a FM modulator of the type in Fig.8 assures phase continuity like the example in Fig.20d. This is a common type of FSK in radio communication and it is abbreviated CPFSK, continuous-phase frequency-shift keying.

All examples in Fig.20 have the bit period T_b set to an integral multiple of the carrier period. This is not a severe restriction as in most new communication systems all oscillator and timing signals are synthesized from the same source. We shall generally adopt the assumption below. The examples may be written in the form,

$$y(t) = \sum_k s_b(t - kT_b), \quad s_b(t) = \begin{cases} s_1(t), & \textit{logical 1 signal}, \\ s_0(t), & \textit{logical 0 signal}. \end{cases} \tag{34}$$

where k sums over all bits in the message. The terms in the sum are non-overlapping, because the two signal waveforms are zero outside a bit-period. The particular signals are,

$$\textit{Binary Signal Waveforms } s_1(t),\ s_0(t) \textit{ for } 0 \le t \le T_b$$
$$s_1(t),\ s_0(t) = 0 \textit{ for } t<0 \textit{ or } T_b < t$$

$$\begin{aligned}
ASK_{OOK}:\ & s_1(t) = A_{c0}\cos(\omega_c t - \tfrac{1}{2}\pi) && s_0(t) = 0, \\
& = A_{c0}\sin\omega_c t && (a) \\
PSK_{PRK}:\ & s_1(t) = A_{c0}\cos(\omega_c t - \tfrac{1}{2}\pi) && s_0(t) = A_{c0}\cos(\omega_c t + \tfrac{1}{2}\pi) \\
& = A_{c0}\sin\omega_c t, && = -A_{c0}\sin\omega_c t, \qquad (b) \\
FSK:\ & s_1(t) = A_{c0}\cos(\omega_c t + \Delta\omega t + \theta_k), && s_0(t) = A_{c0}\cos(\omega_c t - \Delta\omega t + \theta_k), \\
& \omega_c = \tfrac{1}{2}(\omega_1 + \omega_2), && \Delta\omega = \tfrac{1}{2}(\omega_1 - \omega_2). \qquad (c)
\end{aligned} \tag{35}$$

The phase offset of ½π in (a),(b) have been introduced to let the signals agree with the figures. They are drawn continuous at bit boundaries that encompass t=0. The FSK waveform in (c) includes a phase offset θ_k for the initial phase at bit number k. With CPFSK it serves the purpose of accumulating the phase changes throughout all foregoing bits.

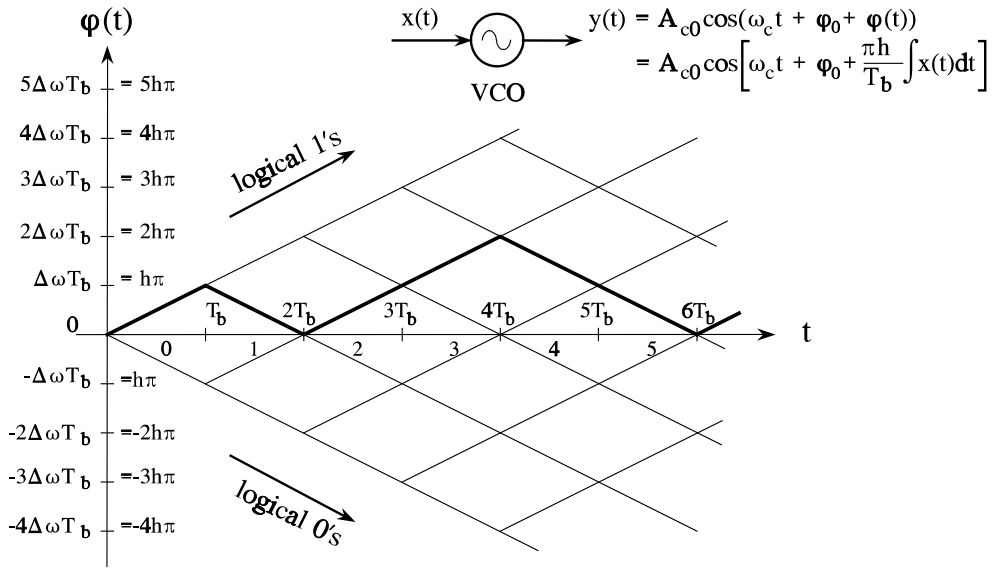


Fig.21 Phase tree in continuous phase frequency shift keying, CPFSK. The heavy line corresponds to the phase time function for the baseband signal in Fig.20a.

Constraining the phase offsets θ_k in FSK by a phase continuity requirement may be examined through the phase tree in Fig.21. With initial phase deviation set to zero, the tree shows how the phase patterns may emerge when the instantaneous frequency - controlled by the baseband signal x(t) - is integrated. Besides the frequency deviation $\Delta\omega=2\pi\Delta f$, the scaling factor in digital FSK is often given through a modulation index h, which is defined by

$$h \equiv 2\,\Delta f\,T_b\,, \tag{36}$$

A rational modulation index, i.e. h=n/m with integers n,m, gives a finite number of modulo 2π phase offsets at bit boundaries. The simplest cases are indicated by Fig.22. It flashes the state of the baseband signal onto an in-phase and quadrature baseband components plane at the bit boundaries, so it is an instant picture of the complex envelope vector from Eq.(27). We shall see below that to get conditions for optimal detection with digital modulations, the receiver must hold or reproduce the transmitted signals exactly in time. In that respect special cases with few initial states are attractable where Fig.22b is known as Sunde´s FSK and Fig.22d makes the foundation of the so-called minimum-shift keying, MSK.

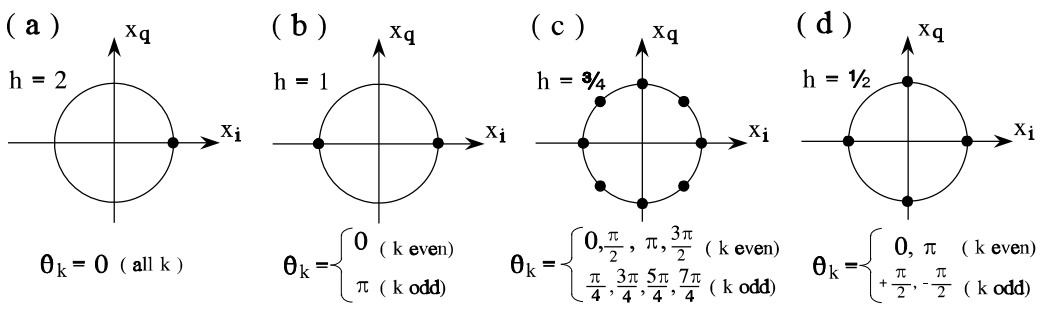


Fig.22 Examples of CPFSK phase-offsets at bit boundaries with $\varphi_0 = 0$ for simple, rational modulation indexes h.

Optimal Detection of Binary Modulated Signals

A hypothetical scheme for demodulation binary digitally modulated signal is given in Fig.23. The incoming signal is applied to a filter with impulse response h(t). The output of the filter is sampled after each bit period and the sampled value is compared to a threshold value Z_{tr} to decide whether a $s_1(t)$ or a $s_0(t)$ signal was received. The sampling instants t_k are assumed to be exactly synchronized to the bit periods in y(t). This is not a straight-forward process, but its details are presently disregarded.

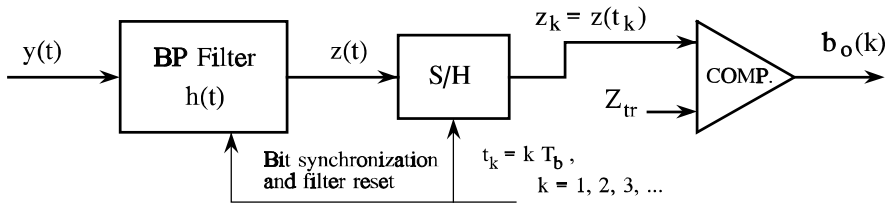


Fig.23 Detection principle for binary digital modulated signals. The operation of the S/H and comparator blocks are summarized by Fig.24 and Fig.25.

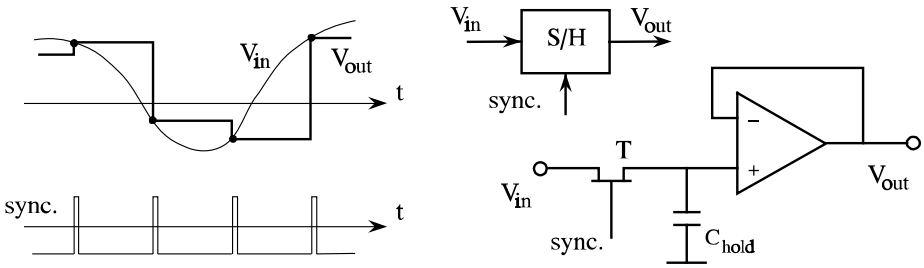


Fig.24 Sample and hold function and circuit. The input voltage is sensed at the synchronization instants where T switches on and charges the capacitor. This voltage is buffered to the output and kept constant while T is off.

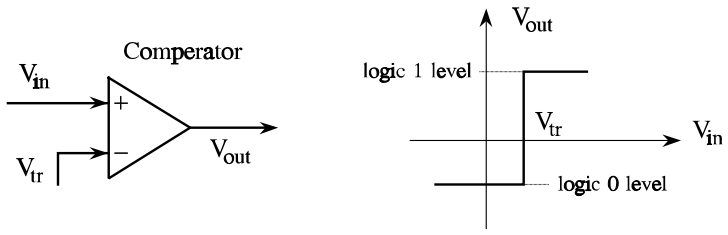


Fig.25 Comparator characteristic.

The response z_b of the filter to input signal s_b at the end of a bit period, where the filter is initially at rest, i.e. all capacitor voltages and inductor currents are zero, is given by the convolution integral,

$$z_b = s_b(t) \star h(t) \equiv \int_0^{T_b} s_b(\lambda)\, h(T_b - \lambda)\, d\lambda = \int_0^{T_b} s_b(T_b - \lambda)\, h(\lambda)\, d\lambda \,, \qquad (37)$$

To make the detector insensitive to transmission disturbances or design approximations, the filter should be devised to maximize the distance between the two possible responses at the end of a period or, equivalently, the square of the difference

$$| z_1 - z_0 |^2 = \left| \int_0^{T_b} [s_1(T_b - \lambda) - s_2(T_b - \lambda)] \, h(\lambda) \, d\lambda \right|^2 . \tag{38}$$

The squared integral complies with Schwartz inequality that states,

$$\left| \int_{-\infty}^{\infty} V(t) \, W^*(t) \, dt \right|^2 \le \int_{-\infty}^{\infty} | V(t) |^2 \, dt \int_{-\infty}^{\infty} | W(t) |^2 \, dt , \tag{39}$$

where V(t) and W(t) are two real or complex valued functions and where equality of the two sides is obtained if and only if W(t) is proportional to V(t). To use this result to the problem in Eq.(38) it is noticed, that the differences in integral bounds get no significance if h(t), like the functions $s_1(t)$ and $s_0(t)$, vanishes outside the bit interval. To maximize the left hand side of Eq.(38), the optimizing impulse response of the filter is a scaled, mirrored, and in principle complex conjugated version of the difference between the two basic signal waveforms. With this choice, the squared distance between the two responses may be expressed[4],

$$| z_1 - z_0 |^2 \le | z_1 - z_0 |^2_{max} = \int_0^{T_b} | s_1(t) - s_0(t) |^2 \, dt \int_0^{T_b} | h_{opt}(t) |^2 \, dt \tag{40}$$

$$h_{opt}(t) = K_0 \left\{ s_1^*(T_b - t) - s_0^*(T_b - t) \right\} . \tag{41}$$

K_0 is an arbitrary gain factor for the filter that is implicitly contained also at the left hand side distances in Eq.(40).

Influence on the maximum detection distance from the characteristics of the original signal waveforms are more clearly revealed by rewriting the integral,

$$\begin{aligned} \int_0^{T_b} | s_1(t) - s_0(t) |^2 dt &= \int_0^{T_b} |s_1(t)|^2 dt + \int_0^{T_b} |s_0(t)|^2 dt - 2 \int_0^{T_b} Re\{ s_1(t) s_0^*(t) \} dt \\ &= E_1 + E_0 - 2 \, Re\{ \rho_{10} \} \sqrt{E_1 E_0} . \end{aligned} \tag{42}$$

Thinking of s_1 and s_0 being voltages across a 1Ω resistor, E_1 and E_0 are recognized as the energy per bit of the two signals. ρ_{10} is the correlation coefficient between the two signals, which is defined by,

---

4) Although the examples in Eq.(35) are all real, we maintain complex notation in basic developments to emphasize their scopes for future uses of the results.

$$\rho_{10} \equiv \frac{\int_0^{T_b} s_1(t)\, s_0^*(t)\,dt}{\sqrt{\int_0^{T_b} |s_1(t)|^2 dt \int_0^{T_b} |s_0(t)|^2 dt}} = \frac{\int_0^{T_b} s_1(t)\, s_0^*(t)\,dt}{\sqrt{E_1\, E_0}} \,. \tag{43}$$

In case the two waveforms have equal amplitudes, their energies are equal, and the detectable difference depends heavily on the correlation coefficient. If it is zero, the two signals are said to be orthogonal. If the correlation is -1, which implies the greatest detection distance, the signals are called antipodal.

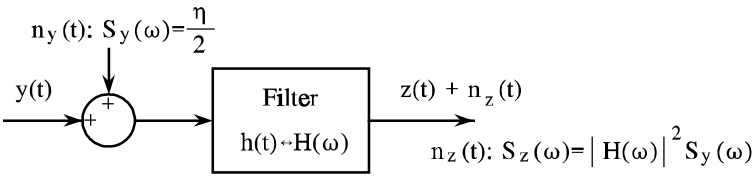


Fig.26 Noise components. The input noise power spectral density $S_y(\omega)$ gives the output noise power spectral density $S_z(\omega)$.

In realistic environments the signal that reaches the detector is contaminated by noise, which, as shown by Fig.26, is modeled by a single input noise source $n_y(t)$ that collects all noise contributions. It is a random process characterized by its spectral density $S_y(\omega)$ and its amplitude probability distribution function. We shall go in more details with noise later, here it suffices to recall that if the input noise to a linear filter is Gaussian distributed so will be the output. The power of the output noise becomes the variance of $n_z(t)$, $\sigma_z^2 = \langle |n_z(t)|^2 \rangle$, but also the integral of the corresponding power distribution function across the entire frequency range. Assuming Gaussian input noise with constant double-sided spectral density ½η[W/Hz], we have a so-called additive white Gaussian noise channel[5], AWGN. Now $n_z(t)$ gets the power spectral density,

$$S_z(\omega) = |H(\omega)|^2 S_y(\omega) = \frac{\eta}{2} |H(\omega)|^2 \,, \tag{44}$$

where H(ω) is the filter transfer function. By this we get

$$\sigma_z^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_z(\omega)\,d\omega = \frac{\eta}{2} \frac{1}{2\pi} \int_{-\infty}^{\infty} |H(\omega)|^2 d\omega = \frac{\eta}{2} \int_0^{T_b} |h(t)|^2 dt \,. \tag{45}$$

The last expression follows from Rayleigh's theorem,

$$f(t) \leftrightarrow F(\omega) \quad \Rightarrow \quad \int_{-\infty}^{\infty} |f(t)|^2\, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(\omega)|^2\, d\omega \,, \tag{46}$$

5) η stands for kT, the product of Boltzman´s constant k=1.3807 10^{-23}[J/K] and the absolute temperature T in Kelvin [K].

and the fact, that H(ω) and the impulse response $h_{opt}(t)$ make a Fourier transform pair. Inserting into Eq.(40) gives

$$\frac{|z_1 - z_0|}{2\sigma_z} \le \sqrt{\frac{1}{2\eta}\left[E_1 + E_0 - 2\,\mathbf{Re}\{\rho_{10}\}\sqrt{E_1 E_0}\right]}\ , \tag{47}$$

where equality is cautioned by the use of $h(t) = h_{opt}(t)$ from Eq.(41) in calculating the noise variance by Eq.(45). We shall see below that this ratio is important for computing detection error probabilities. It gives the maximum detection distance relatively to twice the noise deviation at the output of an optimal filter expressed through the signal and noise conditions at the input to the filter.

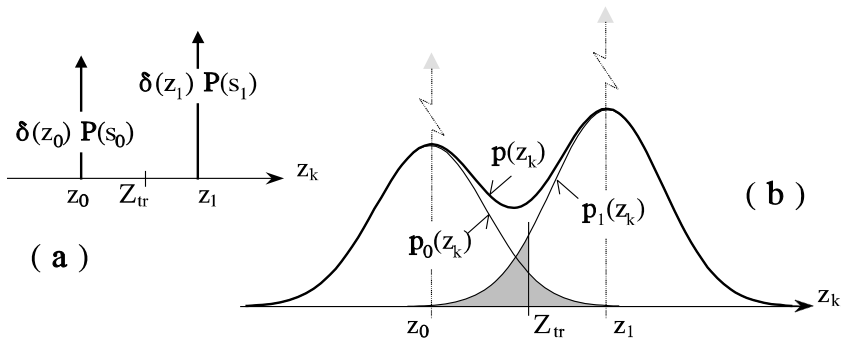


Fig.27 Probability distributions of the sampled signals excluding (a) and including (b) noise. The hatched areas in (b) represent the detection error probability, P_{error} from Eq.(50).

Suppose the probabilities of transmitting a logical 1 and a logical 0 are $P(s_1)$ and $P(s_0) = 1 - P(s_1)$ respectively. Had there been no noise, the probability distribution of the sampled response $z_b(k)$ becomes the two impulses in Fig.27a. There is no possibility of making a wrong detection by the comparator of Fig.23, if the threshold is chosen anywhere in the interval $z_0 < Z_{tr} < z_1$. The presence of noise smears out the probability distribution of z(k) as demonstrated by Fig.27b. It is given by

$$p(z_k) = p_0(z_k) + p_1(z_k) = \frac{1}{\sigma_z}\, p_N\left(\frac{z_k - z_0}{\sigma_z}\right) P(s_0) + \frac{1}{\sigma_z}\, p_N\left(\frac{z_k - z_1}{\sigma_z}\right) P(s_1)\ , \tag{48}$$

where $p_N(x)$ is the Gaussian, normal distribution function of zero mean and unit variance. This is the version that is commonly tabulated, and it relates to the Gaussian distribution with mean value x_0 and variance σ_x^2 through,

$$p_G(x,\, x_0,\, \sigma_x) = \frac{1}{\sigma_x\sqrt{2\pi}}\, e^{\frac{-(x-x_0)^2}{2\sigma_x^2}}\Bigg|_{\substack{x_0=0\\ \sigma_x=1}} = p_N(x) = \frac{1}{\sqrt{2\pi}}\, e^{\frac{-x^2}{2}}\ . \tag{49}$$

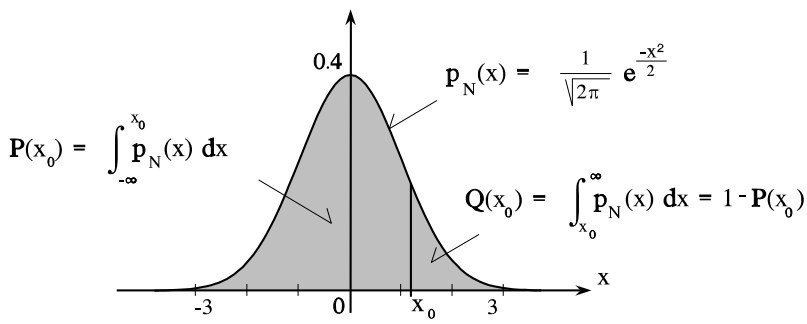


Fig.28 Gaussian probability density function with zero mean value and unit variance. The integrals $P(x_0)$ and $Q(x_0)$ are the probabilities of $x<x_0$ and $x>x_0$ respectively.

The shape of the Gaussian distribution is summarized by Fig.28 together with the two integrals $P(x_0)$, $Q(x_0)$[6], which express the probabilities of the random variable x being $< x_0$ or $> x_0$ respectively. The two cases are exclusive and exhaustive so their sum - the total area beneath the distribution curve - must be one. P(x) and Q(x) cannot be evaluated explicitly but may be looked up in standard mathematical tables, for instance Table 26.1 in ref.[6]. In the first term of Eq.(48), the Gaussian distribution is the density function of the conditional probability for z(k) when a s_0 was sent. Likewise, the distribution in the second term is conditioned by a s_1 signal being send. Since the Gaussian distributions are non-zero in the whole definition range of $-\infty < x < \infty$, the noise causes a non-zero probability of making a detection error for any comparator threshold setting Z_{tr}. This is illustrated by the hatched areas in Fig.27b representing the probabilities of detecting logical 1 if the logical 0 signal s_0 was sent or detecting a logical 0 if the logical 1 signal s_1 was sent. Together, the two situations set the detection error probability that is expressed and evaluated though,

$$P_{error} = P\{\, z_b(k) > Z_{tr} \mid s_0 \; sent \,\} + P\{\, z_b(k) < Z_{tr} \mid s_1 \; sent \,\}$$

$$= \int_{(Z_{tr}-z_0)/\sigma_z}^{\infty} p_N\left(\frac{z_b - z_0}{\sigma_z}\right) \frac{dz_b}{\sigma_z} \; P(s_0) + \int_{-\infty}^{(Z_{tr}-z_1)/\sigma_z} p_N\left(\frac{z_b - z_1}{\sigma_z}\right) \frac{dz_b}{\sigma_z} \; P(s_1) \qquad (50)$$

$$= Q\left(\frac{Z_{tr} - z_0}{\sigma_z}\right) P(s_0) + Q\left(\frac{Z_{tr} - z_1}{\sigma_z}\right) P(s_1) \;,$$

---

6) Another common way of expressing the probabilities uses the error function erf() and the complementary error function erfc(). The two forms are related by

$$\left.\begin{matrix} Q(x)\,|_{x\ge 0} \\ P(x)\,|_{x\le 0} \end{matrix}\right\} = \frac{1}{2}\, erfc\left(\frac{|x|}{\sqrt{2}}\right), \qquad \left.\begin{matrix} P(x)\,|_{x\ge 0} \\ Q(x)\,|_{x\le 0} \end{matrix}\right\} = \frac{1}{2}\left[\, 1 + erf\left(\frac{|x|}{\sqrt{2}}\right)\right].$$

where still $P(s_0)=1-P(s_1)$. The Q functions hold the normalized integral defined in Fig.28. As the probability of error depends upon the threshold setting Z_{tr}, an optimal choice of this level should be the one that minimizes P_{error}. Using the middle expression from Eq.(50), where each integral is differentiated with respect to one of its bounds, we get

$$\frac{dP_{error}}{dZ_{tr}} = -\frac{1}{\sigma_z^2}\, p_N\left(\frac{Z_{tr}-z_0}{\sigma_z}\right) P(s_0) + \frac{1}{\sigma_z^2}\, p_N\left(\frac{Z_{tr}-z_1}{\sigma_z}\right) P(s_1) = 0 \,. \tag{51}$$

When the threshold lies between z_1 and z_0, as in Fig.27, the slope of p_N in the first term is negative while it is positive in the second term. Under this condition 2nd order derivative of P_{error} is always positive and Eq.(51) corresponds to the desired minimum probability. Using the last part of Eq.(49) the result may be manipulated to yield the best threshold level

$$Z_{tr} = \frac{1}{2}\left[\, z_1 + z_0 \,\right] + \frac{\sigma_z^2}{z_1 - z_0} \log \frac{P(s_1)}{P(s_0)}\Bigg|_{P(s_0)=1-P(s_1)} \,. \tag{52}$$

Comparing error probabilities for different types of modulation on equal terms, it is custom to set the two bits in a digital transmission equally likely, i.e. $P(s_1)=P(s_0)=½$. In that case the last term in Eq.(52) becomes zero and we get the plausible result that the threshold level should be centered between the two possible outcomes z_0 and z_1. Due to the symmetry of the Gaussian distribution the two error contributions from Eq.(50) may be collected to one term, so this important case is summarized by

$$P(s_0) = P(s_1) = \tfrac{1}{2} : \quad Z_{tr} = \tfrac{1}{2}\left[\, z_1 + z_0 \,\right] \quad \Rightarrow$$

$$P_{error} = Q\left(\frac{|\, z_1 - z_0 \,|}{2\sigma_z}\right) = Q\left(\sqrt{\frac{1}{2\eta}\left[E_1 + E_0 - 2\,Re\{\rho_{10}\}\sqrt{E_1 E_0}\,\right]}\,\right) , \tag{53}$$

where the last argument to the Q function is taken from the ratio in Eq.(47). Two points of importance for the use of the result should be stressed. First, there are made no assumptions about the particular waveshapes of the bit signals. Second, signals and noise are amplified equally so the filter gain factor K_0 does no influence the error probability. However, the threshold in a practical realization of an optimal filter may depend on the gain. Eqs.(37) and (41) yield,

$$z_1 = K_0 \int_0^{T_b} s_1(t)\left[\, s_1^*(t) - s_0^*(t) \,\right] dt = K_0\, E_1 - K_0\, \rho_{10} \sqrt{E_1 E_0} \,, \tag{54}$$

$$z_0 = K_0 \int_0^{T_b} s_0(t)\left[\, s_1^*(t) - s_0^*(t) \,\right] dt = K_0\, \rho_{10} \sqrt{E_1 E_0} - K_0\, E_0 \,. \tag{55}$$

This result is again independent of the specific bit signal waveshapes of s_1 and s_0.

Errors in Optimal Detection of Binary Modulations

To compare threshold levels and error probabilities in optimal detection with different type of binary modulations, we need to know the two noiseless detection levels, z_1, z_0, the signal energies E_1, E_0, and their correlation coefficient ρ_{10}. In a sinusoidal signal of constant amplitude A_{c0}, voltage or current, the energy normalized to 1Ω becomes,

$$E_{c0} = \int_0^{T_b} A_{c0}^2 \sin^2 \omega_c t\, dt = \frac{A_{c0}^2}{2} \int_0^{T_b} \left[1 - \cos 2\omega_c t \right] dt = \frac{A_{c0}^2}{2} T_b \left[1 - \frac{\sin(2\omega_c T_b)}{2\omega_c T_b} \right] \tag{56}$$

$$= \frac{A_{c0}^2}{2} T_b \qquad for \begin{cases} \omega_c T_b \to \infty \\ \omega_c T_b = p\frac{\pi}{2}, \quad p = \pm 1, \pm 2, \pm ... \end{cases} \tag{57}$$

The mean power of a sine wave is usually given by $\frac{1}{2}A_{c0}^2$. This power is multiplied by the bit duration T_b to give the corresponding energy in the last expression. As seen, the conventional power formula requires, that the mean value is either calculated over long time or - due to the symmetry of the sinewave - over an integral number of quarter cycles. We have previously made the assumption that the bit period is an integral number of carrier cycles, so the last condition applies. In RF communication systems both conditions are commonly met, so Eq.(57) may be used to calculate signal energies in any case.

If logical 1's and 0's are equally likely, the mean energy per bit E_b in the signal reaching the detector becomes,

$$E_b = \tfrac{1}{2}\left[E_1 + E_0 \right]. \tag{58}$$

The two first cases of modulated waves from Eq.(35) are now characterized,

ASK - OOK modulation

$$E_1 = E_{c0}, \quad E_0 = 0, \quad E_b = \tfrac{1}{2}E_{c0}, \quad \rho_{10} = 0, \quad z_1 = 2K_0 E_b, \quad z_0 = 0, \quad Z_{tr} = K_0 E_b, \tag{59}$$

$$BER = P_{error} = Q\left(\sqrt{\frac{E_{c0}}{2\eta}}\right) = Q\left(\sqrt{\frac{E_b}{\eta}}\right),$$

PSK - PRK modulation

$$E_1 = E_{c0}, \quad E_0 = E_{c0}, \quad E_b = E_{c0}, \quad \rho_{10} = -1, \quad z_1 = 2K_0 E_b, \quad z_0 = -2K_0 E_b, \quad Z_{tr} = 0, \tag{60}$$

$$BER = P_{error} = Q\left(\sqrt{\frac{2E_{c0}}{\eta}}\right) = Q\left(\sqrt{\frac{2E_b}{\eta}}\right),$$

where BER stands for bit error rate. It is a common performance reference in comparisons including modulation methods that may extend to simultaneous transmission of more bits.

Noticing, for instance from Fig.28, that the Q-function decreases with growing argument, it is seen above, that the binary PSK-PKR modulations method has better error performance than the ASK-OOK scheme in both versions of P_{error}. The first one corresponds to equal maximum bit energies E_{c0}, i.e. equal amplitudes A_{c0}. This is relevant if it is the transmitter peak-power that imposes limitations. If they are set by regulations on the average power, where nothing is transmitted half the time in ASK-OOK, the s_1 signal amplitude is raised by a factor of $\sqrt{2}$ compared to the PSK modulated transmission, but even here the latter has an advantage. A final point of practical consequences that further favor PSK in this comparison is its zero threshold level that needs no adjustment, if the input energy level changes. With ASK-OOK adjustments in the K_0 factor through automatic gain control, AGC, must be devised.

With FSK there is one more design parameter to consider compared to the cases above, namely the frequency deviation $\Delta\omega$ that determines the separation between the two signals $s_1(t)$ and $s_0(t)$ and controls the correlation coefficient. It is assumed that the energies E_1 and E_0 are equal and expressible by Eq.(57), which also gives the mean energy $E_b = E_1 = E_0$. The correlation between the two signals becomes, cf. Eqs.(3), (35), and (43),

$$\begin{aligned}\rho_{10} &= \frac{1}{E_b}\int_0^{T_b} s_1(t)\, s_0^*(t)\,dt = \frac{A_{c0}^2}{E_b}\int_0^{T_b} \cos(\omega_c t + \Delta\omega t + \theta_k)\,\cos(\omega_c t - \Delta\omega t + \theta_k)\,dt \\ &= \frac{1}{T_b}\int_0^{T_b} \cos(2\omega_c t + 2\theta_k)\,dt + \frac{1}{T_b}\int_0^{T_b} \cos(2\Delta\omega t)\,dt \\ &= \frac{\sin 2\theta_k - \sin(2\omega_c T_b + 2\theta_k)}{2\omega_c T_b} + \frac{\sin(2\Delta\omega T_b)}{2\Delta\omega T_b}\,.\end{aligned} \tag{61}$$

The first term in the last expression is exactly zero if the bit period is an integral multiple of the carrier period, otherwise it will loose significance with $\omega_c T_b \gg 1$. The same premises gave formerly Eq.(57). The last term may also disappear, but here the important cases are determined by the zeros of the sine function. Introducing the modulation index from Eq.(36), the correlation coefficient is expressed,

$$\rho_{10} = \frac{\sin(2\pi h)}{2\pi h}\,, \quad \rho_{10} = 0 \quad \textit{for} \quad h = \frac{p}{2}, \quad p = 1, 2, 3, \ldots\,. \tag{62}$$

Using Eqs.(53) to (55) the properties of frequency-shift keying modulation from Eq.(35) (c) may be summarized,

FSK modulation

$$E_1 = E_{c0}, \quad E_0 = E_{c0}, \quad E_b = E_{c0}, \quad \rho_{10} = \frac{\sin(2\pi h)}{2\pi h}, \quad z_1 = -z_0 = K_0 E_b (1 - \rho_{10}), \quad Z_{tr} = 0, \tag{63}$$

$$BER = P_{error} = Q\left(\sqrt{\frac{E_b}{\eta}\left[1 - \frac{\sin(2\pi h)}{2\pi h}\right]}\right).$$

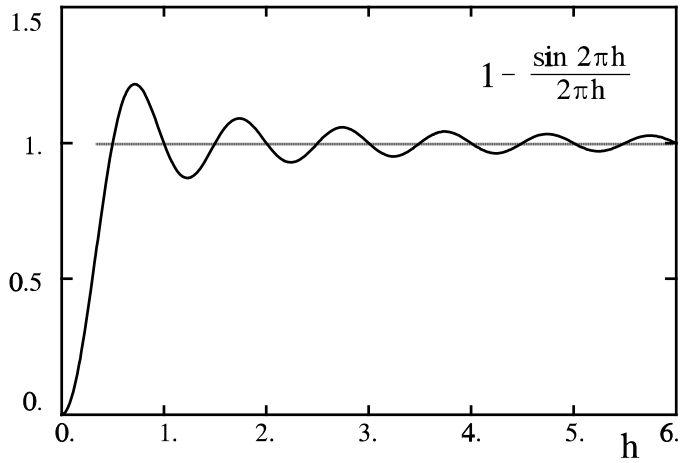


Fig.29 Influence of signal correlation in FSK as function of the modulation index h. Function value 1 corresponds to uncorrelated (orthogonal) signals.

Compared to Eq.(59) it is seen, that orthogonal FSK signals give the same error probability as the ASK-OOK modulation does. The effects of correlated FSK waveforms are demonstrated by Fig.29, which shows the bracket factor that makes the difference to ASK. The probability of error is clearly minimized with the most negative correlation, and this is obtained with non-orthogonal FSK signals having h = 0.715. The bracket factor is here 1.22 corresponding to a correlation of -.22, but this is still far from the value of -1 that characterizes phase-shift keying PSK-PRK. The FSK optimum is often approximated by h=3/4 with a corresponding correlation of -.21. This is an approximation that is rational and could be implemented as a continuous-phase FSK signal with bit boundary offsets as shown in Fig.22(c). However, the considerations in this section concerning optimal detection cannot utilize any phase continuity in the detection. To do this requires memory beyond one bit period, which means a more complex detector. We shall consider an example below when minimum-shift keying, MSK, is discussed. MSK is a binary FSK modulation using h=0.5, where the phase continuity exposes in the demodulation and provides a bit error rate equal to optimal PSK-PRK detection.

Fig.30 shows an overview of the error probabilities of the methods, which were considered in this section, i.e. Eqs.(59),(60), and (63). All the curves are similar but horizontally paralleled to each other as a consequence of the dB scaled arguments. The ratio E_b/η is sometimes denoted the signal-to-noise ratio, SNR, per bit.

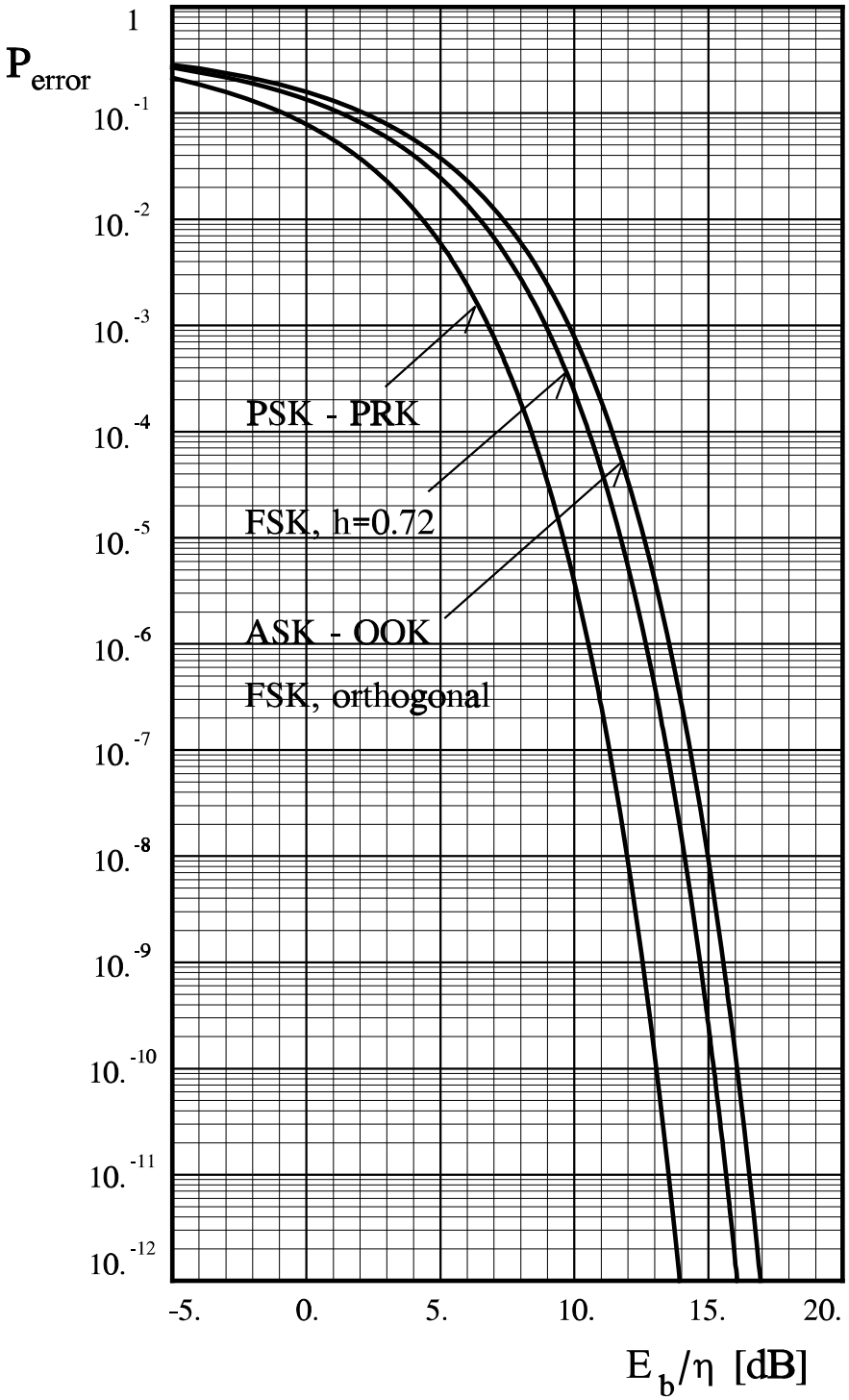


Fig.30 Error probabilities in the basic binary modulations. E_b/η is the signal to noise ratio per bit, with mean bit energy E_b and double-sided noise spectral density η.

Optimal Detectors for Binary Modulations

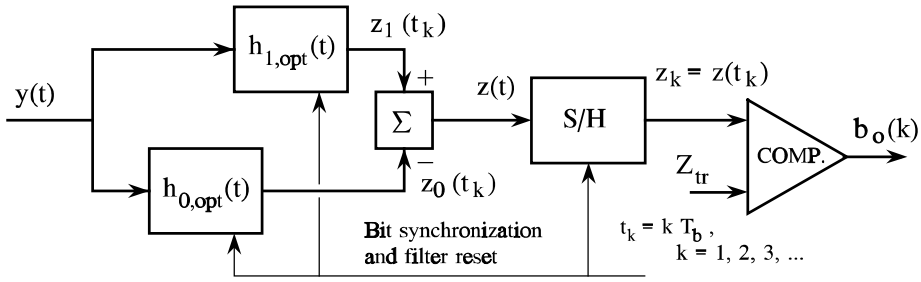


Fig.31 Realization of the optimal detector using two matched filters, one for each binary signal $s_1(1)$ and $s_0(t)$.

Realization of an optimal detector for binary modulation required a filter with impulse response equal to the difference between the two bit-signals s_1, s_0[7]. This function may also be achieved by the scheme in Fig.31, where there is a filter for each bit-signal, i.e.

$$h_{1,opt}(t) = K_0\, s_1(T_b - t)\ , \qquad h_{0,opt}(t) = K_0\, s_0(T_b - t)\ . \tag{64}$$

Taking difference between the outputs from the two filters is equivalent to the basic optimum condition from Eq.(41). The structure is called the matched filter realization of the optimal detector, because the filters are defined by known waveforms the same way the matched filter concept is used elsewhere in signal processing. The outcomes of the two filters at the bit boundaries t_k are expressible through the convolution

$$z_m(t_k) = h_{m,opt}(t) \star s_n(t) = K_0 \int_{(k-1)T_b}^{kT_b} s_n(t)\, s_m(t)\, dt\ , \qquad m,n \in \{1,0\}\ . \tag{65}$$

The detector structure in Fig.32 redo the convolution integral directly. It evaluates like the numerator integral in the correlation coefficient from Eq.(43), so the structure in Fig.32 is called the correlator realization of the optimal detector. The type of filtering applied here, with

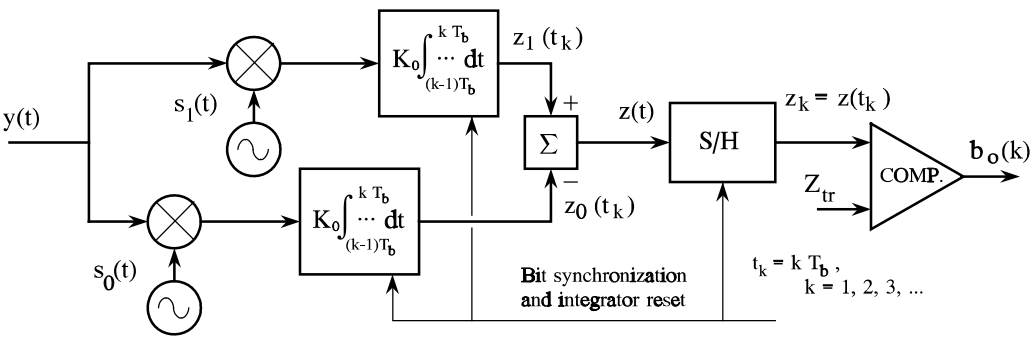


Fig.32 Correlator realization of an optimal detector for binary modulations. Local bit signal sources $s_1(t)$, $s_0(t)$ must be synchronized to their incoming counterparts.

---

7) All signals are real-valued in this section.

integrators that are reset at bit boundaries immediately after their outputs are sensed by the sample and hold circuit, is called "integrate and dump" filtering.

The two realizations above perform identically if the basic optimum criteria in Eq.(64) are met and the necessary synchronization is perfect. In that respect the correlator version clearly emphasizes the requirement of signal waveform coherency in optimal detection as the detector contains synchronized replica of the two binary waveforms. In the matched filter version this property, especially the requirement to synchronization accuracy, is more implicitly contained in requirements to timing precision of the filter resetting and sampling times t_k. The problem may be illustrated by considering the reception of a $s_1(t)$ signal in either ASK or PSK. From Eqs.(35) and (64) we get

$$z(t) = s_1(t) \star h_{1,opt}(t) = K_0 s_1(t) \star s_1(T_b - t) = K_0 \int_{-\infty}^{\infty} s_1(\lambda) s_1(\lambda - t + T_b) d\lambda$$

$$= \begin{cases} 0, & t < 0 \text{ or } t > 2T_b, \\ K_0 A_{c0}^2 \int_0^t \sin\omega_c\lambda \, \sin\left[\omega_c(\lambda - t + T_b)\right] d\lambda, & 0 \le t \le T_b, \\ K_0 A_{c0}^2 \int_{t-T_b}^{T_b} \sin\omega_c\lambda \, \sin\left[\omega_c(\lambda - t + T_b)\right] d\lambda, & T_b \le t \le 2T_b. \end{cases} \tag{66}$$

Fig.33 shows how the integral boundaries are taken from λ-axis ranges, where the functions in the convolution integral differs from zero. The integrals evaluate as, cf.[7] no.2.532,

$$\int \sin\omega_c\lambda \, \sin\left[\omega_c(\lambda - t + T_b)\right] d\lambda = \frac{\lambda}{2} \cos\left[\omega_c(t - T_b)\right] - \frac{\sin\left[\omega_c(2\lambda - t + T_b)\right]}{4\omega_c} \tag{67}$$

In case of $\omega_c T_b \gg 1$ the first term becomes the dominant one. By this and the assumption of T_b being an integral multiple of the carrier period, the non-vanishing part of the filter response to the $s_1(t)$ signal is approximated by,

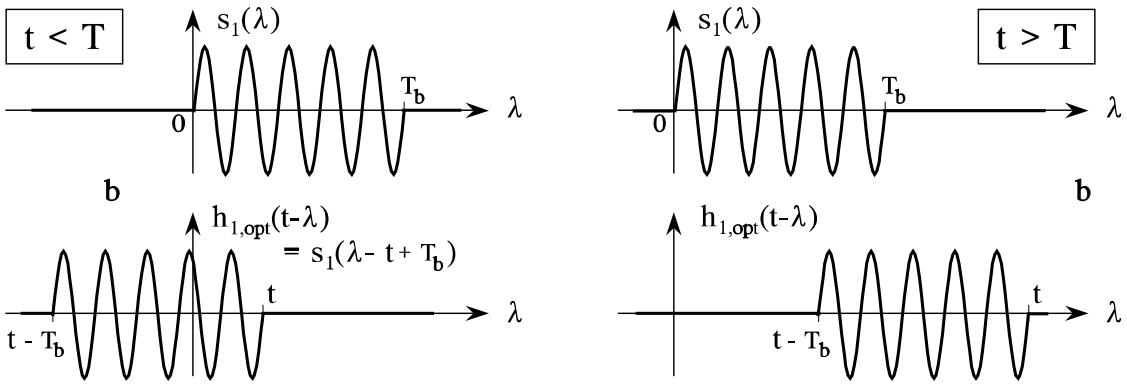


Fig.33 Signal and impulse response for the convolution in Eq.(66). For $t<T_b$ the interval of non-zero product is $0<\lambda<t$, for $t>T_b$ it is $t-T_b<\lambda<T_b$.

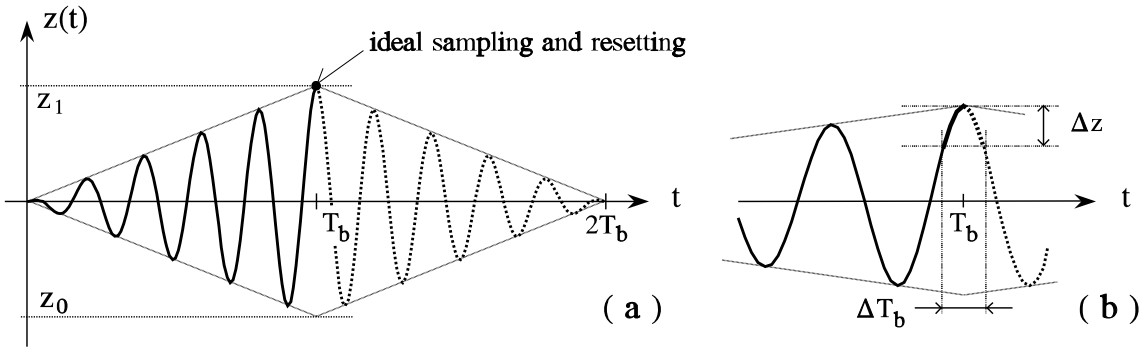


Fig.34 Output from the $h_{1,opt}$ filter in Fig.31. A $s_1(t)$ signal is received. z_1 should be detected, but is diminished Δz by the timing error $\pm\Delta T_b$.

$$z(t) \approx \frac{K_0 A_{c0}^2}{2} T_b \cos\omega_c t \left[1 - \frac{|t - T_b|}{T_b} \right], \qquad 0 \le t \le 2T_b . \tag{68}$$

This result requires that the filter is initially at rest and not reset in the whole interval $0<t<2T_b$, where the convolution differs from zero. As seen in Fig.34, the best sampling instant is the bit boundary T_b, where the response gets the expected maximum of $z_1 = ½K_0A^2_{c0}T_b = 2K_0E_b$, cf.Eqs.(59),(60). After sampling the filter is reset, so the remaining part of the response is not used and therefore shown dotted in the figure. Due to the oscillating nature of the output, the sensitivity of the sampling instant around the boundary follows a $\cos(\theta=\omega_c\Delta T_b)$ relationship. Timing errors do not influence the noise properties in the detector, so the net effect is a reduction of the difference $|z_1-z_0|$ from Eq.(53) by a factor of $\cos\theta$ and, consequently, a greater error probability. Taking PSK-PRK as an example, Eq.(60) should be modified to read,

$$P_{error}\Big|_{PSK-PRK} = Q\left(\cos\theta\sqrt{\frac{2E_b}{\eta}}\right) . \tag{69}$$

The problem of synchronizing and making resets is one of the aspects that favor the correlator method of building an optimal detector in RF circuits. As seen in Fig.35, it is easy to make an integrator reset. Designing a filter with reset might be troublesome. Furthermore,

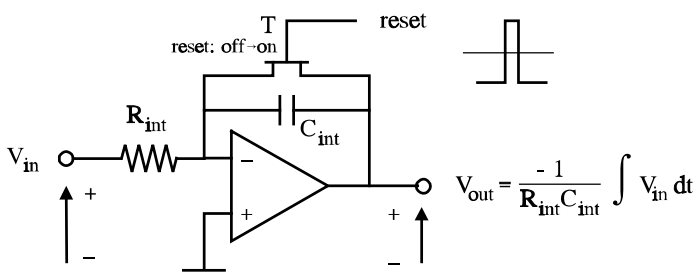


Fig.35 Integrator coupling (inverting) with an operational amplifier. Transistor T is normally open but it short-circuits in the reset period.

a RF filter must be causal to be physically realizable. The filter is defined by the signal waveshape, but most signals - in particular the simpler types we consider - do´nt have this property, so a hardware filter realization cannot be more than an approximation to optimal conditions from the very beginning.

Simplifications in the full optimal detector schemes are possible if the specific bit signal waveforms are constructed from a smaller number of more fundamental waves. In PSK-PRK or ASK-OOK modulation, only one correlator branch is required, because the two bit signals here are weighed versions of one basic signal, the sinusoidal carrier over a bit period. A received $s_1(t)$ signal of amplitude A_{c0} causes the following outputs from the two branches of the detectors in Fig.32,

$$y(t) = s_1(t) = A_{c0}\sin\omega_c t : \quad \begin{cases} z_1(t_k) = K_0 A_{c0}^2 \int_0^{T_b} \sin^2\omega_c t dt = \dfrac{K_0 A_{c0}^2}{2} T_b , \\ z_0(t_k) = K_0 A_{c0}^2 \int_0^{T_b} -\sin^2\omega_c t dt = -\dfrac{K_0 A_{c0}^2}{2} T_b , \end{cases} \tag{70}$$

where it is assumed, that there are an integral number of carrier periods in T_b. The difference between the two outputs becomes positive, twice the magnitude of a single branch. With a $s_0(t)=-s_1(t)$ signal being received in PSK-ASK we get correspondingly,

$$y(t) = s_0(t) = -A_{c0}\sin\omega_c t : \quad z_1(t_k) = -\frac{K_0 A_{c0}^2}{2} T_b , \quad z_0(t_k) = \frac{K_0 A_{c0}^2}{2} T_b, \tag{71}$$

and again the output difference becomes twice the size of a single branch, but now of negative sign. Had we used ASK-PRK the $s_0(t)=0$ provides zero output here. The same results are also achieved using the detector in Fig.36, if we set the amplitude of the single local oscillator $A_{cl} = 2\ A_{c0}$ to make comparisons on equal terms.

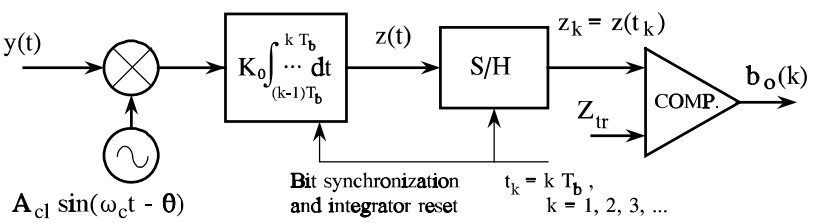


Fig.36 Detector for PSK-PRK or ASK modulation. The local oscillator must be synchronized to the carrier in y(t) with θ=0, cf. Eq.(35).

The local oscillator in Fig.36 must be synchronized to the incoming carrier. To see the effect of a phase synchronization error θ in a correlator detector, consider the integrator output with a s_1 signal being received. It becomes

$$z(t) = 2K_0A_{c0}^2\int_0^t \sin\omega_c t\,\sin(\omega_c t-\theta)\,dt = K_0A_{c0}^2\left[\,t\cos\theta - \frac{\sin(2\omega_c t-\theta)}{4\omega_c}\,\right]. \qquad (72)$$

The integral evaluates like Eq.(67) and again, the first term becomes the dominant one. The detectable maximum value is reduced by the cosθ factor, so the error probability follows Eq(69) in this case too, although the premises are quite different as illuminated by Fig.37. The request for timing accuracy has moved to the phase of the local oscillator by $\theta=\omega_c\Delta t$, whereas bit timing errors now are relatively insignificant.

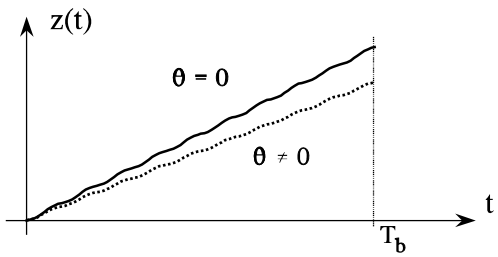


Fig.37 Integrator output from the correlator detector, Eq.(72), using $\omega T_b=10\pi$ like Fig.34. The dotted curve shows the effect of a phase synchronization error θ.

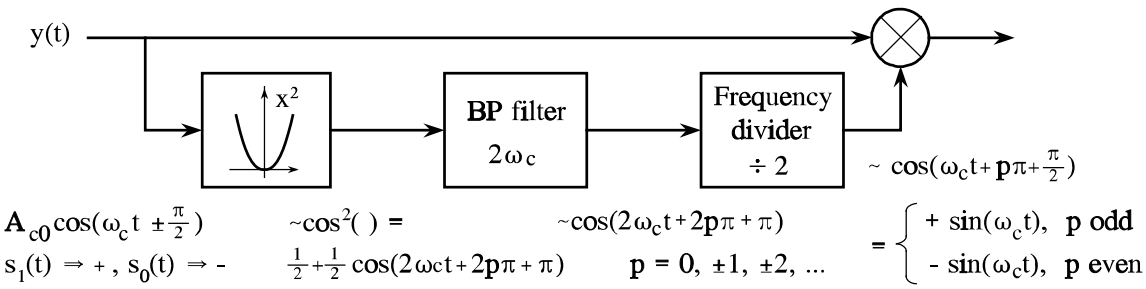


Fig.38 Carrier recovery by squaring in PSK-PRK modulation. The output is synchronized to the incoming carrier with an ambiguity of π.

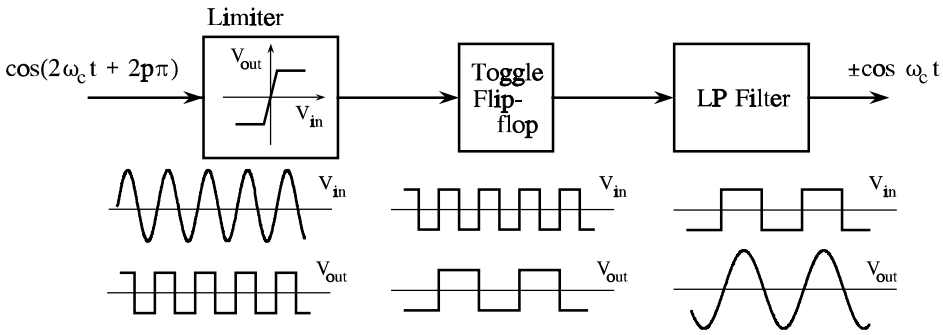
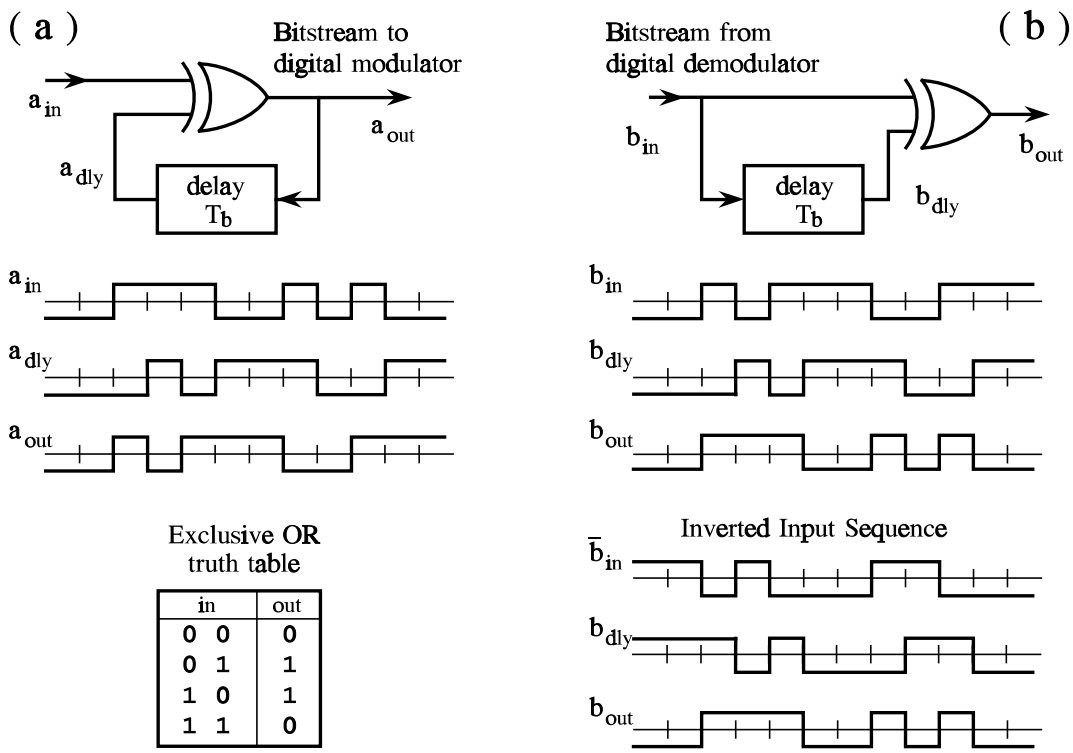


Fig.39 Frequency divider principle. The limiter is a high-gain amplifier with symmetrically restricted output range. The toggle shifts state on positively going pulses.

Synchronization in binary digital modulation causes often a phase ambiguity of π in the locally generated carrier. The problem is illustrated by Fig.38 that shows a simple[8] synchronization method for PSK-PRK. First the incoming signal is squared, which gives a DC component and a double carrier frequency component. The important point of the method is that the sign is independent of the signal being received, i.e. both $s_1(t)$ and the opposite $s_0(t)$ cause the same 2nd harmonic component, which follows the carrier phase. After removal of DC by filtering, the carrier is recovered through a frequency divider. In this operation there are no means to decide whether or not the result will be in phase or 180° out of phase with the original carrier. In plain terms considering the frequency divider method in Fig.39, the toggle that makes the division cannot distinguish between even or odd multiples p of 2π phase offsets from the incoming signal. When p is fixed initially, it should retain its value throughout the rest of the message. However, no synchronization is perfect. Input noise and noise from the circuitry contribute so-called phase or timing jitter and cause non-optimal performance. These effects are disregarded here, but we shall return to that question when more refined methods are discussed.

Although it might be possible to devise schemes that could exactly synchronize to the carrier, the alternative of making the transmission transparent to the 180° phase ambiguity by a method called differential encoding is often preferred. Before a digital message is modulated



in	out
0 0	0
0 1	1
1 0	1
1 1	0

Fig.40 Differential encoding (a) and decoding (b). Output from (a) toggles if a_{in} = 1 and stays constant on -1 (logical 0). The decoder output is independent of b_{in}´s polarity.

8) More elaborate systems use phase-locked loop techniques for this purpose. They are considered later.

on the carrier, it is transformed to another bitstream having the property, that logical ones are transmitted as a shifts in binary states from -1 to 1 or from 1 to -1, while logical zeros keep the binary state constant either 1 or -1. The reverse process of regenerating the original message after the detector will be independent of the polarity of the received bitstream or in turns, the 180° phase ambiguity of the recovered carrier. Fig.40 shows a differential encoder/decoder pair using exclusive or gates. The figure demonstrates, that the output b_{out} is indifferent to the polarity of the detected binary signal, b_{in} or $\overline{b_{in}}$.

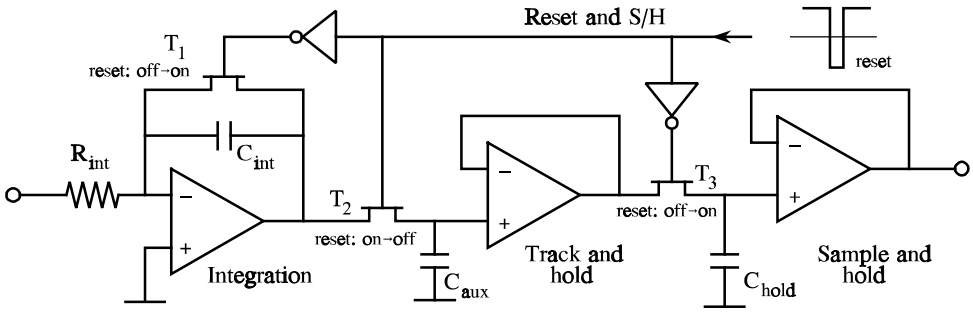


Fig.41 Integrate and dump filter. The tracking stage follows the integrator output until reset. Then T_2 disconnects from the integrator and carries the integrator output voltage to the output S/H stage.

It is important for proper operation of the integrate and dump filter, that the integrator is not reset before its output voltage is carried on to the sample and hold circuit. Fig.41 shows a possible realization where this is secured through an intermediate stage that tracks the integrator in most of the bit-period. Due to the inevitable delays of the inverters in the reset signal path, the tracking stage is uncoupled from the integrator just before the integrator resets. The final integrator output voltage is kept across C_{aux}, while it is sensed by the output sample and hold stage.

I-3 Quadrature Digital Modulations, QPSK and beyond

Among the binary modulations considered so far, FSK has phase excursions from the carrier that needs projections on both the in-phase and the quadrature component of the carrier. To describe PSK-PRK or ASK, only one component was required. This leaves room for extensions with more efficient use of the frequency band around a given carrier.

Quadrature Phase-Shift keying, QPSK

The first step to improve spectral efficiency is to PSK-PRK modulate two bit sequences on the in-phase and quadrature carrier respectively. The two bitstreams may be taken from the same original data sequence as shown by the modulator in Fig.42, where even and odd numbered bits in NRZ form are directed to the in-phase and quadrature multipliers. The modulation is often called quaternary or quadrature phaseshift keying, QPSK, but due to the coincidence between PSK-PRK and AM-DSB modulations with binary data, it may also be considered as a quadrature amplitude modulation.

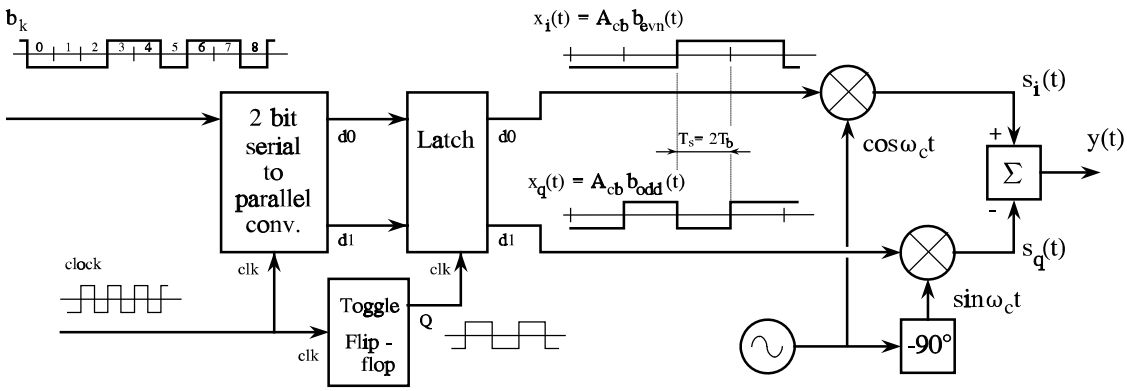


Fig.42 QPSK modulator where even bits from the original message are modulated on the in-phase carrier and odd bits simultaneously on the quadrature carrier.

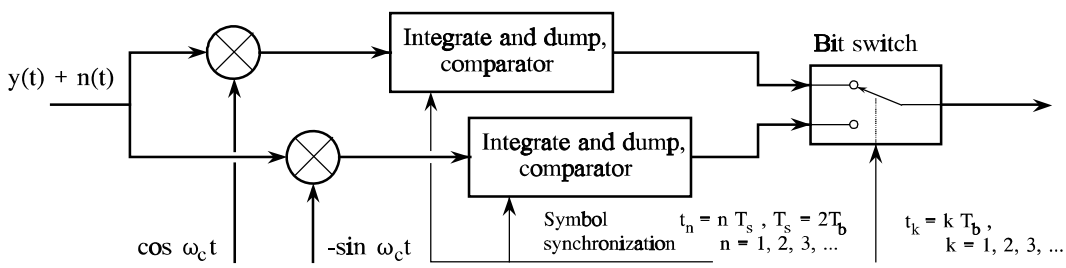


Fig.43 QPSK detection principle. Each branch holds an integration, S/H , and comparator chain of the type presented in Fig.36.

A detector for this type of modulation is shown in Fig.43. Each branch holds an optimal detector for the binary PSK modulated signals like the one in Fig.36. The integrators are reset in intervals of $T_s = 2\,T_b$ corresponding to the duration of bit signals in the baseband

in-phase and quadrature sequences $x_i(t)$, $x_q(t)$. The intervals are twice the bit period T_b of the input sequence a_k, and the final bits switch is operated in intervals of T_b to reconstruct the original sequence. Provided that the synchronization is perfect, this detector is optimal in the same sense as before, i.e. for a given level of white Gaussian noise in addition to the input signal, it has the smallest probability of making a wrong detection. To see this, the deterministic part of the input signal that comes from the modulator may be written,

$$\begin{aligned} y(t) &= x_i(t)\cos\omega_c t - x_q(t)\sin\omega_c t \\ &= \sum_n s_{ib}(t - nT_s) - \sum_n s_{qb}(t - nT_s)\,, \end{aligned} \tag{73}$$

where n sums over all pair of bits - or symbols - in the message. The bit signals for the in-phase and quadrature sequences are considered as projections on two basis signals $\varphi_i(t)$ and $\varphi_q(t)$ along the in-phase and quadrature carrier. They are defined by

$$\varphi_i(t) = \begin{cases} \sqrt{\dfrac{2}{T_s}}\cos\omega_c t, & 0 \le t \le T_s, \\ 0\,, & otherwise, \end{cases} \qquad \varphi_q(t) = \begin{cases} \sqrt{\dfrac{2}{T_s}}\sin\omega_c t, & 0 \le t \le T_s, \\ 0\,, & otherwise. \end{cases} \tag{74}$$

The assumption of the carrier period being and integral multiple of the bit and in turns the symbol period makes the basis signals orthogonal. The scaling is chosen so they become orthonormal too, so they get the properties,

$$\int_{-\infty}^{\infty} \varphi_i(t)\varphi_q(t)\,dt = \int_0^{T_s} \varphi_i(t)\varphi_q(t)\,dt = 2\int_0^{T_b} \varphi_i(t)\varphi_q(t)\,dt = 0\,, \tag{75}$$

$$\int_{-\infty}^{\infty} \varphi_i^2(t)\,dt = \int_0^{T_s} \varphi_i^2(t)\,dt = 1, \qquad \int_{-\infty}^{\infty} \varphi_q^2(t)\,dt = \int_0^{T_s} \varphi_q^2(t)\,dt = 1. \tag{76}$$

Thereby, the bit signals may be written

$$s_{ib}(t) = \pm\sqrt{E_b}\,\varphi_i(t)\,, \qquad s_{qb}(t) = \pm\sqrt{E_b}\,\varphi_q(t)\,, \tag{77}$$

where positive and negative projections represent logical ones and zeros respectively. In units of the basis signals, the projections are the square roots of the bit energy E_b, which we have seen plays an important role for determining error probabilities when input noise is present. Working backwards, the instant amplitude A_{cb} of either bit signal is expressed

$$s_{ib} = \pm A_{cb}\cos\omega_c t \;\Rightarrow\; A_{cb} = \sqrt{\frac{2E_b}{T_s}} \quad or \quad E_b = \frac{A_{cb}^2}{2}\,T_s\,, \tag{78}$$

which is similar to the relationship that was previously used by Eq.(57), taking into account that the energy now corresponds the period $T_s = 2\ T_b$.

It is the orthogonality condition in Eq.(75) that makes the detector in Fig.43 optimal. The input to the two branches are the combined $s_{ib}(t)$ and $s_{qb}(t)$ signals, but with perfect synchronization of the local oscillators the upper and lower branch suppress completely the s_{qb} or s_{ib} respectively. The developments in the forgoing section applied to the lower branch with $s_b = \pm A_{cb}\sin\omega_c t$, but all calculations could equally well have used $s_b = \pm A_{cb}\cos\omega_c t$. Therefore, both branches in Fig.43 are optimal and have equal bit error probabilities or error rates. Since the number of bits received by the complete QPSK demodulator is twice the number of bits in one of the branches, the total bit error rate of the modulator is equal to the rate of a single branch and we have, cf. Eq.(60),

$$\boldsymbol{BER_{QPSK}} = \boldsymbol{BER_{PSK-PRK}} = Q\left\{\sqrt{\frac{2E_b}{\eta}}\right\}. \tag{79}$$

While E_b is the total energy of the binary PSK signal, the total energy of a QPSK signal - called the symbol energy E_s - is twice of that. To acquire the potential double transmission rate of QPSK modulation within the bandwidth of a PSK system, the symbol time in QPSK must be the bit time in PSK. If the two systems should perform equally with respect to bit error rates, the output power P_{sig} of the QPSK transmitter must be twice the power of the PSK-PRK system.

The power, energy and amplitude relationships in a QPSK signal are illuminated by Fig.44, which shows the two common graphical representations of modulated signals. Fig.44b displays the in-phase and quadrature components of the signal. It is a mapping of the complex envelope to the signal that was introduced earlier, cf.p.14. The plot in Fig.44a is called a signal space diagram and shows the possible signal states in basic orthonormal coordinates.

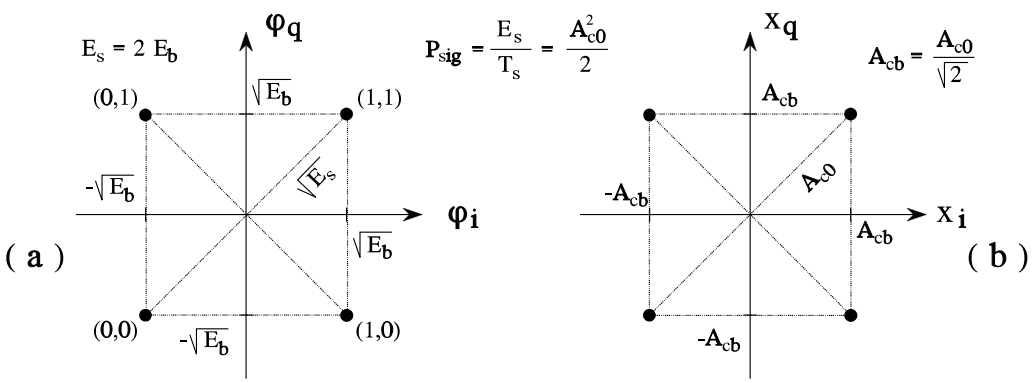


Fig.44 Signal space (a) and in-phase, quadrature component (b) representation of a QPSK modulated signal. The signal state may change instantly along the dashed lines at symbol boundaries.

Offset Quadrature Phase-Shift keying, OQPSK

Constant amplitude signals are preferred in many communication systems. They may ease construction of demodulators and allow employment of nonlinear, high efficiency amplifiers in the system. The QPSK signal has clearly this merit and would retain it through transmission from the modulator to the demodulator, had there been no filtering or other bandwidth limitation in the path between the two. The failure of this illusory assumption has the consequence that abrupt changes in the signal phase cause amplitude variations. When the message state change in a QPSK signal, the phase may instantly exhibit 90° and 180° jumps. The 180° jump is avoided if one of the signal component is offset or staggered by a bit period in a modulation format called offset quaternary phase-shift keying, OQPSK. It is produced by a the modulator in Fig.45.

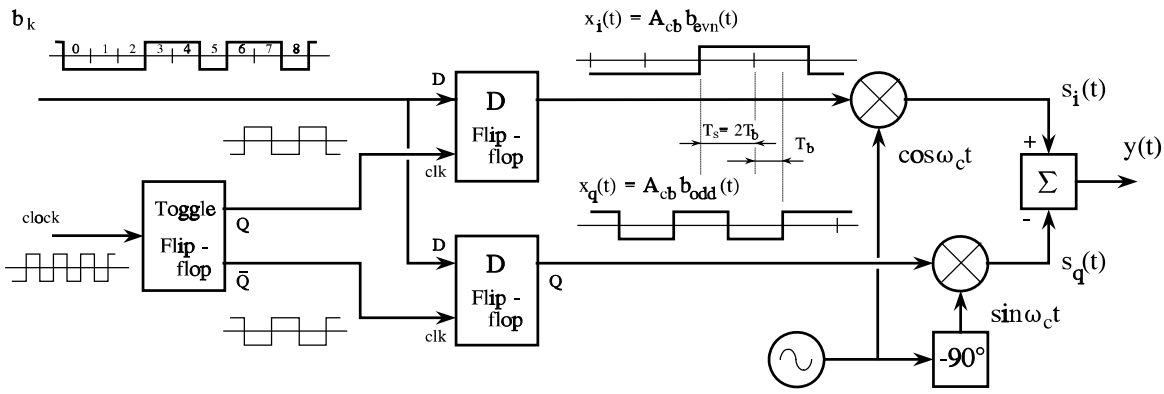


Fig.45 OQPSK modulator. Compared to the QPSK modulator in Fig.42, the $x_q(t)$ sequence is delayed one bit period T_b with respect to the $x_i(t)$ sequence.

If there are no bandwidth restrictions, the properties of a OQPSK system are basically the same as the properties of a QPSK system. Due to the orthogonality that stems from the in-phase and quadrature carriers, the in-phase and quadrature branches could be considered separately before. This is also conceivable with OQPSK, because the orthogonality applies to a bit period so even with offset, the in-phase and quadrature signals do not interfere in demodulation. The only modification to the demodulator in Fig.43 for handling OQPSK signals is, that the two integrator resettings must be offset by a bit period. All other conditions are equal, in particular, the power spectra of the two signals are the same[9]. At a first glance this may seem strange, as the maximum phase jump in the OQPSK signal has halved. However, the effect of the halving is balanced by the fact, that the jumps now occur with double rate because the modulated wave may change state in intervals of T_b instead of $T_s=2T_b$ as seen in the OPQSK example of Fig.46.[10]

---

9) The power spectra are presented on page 47.

10) Fig.46, and later Fig.47, are drawn with a carrier periods equal to one bit interval, i.e. $\omega_c=2\pi/T_b$. This is sufficient to meet all assumptions about orthogonality and it keeps the figures clear. In practical narrowband applications there must be many more carrier periods per bit interval.

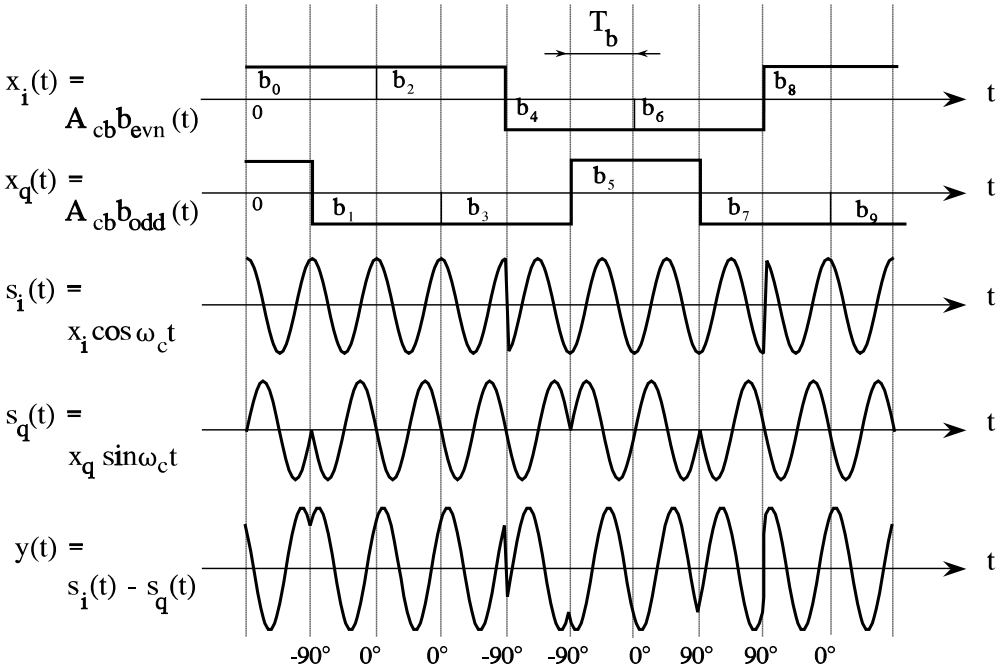


Fig.46 Waveshapes in OQPSK signal generation by the modulator in Fig.45.

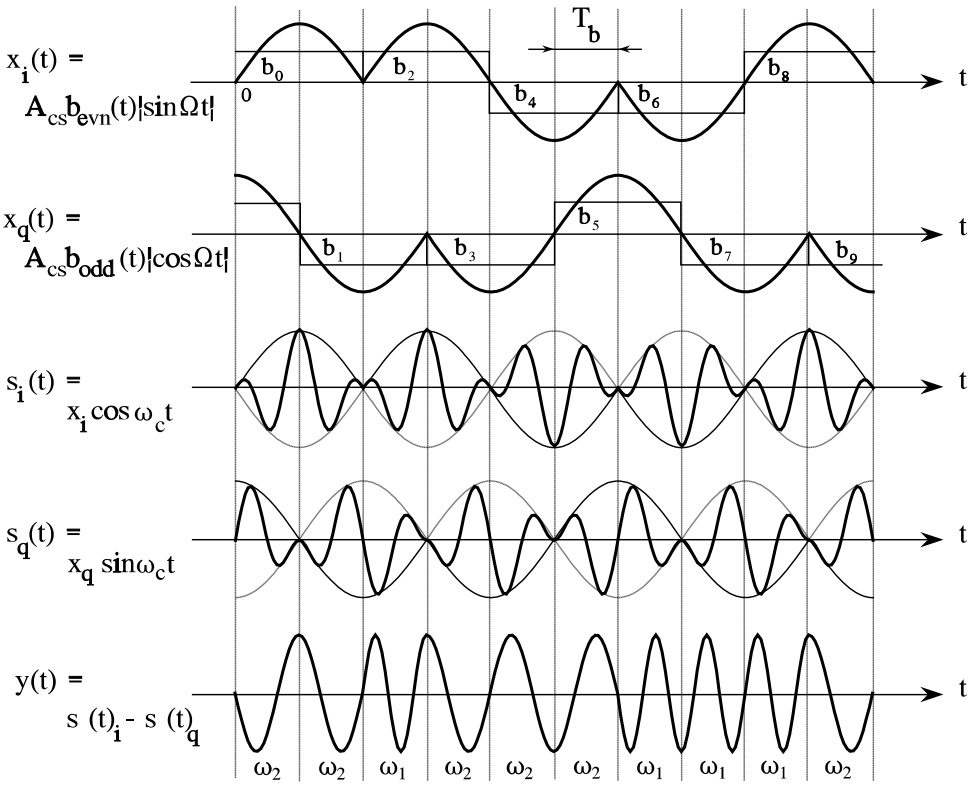


Fig.47 Shaped OQPSK modulation from the modulator in Fig.48. Equivalency to MSK is illuminated by the continuous phase and the two frequencies ω_1 and ω_2 in the output.

The rest of the chapter can be requested at vizh@dtu.dk