AMPLITUDE MODULATION AND DEMODULATION

Modulation is a technique to transmit information via radio carrier waveform. It is a non-linear process that generates additional frequencies, as we will see. **Amplitude Modulation** (**AM**) works by varying the amplitude ("strength") of a carrier signal, in proportion to amplitude of a modulation signal that is to be transmitted to one or more receivers.

The carrier signal *c(t)* is just a sinusoid with a frequency f_c and an amplitude A_c. This is mathematically expressed as

$$c(t) = A_c \cdot \cos(2\pi \cdot f_c t) = A_c \cdot \cos(\omega_c t)$$

where the angular frequency ω (in radians/sec) is 2π times the ordinary frequency f (in hertz).

Let's assume that the modulation signal (a.k.a. baseband signal) is a constant audio tone. That is, a sinusoid with frequency f_m and amplitude A_m:

$$m(t) = A_m \cdot \cos(2\pi \cdot f_m t) = A_m \cdot \cos(\omega_m t),$$

To simplify the following analysis, without changing the essence, we choose $A_c = 1$ and impose the condition that $|A_m| < 1$, hence $|m(t)| < 1$.

The AM modulation process consists of adding a positive DC offset (bias) to the modulation signal, and multiplying the result with the carrier signal. For a DC offset equal to 1 (again, for convenience), the resulting AM modulated signal *am(t)* is:

$$\begin{aligned} am(t) &= [1 + m(t)] \cdot c(t) \\ &= [1 + A_m \cdot \cos(\omega_m t)] \cdot \cos(\omega_c t) \\ &= \cos(\omega_c t) + A_m \cdot \cos(\omega_m t) \cdot \cos(\omega_c t) \end{aligned}$$

Figure 1 illustrates the process:

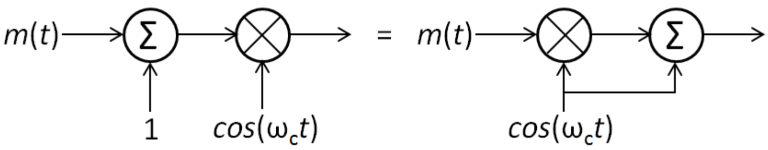

Figure 1: the AM modulation process

The DC offset is required: without it, the transmitted AM signal cannot be demodulated with a simple envelope-detector receiver. Instead, a more complicated synchronous (coherent) detector is required, with a local oscillator frequency that *exactly* matches the carrier frequency.

Let's use the following standard trigonometry product-to-sum identity (*that we all learned in school, right?*), to expand the product term in the above equation:

$$\cos(x) \cdot \cos(y) = \frac{1}{2} \cdot [\cos(x + y) + \cos(x - y)]$$

Hence,

$$\begin{aligned} am(t) &= \cos(\omega_c t) + A_m \cdot \cos(\omega_c t) \cdot \cos(\omega_m t) \\ &= \cos(\omega_c t) + \frac{A_m}{2} \cdot \cos(\omega_c t + \omega_m t) + \frac{A_m}{2} \cdot \cos(\omega_c t - \omega_m t) \end{aligned}$$

Clearly, the frequency spectrum of this signal consists of three discrete spectral lines (on the positive half of the frequency axis). The frequencies are f_c, $f_c + f_m$, and $f_c - f_m$. That is, the original carrier frequency, as well as the sum and difference frequency. The latter two are known as the lower sideband (LSB) and upper sideband (USB). They are located symmetrically about the carrier frequency. The AM signal is also known as DSB-FC or DSB-LC: Double Sideband – Full (or Large) Carrier, as opposed to DSB-SC (Double Sideband - Suppressed Carrier). Normally, $f_c >> f_m$.

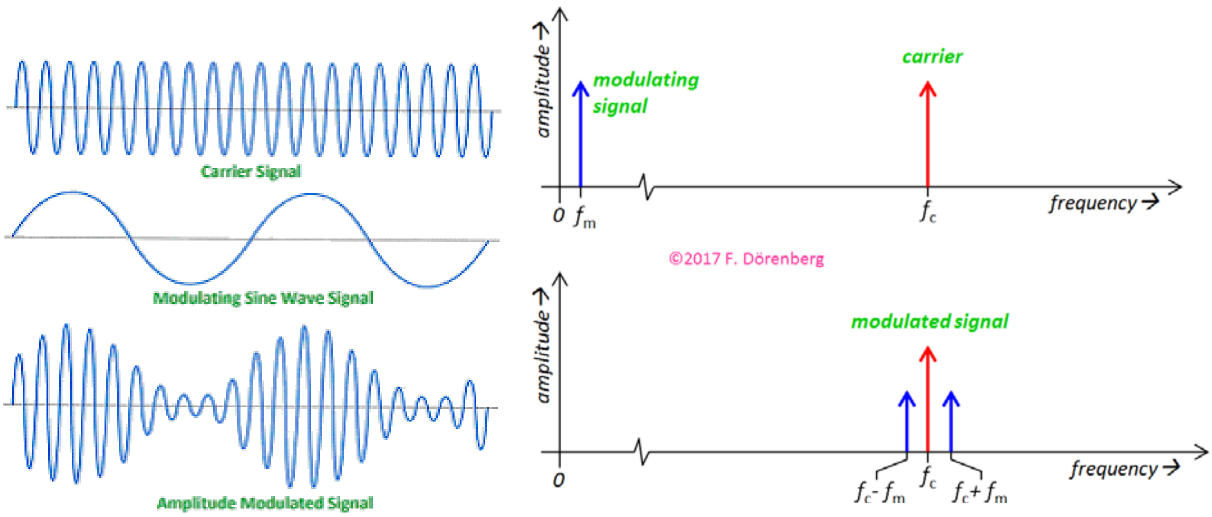

Figure 2: time domain signals and the equivalent frequency spectrum

The equation above shows that the height of the sideband lines is different from that of the residual carrier. The relative height is 20log(M) dB, where M is the non-negative Modulation Index (a.k.a. modulation depth). This is the ratio of the modulation amplitude and the amplitude of the unmodulated carrier. For $M = 1$ (i.e., 100% modulation), the envelope of the modulated signal periodically reaches zero, and the sidebands are 6 dB below the carrier. For smaller M, the sidebands are lower (smaller) than that (i.e., by more than 6 dB). For very small M, most of the transmitter power is used for the carrier frequency, rather than the transmitted information (i.e., the modulation). For $M > 1$, the modulated signal is distorted whenever the modulating amplitude exceeds the carrier amplitude.

At the receiver, the received AM signal must be demodulated. That is, the original modulating (baseband) signal must be recovered (reconstructed). There are several basic ways to do this:

- envelope detection
- Square-Law demodulation
- synchronous product detection
- asynchronous product detection

A simple **envelope detector** consists of a full-wave or half-wave diode rectifier, followed by a low-pass filter. The filter cuts off the carrier frequency, but not the modulating frequency. The output of the filter is the envelope of the AM signal plus a DC offset. The latter is proportional to the amplitude of the received carrier signal, and is removed with a blocking capacitor (or transformer) in series with the filter output. As diodes are inherently non-linear, the output signal of the detector has some distortion on it. Envelope detection using the Hilbert Transform is beyond the scope of this discussion.

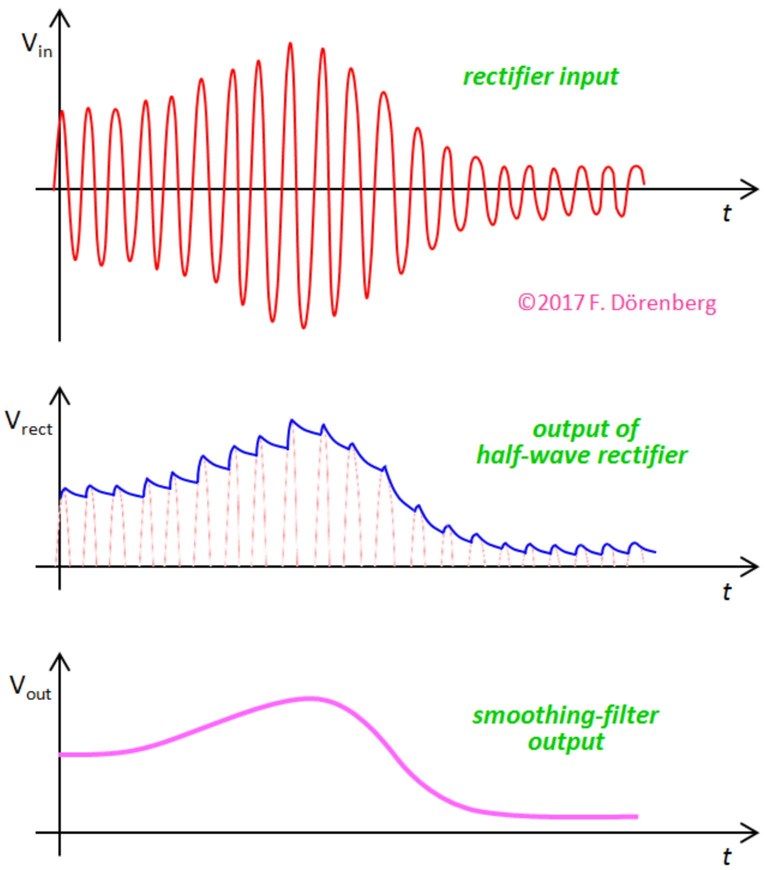

Figure 3: input & output of a diode-rectifier envelope detector

---

As the name suggests, with **Square-Law** demodulation, the received *am(t)* signal is squared, i.e., multiplied by itself. The resulting signal *sqam*(*t*) is the square of the modulating signal 1+ *m(t)*, plus that same squared signal times the second harmonic (= double frequency) of the carrier signal:

$$sqam(t) = am^2(t) = [1 + m(t)]^2 \cdot \cos^2(\omega_c t)$$

Again, using the product-to-sum identity

$$\cos^2(x) = \cos(x)\cdot\cos(x) = \frac{1}{2}\cdot[\cos(0) + \cos(2x)] = \frac{1}{2}\cdot[1 + \cos(2x)]$$

we obtain

$$sqam(t) = [1 + m(t)]^2 \cdot \cos^2(\omega_c t) = [1 + m(t)]^2 \cdot \frac{1}{2}\cdot[1 + \cos(2\omega_c t)]$$
$$\propto [1 + m(t)]^2 + [1 + m(t)]^2 \cdot \cos(2\omega_c t)$$

The latter high frequency signal $2\omega_c t$ is filtered out, and the remaining squared modulating signal is passed through a "square root" circuit. After removing the DC content, *m(t)* is obtained.

A **synchronous detector** is a product detector (mixer) with a local oscillator (LO) signal that is synchronized to the carrier of the received signal (both frequency and phase, i.e., **coherent**). The multiplication performed by the mixer is referred to as "heterodyning", or "beating". The LO signal is derived from the received carrier signal itself, e.g. with a Phase-Locked Loop circuit (PLL). The mixer output *mo(t)* of this detector consists of a DC signal, the modulating signal, and the second-order harmonic (= double) of the carrier frequency:

$$mo(t) \propto am(t) \cdot A_{LO} \cdot \cos(\omega_c t) = [1 + A_m \cdot \cos(\omega_m t)] \cdot A_c A_{LO} \cdot \cos^2(\omega_c t)$$

Using

$$\cos^2(\omega_c t) = \frac{1}{2} \cdot [\cos(\omega t - \omega t) + \cos(\omega t + \omega t)] = \frac{1}{2} \cdot [1 + \cos(2\omega t)]$$

we obtain

$$\begin{aligned} mo(t) &\propto [1 + A_m \cdot \cos(\omega_m t)] \cdot A_c A_{LO} \cdot \cos^2(\omega_c t) \\ &\propto [1 + A_m \cdot \cos(\omega_m t)] \cdot [1 + \cos(2\omega_c t)] \\ &= [1 + A_m \cdot \cos(\omega_m t)] + [1 + A_m \cdot \cos(\omega_m t)] \cdot \cos(2\omega_c t) \end{aligned}$$

That is, a DC term plus the modulating signal f_m, plus the complete AM signal, shifted from f_c to $2f_c$:

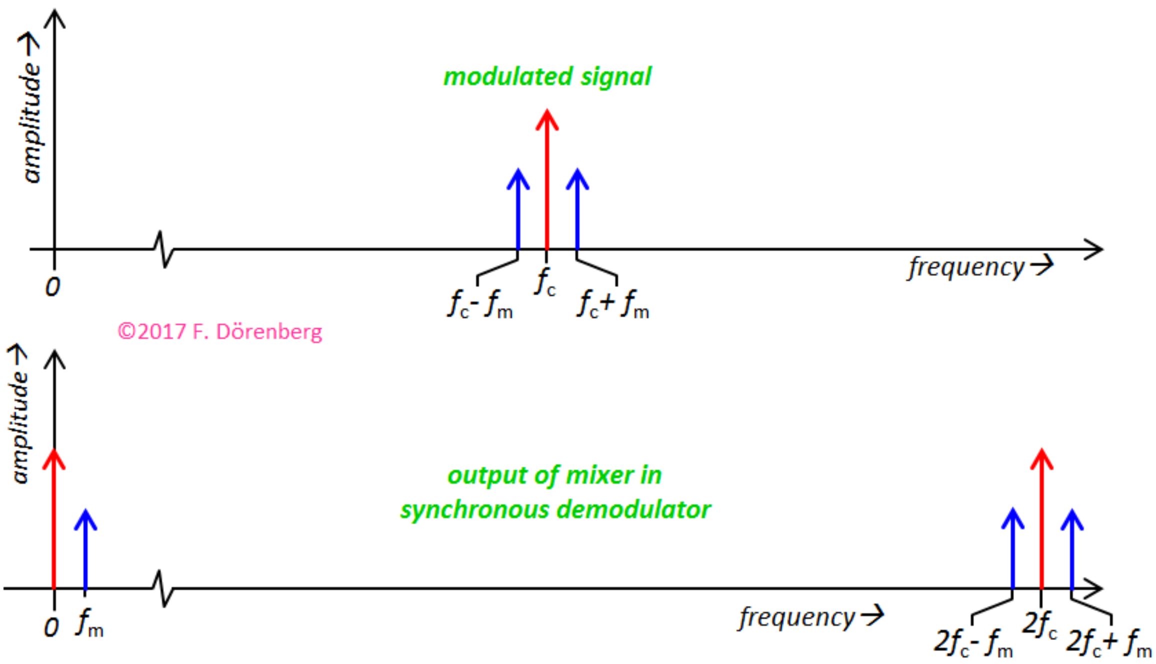

Figure 4: AM modulation and synchronous de-modulation

The higher-frequency product signals are filtered out with a low-pass filter that has a cut-off frequency between the modulating frequency f_m and $2f_c - f_m$. Note that with *synchronous* detection, the DC-term in *mo(t)* serves no purpose, and injection of the DC bias during the modulation process is also not necessary.

An **asynchronous detector** that uses a mixer to multiply (= heterodyne) the AM signal with a sinusoidal signal from a local oscillator (LO). Unlike the much more complicated synchronous detector, the LO of this detector is now <u>not</u> synchronized to the carrier signal. I.e., it has a different frequency and, hence, different phase. This causes the carrier of the received modulated signal not to be shifted to zero frequency ($f_c - f_{LO} = 0$), but to the non-zero difference between that carrier frequency and the LO frequency ($f_c - f_{LO} = \Delta f$):

$$
\begin{aligned}
mo(t) &\propto am(t)\cdot A_{LO}\cdot\cos(\omega_{LO}t)\\
&= [1+A_m\cdot\cos(\omega_m t)]\cdot A_c\cdot\cos(\omega_c t)\cdot A_{LO}\cdot\cos(\omega_{LO}t)\\
&\propto [1+A_m\cdot\cos(\omega_m t)]\cdot\cos(\omega_c t)\cdot\cos(\omega_{LO}t)\\
&= [1+A_m\cdot\cos(\omega_m t)]\cdot\frac{1}{2}\{\cos[(\omega_c-\omega_{LO})t]+\cos[(\omega_c+\omega_{LO})t]\}\\
&\propto [1+A_m\cdot\cos(\omega_m t)]\cdot\cos[(\omega_c-\omega_{LO})t]+[1+A_m\cdot\cos(\omega_m t)]\cdot\cos[(\omega_c+\omega_{LO})t]\\
&= am_1(t)+am_2(t)
\end{aligned}
$$

Hence, the associated frequency spectrum shows two complete AM signals (carrier plus sidebands): one centered on the *difference* between the carrier and LO frequency (i.e., at $f = f_c - f_{LO}$), the other centered on the *sum* of the carrier and LO frequency (i.e., at $f = f_c + f_{LO}$):

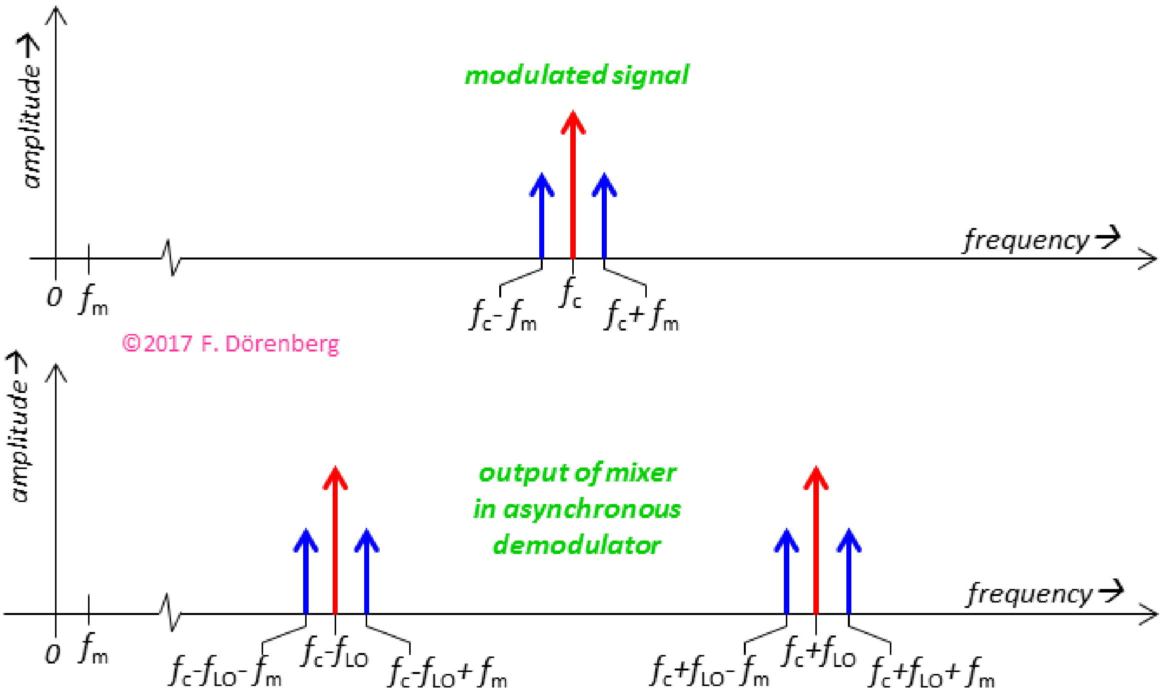

Figure 5: AM modulation and asynchronous de-modulation

If f_{LO} is large compared to f_c, then the AM signal at $f = f_c + f_{LO}$ can be filtered out. The filter output can then be passed through envelope detector, to recover the original modulating signal. An AM-receiver comprising an asynchronous mixer and an envelope detector is called a "superheterodyne" receiver, or "superhet" for short.

For the special case where $f_{LO} = f_c$, we basically obtain the synchronous detection case (besides the effects of phase difference).

As *cos*(*x*) = *cos*(-*x*), it basically does not matter on which side of f_c the f_{LO} is located (through there may be practical reasons to select a specific side). If the LO frequency f_{LO} is significantly higher than the modulation frequency f_m, (e.g., relatively close to the carrier frequency f_c), then the second term (with the sum "beat" frequency $f = f_c + f_{LO}$) can easily be filtered out. If f_{LO} is chosen properly, the remaining filtered AM signal can then be demodulated with an envelope detector as described above.

Why heterodyne the modulated signal down to an intermediate frequency (IF)? This allows sharp filters – *with a fixed center frequency* - placed *after* the detector, instead of sharp filters - *with an adjustable center frequency that must track the LO frequency* - placed in *front of* the multiplier (mixer).

What happens if we have not one, but two AM radio transmitters: each with a 1-tone modulating signal (f_{m1} and f_{m2}), and with carrier frequencies (f_{c1} and f_{c2}, respectively) that are spaced by less than the HF, IF, and audio bandwidth of the radio receiver. I.e., what if the radio can receive both AM signals simultaneously (in broadcast radio, this would be called “interference”)? The resulting transmitted RF spectrum simply comprises two independent AM signals:

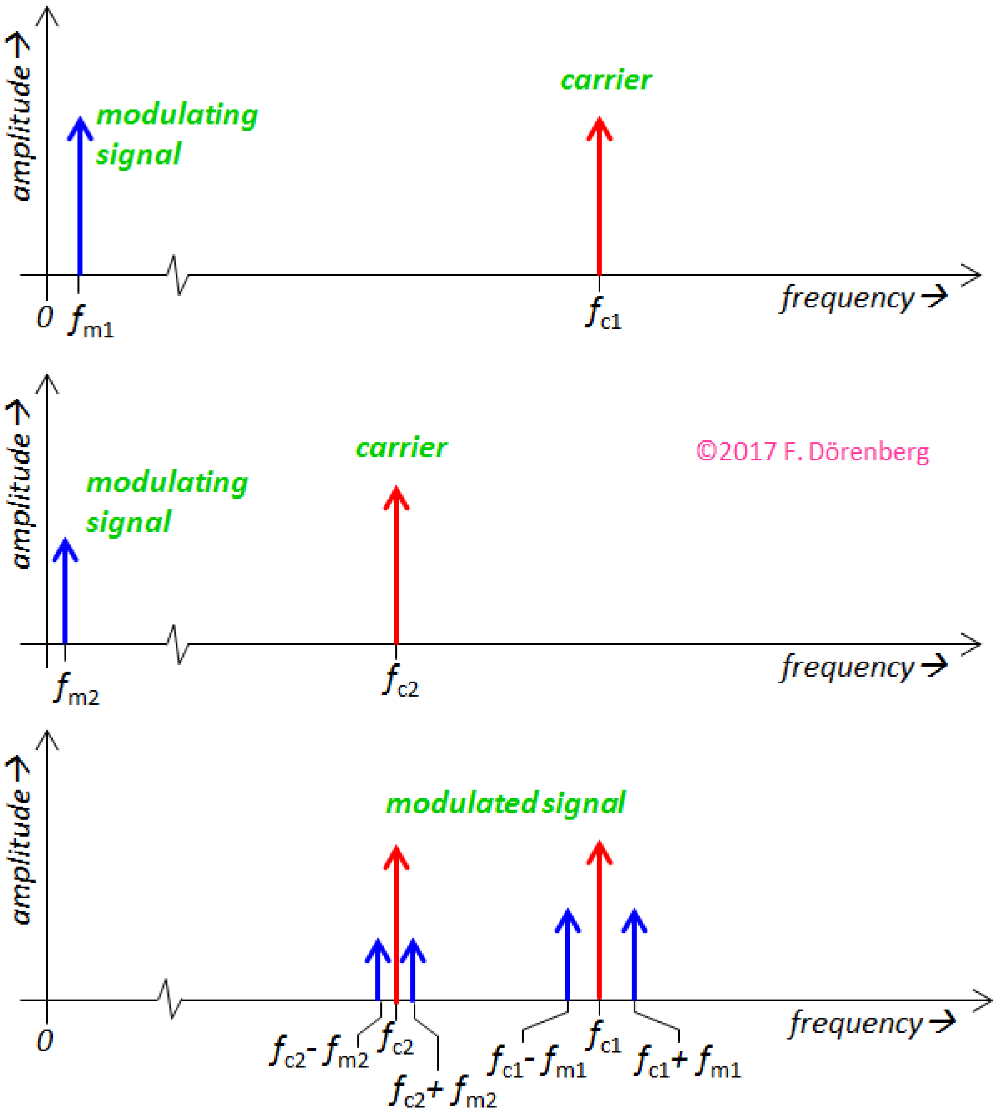

Figure 6: spectrum of two separate carriers, each AM-modulated with a constant tone

Note that this is <u>not</u> the same spectrum as for a single carrier that is AM modulated with two separate constant tones:

AMPLITUDE MODULATION AND DEMODULATION

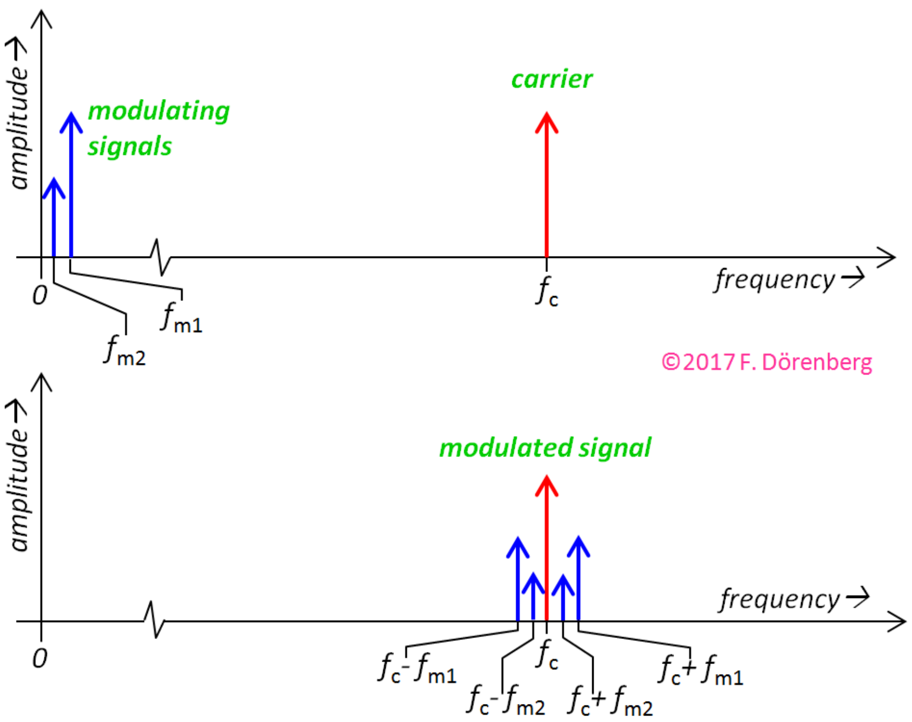

Figure 7: spectrum of a single carrier, AM-modulated with two constant tones

If the receiver has an *asynchronous* product detector, then both of the received AM signals in Figure 6 will be shifted downward (i.e., moved to the left on the frequency scale) by an amount equal to f_{LO}, where $f_{LO} \neq f_c$. At the same time, two additional AM signals appear: the same received AM signals, but now both shifted upward by an amount equal to f_{LO}:

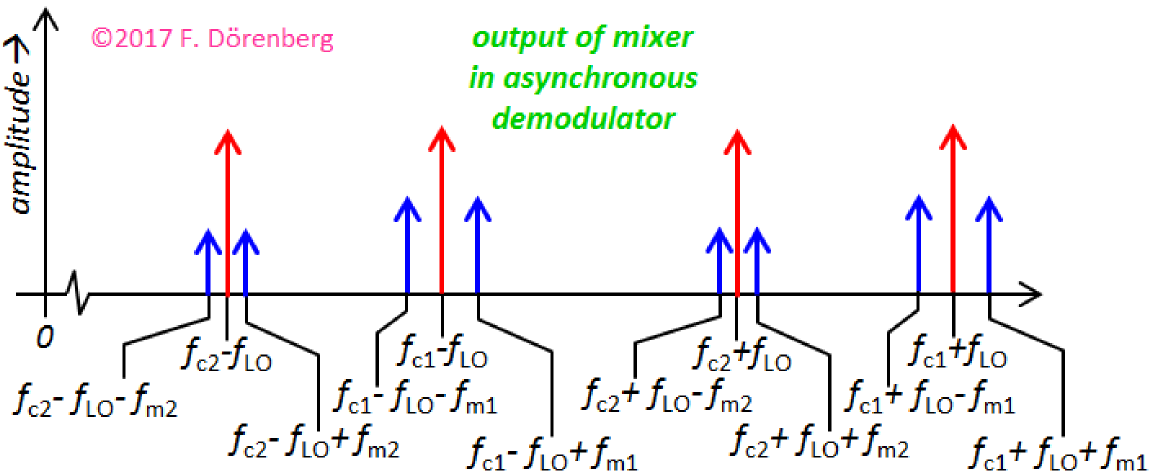

Figure 8: output of the mixer of an asynchronous product detector

If f_{LO} is large compared to f_{c1} and f_{c2}, then the AM signals at f_{c1} + f_{LO} and f_{c1} + f_{LO} can be filtered out. The base signals at f_{m1} and f_{m2} can then be recovered with an envelope detector. Note that this approach does not work with a *synchronous* demodulator: f_{LO} cannot be equal to f_{c1} and to f_{c2} at the same time. E.g., if the LO is synchronized to f_{c2} (i.e., $f_{LO} = f_{c2}$), then the mixer will shift the AM signal at f_{c1} to $f_{c1} - f_{LO} \neq 0$. Hence, the latter AM signal will not be demodulated correctly: the baseband modulating tone at f_{m1} will sound like a tone at $f = f_{c1} - f_{LO} - f_{m1} \neq f_{m1}$.

For baseband signals other than a single tone, the story is exactly the same. But instead of a discrete spectral line at a distance f_m on either side of the carrier, there will be sidebands with a certain bandwidth:

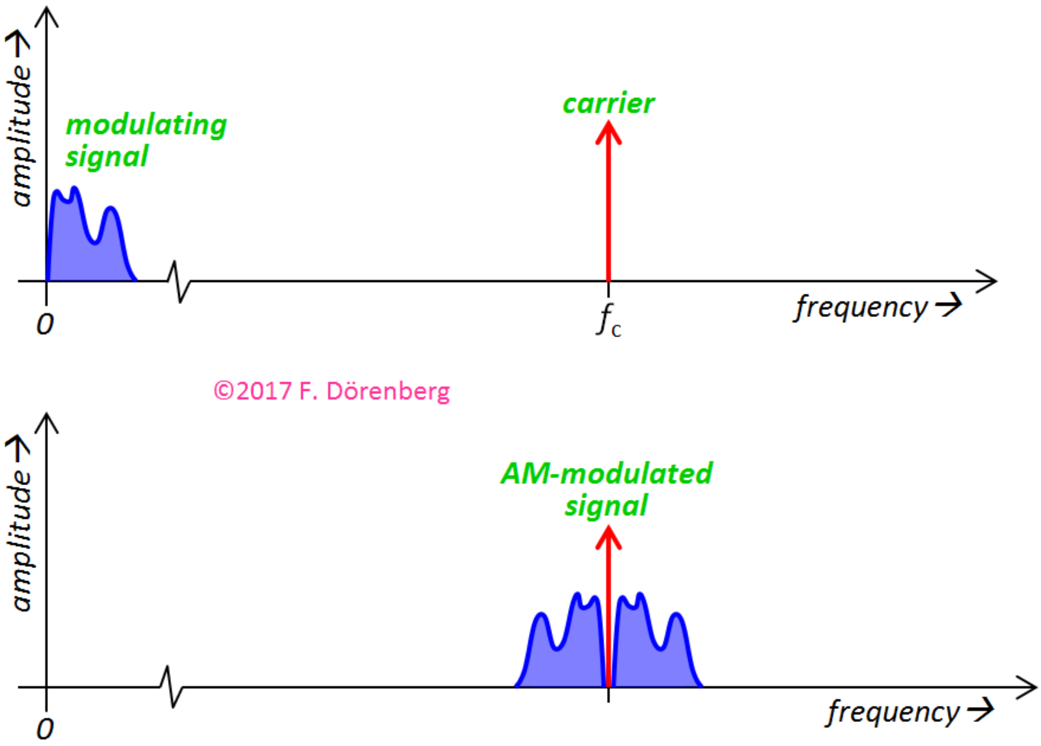

Figure 9: output AM spectrum for modulating signal that is not a single tone