Chapter 5

Traditional Analog Modulation Techniques

Mikael Olofsson — 2002–2007

Modulation techniques are mainly used to transmit information in a given frequency band. The reason for that may be that the channel is band-limited, or that we are assigned a certain frequency band and frequencies outside that band is supposed to be used by others. Therefore, we are interested in the spectral properties of various modulation techniques.

The modulation techniques described here have a long history in radio applications. The information to be transmitted is normally an analog so called baseband signal. By that we understand a signal with the main part of its spectrum around zero. Especially, that means that the main part of the spectrum is below some frequency W, called the bandwidth of the signal.

We also consider methods to demodulate the modulated signals, i.e. to regain the original signal from the modulated one. Noise added by the channel will necessarily affect the demodulated signal. We separate the analysis of those demodulation methods into one part where we assume an ideal channel that does not add any noise, and another part where we assume that the channel adds white Gaussian noise.

5.1 Amplitude Modulation

Amplitude modulation, normally abbreviated AM, was the first modulation technique. The first radio broadcasts were done using this technique. The reason for that is that AM signals can be detected very easily. Essentially, all you need is a nonlinearity. Actually, almost any nonlinearity will suffice to detect AM signals. There have even been reports of people hearing some nearby radio station from their stainless steel kitchen sink. And some (including the author) have experienced that with a guitar amplifier. The crystal receiver is a demodulator for AM that can be manufactured at a low cost, which helped making radio broadcasts popular.

In Chapter 3, Theorem 9, we noted that a convolution in the time domain corresonds to a multiplication in the frequency domain. In fact, the opposite is also true.

Theorem 10 (Fourier transform of a multiplication)
Let $a(t)$ and $b(t)$ be signals with Fourier transforms $A(f)$ and $B(f)$. Then we have

$$\mathcal{F}\left\{a(t)b(t)\right\} = (A * B)(f).$$

Proof: The proof is along the same line as the proof of Theorem 8, but starting with the inverse transform of the suggested spectrum. Based on our definitions, we have

$$\mathcal{F}^{-1}\left\{(A * B)(f)\right\} = \int_{-\infty}^{\infty} (A * B)(f) e^{j2\pi ft}\, df = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(\phi)B(f-\phi)\, d\phi\, e^{j2\pi ft}\, df.$$

We can rewrite the expression above as

$$\mathcal{F}^{-1}\left\{(A * B)(f)\right\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(\phi)B(f-\phi)e^{j2\pi ft}\, d\phi\, df.$$

Now, set $\lambda = f - \phi$, and we get

$$\mathcal{F}^{-1}\left\{(A * B)(f)\right\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(\phi)B(\lambda)e^{j2\pi(\lambda+\phi)t}\, d\tau\, d\lambda = \int_{-\infty}^{\infty} A(\phi)e^{j2\pi\phi t}\, d\phi \int_{-\infty}^{\infty} B(\lambda)e^{j2\pi\lambda t}\, d\lambda.$$

Finally, we identify the last two integrals as the inverse Fourier transforms of $A(f)$ and $B(f)$, and we get

$$\mathcal{F}^{-1}\left\{(A * B)(f)\right\} = a(t)b(t).$$

□

So, multiplying in the time domain corresponds to a convolution in the frequency domain.

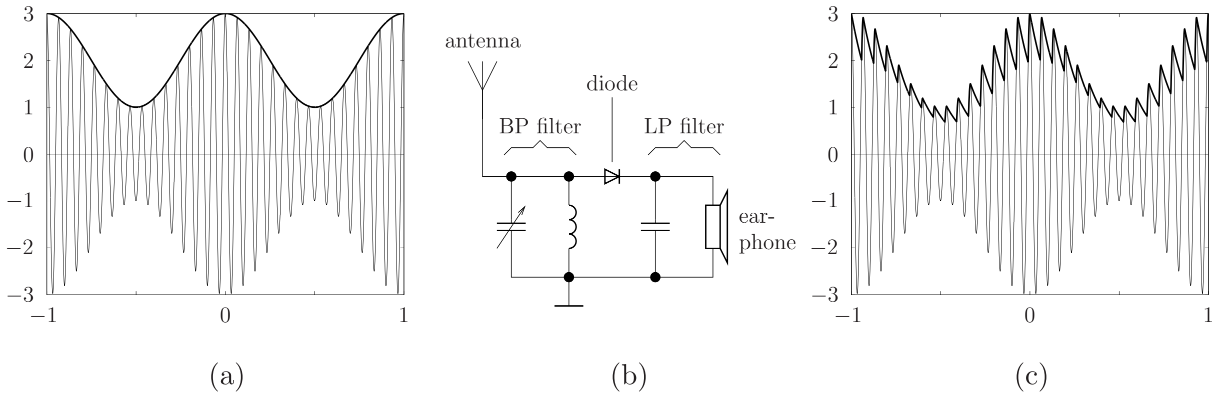

Figure 5.1: (a) A standard AM signal for the message $m(t) = \cos(2\pi t)$ with $C = 2$ and $A = 1$. The dark line is $C + m(t)$. (b) Principle of an envelope detector. (c) The corresponding output from an envelope detector.

Standard AM

An AM signal, $x(t)$, corresponding to the message signal, $m(t)$, is given by the equation

$$x(t) = A(C + m(t)) \cos(2\pi f_c t),$$

where f_c is referred to as the *carrier frequency*, A is some non-zero constant, and where the constant C is chosen such that $|m(t)| < C$ holds for all t. In Figure 5.1a a standard AM signal is presented together with the message, which in this particular example is a cosine signal.

We mentioned that AM signals can be detected using a nonlinearity. The first AM receiver was the so called *crystal receiver*. It consists of an antenna, a resonance circuit (bandpass filter), a diode and a simple low-pass filter. It extracts the envelope $C + m(t)$ from $x(t)$, and is therefore often called an *envelope detector*. The diode in Figure 5.1b is the non-linearity that makes the detection possible. The few simple components makes it possible to manufacture the receiver at a low cost. In addition to that, it doesn't even need a power source of its own. The power is taken directly from the antenna. The output power is of course very small, and only one listener could use the small earphone that was used. In Figure 5.1c, the AM signal is presented together with the output of an envelope detector. Note that the output is very similar to the original message. The mechanical parts in the earphone, and the ear will further low-pass filter the output, so the listener will hear almost the same signal as the one transmitted. Modern envelope detectors have amplifiers in various places and may be implemented digitally, but the basic construction is still a bandpass filter, a diode (or some other rectifier) followed by a lowpass filter. An advantage of envelope detectors is that the BP filter that filters out everything except the intended

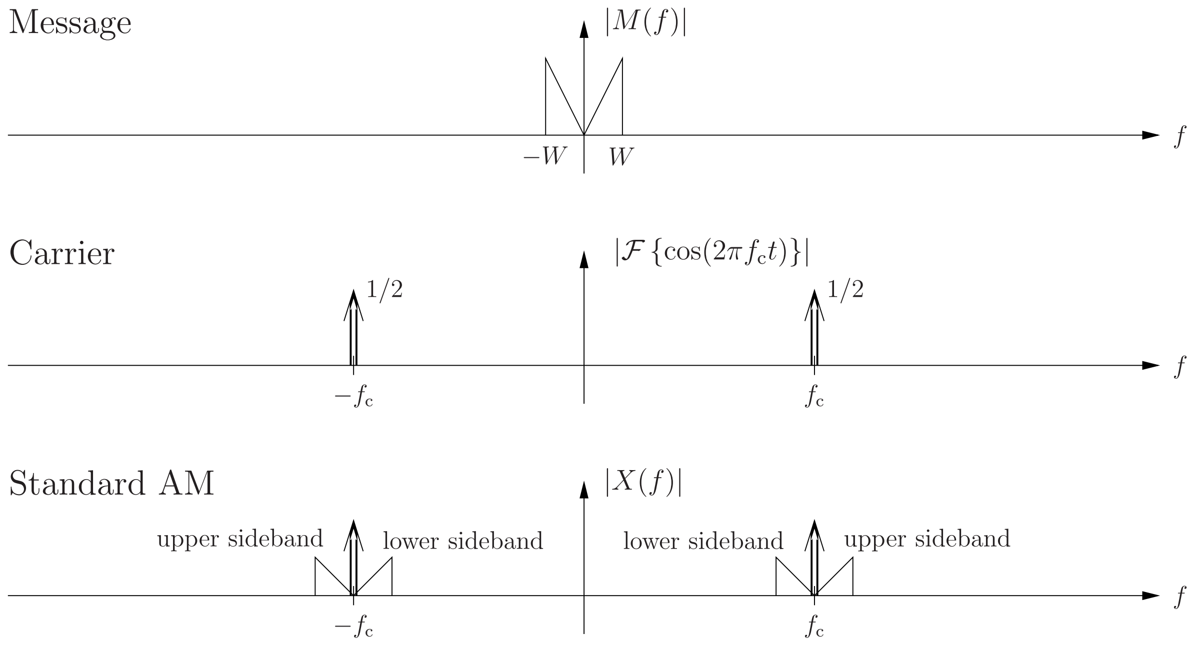

Figure 5.2: Spectrum for standard AM signals.

frequency band is not critical. It is enough if its center frequency is approximately correct. In other words, it does not need to know the carrier frequency exactly, or the carrier phase for that matter.

We wish to study the spectrum of AM signals. Since an AM signal is the product of a message and a carrier, that is easiest done based on Theorem 10. Thus, we need to find the Fourier transform of $\cos(2\pi f_\mathrm{c} t)$. First consider

$$\mathcal{F}^{-1}\left\{\delta(f - f_\mathrm{c})\right\} = \int_{\infty}^{\infty} \delta(f - f_\mathrm{c}) e^{j2\pi f t}\, df = e^{j2\pi f_\mathrm{c} t},$$

where the first equality is given by the definition of the inverse Fourier transform, and where the last equality is given by the definition of the unit impulse. So, we have

$$\mathcal{F}\left\{\cos\left(2\pi f_\mathrm{c} t\right)\right\} = \mathcal{F}\left\{\frac{e^{j2\pi f_\mathrm{c} t} + e^{-j2\pi f_\mathrm{c} t}}{2}\right\} = \frac{1}{2}\left(\delta(f - f_\mathrm{c}) + \delta(f + f_\mathrm{c})\right).$$

Now we are ready to apply Theorem 10 on $x(t)$. Let $M(f)$ be the spectrum of $m(t)$ and let $X(f)$ be the spectrum of $x(t)$. Then we get

$$\begin{aligned} X(f) &= \mathcal{F}\left\{AC\cos\left(2\pi f_\mathrm{c} t\right)\right\} + \mathcal{F}\left\{Am(t)\cos\left(2\pi f_\mathrm{c} t\right)\right\} \\ &= \frac{AC}{2}\left[\delta(f - f_\mathrm{c}) + \delta(f + f_\mathrm{c})\right] + \frac{A}{2}\left[M(f - f_\mathrm{c}) + M(f + f_\mathrm{c})\right]. \end{aligned}$$

It is left as an exercise to verify that the last equality holds. The involved spectra are displayed in Figure 5.2. Here $\frac{AC}{2}\left(\delta(f-f_c)+\delta(f+f_c)\right)$ is referred to as the *carrier*, since that term corresponds to $AC\cos\left(2\pi f_c t\right)$. The other part of the spectrum, $\frac{A}{2}\left(M(f-f_c)+M(f+f_c)\right)$, is referred to as the sidebands. Those sidebands are the only parts of the spectrum that depend on the message $m(t)$. The sidebands are called upper and lower sidebands based on where they are compared to the carrier frequency, according to the following.

- *Upper sideband*: $\frac{A}{2}\left(M(f-f_c)+M(f+f_c)\right)$ for $|f|>f_c$.
- *Lower sideband*: $\frac{A}{2}\left(M(f-f_c)+M(f+f_c)\right)$ for $|f|<f_c$.

Because of those two sidebands, this type of AM modulation is often called *double sideband* AM, abbreviated AM-DSB. There are also other versions of AM, but they cannot be detected using an envelope detector.

Suppressed Carrier Modulation

All the information about the message $m(t)$ in standard AM is in the sidebands. The carrier itself does not carry any information, and in that respect the carrier corresponds to unnecessary power dissipation. One version of AM that cannot be detected using an envelope detector is called AM-SC or AM-DSB-SC, where SC should be interpreted as *Suppressed Carrier*. For this type of modulation, the constant C is simply set to zero, i.e. we have

$$x(t)=Am(t)\cos\left(2\pi f_c t\right),$$

and the corresponding spectrum is

$$X(f)=\frac{A}{2}\left(M(f-f_c)+M(f+f_c)\right).$$

So the carrier is removed from the spectrum, as the name suggests. In Figure 5.3, an AM-SC signal is presented together with the message and the corresponding envelope detector output, as well as the absolute value of the message. Note that the output from an envelope detector in this case is close to the absolute value of the message. Thus, an envelope detector cannot be used to receive AM-SC.

Demodulation of AM-SC can instead be done by modulating once more. Let $y(t)$ with spectrum $Y(f)$ be the output of that modulation. Then we have, similarily as above,

$$y(t)=x(t)\cos\left(2\pi f_c t\right)=Am(t)\cos^2\left(2\pi f_c t\right)=\frac{A}{2}m(t)\left(1+\cos(4\pi f_c t)\right),$$

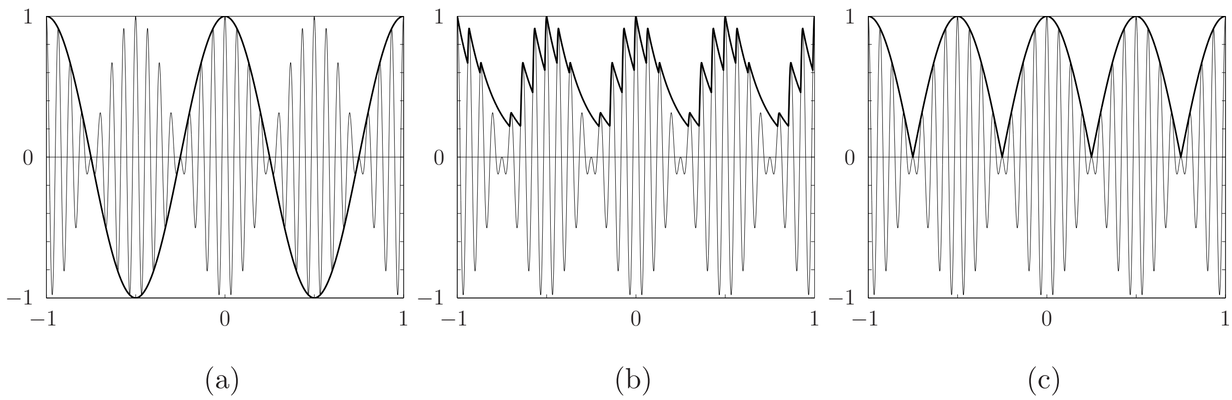

Figure 5.3: (a) An AM-SC signal for the message $m(t) = \cos(2\pi t)$ with $A = 1$. The thick line is $m(t)$. (b) The corresponding output from an envelope detector. (c) The AM-SC signal together with $|m(t)|$ for comparison.

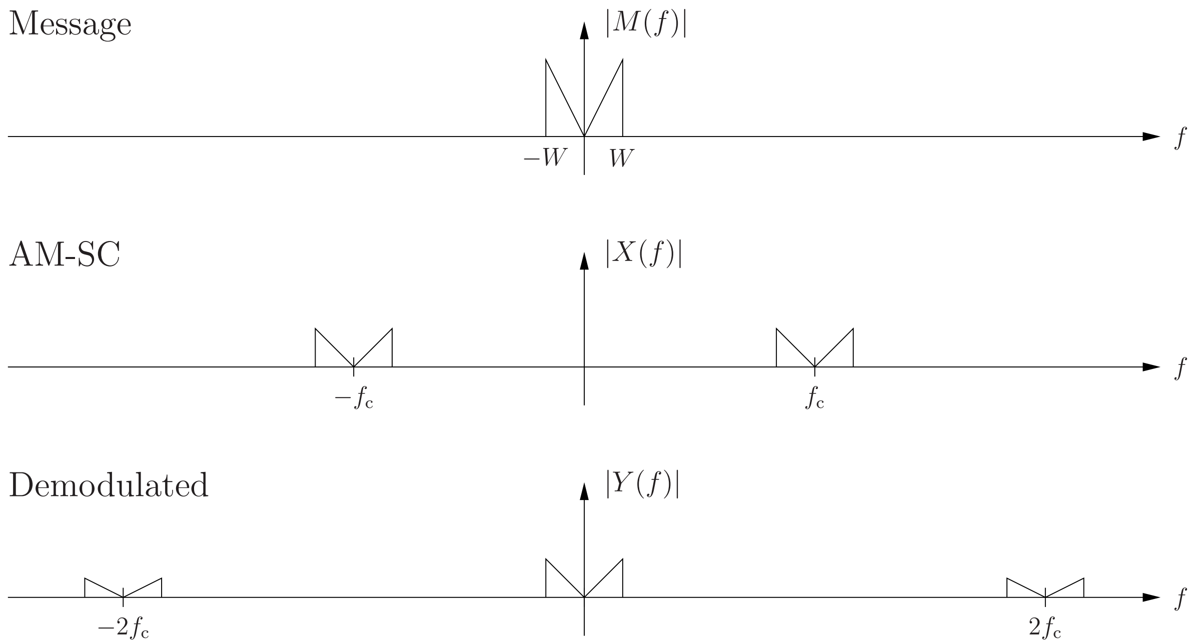

Figure 5.4: Modulation of AM-SC and demodulation by modulating again.

and the corresponding spectrum is

$$Y(f) = \frac{A}{2}\left(X(f-f_\mathrm{c}) + X(f+f_\mathrm{c})\right) = \frac{A}{2}M(f) + \frac{A}{4}\left(M(f-2f_\mathrm{c}) + M(f+2f_\mathrm{c})\right).$$

So, we have regained $M(f)$, but we also have copies of $M(f)$ centered around $\pm 2f_\mathrm{c}$. The involved spectra are displayed in Figure 5.4. If W, the bandwidth of the message $m(t)$, is smaller than f_c, which normally is the case, then those copies do not overlap with the original spectrum. Thus, we can use a suitable low-pass filter to remove the unwanted copies. The further away the unwanted copies are in the frequency domain, the simpler that filter can be. It should be noted that standard AM can also be demodulated using this method.

Single Sideband Modulation

Since there is a one-to-one relation between $M(f)$ and $M(-f)$, there is also a one-to-one relation between the two sidebands, at least if W is smaller than f_c. So, in both standard AM and AM-SC, we actually transmit our data twice in the frequency domain. No information is lost if we only transmit one of the sidebands. This type of AM is referred to as SSB, which should be interpreted as *Single SideBand*. There are SSB versions of both standard AM and AM-SC, and they can be obtained by first generating standard AM or AM-SC, and then using a suitable band-pass filter to remove the unwanted sideband. Spectra of AM-SSB and AM-SSB-SC are displayed in Figure 5.5. Obviously, SSB modulation only needs half the bandwidth compared to original AM or AM-SC. SSB-modulated signals can also be demodulated by modulating again using AM-SC, and we still get copies near $\pm 2f_\mathrm{c}$, that has to be removed by a low-pass filter. However these copies now contain only one sideband.

Synchronization for AM Demodulation

Demodulation by remodulation as described above is a method that can be used for detection of all variants of AM. However, that demands that we have a correct carrier available in the demodulator, with both correct frequency and at least approximately correct phase. For standard AM and for AM-SSB, where the carrier is available in the signal, it can easily be extracted from the received signal using a narrow BP-filter with center frequency f_c. For AM-SC, and for AM-SSB-SC, the absence of a carrier makes it impossible to extract the carrier in that way. One way for the receiver to extract a carrier signal from an AM-SC signal

$$x(t) = Am(t)\cos\left(2\pi f_\mathrm{c} t\right),$$

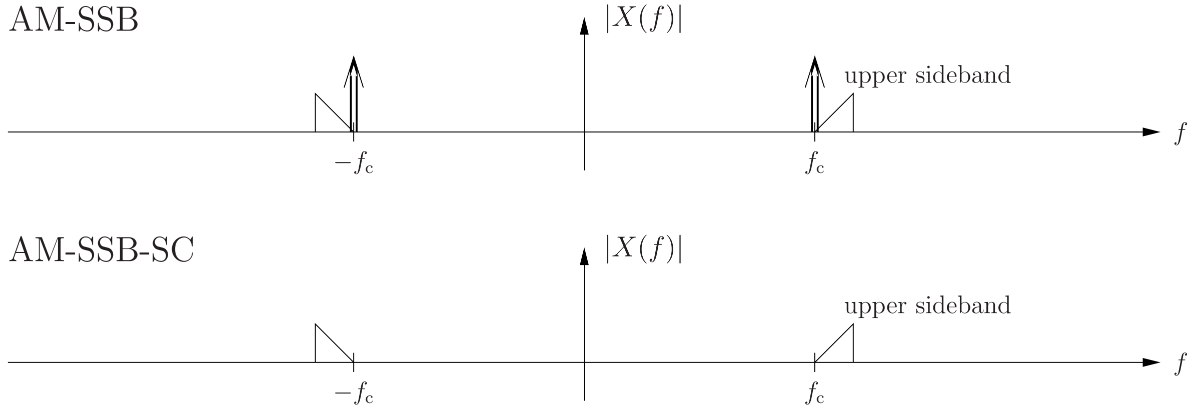

Figure 5.5: Spectra for AM-SSB and AM-SSB-SC.

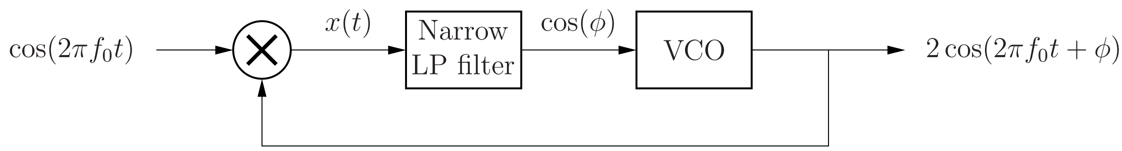

Figure 5.6: A phase-locked loop for generation of a well-defined carrier signal. The signal $x(t)$ is given by $x(t) = \cos(2\pi f_0 t)\cos(2\pi f_0 t + \phi) = \frac{1}{2}\left(\cos(4\pi f_0 t + \phi) + \cos(\phi)\right)$. The device labelled VCO is a voltage controlled oscillator.

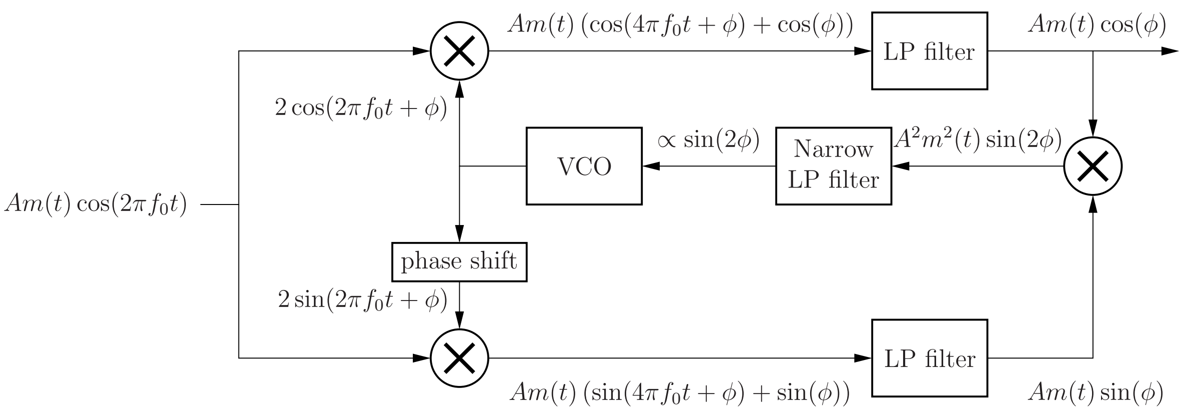

Figure 5.7: A Costas loop for detection of AM-SC.

is to produce the square

$$x^2(t) = A^2 m^2(t) \cos^2(2\pi f_c t) = \frac{A^2 m^2(t)}{2} (1 + \cos(4\pi f_c t))$$

When we send information, the average of $m^2(t)$ is non-zero, which means that a scaled version of $\cos(4\pi f_c t)$ can be extracted from $x^2(t)$, again using a narrow BP-filter, but with center frequency $2f_c$. Extracting carriers in those ways will produce signals with a frequency that is the correct carrier frequency in the AM or AM-SSB case and twice the carrier frequency in the AM-SC case, but the amplitude can vary from time to time depending on the actual behaviour of the channel or the statistics of the information. Also, the extracted signal may include noise and parts of the sideband(s).

A clean carrier with both the correct frequency and a well defined amplitude, without any noise or residues from the sidebands, can be obtained from the extracted signal using a *phase-locked loop* (PLL). There are several variants of phase-locked loops in use, and a simple variant is displayed in Figure 5.6. The signals given in Figure 5.6 assume that the input is already a clean sinusoid. In practice, the input is an extracted approximate carrier which, as noted above, is polluted with noise and residues from the sidebands. A phase-locked loop is a contol loop that produces a sinusoid with constant amplitude, the correct frequency f_0 and an approximate phase ϕ. The *voltage-controlled oscillator* (VCO) is chosen such that the wanted carrier frequency corresponds to zero input. The frequency $f_{\text{VCO}}(V_{\text{in}})$ of the output of the VCO is a function of the input voltage V_{in}, such that the derivative of that function is positive, i.e. we have $\frac{d}{dV_{\text{in}}} f_{\text{VCO}}(V_{\text{in}}) > 0$.

The carrier frequency in use may differ slightly from the wanted carrier frequency, and the phase-locked loop follows the carrier frequency, which means that the input to the VCO may differ slightly from zero. In Figure 5.6 that means that the signal produced by the PLL has phase ϕ for which $\cos(\phi) \approx 0$ holds, in a point where $\cos(\phi)$ has positive derivative. In other words, we have $\phi \approx -\frac{\pi}{2} + k \cdot 2\pi$ for some integer k. We may assume any integer value of k, since the produced signal is the same for different values of k. So, we can for instance say that the signal has the phase $\phi \approx -\frac{\pi}{2}$.

For AM-SC, where we have squared the signal, and where the frequency of the extracted signal is twice the carrier frequency, the feedback is equipped with a frequency doubler which can for instance be a squarer. The resulting output is then a sinusoid with the correct carrier frequency and phase approximately $-\frac{\pi}{4}$.

A special type of phase-locked loop that is especially well suited for detection of AM-SC signals is the *Costas loop*, given in Figure 5.7. It extracts the carrier directly from the signal. Again, the VCO is chosen such that the wanted carrier frequency corresponds to zero input. Thus, the loop produces a sinusoid whose frequency is the carrier frequency with phase ϕ for which $\sin(2\phi) \approx 0$ holds, in a point where $\sin(\phi)$ has positive derivative. The resulting phase is therefore $\phi \approx 0$. The output of the Costas loop is the message $m(t)$ scaled by $\cos(\phi)$, but since we have $\phi \approx 0$, we also have $\cos(\phi) \approx 1$.

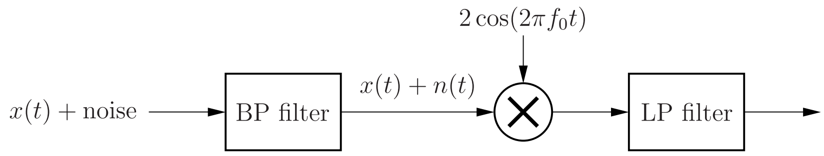

Figure 5.8: Demodulation of AM signals in the presence of noise.

The reason that the Costas loop works is the presence of both sidebands, and the one-to-one mapping between the two sidebands. The two sidebands point at the carrier frequency, and the phase information in the sidebands gives us the carrier phase. The Costas loop can therefore not be used for AM-SSB-SC. There is simply no way to extract the carrier from an AM-SSB-SC signal, due to the fact that the carrier is not available in the signal, and nothing in the spectrum gives any hint about the carrier frequency. That is the price we have to pay for suppressing both the carrier and one of the sidebands. Therefore, there are a number of modifications of AM-SSB-SC, that makes it possible to extract a carrier anyway. The most simple method is not to suppress the carrier completely. Then the – however weak – carrier can be extracted from the signal. Another method is to send a short carrier burst, and let a PLL lock on to that burst. After the burst, the oscillator continues producing an internal carrier based on that burst. Of course, the oscillator will most probably diverge from the used carrier eventually. Therefore the carrier burst is repeated regularly. A third possibility is to keep a small part of the removed sideband, and in the receiver filter the other sideband similarily, and then extract a carrier using one of the methods above.

Impact of Noise in AM Demodulation

We would like to analyze the impact of noise on demodulation of AM signals. For this analysis we need to make some assumptions about the noise and about the demodulation. The first assumption is that the noise is dominated by thermal noise, and that it is independent of the message. As mentioned in Section 4.1, such noise can be modeled as white Gaussian noise. We assume that the received signal is filtered by an ideal BP filter that exactly matches the bandwidth of the AM signal before demodulation. We assume that the demodulation is done by remodulation by $2 \cdot \cos(2\pi f_0 t)$ as in the Costas loop. We also assume that the demodulated signal is filtered by an ideal LP filter that exactly matches the bandwidth of the message. See Figure 5.8.

We need to introduce some notation. Let W denote the bandwidth of the message. Let N_0 denote the one-sided power spectral density of the assumed white Gaussian noise, and let $n(t)$ denote the noise *after* the BP filter. Also, introduce the following notation for the involved powers.

- P: The (expected) power of the message $m(t)$.
- $P_{\mathrm{m-mod}}$: The (expected) power of the *received* modulated signal $x(t)$. Note that this means that A includes impacts of the channel.
- P_{m}: The (expected) power of the message after demodulation and LP filter.
- $P_{\mathrm{n-mod}}$: The expected power of the ideally BP-filtered noise $n(t)$ before demodulation.
- P_{n}: The expected power of the demodulated and LP-filtered noise.

We define the signal-to-noise ratio $P_{\mathrm{m}}/P_{\mathrm{n}}$ after demodulation. We will compare this signal-to-noise ratio for DSB and SSB modulation using the same sent power $P_{\mathrm{m-mod}}$ transmitted over a channel with the same N_0.

First we consider AM-SC. Then we have the signal

$$x(t) = Am(t)\cos(2\pi f_{\mathrm{c}} t)$$

with bandwidth $2W$ and expected power $P_{\mathrm{m-mod}} = A^2P/2$ since the carrier $\cos(2\pi f_{\mathrm{c}} t)$ has average power $1/2$. After demodulation, we regain $Am(t)$, which means that we have $P_{\mathrm{m}} = A^2P$. For the noise, we have $P_{\mathrm{n-mod}} = 2WN_0$. The demodulated noise

$$n(t) \cdot 2\cos(2\pi f_{\mathrm{c}} t)$$

has expected power $2P_{\mathrm{n-mod}}$, since the carrier $2\cos(2\pi f_{\mathrm{c}} t)$ has average power 2. Half of that expected power is in the frequency interval $|f| < W$, while the other half is in the frequency interval $2f_0 - W < |f| < 2f_0 + W$. The latter part is removed by the LP filter, leaving us with $P_{\mathrm{n}} = P_{\mathrm{n-mod}}$. Finally, that gives us the signal-to-noise ratio

$$\frac{P_{\mathrm{m}}}{P_{\mathrm{n}}} = \frac{A^2P}{2WN_0}$$

For AM-SSB-SC, one of the sidebands from AM-SC is removed, which means that the power $P_{\mathrm{m-mod}}$ is reduced to half that of AM-SC. To produce an SSB signal with the same power as in the DSB case, we therefore need to amplify the signal by $\sqrt{2}$. So, we start with

$$x(t) = \sqrt{2}Am(t)\cos(2\pi f_{\mathrm{c}} t),$$

and filter out one of the sidebands. Then we have the same sent power $P_{\mathrm{m-mod}} = A^2P/2$. After demodulation, we get a scaled version of the message. More precisely, the output is $\frac{A}{\sqrt{2}}m(t)$, which has power $P_{\mathrm{m}} = A^2P/2$. For the noise, we have $P_{\mathrm{n-mod}} = WN_0$, since the bandwidth is W. The demodulated noise

$$n(t) \cdot 2\cos(2\pi f_{\mathrm{c}} t)$$

still has expected power $2P_{\mathrm{n-mod}}$, since the carrier $2\cos(2\pi f_c t)$ has average power 2. Half of that expected power is in the frequency interval $|f| < W$, while the other half is in the frequency interval $2f_0 - W < |f| < 2f_0 + W$. Actually, the other half of the power is in the interval $2f_0 - W < |f| < 2f_0$ if the lower sideband is used, or in the interval $2f_0 < |f| < 2f_0 + W$ if the upper sideband is used. In any case, the part of the spectrum that is near $2f_0$ is removed by the LP filter, leaving us with $P_{\mathrm{n}} = P_{\mathrm{n-mod}}$. Finally, that gives us the signal-to-noise ratio

$$\frac{P_{\mathrm{m}}}{P_{\mathrm{n}}} = \frac{A^2 P}{2WN_0},$$

i.e. the same signal-to-noise ratio as for DSB.

5.2 Angle Modulation

Angle modulation is the common name for a class of modulation techniques, with that in common that the bandwidth of the modulated signal is not given only by the bandwidth of the message, but also by a parameter called the modulation index. By setting this modulation index to a suitable number, we can decide what bandwidth to use, and the larger that index is, the better is the obtained quality. These methods, and combinations of them are used in radio broadcasts in the FM-band (88–108 MHz).

The idea that angle modulation is based on is to let a function $\phi(m(t))$ of the message $m(t)$ be the phase of a carrier, i.e. the sent signal is

$$x(t) = A \cdot \cos\left(2\pi f_c t + \phi(m(t))\right),$$

where A is some non-zero constant, and where f_c again is referred to as the carrier frequency. We say that $\phi(m(t))$ is the momentary phase of $x(t)$. Then the phase deviation $\phi_{\mathrm{d}}(t)$ is the difference between the momentary phase and the average of the momentary phase. Typically, $m(t)$ has average zero, and the function f is chosen such that $\phi(m(t))$ also has average 0. Then we have

$$\phi_{\mathrm{d}}(t) = \phi(m(t)),$$

and the peak phase deviation is defined as

$$\phi_{\mathrm{d,max}} = \max|\phi_{\mathrm{d}}(t)|.$$

The peak phase deviation is also called the phase modulation index, and is denoted μ_{p}.

An alternative interpretation of the varying phase of the signal $x(t)$, is to say that $x(t)$ has varying frequency. We define the momentary frequency as

$$f_{\mathrm{mom}}(t) = \frac{1}{2\pi} \cdot \frac{d}{dt}\left(2\pi f_c t + \phi(m(t))\right) = f_c + \frac{1}{2\pi} \cdot \frac{d}{dt}\phi(m(t)).$$

Note, that this frequency is a function of time, just as the momentary phase also is a function of time. The frequency deviation $f_{\mathrm{d}}(t)$ is the difference between the momentary frequency and the carrier frequency, i.e.

$$f_{\mathrm{d}}(t) = f_{\mathrm{mom}}(t) - f_{\mathrm{c}},$$

and the peak frequency deviation is defined as

$$f_{\mathrm{d,max}} = \max |f_{\mathrm{d}}(t)|.$$

The frequency modulation index μ_{f} is defined as

$$\mu_{\mathrm{f}} = \frac{f_{\mathrm{d,max}}}{W},$$

where W is the bandwidth of the message $m(t)$, or rather the highest frequency component in $m(t)$. If $m(t)$ is a stationary sine, then the two modulation indices are equal.

Spectrum of Angle Modulation

The spectra of angle modulated signals are generally hard to determine, due to the fact that angle modulation is non-linear. For the simple example

$$x(t) = A \cdot \cos\left(2\pi f_{\mathrm{c}} t + \mu \sin(2\pi f_{\mathrm{m}} t)\right),$$

for $f_{\mathrm{m}} \ll f_{\mathrm{c}}$, it is easily shown that μ is both the phase modulation index and the frequency modulation index of that signal. It can also be shown that we have

$$x(t) = \sum_{n=-\infty}^{\infty} A \cdot J_n(\mu) \cos\left(2\pi(f_{\mathrm{c}} + n f_{\mathrm{m}})t\right),$$

where $J_n(\mu)$ is the Bessel function of order n. We will not at all try to perform that proof. The spectrum of that signal is

$$X(f) = \sum_{n=-\infty}^{\infty} \frac{A \cdot J_n(\mu)}{2} \left[\delta(f + f_{\mathrm{c}} + n f_{\mathrm{m}}) + \delta(f - f_{\mathrm{c}} - n f_{\mathrm{m}})\right].$$

The Bessel function of order n is given by

$$J_n(\mu) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(n+k)!} \left(\frac{\mu}{2}\right)^{n+2k}$$

for positive integers n. It can also be written as

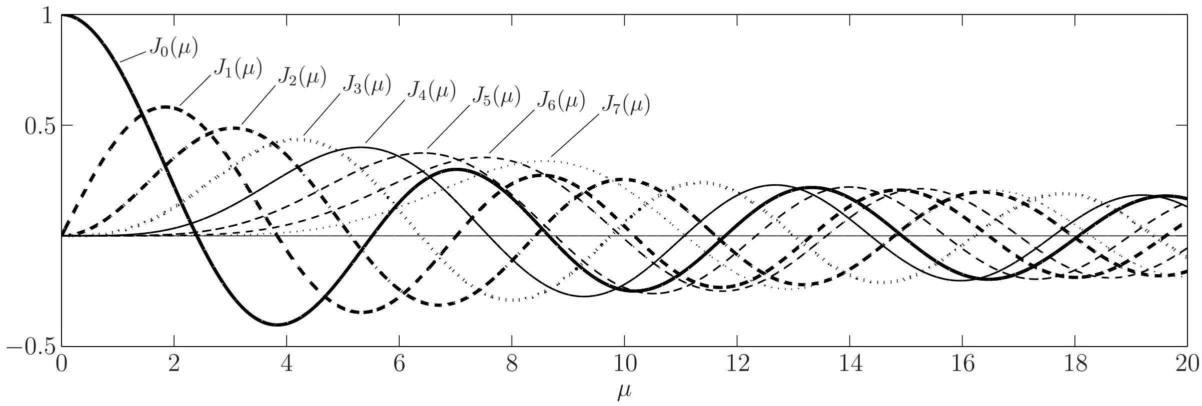

Figure 5.9: Bessel functions $J_n(\mu)$ for n up to 7.

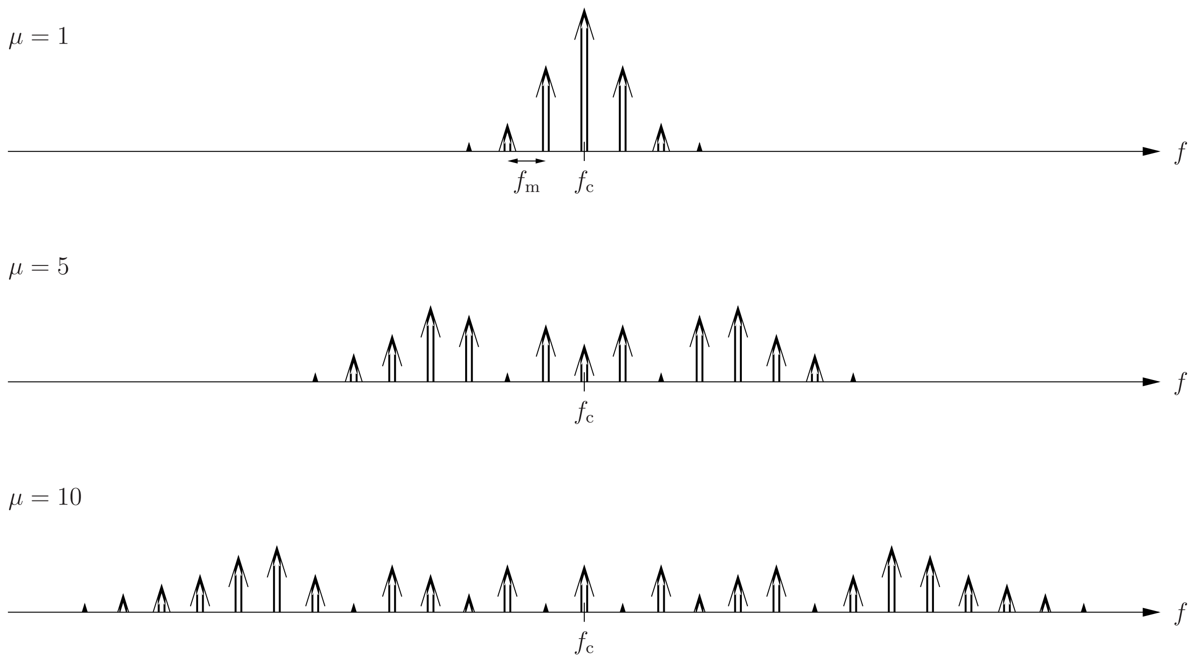

Figure 5.10: Spectrum of an angle modulated signal with modulation index μ, carrier frequency f_c, a cosine message with frequency f_m. The hights of the arrows denoting impulses represent the amplitude of the corresponding frequencies.

$$J_n(\mu) = \frac{1}{\pi}\int_0^{\pi} \cos(\mu\sin(\phi) - n\phi)\,d\phi,$$

still for positive integers n. For negative integers n, we have

$$J_n(\mu) = (-1)^n J_{-n}(\mu).$$

Formally, the bandwidth of this signal is infinite. However, the coefficients $J_n(\mu)$ decrease rapidly towards 0 for $|n| > \mu$, see Figure 5.9. Thus, we can state that the bandwidth of the signal is approximately $2\mu f_{\mathrm{m}}$. A common approximation of the bandwidth is $2(\mu+1)f_{\mathrm{m}}$, which is known as Carson's rule. Using the identity $\mu = f_{\mathrm{d,max}}/f_{\mathrm{m}}$, we can express the bandwidth as $2\left(1+\frac{1}{\mu}\right)f_{\mathrm{d,max}}$. In Figure 5.10, we have plotted the spectra for three different values of the modulation index μ, with a sine shaped message. Note how the bandwidth grows with the modulation index.

Phase Modulation

Phase modulation is normally abbreviated PM or PhM. The message, $m(t)$, is in this technique used directly to determine the momentary phase, i.e. we have

$$\phi(m(t)) = a\cdot m(t),$$

where a is some constant. The modulated signal, $x(t)$, is thus given by

$$x(t) = A\cdot\cos\left(2\pi f_{\mathrm{c}}t + a\cdot m(t)\right).$$

The momentary frequency for this modulation is

$$f_{\mathrm{mom}}(t) = \frac{1}{2\pi}\cdot\frac{d}{dt}\left(2\pi f_{\mathrm{c}}t + a\cdot m(t)\right) = f_{\mathrm{c}} + \frac{a}{2\pi}\cdot\frac{d}{dt}m(t),$$

the frequency deviation is

$$f_{\mathrm{d}}(t) = \frac{a}{2\pi}\cdot\frac{d}{dt}m(t).$$

and the peak frequency deviation is

$$f_{\mathrm{d,max}} = \frac{a}{2\pi}\cdot\max\left|\frac{d}{dt}m(t)\right|.$$

Finally, the frequency modulation index is given by

$$\mu_{\mathrm{f}} = \frac{a}{2\pi W}\cdot\max\left|\frac{d}{dt}m(t)\right|.$$

We notice that the peak frequency deviation depends on $a\cdot\max\left|\frac{d}{dt}m(t)\right|$. Hence, the bandwidth of $x(t)$ depends on the bandwidth of $m(t)$, but also on the amplitude of $m(t)$. In Figure 5.11a, a PM signal is presented together with the corresponding message.

Frequency Modulation

Frequency modulation is normally abbreviated FM. As for PM, the message, $m(t)$, determines the phase of the carrier, but not directly. Instead, the derivative of the phase is proportional to $m(t)$, i.e. the phase is a scaled indefinite integral of $m(t)$. More precisely, we have

$$\phi(m(t)) = a \int m(t)\, dt,$$

where a is some constant. The modulated signal, $x(t)$, is then given by

$$x(t) = \cos\left(2\pi f_{\mathrm{c}} t + a \int m(t)\, dt\right),$$

and the momentary frequency is given by

$$f_{\mathrm{mom}}(t) = \frac{1}{2\pi} \cdot \frac{d}{dt}\left(2\pi f_{\mathrm{c}} t + a \int m(t)\, dt\right) = f_{\mathrm{c}} + \frac{a}{2\pi} \cdot m(t).$$

Thus, the momentary frequency is directly given by the message. Note that any indefinite integral of $m(t)$ can be used as the phase. A natural choice is

$$\phi(m(t)) = a \int_{t_0}^{t} m(\tau)\, d\tau,$$

where t_0 is the time instance when the communication starts. The frequency deviation is

$$f_{\mathrm{d}}(t) = \frac{a}{2\pi} \cdot m(t).$$

and the peak frequency deviation is

$$f_{\mathrm{d,max}} = \frac{a}{2\pi} \cdot \max |m(t)|.$$

Finally, the frequency modulation index is given by

$$\mu_{\mathrm{f}} = \frac{a}{2\pi B} \cdot \max |m(t)|.$$

We notice that the frequency deviation depends on $\max |m(t)|$. Hence, the bandwidth of $x(t)$ depends on the amplitude of $m(t)$, but not on the bandwidth of $m(t)$. In Figure 5.11b, an FM signal is presented together with the corresponding message.

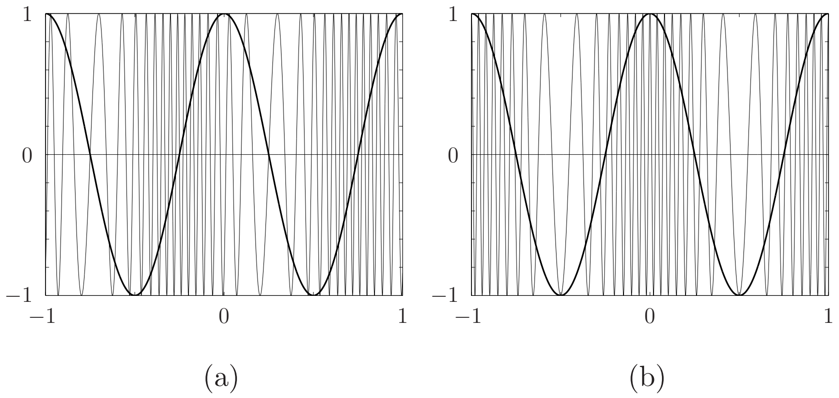

Figure 5.11: (a) A PM signal (thin line) for the message $m(t) = \cos(2\pi t)$ (thick line). (b) An FM signal (thin line) for the message $m(t) = \cos(2\pi t)$ (thick line). The modulation index is in both cases 10 and we have $A = 1$.

Demodulation of PM and FM

Recall that the sent signal is

$$x(t) = A \cdot \cos\left(2\pi f_c t + \phi(m(t))\right).$$

This signal can be demodulated by determining the derivative of the signal,

$$\frac{d}{dt}x(t) = -A\left(2\pi f_c + \frac{d}{dt}\phi(m(t))\right)\sin\left(2\pi f_c t + \phi(m(t))\right).$$

This gives us a signal for which the amplitude depends on the message $m(t)$ in a way similar to AM-DSB, but its carrier has varying phase. This signal can then be demodulated using an envelope detector, which gives us the envelope

$$A\left(2\pi f_c + \frac{d}{dt}\phi(m(t))\right).$$

The constant term can be removed using a BP or HP filter, leaving us with the signal $A\frac{d}{dt}\phi(m(t))$. For PM, we have $\phi(m(t)) = am(t)$. This means that we nead to integrate the signal $A\frac{d}{dt}\phi(m(t))$ to get the wanted message, i.e. we must produce

$$\int_{t_0}^{t} A\frac{d}{d\tau}\phi(m(\tau))\,d\tau = Aa \cdot (m(t) - m(t_0)),$$

where t_0 is the time instance when the communication started. For FM, we have

$$\phi(m(t)) = a\int m(t)\,dt.$$

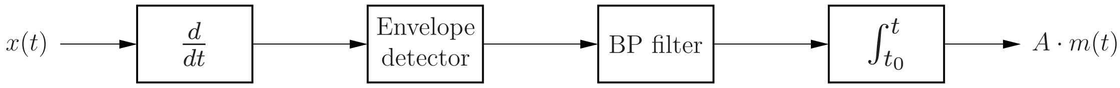

Figure 5.12: Demodulation of PM.

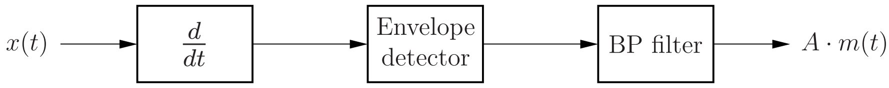

Figure 5.13: Demodulation of FM.

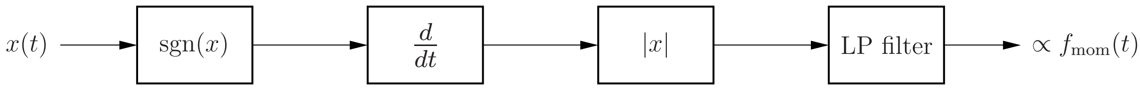

Figure 5.14: Detection of momentary frequency in angle modulated signals using zero crossings.

Then we have the signal

$$A\frac{d}{dt}\phi(m(t)) = A\frac{d}{dt}a\int m(t)\,dt = Aa \cdot m(t).$$

Therefore, demodulation of PM and FM can be done as indicated in Figures 5.12 and 5.13, respectively.

Alternatively, PM and FM can be demodulated by extracting the momentary frequency of the modulated signal $x(t)$. That can be done by detecting the zero crossings of $x(t)$. The demodulator in Figure 5.14 is based on this approach. The first block outputs the sign of its input, i.e. its output is $\mathrm{sgn}(x(t))$. The sgn function is defined as

$$\mathrm{sgn}(x) = \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ \text{-}1, & x < 0. \end{cases}$$

In practice, this block is an amplifier with very high gain. The second block produces the derivative of its input. The result is that the output of the second block is a positive (negative) impulse when $x(t)$ passes zero with positive (negative) derivative. After the third block, which is a rectifier, all those impulses are positive, with momentary frequency that is twice the momentary frequency of $x(t)$. All that is left to produce the message is an LP filter, which is the last block in Figure 5.14. This produces a signal that is proportional to the momentary frequency $f_{\mathrm{mom}}(t)$ of $x(t)$. For FM this is essentially the message $m(t)$, and for PM this is essentially $\frac{d}{dt}m(t)$. All we have left to do is to remove the DC component that originates from the carrier frequency, using a HP filter, or by replacing the LP filter in Figure 5.14 by a BP filter. For PM, we also have to integrate the output to get the message.

Impact of Noise in PM and FM Demodulation

We assume that the demodulation is carried out as described above, and we assume that the noise is dominated by white Gaussian noise with one-sided power spectral density N_0. We use the same notation for the involved powers as we did for AM demodulation.

- P: The (expected) power of the message $m(t)$.
- $P_{\mathrm{m-mod}}$: The average power of the modulated signal $x(t)$.
- P_{m}: The (expected) power of the message after demodulation and LP filter.
- $P_{\mathrm{n-mod}}$: The expected power of the ideally BP-filtered noise before demodulation.
- P_{n}: The expected power of the demodulated and LP-filtered noise.

The signal $x(t)$ is a cosine with amplitude A. The average power $P_{\mathrm{m-mod}}$ of the modulated signal $x(t)$ is therefore given by $P_{\mathrm{m-mod}} = A^2/2$. The varying phase - or frequency for that matter - is irrelevant in this respect. For both PM and FM, the demodulated signal is $Aa \cdot m(t)$, which gives us $P_{\mathrm{m}} = A^2 a^2 P$. The bandwidth of $x(t)$ is approximately $2f_{\mathrm{d,max}}$. The expected power $P_{\mathrm{n-mod}}$ of the ideally BP-filtered noise before demodulation is therefore $P_{\mathrm{n-mod}} = 2f_{\mathrm{d,max}} N_0/2 = f_{\mathrm{d,max}} N_0$.

Angle modulation methods are non-linear. That makes the analysis of detection in the presence of noise a lot more complicated than for AM. We simply skip that analysis and state the noise power for the two cases, under the assumption that the signal-to-noise ratio on the channel $P_{\mathrm{m-mod}}/P_{\mathrm{n-mod}}$ is high.

We start with PM. Then it can be shown that the noise power is given by

$$P_{\mathrm{n}} = 2WN_0.$$

This gives us the signal-to-noise ratio

$$\frac{P_{\mathrm{m}}}{P_{\mathrm{n}}} = \frac{A^2 a^2 P}{2WN_0}.$$

We would like to express this signal-to-noise ratio using the phase modulation index μ_{p}. We get

$$\mu_{\mathrm{p}} = \max|\phi(m(t))| = a \cdot \max|m(t)|,$$

from which we get

$$a = \frac{\mu_{\mathrm{p}}}{\max|m(t)|}.$$

We use this relation to rewrite the signal-to-noise ration as

$$\frac{P_{\mathrm{m}}}{P_{\mathrm{n}}} = \left(\frac{\mu_{\mathrm{p}}}{\max|m(t)|}\right)^2 \frac{A^2 P}{2WN_0}.$$

As we can see, we get increased signal-to-noise ratio with increased phase modulation index.

Now we turn to FM. It can be shown that the noise power is given by

$$P_{\mathrm{n}} = \frac{2W^3 N_0}{3}.$$

This gives us the signal-to-noise ratio

$$\frac{P_{\mathrm{m}}}{P_{\mathrm{n}}} = \frac{A^2 a^2 P}{2W^3 N_0/3}.$$

Now we would like to express the signal-to-noise ratio using the frequency modulation index μ_{f}. We have already noted that we have

$$\mu_{\mathrm{f}} = \frac{a \cdot \max|m(t)|}{2\pi W},$$

from which we get

$$\frac{a}{W} = \frac{2\pi\mu_{\mathrm{f}}}{\max|m(t)|}.$$

We use this relation to rewrite the signal-to-noise ratio as

$$\frac{P_{\mathrm{m}}}{P_{\mathrm{n}}} = 12\pi^2 \left(\frac{\mu_{\mathrm{f}}}{\max|m(t)|}\right)^2 \frac{A^2 P}{2WN_0}.$$

Here we get increased signal-to-noise ratio with increased frequency modulation index.

Pre-emphasized FM

The resulting noise after demodulating PM signals is evenly distributed over frequencies from 0 to W, which resembles the situation for AM. That is not the case for FM, where instead the resulting noise is dominated by high frequencies (near W). More precisely, the power spectral density of the resulting noise after demodulating FM as described above is $2N_0 f^2$, where f is frequency. Therefore, if we could combine PM and FM in such a way that PM is used for high frequencies in the message $m(t)$ and FM is used for low frequencies, we could hope for reduced noise compared to any of the two methods by themselves. Such a combination exists and is called pre-emphasized FM. Then the message is first filtered using a pre-emphasis filter with frequency response

$$H_1(f) = 1 + jf/f_0.$$

The output of that filter is then frequency modulated. In the receiver, after ordinary demodulation of the FM signal, the result is filtered with an inverse filter of $H_1(f)$ called a de-emphasis filter. That filter has frequency response

$$H_2(f) = \frac{1}{H_1(f)} = \frac{1}{1 + jf/f_0}.$$

The result is that we regain the original signal, and that the resulting noise is smaller than if we would have used ordinary PM or FM. The transmissions on the so called FM band (88-108 MHz) are done using pre-emphasized FM with $f_0 = 2122$ Hz.